Online Learning and Information Exponents: The Importance of Batch Size & Time / Complexity Tradeoffs

Luca Arnaboldi [1] **Yatin Dandi** [1 2] **Florent Krzakala** [1] **Bruno Loureiro** [3] **Luca Pesce** [1] **Ludovic Stephan** [1]

Abstract

We study the impact of the batch size n_b on the iteration time T of training two-layer neural networks with one-pass stochastic gradient descent (SGD) on multi-index target functions of isotropic covariates. We characterize the optimal batch size minimizing the iteration time as a function of the hardness of the target, as characterized by the information exponents. We show that performing gradient updates with large batches $n_b \lesssim d^{\ell/2}$ minimizes the training time without changing the total sample complexity, where ℓ is the information exponent of the target to be learned (Ben Arous et al., 2021) and d is the input dimension. However, larger batch sizes than $n_b \gg d^{\ell/2}$ are detrimental for improving the time complexity of SGD. We provably overcome this fundamental limitation via a different training protocol, *Correlation loss SGD*, which suppresses the auto-correlation terms in the loss function. We show that one can track the training progress by a system of low-dimensional ordinary differential equations (ODEs). Finally, we validate our theoretical results with numerical experiments.

1. Introduction

Descent-based algorithms, such as Stochastic Gradient Descent (SGD) and its variations, are the backbone of contemporary machine learning. Their simplicity in implementation, efficiency in operation, and notably effective performance in practice highlight their importance. A mathematical understanding of SGD's effectiveness remains a key focus in the field. Recent progress has been particularly noteworthy in the realm of shallow neural networks. A sequence of works demonstrated that optimizing large width two-layer neural networks can be mapped into a convex optimization problem over the space of probability measures of weights, the so-called mean-field analysis (Mei et al., 2018; Chizat and Bach, 2018; Rotskoff and Vanden-Eijnden, 2022; Sirignano and Spiliopoulos, 2020). Following this breakthrough, a large part of the theoretical effort has shifted to describing what class of functions can be efficiently learned by SGD, i.e. time and computational complexities required to learn a given class of functions. This has been, in particular, thoroughly analyzed in a series of recent works focusing on isotropic distributions (e.g. Gaussian, spherical or in the hypercube) and targets depending only on a few relevant directions (a.k.a. *multi-index models*). A key result from this literature is that the time complexity of SGD scales with the covariates dimension according to the so-called *information exponent* (Ben Arous et al., 2021) for single-index and *leap complexity* (Abbe et al., 2021; 2023) for multi-index targets, sparking increasing interest from the theoretical machine learning community over the last few months (Damian et al., 2022; 2024; Dandi et al., 2023; Bietti et al., 2023; Ba et al., 2024; Moniri et al., 2023; Mousavi-Hosseini et al., 2023; Zweig and Bruna, 2023).

Our work follows this thread, focusing instead on the effect of batch size n_b, parallelization, and sample-splitting into the overall complexity required to learn a multi-index target. Instead of looking at data one-by-one, as is common in theoretical studies, we investigate the finite n_b problem, and characterize the time/complexity tradeoff when learning with one-pass SGD. Our central goal is to paint a complete picture of how fast generalized linear models and two-layer neural networks adapt to the features of training data as a function of n_b, and the structure of the target function.

Our analysis sheds light on a fundamental limitation of one-pass (or online) SGD, namely that for batch sizes larger than the input dimension, the dynamics of the training algorithm is dominated by negative feedback terms that do not permit to reduce the time iterations needed to learn the target. Therefore, we provide a rigorous solution to this fundamental limitation of SGD by considering gradient updates

[1] Information Learning and Physics Laboratory, École Polytechnique Fédérale de Lausanne (EPFL) [2] Statistical Physics Of Computation Laboratory, École Polytechnique Fédérale de Lausanne (EPFL) [3] Département d'Informatique, École Normale Supérieure, Paris, France. Correspondence to: Luca Arnaboldi <luca.arnaboldi@epfl.ch>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

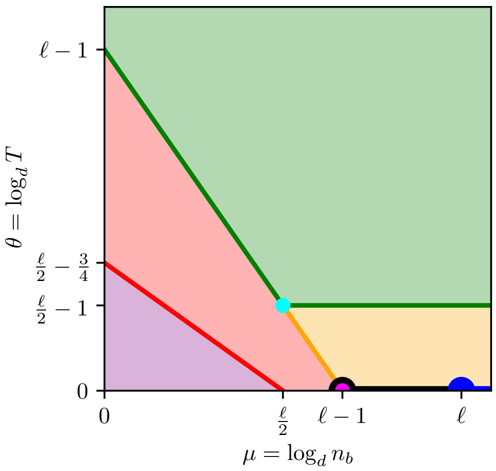

Figure 1. **Time / Batch size tradeoff for weak recovery:** Phase diagram illustrating different SGD learning regimes as a function of the batch size exponent $\mu=\log_d n_b$ and weak recovery time exponent $\theta=\log_d T$. The analysis is dependent on the target's information exponent ℓ, this particular plot is valid when $\ell\geq 3$. **Not correlating region:** SGD is not able to achieve weak recovery. **Self-interaction regime:** SGD is not able to perform weak recovery, but Correlation loss SGD overcomes this limitation. **Weak recovery region:** SGD successfully achieve weak recovery. Note that it exists an **optimal choice** at batch size $n_b=O(d^{\ell/2})$ that minimizes the number of iterations needed by SGD, and another **optimal point** at $n_b=O(d^{\ell-1})$ for *Correlation Loss*. The critical line where $n_b=\Omega(d^{\ell-1})$ is not addressed by our formal . See details about the other two regions (**Polylog Regime** and **One-step regime** (Dandi et al., 2023)) in Appendix D.

on the correlation loss. Our approach, drawing inspiration from the *summary statistics* method employed by (Saad and Solla, 1995a; Ben Arous et al., 2021; 2022), concentrates on the overlaps of neurons with the target subspace and their norms. This differs from recent studies, such as those by (Abbe et al., 2022) and (Damian et al., 2022), which focus on the full gradient vector.

2. Setting, Contributions, and Related Works

Consider a two-layer neural network with activation function σ and first and second layer weights given by $W\in\mathbb{R}^{p\times d}$ and $\boldsymbol{a}\in\mathbb{R}^p$ respectively:

$$f(\boldsymbol{z})=\frac{1}{p}\sum_{j=1}^{p}a_j\sigma(\langle\boldsymbol{z},\boldsymbol{w}_j\rangle)\,. \tag{1}$$

We are interested in studying the capacity of f to learn from training data $\mathcal{D}=\{(\boldsymbol{z}^\nu,y^\nu)_{\nu\in[N]}\in\mathbb{R}^{d+1}\}$. In the following, we work under the following setting.

Data model — As hinted in the introduction, we focus on the case the (noisy) labels depend on the covariates only through a projection over a k-dimensional subspace:

$$y^\nu=h^\star(W^\star\boldsymbol{z}^\nu)+\sqrt{\Delta}\xi^\nu,\quad \boldsymbol{z}^\nu\sim\mathcal{N}(0,I_d) \tag{2}$$

where $W^\star=\{\boldsymbol{w}_r^\star\}_{r\in[k]}\in\mathbb{R}^{k\times d}$ are the target weights, $h^\star:\mathbb{R}^k\to\mathbb{R}$ is a non-linear activation function, $\xi^\nu\sim\mathcal{N}(0,1)$ is the label noise with variance given by $\Delta\geq 0$. We focus on the case where $k=O(1)$ and d is large, i.e. the label only depends on a few directions of a high-dimensional ambient space. The target function $f^\star(\boldsymbol{z})=h^\star(W^\star z)$ is often refereed in the literature as a *multi-index model*.

Note that the setting above where we assume a generative model for the data and study the capacity of a model to learn is also known as teacher-student model in the literature. We adopt this terminology and refer to $f^\star$ and f as the *teacher* and the *student* functions, respectively. Similarly, we refer to $W^\star$ and W as the *teacher* and *student* weights.

Hardness of the learning task — Characterizing what class of targets are efficiently learned by two-layer networks is arguably one of the key question in theoretical machine learning. The pivotal work of (Ben Arous et al., 2021) provably describes that for $k=1$, the hardness of the learning task is encoded by a single number, the information exponent ℓ. More precisely, given the activation $h^\star$ in (2), ℓ is the lowest degree of the Hermite polynomials $\{\mathrm{He}_j\}_{j\in\mathbb{N}}$ appearing in the Hermite expansion of $h^\star$. This notion generalizes direction-wise for multi-index models ($k>1$), where ℓ is known as leap complexity (Abbe et al., 2023).

Definition 2.1 (Information Exponent (Ben Arous et al., 2021))**.**

$$\ell=\min\{j\in\mathbb{N}:\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[h^\star(\xi)\mathrm{He}_j(\xi)\neq 0\right]\} \tag{3}$$

Training algorithm — Given the training data $\mathcal{D}$, we consider the training of $(W,\boldsymbol{a})$ under a *sample splitting* scheme: the data is partitioned $\mathcal{D}=\bigcup_{t=1}^{T}\mathcal{D}_t$ into $T=\lfloor N/n_b\rfloor$ disjoint batches $\mathcal{D}_t$ of size n_b, which are used, one every iteration, to train the network. We consider a common assumption for the training algorithm that is to decouple the training of the hidden weights W and the second layer weights $\boldsymbol{a}$. By keeping fixed the second layer weights at initialization $\boldsymbol{a}=\boldsymbol{a}_0$, the hidden layer weights W are estimated using (projected) SGD:

$$\boldsymbol{w}_{j,t+1}=\frac{\boldsymbol{w}_{j,t}-\gamma\nabla_{\boldsymbol{w}_{j,t}}\ell_t}{\left\|\boldsymbol{w}_{j,t}-\gamma\nabla_{\boldsymbol{w}_{j,t}}\ell_t\right\|}\qquad\forall t\in[T],\,\forall j\in[p] \tag{4}$$

where:

$$\ell_t=\frac{1}{2n_b}\sum_{\nu=1}^{n_b}(y^\nu-f(\boldsymbol{z}^\nu))^2,\qquad\forall t\in[T] \tag{5}$$

is the empirical risk over a batch of data. Two comments are in place. First, the gradient at each step is computed using the empirical loss given by fresh, previously, unseen samples coming from $\mathcal{D}_t$. Each gradient is thus an unbiased estimator of the true gradient, which means that on average this algorithm minimizes the population risk over W:

$$\mathcal{R} = \mathbb{E}_{(\boldsymbol{z},y)}\left[\frac{1}{2}(y - f(\boldsymbol{z}))^2\right] \tag{6}$$

Second, the spherical projection allow us to focus just on the direction learned by the network, putting aside the effect of the change of the norm of the weights. Note that in equation (4) we have kept the read-out layer $\boldsymbol{a}$ fixed. Eventually, the second layer could also be trained with SGD, as the first layer, or even with the Moore-Penrose pseudo-inverse solution; In this paper, however, we consider it fixed and focus on the *feature learning* step, i.e., the recovery of the low-dimensional space spanned by $W^\star$.

High-dimensional regime — We focus in the high-dimensional regime where $d \to \infty$. Of particular interest is the case where the batch size n_b scales with the dimension d. Indeed, in modern machine learning, and in particular in the realm of distributed and federated learning, scenarios with large batches, a single pass, and few iterations often becomes the norm (Goyal et al., 2017; Li et al., 2020) (as for instance when training large language models), further underlining the relevance of this scenario. More precisely, we assume a scaling of the relevant parameters, i.e., learning rate and the batch size, with d, as follows:

$$\gamma = \gamma_0 d^{-\delta} \quad \text{and} \quad n_b = n_0 d^{\mu}. \tag{7}$$

with $\mu \geq 0$ and δ could possibly be any real value. The exponents (δ, μ) characterize the Time / Complexity tradeoff illustrated in the phase diagram (Fig. 1). More precisely, the figure shows the time complexity $T = T_0 d^\theta$ as a function of the batch size exponent μ. The time exponent (θ) is linked to the learning rate one (δ) and the information exponent (ℓ), and determining this relation is the main object of analysis of the following sections.

Weak recovery of the target — The central object of our analysis is to characterize the time iterations needed for the SGD dynamics defined in eq. (5) to learn the low-dimensional features $W^\star$. More precisely, we are interested in studying the number of steps to achieve order one correlation with the target weights $W^\star$. We refer to this condition as *weak recovery* of the target subspace, formalized in the following definition.

Definition 2.2 (Weak recovery)**.** The *target subspace* $V^\star$ is defined as the span of the rows of the target weights $W^\star$:

$$V^\star = \operatorname{span}(\boldsymbol{w}_1^\star, \dots, \boldsymbol{w}_k^\star) \tag{8}$$

We define the following weak recovery stopping time for a parameter $\eta \in (0,1)$ independent from d:

$$t_\eta^+ = \min\{t \geq 0 : \|WW^{\star\top}\|_F \geq \eta\} \tag{9}$$

Our key objective is to characterize the largest affordable batch size n_b to achieve weak recovery of the relevant target subspace $V^\star$ while minimizing the training time iterations T. Indeed, the updates of one-pass SGD in eq. (5) consist of sums of independent terms that can be parallelized efficiently with decentralized learning protocols.

Our **main contributions** in this paper are the following:

- We study how the batch size influences the number of steps required to learn a target function, for different information exponents of the problem. We introduce a schematic phase diagram describing the different learning regimes, see Fig. 1.
- We show that performing gradient updates with large batch sizes can reduce the training time without changing the total sample complexity to weakly recover the teacher subspace only up to $n_b \lesssim \Psi(\ell)$ samples per steps, with d the data dimension and ℓ the information exponent of the target. Beyond this limit, larger batch sizes are detrimental for one-pass SGD.
- We characterize that it is possible to improve over this fundamental limitation of one-pass SGD by using gradient updates on the correlation loss, namely *Correlation loss SGD*. We provably show that the number of steps needed to weakly correlate with the target with this new training protocol can then be pushed down to $T = \operatorname{polylog}(d)$ when using batch sizes $n_b = O(d^{\ell-1})$, with ℓ the information exponent. Additionally, we provide sharp prescription on how to scale the learning rate with batch size and input dimension, in order to achieve the best time-memory tradeoff.
- We show that the asymptotic training dynamics is described by a system of Ordinary Differential Equations (ODEs) that can be solved exactly. We leverage on the ODE description to characterize the different learning phases of two-layer networks when intialized with non-vanishing initial correlation with the target direction to be learned (warm starts). We also discuss finite d corrections to the asymptotic dimension-free description.
- Finally, we validate and illustrate our theoretical results with numerical experiments.

The code to reproduce representative figures are available in the Github repository https://github.com/IdePHICS/batch-size-time-complexity-tradeoffs. We refer to App. E for details on the numerical implementations while the rigorous proofs of the main results are detailed in App. A.

Other related works — The dynamics of Stochastic Gradient Descent (SGD) in two-layer neural networks, particularly when trained on synthetic Gaussian data, have been a topic of interest since the seminal works in the mid-1990s (Saad and Solla, 1995a;b; Biehl and Schwarze, 1995; Riegler and Biehl, 1995). This area has experienced a resurgence in recent years (Tan and Vershynin, 2023; Goldt et al., 2019; Veiga et al., 2022; Arnaboldi et al., 2023a;b; Berthier et al., 2023; Ben Arous et al., 2021; Paquette et al., 2022; Collins-Woodfin et al., 2023; Martin et al., 2024).

Many theoretical efforts highlighted the class of functions that are efficiently learned by two layer neural networks. In the context of single-index targets, (Ben Arous et al., 2021) introduces the notion of information exponent to quantify the hardness of the learning task. Similarly, for multi-index models, (Abbe et al., 2022; 2023), building on their earlier work (Abbe et al., 2021), demonstrated how the leap complexity of target functions dictates the amount of training samples needed from two-layer networks in the mean-field limit to learn the target. Note that (Abbe et al., 2023) also considered the case of $n_b \lesssim O(d)$. A large number of theoretical studies devoted to the understanding of the feature learning regime in two-layer networks often assume an asymptotically vanishing initialization for the second layer weights $\boldsymbol{a}_0$ in eq. (1), see e.g. (Abbe et al., 2022; Berthier et al., 2023; Abbe et al., 2023). Although this assumption is amenable for theoretical characterizations, our analysis provably shows that a careful reasoning on the second layer magnitude is needed to offer a complete portrait of the learning dynamics of SGD. More precisely, we describe a sharp divergence when the batch size $n_b \gg d^{\ell/2}$ between the dynamics of SGD when optimizing the MSE loss (vanilla SGD), in contrast to the correlation loss $\tilde{\ell} = \frac{1}{n_b}\sum_{\nu\in[n_b]} 1 - y^\nu f(\boldsymbol{z}^\nu)$ (Correlation loss SGD). The latter training protocol is indeed equivalent to consider an asymptotically vanishing second layer weights $\boldsymbol{a}_0$ at initialization in the optimization routine, e.g. see (Damian et al., 2024).

Closer to us, the analysis of the *first* gradient descent step with large n_b has been discussed in detail in recent papers (Ba et al., 2022; Damian et al., 2022; Dandi et al., 2023). (Ba et al., 2022) showed that a single large learning rate gradient step allows to beat kernel methods when the number of training samples is proportional to the input dimension . While their results are limited to single-index target and to a single gradient step, (Damian et al., 2022) further showed that with $n = \omega(d^2)$ samples, two-layer nets can learn multi-index target function with zero first Hermite coefficient (ℓ=2). (Dandi et al., 2023) extended their conditions on the sample complexity to general $\ell \geq 1$, showed this complexity is optimal for single-step learning, and extended the results to higher information exponents. Although motivated from different objectives, (Sclocchi and Wyart, 2024) heuristically sketch a phase diagram for the performance of SGD on realistic datasets as a function of the algorithm's relevant parameters, i.e. batch size and learning rate.

A common assumption in theoretical studies is to consider sample-splitting schemes for the training protocol. At each iteration, the optimization algorithm is ran using a fresh batch of observations of the model, drawn independently of past iterations; this routine has been used extensively in the analysis of iterative algorithms (see e.g. (Chandrasekher et al., 2021; Jain et al., 2013; Hardt and Wootters, 2014; Jain and Netrapalli, 2015; Kwon et al., 2019)).

3. Time / Complexity Tradeoffs

In this section, we characterize the intertwined dependence between the batch size and the hardness of the learning task in determining the number of one-pass SGD iterations needed to achieve weak recovery of the teacher subspace as in Definition 2.2. We offer a detailed picture of the tradeoffs to consider in order to minimize the training iteration time T, compactly illustrated in the phase diagram in Fig. 1.

Network initialization — We consider random initialization for the hidden layer weights of the network (1), while the second layer weights are kept fixed:

$$\boldsymbol{w}_{j,0} \sim \mathrm{Unif}(\mathbb{S}^{d-1}), \quad a_{j,0} = 1 \qquad j \in [p]. \tag{10}$$

We will refer to this situation as *cold start*, since the initial network correlation with the target directions is vanishing when $d \to +\infty$.

Generalized Linear Models — The seminal work of (Ben Arous et al., 2021) studies the weak recovery problem for Generalized Linear Models (GLMs), i.e. $p = 1$, when learning single-index targets ($k = 1$). Starting from randomly initialized networks as defined in (1), the time iterations needed for one-pass SGD (with one sample per batch) to achieve weak recovery of the target direction respects:

$$I(\ell) = \begin{cases} O(d^{\ell-1}) & \text{if } \ell > 2 \\ O(d\log d) & \text{if } \ell = 2 \\ O(d) & \text{if } \ell = 1 \end{cases} \tag{11}$$

where ℓ is the information exponent of the target $f_\star$.
As far as weak recovery of the target subspace is concerned, the characterization of multi-index targets follows the same lines of thought just replacing the information exponent by the leap index of the target, e.g. see Definition **3** of (Dandi et al., 2023) or Definition **1** of (Abbe et al., 2023). Similarly to the information exponent definition, the leap index is the lowest rank of the tensors appearing in the Hermite expansion of the target $f^\star$. Therefore, we choose to study in the following the training dynamics for the $p = k = 1$

scenario for general batch sizes n_b. This assumption is useful to provide rigorous guarantees as it largely reduces the complexity of the projected SGD dynamics. However, we argue (supported by numerical simulations in Appendix E.1) that the same phenomenology will hold for larger values of p and k.

3.1. Weak recovery with one-pass SGD

Consider the gradient descent dynamics defined on the hidden layer weights by eq. (4). We focus on the description of the time evolution of the correlation between the network's hidden layer weight and the target direction:

$$m_t = \langle \boldsymbol{w}_t, \boldsymbol{w}^\star \rangle \tag{12}$$

Our first main result is to characterize the time to achieve weak recovery of the target direction $\boldsymbol{w}^\star$ as a function of the batch size and the information exponent of the target. We make very weak assumptions on the activation and labeling functions, namely only assuming a sub-polynomial growth:

Assumption 3.1 (Polynomial growth)**.** The activation function σ is differentiable everywhere, except maybe at a finite set of points. Both σ' and $f^\star$ are sub-polynomial, i.e. there exists a $k > 0$ and a constant C such that for any $x \in \mathbb{R}$

$$|\sigma'(x)| \leq C(1+x)^k \quad \text{and} \quad |f^\star(x)| \leq C(1+x)^k \tag{13}$$

Assumption 3.2 (Well-posedness)**.** Let $(c_k)_{k\geq 0}$ and $(c_k^\star)_{k\geq 0}$ be the Hermite coefficients of σ and $h^\star$, respectively. Then $c_\ell \neq 0$, and if ℓ is even, then $c_\ell c_\ell^\star > 0$.

Assumption 3.3 (Initialization)**.** There exists a $\kappa > 0$ such that $m_0 > \kappa/\sqrt{d}$. Further, if ℓ is odd, then m_0 is such that

$$c_\ell c_\ell^\star m_0 > 0$$

Assumption 2 ensures that the optimization problem is achievable for gradient flow on the population loss $\mathcal{R}$. Indeed, one can show that when $m \approx 0$,

$$\mathcal{R} = 2(1 - c_\ell c_\ell^\star m^\ell) + o(m^\ell);$$

as a result, if ℓ is even and $c_\ell c_\ell^\star < 0$, then $m = 0$ is a local maximum of $\mathcal{R}$ and weak recovery is impossible. When ℓ is odd, the point $m = 0$ is always a strict saddle, so Assumption 3.3 that we start on the correct side of the saddle. Under the initialization scheme described by Equation 10, the first condition is satisfied with arbitrarily high probability upon decreasing κ, while the second is a $1/2$-probability event.

We are now in the position to formally state the result:

Theorem 3.4 (Projected SGD weak recovery)**.** *Consider the projected SGD algorithm with square loss (Eqs. (4), (5)), and suppose that Assumptions 3.1-3.3 hold. There exist absolute constants c_γ, C_γ such that if*

$$\gamma \leq c_\gamma \min\left(1, n_b d^{-\left(\frac{\ell}{2} \vee 1\right)} \log(d)^{-C_\gamma}\right),$$

then for large enough d we have with probability $1 - ce^{-c\log(n)^2}$

$$t_\eta^+ \leq C\gamma^{-1} d^{\left(\frac{\ell}{2}-1\right)\vee 0} \log(d). \tag{14}$$

3.2. Illustration of Theorem 3.4

The phase diagram in Fig. 1 exemplifies Theorem 3.4. We identify three *learning phases*: SGD learning, Correlation Loss SGD learning, and SGD impossible. These regions are explored by varying the batch size and learning rate exponents δ, μ. Our theory characterizes the optimal learning rate to achieve the lowest possible time iterations of SGD to weakly recover the target direction $\boldsymbol{w}^\star$ when the batch size respects $n_b = o(d^{\ell-1})$:

$$\delta^\star(\mu) = \begin{cases} \frac{\ell}{2} - \mu & \text{if } \mu < \ell/2 \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

Weak recovery region — In the region $n_b \lesssim d^{\ell/2}$ there is a net benefit in using larger batch sizes in the SGD optimization. This section shows a similar phenomenology to (Ben Arous et al., 2021): if we optimally choose the learning rate exponent $\delta^\star(\mu)$, the number of time iterations needed to weakly recover the teacher direction $\boldsymbol{w}^\star$ is simply $T(n_b) = I(\ell)/n_b$, rescaling straightforwardly the time complexity of the $n_b = 1$ case in eq. (11). By considering higher values for the learning rate ($\delta < \delta^\star(\mu)$) SGD is not able to weakly recover the signal as the dynamics is dominated by terms contracting the network / target correlation to zero, defining the SGD impossible region. Vice versa, if one takes into account lower learning rates ($\delta > \delta^\star(\mu)$), it is certainly possible to weakly-recover the target, but at a higher time complexity cost.

Self-interaction regime — Conversely, the region $d^{\ell/2} \ll n_b \lesssim d^{\max(\ell-1,1)}$ does not adhere to the same straightforward paradigm. Indeed, standard SGD is not able to achieve weak recovery of the teacher direction using $T(n_b) = I(\ell)/n_b$ time iterations, but a simple modification of it - that we call *Correlation Loss SGD* - is able to. The number of steps needed to weakly recover the target with this new training protocol can then be pushed down to $T = \mathrm{polylog}(d)$ when using batch sizes $n_b = O(d^{\max(\ell-1,1)})$. We refer to the next section for a detailed analysis of this regime.

One step regime — Recent works have discussed the role of one large learning rate gradient descent step (*giant step*) when training of two-layer networks (Ba et al.,

	SGD with $n_b \lesssim d^{\ell/2}$	SGD with $d^{\ell/2} \ll n_b \lesssim d^{\max(\ell-1,1)}$	Correlation loss SGD with $n_b = o\left(d^{\max(\ell-1,1)}\right)$	One step with $n_b = O(d^\ell)$
$\ell=1$	$T=O(d/n_b), N=O(d)$	$T=O(1), N=O(d)$	$T=O(d/n_b), N=O(d)$	$T=1, N=O(d)$
$\ell=2$	$T=O(d\log d/n_b), N=O(d\log d)$	$T=O(\log d), N=O(d\log d)$	$T=O(d\log d/n_b), N=O(d\log d)$	$T=1, N=O(d^2)$
$\ell>2$	$T=O(d^{\ell-1}/n_b), N=O(d^{\ell-1})$	$T=O(d^{\ell/2-1}), N=O(n_b d^{\ell/2-1})$	$T=O(d^{\ell-1}/n_b), N=O(d^{\ell-1})$	$T=1, N=O(d^\ell)$

Table 1. **Time / Complexity tradeoffs:** Number of iterations T and the total number of samples N needed to achieve weak recovery of the target for different training protocols in high dimensions. **Left:** One-pass SGD of batch size $n_b = d^{\ell/2}$, in this regime the optimal time complexity is obtained rescaling by n_b the result of (Ben Arous et al., 2021) for $n_b = 1$, i.e. by choosing the optimal learning rate $\gamma = O(n_b d^{-\ell/2})$. **Center-left:** One-pass SGD with batch size $d^{\ell/2} \ll n_b \lesssim d^{\max(\ell-1,1)}$, for hard problems ($\ell > 2$) the sample complexity is significantly increased with respect to the $n_b = 1$ case up to $N = O(n_b d^{\ell/2-1})$. The learning rate cannot be increased proportionally to n_b in this region, fixed to be $\gamma = O(1)$. **Center-Right:** Correlation loss SGD with $n_b = o(d^{\max(\ell-1,1)})$, this training protocol overcomes the limitation of SGD when $n_b \gg d^{\ell/2}$ and $\ell > 2$; the total sample complexity is $N = O(d^{\ell-1})$. The learning rate is fixed again to be proportional to the batch size $\gamma = O(n_b d^{-\ell/2})$. **Right:** The target is weakly recovered with one GD step of $n_b = O(d^\ell)$ batch. The learning rate is chosen as $\gamma = O(d^{(\ell\,-\,1)/2})$ for the One Step routine (Dandi et al., 2023).

2022; Damian et al., 2022; Dandi et al., 2023). More precisely, (Dandi et al., 2023) sharply characterizes the section $n_b = \Omega(d^\ell)$ where it is possible to learn the teacher direction in just one step by setting the learning rate to $\delta_{\text{giant-step}}(\mu) = \frac{1-\ell}{2}$.

3.3. The self-interaction regime

Surprisingly, when the learning rate becomes extensive ($\gamma = \omega(1)$), the usual SGD algorithm struggles to achieve weak recovery. This can be explained by writing the gradient update as

$$\boldsymbol{w}_t + \gamma \boldsymbol{g}_t = (1 - \gamma\langle \boldsymbol{g}_t, \boldsymbol{w}_t\rangle)\boldsymbol{w}_t + \boldsymbol{g}_t^\perp,$$

where $\boldsymbol{g}_t, \boldsymbol{g}_t^\perp$ are the gradient at time t and its component orthogonal to $\boldsymbol{w}_t$, respectively. As a result, projected gradient descent can be seen as a version of spherical SGD with a random weight decay $\gamma\langle \boldsymbol{g}_t, \boldsymbol{w}_t\rangle$. When $\gamma = \omega(1)$, this weight decay also becomes of order $\omega(1)$, which leads to very unpredictable behavior of the process $(\boldsymbol{w}_t)_{t\geq 0}$.

In this section, we study a modified version for the training protocol, in which the self-interaction term $\langle \boldsymbol{g}_t, \boldsymbol{w}_t\rangle$ is much smaller; we will refer to this new algorithm as *Correlation loss SGD* (see e.g. (Damian et al., 2024)), as it effectively amounts to gradient updates on the correlation loss:

$$\tilde{\ell} = \frac{1}{n_b} \sum_{\nu\in[n_b]} 1 - y^\nu f(\boldsymbol{z}^\nu) \tag{16}$$

The above-described protocol is equivalent to consider a vanishing initialization scale for the second layer weights $\boldsymbol{a}_0$ of the network 1. Such assumptions are often considered in different theoretical efforts (see e.g. (Abbe et al., 2022; Berthier et al., 2023; Abbe et al., 2023)). However, Fig. 1 illustrates that a careful analysis of the initialization scale $\boldsymbol{a}_0$ is needed to paint an exhaustive description of the SGD dynamics. Indeed, considering Correlation loss SGD allows to overcome the limitations highlighted by Theorem 3.4 for projected SGD. In particular, Correlation loss SGD is able to access the yellow region depicted in Fig. 1 where the time complexity can be reduced again to $\tilde{T}(n_b) = {}^{I(\ell)}/_{n_b}$ using the optimal learning rate $\tilde{\delta}^\star(\mu) = {}^{\ell}/_{2} - \mu$ even for $\mu > {}^{\ell}/_{2}$. This is precisely stated in the following theorem.

Theorem 3.5 (*Correlation Loss SGD* weak recovery)**.** *Consider the projected SGD algorithm with correlation loss (eqs.* (4)*,* (16)*), and suppose that Assumptions 3.1-3.3 hold. There exists absolute constants c_γ, C_γ such that if*

$$\gamma \leq c_\gamma \log(d)^{-C_\gamma} \min\left(n_b d^{-\left(\frac{\ell}{2}\vee 1\right)}, \sqrt{\frac{n_b}{d}}\right)$$

Then if d is large enough, we have with probability $1 - ce^{-c\log(n)^2}$

$$t_\eta^+ \leq C\max\left(1, \gamma^{-1} d^{\left(\frac{\ell}{2}-1\right)\vee 0}\log(d)\right). \tag{17}$$

The derivation of Theorems 3.4 and 3.5 generalizes (Ben Arous et al., 2021) which studies the $n_b = 1$ case. Informally, the result is obtained by analyzing the stability of the equation for the correlation m_t, along with the requirement on the step-size for the suppression of the effects of the noise across time. However, there is a major difficulty introduced by the large stepsize regime: when the gradient updates become larger, the Taylor-inspired bounds used in (Ben Arous et al., 2021) become vacuous. We work around this problem by showing that in this regime, there is a *one-step improvement* which jumps directly to meaningful correlation with the target vector. All details can be found in App. A. We provide in Table 1 a representative summary of the results in Thms (3.4. 3.5) characterizing the time/complexity tradeoffs to achieve weak recovery of general single index target $f^\star$.

The theoretical predictions of Thm. 3.5 are evaluated in Fig. 2. The plot compares the student-teacher weight correlation ($m_t = \langle \boldsymbol{w}_t, \boldsymbol{w}^\star \rangle$) achieved by vanilla projected SGD and *Correlation Loss SGD* as a function of time. The teacher activation $h^\star$ is fixed to be the third Hermite polynomial ($\ell = 3$), and the batch size varies, effectively changing the region of the phase diagram considered. In agreement with Theorem 3.5, the figure shows that *Correlation Loss SGD* is always able to achieve faster weak recovery with respect to SGD. Furthermore, the batch size that can be used with *Correlation Loss SGD* in combination with the optimal learning rate $\tilde{\delta}^\star(\mu) = \ell/2 - \mu$ is larger, as presented in the phase diagram of Figure 1.

Remark 3.6. Theorem 3.5 does not claim superiority of *Correlation Loss SGD* with respect to plain SGD when trying to fully learn the target, but only for achieving weak-correlation faster (Definition 2.2). As Figure 2 shows, *Correlation Loss SGD* escapes the initial dynamical plateau faster, but is then limited by a loss function not designed properly to reach the global minimum. In Appendix E we investigate the possibility to combine both the algorithms sketched in Fig. 2, namely escaping the initialization plateau with Correlation loss SGD and then learn the function with SGD; we refer to this protocol as *Adaptive SGD*. Moreover, as Fig. 1 and Table 1 illustrate, the benefits of using Correlation loss SGD are limited to settings in which $\ell > 2$. Indeed, the Self-interaction regime (depicted in yellow in Fig. 1) is not present for $\ell \leq 2$.

4. Exact Asymptotic Description

We now characterize the exact asymptotic description of the dynamics of two-layer networks trained with SGD. In Fig. 3 we sketch a representative phase diagram as a function of the relevant parameter of the algorithm, i.e. the learning rate and the batch size. The plot identifies different regions of parameters defining the network's learning efficiency.

Sufficient statistics — Our study, like many other efforts (Ben Arous et al., 2022; Saad and Solla, 1995a), is based on the concentration of the neurons' overlaps with the target subspace and their norms. This approach only requires the knowledge for every optimization step $t \in [T]$ of the above defined overlaps, often referred to as *sufficient statistics*. Let the pre-activations be defined as:

$$\boldsymbol{\lambda}_t = W_t \boldsymbol{z} \qquad \text{and} \qquad \boldsymbol{\lambda}^\star = W^\star \boldsymbol{z} \tag{18}$$

Thanks to the Gaussian nature of the data, the pre-activations at any time step t are jointly Gaussian vectors $(\boldsymbol{\lambda}_t, \boldsymbol{\lambda}^\star) \sim \mathcal{N}(\mathbf{0}_{p+k}, \Omega_t)$ with covariance $\Omega_t \in \mathbb{R}^{(p+k)\times(p+k)}$:

$$\Omega_t := \begin{pmatrix} Q_t & M_t \\ M_t^\top & P \end{pmatrix} = \begin{pmatrix} W_t W_t^\top & W_t {W^\star}^\top \\ W^\star W_t^\top & W^\star {W^\star}^\top \end{pmatrix} \tag{19}$$

We refer to M_t, Q_t as the *order parameters*.

4.1. Closed form equations

We are now in the position to state our proposition that provides a set of deterministic ODEs to describe one-pass SGD in high-dimensions. This portrayal depends ultimately only on the values of the values of the learning rate and the batch size, as quantified by the exponents (δ, μ).

Proposition 4.1. *Consider $\bar{\Omega}(t)$ the solution of the system of ordinary differential equations*

$$\begin{aligned} \frac{\mathrm{d}M_{jr}}{\mathrm{d}\tau} &= \Psi_{jr}(\Omega) - \frac{M_{jr}}{2}\Phi_{jj}(\Omega) \\ \frac{\mathrm{d}Q_{jl}}{\mathrm{d}\tau} &= \Phi_{jl}(\Omega) - \frac{Q_{jl}}{2}\left(\Phi_{jj}(\Omega) + \Phi_{ll}(\Omega)\right) \end{aligned} \tag{20}$$

where we introduced:

$$\begin{aligned} \Psi_{jr}(\Omega) &= \mathbf{1}_{\{\delta \geq 0 \bigcap 2\delta+\mu \geq 1\}} \frac{\gamma_0}{p} a_j \psi_{jr} \\ \Phi_{jl}(\Omega) &= \mathbf{1}_{\{\delta \geq 0 \bigcap 2\delta+\mu \geq 1\}} \frac{\gamma_0}{p} \left(a_j^t \phi_{jl}^{\mathrm{GF}} + a_l^t \phi_{lj}^{\mathrm{GF}}\right) \\ &\quad + \mathbf{1}_{\{\delta+\mu \geq 1 \bigcap 2\delta+\mu \leq 1\}} \frac{\gamma_0^2}{p^2 n_0} a_j^t a_l^t \phi_{jl}^{\mathrm{HD}} \end{aligned} \tag{21}$$

and auxiliary integrals bearing expectations over $\mathcal{N}(\mathbf{0}, \Omega)$:

$$\begin{aligned} \psi_{jr} &= \mathbb{E}\left[\sigma'(\lambda_j)\lambda_r^\star \mathcal{E}\right] \\ \phi_{jl}^{\mathrm{GF}} &= \mathbb{E}\left[\sigma'(\lambda_j)\lambda_l \mathcal{E}\right] \\ \phi_{jl}^{\mathrm{HD}} &= \mathbb{E}\left[\sigma'(\lambda_j)\sigma'(\lambda_l)\mathcal{E}^2\right] \\ \mathcal{E} &= g_\star(\boldsymbol{\lambda}^\star) - \frac{1}{p}\sum_{j=1}^{p} a_j \sigma(\boldsymbol{\lambda}) \end{aligned}$$

Then, there exists a constant C independent from the input data dimension, such that the discrete stochastic process for the covariance $\{\Omega_t\}_{t\in\mathbb{N}}$ in eq. (19) induced by projected SGD dynamics is approximated by the deterministic covariance matrix $\bar{\Omega}(t)$ with precision:

$$\mathbb{E}\left\|\Omega_t - \bar{\Omega}(t\Delta\tau)\right\| \leq e^{Ct}\sqrt{\Delta\tau} \tag{22}$$

with $\Delta\tau = d^{\max(-\delta, -2\delta+1-\mu)}$.

We refer to Appendix B for the informal derivation of the above result.

In Fig. 3 (left) we summarize the results of Prop. 4.1 in a compact phase diagram. The following dynamical regimes appear:

- **Population Flow**: The dynamics of the sufficient statistics described by a deterministic set of ODEs (20) is equivalent to population gradient flow.
- **Noise learning**: The dynamic is dominated by high-dimensional noise, and consequently the algorithm does not learn the target; the behavior is reflected in the ODEs.

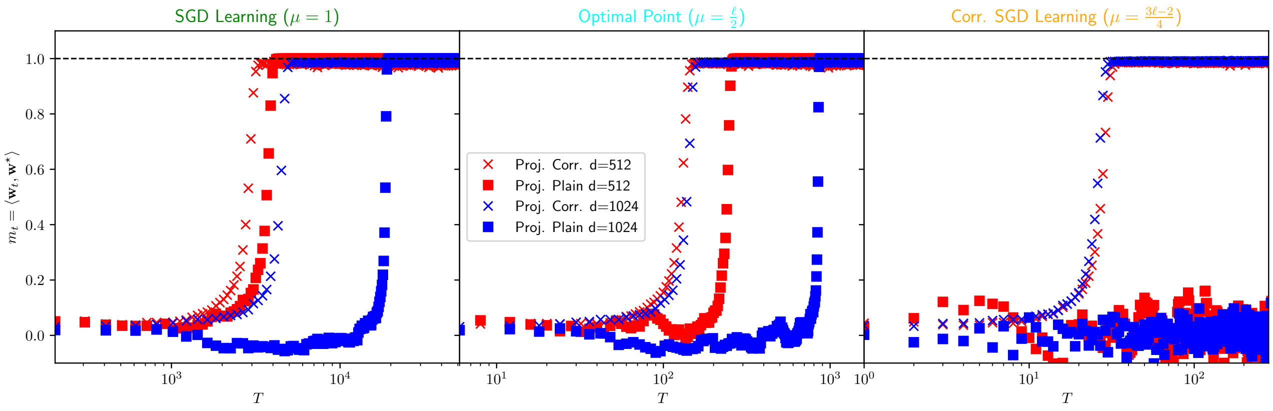

Figure 2. **Correlation Loss SGD weak recovery:** Comparison between the performance of plain SGD and the Correlation Loss SGD, in different regions of the phase diagram, and for different sizes d. The plot shows the test error as a function of the optimization steps. Both the teacher and the student activation functions are fixed to $\sigma = h^\star = \mathrm{He}_3$, so the information exponent is $\ell = 3$. In all the three plots we vary the value of μ, while $\delta = \mu - \ell/2$. Theorem 3.5 predicts that the *Correlation Loss SGD* weakly recovers the target direction while SGD fails when $\delta < 0$, in accordance to what is shown in the plot. Note that the numbers of steps needed for the target recovery drastically decrease when μ becomes large in accordance with Theorems 3.4,3.5.

- **Saad&Solla line**: The ODE description in Prop. 4.1 is equivalent to the pivotal work on 2LNNs (Saad and Solla, 1995a). In particular, the original work corresponds to the point $(\delta, \mu) = (1, 0)$. The learning dynamics is blocked on a plateau characterized by the noise variance in the labels.

- **Dynamics not defined**: For a broad range of values of (δ, μ) the SGD dynamics is not effectively described by a set of low-dimensional deterministic ODEs.

In the right panel of Figure 3 we present a numerical investigation of three particular instances of the regimes presented above. The plot shows a comparison of numerical simulations versus the low-dimensional exact asymptotic characterization given in eqs. (20). The values of the learning rate and the batch size used for SGD training are varied to probe different regions of phase diagram 3.

Remark 4.2. When the target's leap index is $\ell > 1$, the dynamic of SGD is dominated by a first extensive search phase to achieve weak recovery of the teacher direction (Thm. 3.4). Therefore, in order to probe interesting dynamical regimes for general single index teachers, we assume non-vanishing initial correlations of the network's hidden layer weights with the teacher's ones when $d \to +\infty$. In App. E we study the tightness of the exponential bound eq. (22); we argue supported by numerical illustrations that (on the practical side) the low dimensional ODE description is valid well beyond the extents of Prop. 4.1, as already observed by other works (Goldt et al., 2019; Veiga et al., 2022) in different context.

Non asymptotic corrections — Proposition 4.1 unveils a surprising result for the exact asymptotic description of two-layer networks. Indeed, the ODEs written in eq. (20) coincide with the analogous ones for the single sample per batch case ($n_b = 1$), modulo trivial rescaling of the parameters (Veiga et al., 2022). However, a careful consideration of the “intra-batch correlations” in the gradient is needed for correctly describing the low-dimensional process of the order parameters:

$$\sum_{\nu'=1,\nu'\neq\nu}^{n_b} \sigma'(\lambda_j^\nu)\sigma'(\lambda_l^{\nu'})\mathcal{E}^\nu\mathcal{E}^{\nu'}\langle \boldsymbol{z}^\nu, \boldsymbol{z}^{\nu'} \rangle \tag{23}$$

The asymptotic form of this term can be exactly computed to be (using Prop. 4.1 notations):

$$\phi_{jl}^{\mathrm{BC}} = \mathbb{E}\left[\sigma'(\lambda_j)\mathcal{E}\left(\lambda^\star\right)^\top\right] P^{-1}\mathbb{E}\left[\sigma'(\lambda_l)\mathcal{E}\lambda^\star\right] + \tag{24}$$

$$\mathbb{E}\left[\sigma'(\lambda_j)\mathcal{E}(\lambda^\perp)^\top\right]\left(Q^\perp\right)^{-1}\mathbb{E}\left[\sigma'(\lambda_l)\mathcal{E}\lambda^\perp\right] \tag{25}$$

with $\boldsymbol{\lambda}^\perp = \boldsymbol{\lambda} - MP^{-1}\boldsymbol{\lambda}^\star$ and $Q^\perp = Q - MP^{-1}M^\top$. Although the contribution of the above term is asymptotically vanishing in the ODE description (20) when $d \to \infty$, any theoretical description at finite d will effectively depend on ϕ_{jl}^{BC}. In App. E we provide additional numerical investigation on the importance of (24) and the role of large batch sizes for non-asymptotic corrections to the characterization in Prop. 4.1. Moreover, we note that taking into account the presence of large batch size is pivotal to illustrate the time / complexity tradeoffs for weak recovery of the target subspace, as thoroughly discussed in Section 3.

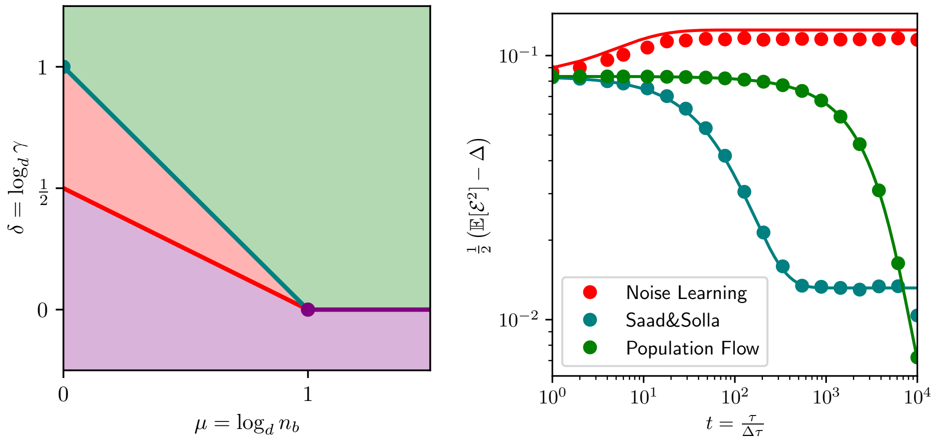

Figure 3. **Exact asymptotic description:** Exact asymptotic characterization of the dynamics of two-layer networks trained with SGD as a function of the batch size (n_b) and the learning rate (γ). **Left:** Illustration of the different dynamical regimes in a compact phase diagram. **Population flow region:** The dynamics is equivalent to population gradient flow. **Noise learning region:** The high-dimensional noise terms dominate the dynamics. **Saad&Solla line:** The learning dynamics attains a plateau characterized by the noise variance (Saad and Solla, 1995a). **Dynamics not defined:** The deterministic low-dimensional description of the eq. (20) is not valid. **Right:** The figure shows a comparison of numerical simulations (dots) and theoretical prediction (continuous lines) for three instances (δ, μ) associated with different learning regimes (identified by the corresponding colors). For both theory and simulations, the test error is plotted as a function of SGD iterations. We consider a matching architectures problem, i.e. $h^\star = \sigma = \text{erf}$ activation, and hidden units $p = 2, k = 2$.

5. Conclusions

In this manuscript, we have explored the intricate relationship between batch size and the efficiency of learning multi-index targets using one-pass SGD on high-dimensional input data. Our findings defies the conventional belief that larger batch sizes invariably lead to better results and reveals a critical batch size threshold, beyond which the advantages of larger batches wane in terms of computational complexity. Applying gradient updates on the correlation loss one may, however, navigate this limitation. Finally, we also provide a system of low-dimensional ODE to describe the exact asymptotic of the SGD dynamics with arbitrary batch-sizes. Moving forward, we hope this research paves the way for deeper inquiries into the optimization behaviors of learning algorithms, prompting further examination of deeper networks and alternative loss functions.

Acknowledgement

This work was supported by the Swiss National Science Foundation under grant OperaGOST (grant number 200390) & the Choose France - CNRS AI Rising Talents program.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *The Journal of Machine Learning Research*, 22(1):4788–4838, 2021.

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.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75(9):1889–1935, 2022. doi: https://doi.org/10.1002/cpa.22074.

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.

Emmanuel Abbe, Enric Boix-Adsera, Matthew S Brennan, Guy Bresler, and Dheeraj Nagaraj. The staircase property: How hierarchical structure can guide deep learning.

Advances in Neural Information Processing Systems, 34: 26989–27002, 2021.

Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2552–2623. PMLR, 2023.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. 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 5413–5452. PMLR, 02–05 Jul 2022.

Alex Damian, Eshaan Nichani, Rong Ge, and Jason D Lee. Smoothing the landscape boosts the signal for sgd: Optimal sample complexity for learning single index models. *Advances in Neural Information Processing Systems*, 36, 2024.

Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. How two-layer neural networks learn, one (giant) step at a time. *arXiv preprint arXiv:2305.18270*, 2023.

Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning gaussian multi-index models with gradient flow. *arXiv preprint arXiv:2310.19793*, 2023.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, and Denny Wu. Learning in the presence of low-dimensional structure: a spiked random matrix perspective. *Advances in Neural Information Processing Systems*, 36, 2024.

Behrad Moniri, Donghwan Lee, Hamed Hassani, and Edgar Dobriban. A theory of non-linear feature learning with one gradient step in two-layer neural networks. *arXiv preprint arXiv:2310.07891*, 2023.

Alireza Mousavi-Hosseini, Denny Wu, Taiji Suzuki, and Murat A Erdogdu. Gradient-based feature learning under structured data. *Advances in Neural Information Processing Systems*, 36:71449–71485, 2023.

Aaron Zweig and Joan Bruna. Symmetric single index learning. *arXiv preprint arXiv:2310.02117*, 2023.

David Saad and Sara A. Solla. On-line learning in soft committee machines. *Physical Review E*, 52(4):4225–4243, October 1995a. doi: 10.1103/PhysRevE.52.4225.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for sgd: Effective dynamics and critical scaling. *Advances in Neural Information Processing Systems*, 35:25349–25362, 2022.

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 *Conference on Learning Theory*, pages 4782–4887. PMLR, 2022.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Li Li, Yuxi Fan, Mike Tse, and Kuo-Yi Lin. A review of applications in federated learning. *Computers & Industrial Engineering*, 149:106854, 2020.

David Saad and Sara Solla. Dynamics of on-line gradient descent learning for multilayer neural networks. *Advances in neural information processing systems*, 8, 1995b.

M Biehl and H Schwarze. Learning by on-line gradient descent. *Journal of Physics A: Mathematical and General*, 28(3):643, feb 1995. doi: 10.1088/0305-4470/28/3/018. URL https://dx.doi.org/10.1088/0305-4470/28/3/018.

P Riegler and M Biehl. On-line backpropagation in two-layered neural networks. *Journal of Physics A: Mathematical and General*, 28(20):L507, oct 1995. doi: 10.1088/0305-4470/28/20/002. URL https://dx.doi.org/10.1088/0305-4470/28/20/002.

Yan Shuo Tan and Roman Vershynin. Online stochastic gradient descent with arbitrary initialization solves non-smooth, non-convex phase retrieval. *Journal of Machine Learning Research*, 24(58):1–47, 2023.

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. *Advances in neural information processing systems*, 32, 2019.

Rodrigo Veiga, Ludovic Stephan, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 23244–23255. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/939bb847ebfd14c6e4d3b5705e562054-Paper-Conference.pdf.

Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, and Bruno Loureiro. From high-dimensional & mean-field dynamics to dimensionless odes: A unifying approach to sgd in two-layers networks. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 1199–1227. PMLR, 12–15 Jul 2023a. URL https://proceedings.mlr.press/v195/arnaboldi23a.html.

Luca Arnaboldi, Florent Krzakala, Bruno Loureiro, and Ludovic Stephan. Escaping mediocrity: how two-layer networks learn hard single-index models with sgd. *arXiv preprint arXiv:2305.18502*, 2023b.

Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks. *arXiv preprint arXiv:2303.00055*, 2023.

Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Homogenization of sgd in high-dimensions: Exact dynamics and generalization properties. *arXiv preprint arXiv:2205.07069*, 2022.

Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi. Hitting the high-dimensional notes: An ode for sgd learning dynamics on glms and multi-index models. *arXiv preprint arXiv:2308.08977*, 2023.

Simon Martin, Francis Bach, and Giulio Biroli. On the impact of overparameterization on the training of a shallow neural network in high dimensions. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3655–3663. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/martin24a.html.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 37932–37946. Curran Associates, Inc., 2022.

Antonio Sclocchi and Matthieu Wyart. On the different regimes of stochastic gradient descent. *Proceedings of the National Academy of Sciences*, 121(9):e2316301121, 2024. doi: 10.1073/pnas.2316301121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2316301121.

Kabir Aladin Chandrasekher, Ashwin Pananjady, and Christos Thrampoulidis. Sharp global convergence guarantees for iterative nonconvex optimization: A gaussian process perspective. *arXiv preprint arXiv:2109.09859*, 2021.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674, 2013.

Moritz Hardt and Mary Wootters. Fast matrix completion without the condition number. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 638–678, Barcelona, Spain, 13–15 Jun 2014. PMLR. URL https://proceedings.mlr.press/v35/hardt14a.html.

Prateek Jain and Praneeth Netrapalli. Fast exact matrix completion with finite samples. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1007–1034, Paris, France, 03–06 Jul 2015. PMLR. URL https://proceedings.mlr.press/v40/Jain15.html.

Jeongyeol Kwon, Wei Qian, Constantine Caramanis, Yudong Chen, and Damek Davis. Global convergence of the em algorithm for mixtures of two component linear regression. 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 2055–2110. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/kwon19a.html.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, 1991. ISBN 9780387520131. Google-Books-ID: juC1QgAACAAJ.

Aad van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media, March 1996. ISBN 9780387946405. Google-Books-ID: OCenCW9qmp4C.

Chris Junchi Li and Michael Jordan. Stochastic approximation for online tensorial independent component analysis. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3051–3106. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/li21a.html.

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

A. Proof of Theorems 3.4 and 3.5

A.1. Preliminaries

Notations and definitions We denote by $\operatorname{polylog} x$ any polynomial in $\log x$ with degree > 1. Since the case where $n_b = O(1)$ is already covered by the results in (Ben Arous et al., 2021), we shall always assume that $\mu > 0$. The Hermite coefficients of σ and $f^\star$ will be denoted by $(c_k)_{k\geq 0}$ and $(c_k^\star)_{k\geq 0}$, respectively. To break the symmetry between m and $-m$ inherent to the problem, we assume without loss of generality that

$$m_0 > 0 \quad \text{and} \quad c_\ell c_\ell^\star > 0.$$

Throughout this section, the update process on $\boldsymbol{w}_t$ will be written as

$$\boldsymbol{w}_{t+1} = \frac{\boldsymbol{w}_t - \gamma\, \boldsymbol{g}_t}{\|\boldsymbol{w}_t - \gamma\, \boldsymbol{g}_t\|}, \tag{26}$$

where $\boldsymbol{g}_t$ is the gradient at time t: $\boldsymbol{g}_t = \nabla_{\boldsymbol{w}_t} \ell_t$, and ℓ_t is the empirical loss at time t, that can be either the correlation or the square loss. When considering the update of the process $(\boldsymbol{w}_t)_{t\geq 0}$, it will be useful to distinguish between the randomness in $\boldsymbol{w}_t$ and the one introduced by the batch drawn at time t. To this end, we introduce the filtration $(\mathcal{F}_t)_{t\geq 0}$ adapted to the process $\boldsymbol{w}_t$, and we shall denote by $\mathbb{P}_t$ (resp. $\mathbb{E}_t$) the probability (resp. expectation) conditioned on $\mathcal{F}_t$.

Concentration in Orlicz spaces We first recall some fact about Orlicz spaces that will be useful for our concentration bounds.

Definition A.1. For any $\alpha \in \mathbb{R}$, let $\psi_\alpha(x) = e^{x^\alpha} - 1$. Let X be a real random variable; the *Orlicz norm* $\|X\|_{\psi_\alpha}$ is defined as

$$\|X\|_{\psi_\alpha} = \inf\left\{ t > 0 \,:\, \mathbb{E}\left[\psi_\alpha\left(\frac{|X|}{t}\right)\right] \leq 1 \right\} \tag{27}$$

It can be checked that $\|\cdot\|_{\psi_\alpha}$ is a well-defined norm on random variables for $\alpha \geq 1$, and can be slightly modified into a norm when $\alpha < 1$; see (Ledoux and Talagrand, 1991; van der Vaart and Wellner, 1996) for more information. We say that a random variable is sub-gaussian (resp. sub-exponential) if its ψ_2 (resp. ψ_1) norm is finite. The main use of this definition is the following concentration inequality:

Lemma A.2. *Let X be a random variable with finite ψ_α-norm for some $\alpha > 0$. Then*

$$\mathbb{P}\left[|X - \mathbb{E}X| > t\|X\|_{\psi_\alpha}\right] \leq 2e^{-t^\alpha}. \tag{28}$$

As a result, any ψ_α-norm bound yields exponential convergence tails. Orlicz norms are also well-behaved with respect to products:

Lemma A.3. *Let X and Y be two random variables such that $\|X\|_{\psi_\alpha}$ and $\|Y\|_{\psi_\beta}$ are finite for some $\alpha, \beta > 0$. Then*

$$\|XY\|_{\psi_\lambda} \leq \|X\|_{\psi_\alpha} \|Y\|_{\psi_\beta},$$

where λ is the number satisfying $\frac{1}{\alpha} + \frac{1}{\beta} = \frac{1}{\lambda}$.

Proof. Assume without loss of generality that $\|X\|_{\psi_\alpha} = \|Y\|_{\psi_\beta} = 1$. We use the following Young inequality: for any $a, b > 0$, and p, q such that $\frac{1}{p} + \frac{1}{q} = 1$,

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}$$

Applying this inequality to $p = \alpha/\lambda$, $q = \beta/\lambda$, $a = X^\lambda$, $b = Y^\lambda$, we get

$$(XY)^\lambda \leq \frac{\lambda X^\alpha}{\alpha} + \frac{\lambda Y^\beta}{\beta}.$$

Then

$$\exp((XY)^\lambda) \le \exp\left(\frac{\lambda X^\alpha}{\alpha}\right) + \exp\left(\frac{\lambda Y^\beta}{\beta}\right)$$
$$\le \frac{\lambda}{\alpha}\exp(X^\alpha) + \frac{\lambda}{\beta}\exp(Y^\beta),$$

where at the last line we used Young's inequality again with the same p and q. The result ensues from taking expectations on both sides, and noticing that $\lambda/\alpha + \lambda/\beta = 1$ by definition. □

Finally, we shall use the following theorem:

Theorem A.4 (Theorem 6.2.3 in (Ledoux and Talagrand, 1991))**.** *Let $X_1, \dots, X_n$ be n independent random variables with zero mean and second moment $\mathbb{E}X_i^2 = \sigma_i^2$. Then,*

$$\left\| \sum_{i=1}^n X_i \right\|_{\psi_\alpha} \le K_\alpha \log(n)^{1/\alpha} \left(\sqrt{\sum_{i=1}^n \sigma_i^2} + \max_i \|X_i\|_{\psi_\alpha} \right) \tag{29}$$

A.2. Computing the gradient at time t

Throughout this section, the update process on $\boldsymbol{w}_t$ will be written as

$$\boldsymbol{w}_{t+1} = \frac{\boldsymbol{w}_t - \gamma\, \boldsymbol{g}_t}{\|\boldsymbol{w}_t - \gamma\, \boldsymbol{g}_t\|}, \tag{30}$$

where $\boldsymbol{g}_t$ is the gradient at time t: $\boldsymbol{g}_t = \nabla_{\boldsymbol{w}_t}\ell_t$, and ℓ_t is the empirical loss t, that can be either the correlation or the square loss. A direct computation of both gradients implies the following lemma:

Lemma A.5. *Define*

$$\boldsymbol{g}_t^\star = \frac{1}{n_b}\sum_{\nu=1}^{n_b} f^\star(\langle \boldsymbol{w}^\star, \boldsymbol{z}^\nu\rangle)\sigma'(\langle \boldsymbol{w}_t, \boldsymbol{z}^\nu\rangle)\boldsymbol{z}^\nu \quad \textit{and} \quad \hat{\boldsymbol{g}}_t = \frac{1}{n_b}\sum_{\nu=1}^{n_b} \sigma(\langle \boldsymbol{w}_t, \boldsymbol{z}^\nu\rangle)\sigma'(\langle \boldsymbol{w}_t, \boldsymbol{z}^\nu\rangle)\boldsymbol{z}^\nu, \tag{31}$$

Then the gradient of the correlation loss ℓ^{corr} is $-\boldsymbol{g}_t^\star$, while the gradient of the square loss ℓ^{sq} is $\hat{\boldsymbol{g}}_t - \boldsymbol{g}_t^\star$.

Hence, the main difference between the gradients of the correlation and square loss is a so-called *interaction term* $\hat{\boldsymbol{g}}$, that only depends on the learned vector $\boldsymbol{w}_t$. Notice that $\boldsymbol{g}_t^\star$ is an average of n_b independent variables of the form

$$\boldsymbol{g}_t^{\star\nu} := f^\star(\langle \boldsymbol{w}^\star, \boldsymbol{z}^\nu\rangle)\sigma'(\langle \boldsymbol{w}_t, \boldsymbol{z}^\nu\rangle)\boldsymbol{z}^\nu, \tag{32}$$

and we define $\hat{\boldsymbol{g}}_t^\nu$ and $\boldsymbol{g}_t^\nu$ in the same way. By Assumption 3.1 and Lemma A.3, each variable $\boldsymbol{g}_t^{\star\nu}$ (resp. $\hat{\boldsymbol{g}}_t^\nu, \boldsymbol{g}_t^\nu$) has finite ψ_α-norm for some $\alpha > 0$, and hence Proposition 2 of (Dandi et al., 2023) holds up to $\mathrm{polylog}(n)$ factors.

We can also compute the conditional expectation of the gradient $\boldsymbol{g}_t$:

Lemma A.6. *For any $t \ge 0$,*

$$\mathbb{E}_t\left[\boldsymbol{g}_t^\star\right] = \phi(m_t)\boldsymbol{w}_t^\star + \psi^{\mathrm{corr}}(m_t)\boldsymbol{w}_t \tag{33}$$

where $\phi(m_t)$ and ψ^{corr} are two functions with Taylor expansion

$$\phi(m) = \sum_{k=0}^\infty c_{k+1}c_{k+1}^\star m^k \quad \textit{and} \quad \psi^{\mathrm{corr}}(m) = \sum_{k=0}^\infty c_{k+2}c_k^\star m^k. \tag{34}$$

Further, we have

$$\mathbb{E}_t\left[\hat{\boldsymbol{g}}_t\right] = c^{\mathrm{sq}}\boldsymbol{w}_t \quad \textit{with} \quad c^{\mathrm{sq}} = \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[z\sigma(z)\sigma'(z)\right]. \tag{35}$$

Proof. The expectation of $\boldsymbol{g}_t^\star$ is a specialization of Lemma 4 from (Dandi et al., 2023) to $r = 1$. By the independence properties of Gaussians, $\mathbb{E}_t\left[\sigma(\langle \boldsymbol{w}_t, \boldsymbol{z}\rangle)\sigma'(\langle \boldsymbol{w}_t, \boldsymbol{z}\rangle)\langle \boldsymbol{z}, \boldsymbol{w}'\rangle\right] = 0$ as soon as $\boldsymbol{w}'$ is orthogonal to $\boldsymbol{w}_t$, hence the expectation of $\hat{\boldsymbol{g}}_t$ lies along $\boldsymbol{w}_t$, and the second result follows. □

In the following, we will denote $\psi^{\mathrm{sq}}(x) = \psi^{\mathrm{corr}}(x) - c^{\mathrm{sq}}$. A generic ψ will be used when no specialization is necessary, so that

$$\mathbb{E}_t\left[\boldsymbol{g}_t\right] = -\phi(m_t)\boldsymbol{w}^\star - \psi(m_t)\boldsymbol{w}_t. \tag{36}$$

A.3. A differential inequality for m_t

The structure of the proof is similar to the one of (Ben Arous et al., 2021). We define the following stopping times for $\zeta > 0$:

$$t_\zeta^+ = \min\{t \geq 0 : m_t \geq \zeta\}, \qquad\qquad t_\zeta^- = \min\{t \geq 0 : m_t \leq \zeta\}, \tag{37}$$

and the following γ-dependent time:

$$\tilde{t}_{\gamma,\zeta}^+ = \min\{t \geq 0 : \gamma m_t^{\ell-1} \geq \zeta\} \tag{38}$$

Our first goal is to show the following high-probability inequality:

Proposition A.7. *Define*

$$t_{\max} = \frac{n_b}{C_{\max} d \log(d)^{C_{\max}} \gamma^2}, \tag{39}$$

for some sufficiently large $C_{\max} > 0$. Then, for a sufficiently small choice of c_γ:

1. *If we are using the square loss, and $\gamma \leq c_\gamma(n_b d^{-\ell/2} \wedge 1)$, there exists $c, \eta > 0$ such that*

$$\mathbb{P}\left(m_t \geq \frac{3}{4}m_0 + c\gamma\sum_{s=0}^{t-1} m_s^{\ell-1} \quad \forall t \leq t_\eta^+ \wedge t_{\max}\right) \geq 1 - ce^{-c\log(n)^2}. \tag{40}$$

2. *If we are using the correlation loss, and $\gamma \leq c_\gamma n_b d^{-\ell/2}$, there exist $c, \varepsilon > 0$ such that*

$$\mathbb{P}\left(m_t \geq \frac{3}{4}m_0 + c\gamma\sum_{s=0}^{t-1} m_s^{\ell-1} \quad \forall t \leq t_\eta^+ \wedge \tilde{t}_{\gamma,\varepsilon}^+ \wedge t_{\max}\right) \geq 1 - ce^{-c\log(n)^2}. \tag{41}$$

The rest of this section is devoted to show Proposition A.7. We define the following "good" event at time t:

$$\mathcal{E}_t := \left\{|\langle \boldsymbol{g}_t, \boldsymbol{w}_t\rangle| \leq \frac{1}{2\gamma}\right\}$$

A (almost) deterministic update inequality We first expand the projection step to obtain a difference inequality for the process m_t. We write

$$\boldsymbol{g}_t = \langle \boldsymbol{g_t}, \boldsymbol{w_t}\rangle \boldsymbol{w}_t + \boldsymbol{g_t}^\perp, \tag{42}$$

where $\boldsymbol{g}_t^\perp$ is orthogonal to $\boldsymbol{w_t}$. Similarly to Equation 36, we can compute the expectation of $\boldsymbol{g}_t^\perp$:

$$\mathbb{E}_t\left[\boldsymbol{g}_t\right] = -\phi(m_t)(\boldsymbol{w}_t^\star - m_t\boldsymbol{w}_t) \tag{43}$$

Lemma A.8. *For any $t \geq 0$, there exists a (random) constant c_t such that the following inequality holds:*

$$m_{t+1} \geq m_t - \gamma c_t\langle \boldsymbol{w}^\star, \boldsymbol{g}_t^\perp\rangle - \frac{\gamma^2 c_t^2 m_t \left\|\boldsymbol{g}_t^\perp\right\|^2}{2} - \frac{1}{2}\gamma^3 c_t^3 |\langle w^\star, \boldsymbol{g}^\perp\rangle| \left\|\boldsymbol{g}_t^\perp\right\|^2. \tag{44}$$

Further, under the event $\mathcal{E}_t$, we have $1/2 \leq c_t \leq 2$.

Proof. We use the decomposition of eq. (42) and write

$$\boldsymbol{w}_t - \gamma\boldsymbol{g}_t = (1 - \gamma\langle \boldsymbol{g_t}, \boldsymbol{w_t}\rangle)\boldsymbol{w}_t + \gamma\boldsymbol{g_t}^\perp$$

As a result, if we define

$$c_t := \frac{1}{1 - \gamma\langle \boldsymbol{g_t}, \boldsymbol{w_t}\rangle},$$

we have

$$\boldsymbol{w}_{t+1} = \frac{\boldsymbol{w}_t - \gamma c_t \boldsymbol{g}_t^\perp}{\left\|\boldsymbol{w}_t - \gamma c_t \boldsymbol{g}_t^\perp\right\|}$$

since the update equation (26) is invariant w.r.t scaling. Taking the scalar product of the above with $\boldsymbol{w}^\star$, we have

$$m_{t+1} = \frac{m_t - \gamma c_t \langle \boldsymbol{w}^\star, \boldsymbol{g}_t^\perp \rangle}{\left\|\boldsymbol{w}_t - \gamma c_t \boldsymbol{g}_t^\perp\right\|}. \tag{45}$$

Expanding the norm in the denominator, and using that $\|\boldsymbol{w}_t\| = 1$ and $\langle \boldsymbol{w}_t, \boldsymbol{g}_t^\perp \rangle = 0$:

$$\|\boldsymbol{w}_t - \gamma\, \boldsymbol{g}_t\| = \sqrt{1 + \gamma^2 c_t^2 \left\|\boldsymbol{g}_t^\perp\right\|^2}$$

By the convexity inequality $(1+x)^{-1/2} \geq 1 - x/2$, valid for all $x \geq 0$, Equation (45) becomes

$$m_{t+1} \geq \left(m_t - \gamma c_t \langle \boldsymbol{w}^\star, \boldsymbol{g}_t^\perp \rangle\right) \left(1 - \frac{\gamma^2 c_t^2}{2} \left\|\boldsymbol{g}_t^\perp\right\|^2\right).$$

The lemma ensues upon expanding and rearranging the terms. □

The expansion in Lemma A.8 can be decomposed in two terms: the term linear in γ is a noisy *drift term*, that will drive the dynamics, and that we will decompose as a sum of a deterministic process and a martingale. All other terms in γ^2 or γ^3 are corrections that we bound with high probability.

The linear term We first control the term linear in γ. We can write

$$\langle \boldsymbol{w}^\star, \boldsymbol{g}_t^\perp \rangle = \langle \boldsymbol{w}^\star, \mathbb{E}_t \left[\boldsymbol{g}_t^\perp\right] \rangle + Z_t, \tag{46}$$

where $(Z_t)_{t\geq 0}$ is by definition a martingale difference sequence for the filtration $(\mathcal{F}_t)_{t\geq 0}$. The expectation term is straightforward to compute using (43):

$$\langle \boldsymbol{w}^\star, \mathbb{E}_t \left[\boldsymbol{g}_t^\perp\right] \rangle = -(1 - m_t^2)\phi(m_t). \tag{47}$$

The contribution of the terms Z_t is bounded by the following lemma:

Lemma A.9. *There exists constants $c, C > 0$ such that with probability $1 - ce^{-c\log(n)^2}$,*

$$\sup_{1\leq t\leq T} \sum_{s=1}^{t} Z_s \leq \frac{C \log(d)^C \sqrt{T}}{\sqrt{n_b}} \tag{48}$$

Proof. The martingale increment Z_t is an average of n_b independent terms Z_t^ν, that satisfy $\|Z_t^\nu\|_{\psi_\alpha} \leq C$ for some $\alpha, C > 0$ by Assumption 3.1. As a result, if we define

$$B_\alpha = \sup_{t\geq 0} \|Z_t\|_{\psi_\alpha},$$

Theorem A.4 implies that

$$B_\alpha = \frac{\text{polylog}(d)}{\sqrt{n_b}}.$$

We now apply Theorem F.1 in (Li and Jordan, 2021) with $z = \log(d)^{\frac{2(\alpha+2)}{\alpha}} \sqrt{T} B_\alpha$, which yields the exact bound needed. □

Bounding the corrections Our next step is to handle the higher-order corrections. We show the following lemma:

Lemma A.10. *Let $T \geq 0$, and $\eta < 1$. There exists a constant $C > 0$ such that for any $t \leq T$*

$$\mathbb{P}\left(\left\|\boldsymbol{g}_t^\perp\right\|^2 \leq C\left(\phi(m_t)^2(1-m_t)^2 + \frac{d\log(d)^C}{n_b}\right) \quad \forall t \leq t_\eta^+ \wedge T\right) \geq 1 - cTe^{-c\log(d)^2} \tag{49}$$

Proof. Fix some $t \in [T]$. We can write

$$\boldsymbol{g}_t^\perp = a(\boldsymbol{w}^\star - m\boldsymbol{w}) + \frac{1}{n}\sum_{\nu=1}^n f^\star(\langle \boldsymbol{w}^\star, \boldsymbol{z}^\nu\rangle)\sigma'(\langle \boldsymbol{w}_t, \boldsymbol{z}^\nu\rangle)\boldsymbol{z}^{\nu\perp}$$

where each $\boldsymbol{z}^{\nu\perp}$ is orthogonal to both $\boldsymbol{w}$ and $\boldsymbol{w}^\star$. From Lemmas 9 and 11 in (Dandi et al., 2023), with probability $1 - ce^{-c\log(d)^2}$,

$$\frac{1}{n}\sum_{\nu=1}^n f^\star(\langle \boldsymbol{w}^\star, \boldsymbol{z}^\nu\rangle)\sigma'(\langle \boldsymbol{w}_t, \boldsymbol{z}^\nu\rangle)\boldsymbol{z}^{\nu\perp} \leq C\frac{d\log(d)^C}{n_b}.$$

Now, we have

$$\langle \boldsymbol{g}_t^\perp, \boldsymbol{w}_t^\star\rangle^2 = a^2(1-m_t^2)^2 \quad \text{and} \quad \left\|a(\boldsymbol{w}^\star - m\boldsymbol{w})\right\|^2 = a^2(1-m_t^2),$$

hence

$$\left\|\boldsymbol{g}_t^\perp\right\|^2 \leq \frac{1}{1-m_t^2}\langle \boldsymbol{g}_t^\perp, \boldsymbol{w}_t^\star\rangle^2 + C\frac{d\log(d)^C}{n_b}.$$

It remains to notice that

$$\langle \boldsymbol{g}_t^\perp, \boldsymbol{w}_t^\star\rangle^2 = (\langle \boldsymbol{w}^\star, \mathbb{E}_t\left[\boldsymbol{g}_t^\perp\right]\rangle + Z_t)^2 \leq 2(\langle \boldsymbol{w}^\star, \mathbb{E}_t\left[\boldsymbol{g}_t^\perp\right]\rangle^2 + Z_t^2) \leq \phi(m_t)(1-m_t^2)^2 + O\left(\frac{d}{n_b}\right).$$

□

Putting it all together We now combine all the previous bounds into a unique proposition.

Proposition A.11. *Let $T \geq 0$. There exists constants $c, C > 0$ such that*

$$\mathbb{P}\left(m_t \geq m_0 + \sum_{s=0}^{t-1}\Phi_{\text{drift}}(m_s) - C\Phi_{\text{noise}}(m_s) - CK(T) \quad \forall t \leq T \quad \middle| \quad \bigcap_{t\leq T}\mathcal{E}_t\right) \geq 1 - Te^{-c\log(n)^2} \tag{50}$$

where Φ_{drift} and Φ_{noise} are given by

$$\Phi_{\text{drift}}(m) = \gamma(1-m^2)\phi(m), \tag{51}$$

$$\Phi_{\text{noise}}(m) = \gamma^2 m(1-m^2)\phi(m)^2 + \gamma^2 m\frac{d\log(d)^C}{n_b} + \gamma^3(1-m^2)^{3/2}\phi(m)^3, \tag{52}$$

$$K(T) = \frac{\gamma\log(d)^C\sqrt{T}}{\sqrt{n_b}} + \gamma^3 T\frac{d\log(d)^C}{n_b}. \tag{53}$$

Proof. By summing the inequality of Lemma A.8 for $0 \leq s \leq t-1$, we get

$$m_{t+1} \geq m_0 - \sum_{s=0}^{t-1}\gamma c_s\langle \boldsymbol{w}^\star, \boldsymbol{g}_s^\perp\rangle - \frac{\gamma^2 c_s^2 m_s\left\|\boldsymbol{g}_s^\perp\right\|^2}{2} - \frac{1}{2}\gamma^3 c_s^3\left\|\boldsymbol{g}_s^\perp\right\|^3. \tag{54}$$

The linear term is handled using the martingale decomposition (46) combined with the expectation computation of (47) and the bound of Lemma A.9 on the martingale contribution. The terms in γ^2 and γ^3 follow from Lemma A.10, as well as Lemma A.9 with $T = 1$, which implies that

$$|\langle \boldsymbol{w}^\star, \boldsymbol{g}_t^\perp\rangle \leq C\log(n)^C$$

for some $C > 0$. Finally, under the events $\mathcal{E}_s$ for $s \leq t$, we can replace every occurence of c_s by either $1/2$ or 2 depending on the sign of the corresponding term. □

Proof of Proposition A.7 The expressions of Lemma A.6 imply the following expansions near 0:

$$\phi(m) = c_\ell c_\ell^\star m^{\ell-1} + O(m^\ell) \qquad \psi^{\mathrm{corr}}(m) = O(m^\ell) \qquad \psi^{\mathrm{sq}}(m) = -c^{\mathrm{sq}} + O(m^\ell) \tag{55}$$

As a result, there exists an $\eta > 0$ and constants $C, c > 0$ such that for any $m \leq \eta$,

$$cm^{\ell-1} \leq \phi(m) \leq Cm^{\ell-1} \qquad |\psi^{\mathrm{corr}}(m)| \leq Cm^\ell \qquad |\psi^{\mathrm{sq}}(m)| \leq C.$$

We first lower bound the drift inequality of Proposition A.11. Whenever $m_t \geq \eta$, we have

$$\Phi_{\mathrm{drift}}(m_t) \geq p(\gamma m_t^{\ell-1}) - C\gamma^2 \frac{d}{n_b},$$

where $p(x) = x - C(x^2 + x^3)$. Define $\varepsilon > 0$ such that $p(x) > x/2$ on $[0, \varepsilon]$, so when $t \leq t_\eta^+ \wedge \tilde{t}_{\gamma,\varepsilon}^+$

$$\Phi_{\mathrm{drift}}(m_t) \geq c\gamma m_t^{\ell-1} - C\gamma^2 \frac{d}{n_b} m.$$

When $\gamma \leq c_\gamma n_b d^{-\ell/2} \log(d)^{-C_\gamma}$,

$$\gamma \frac{d \log(d)^C}{n_b} m \leq c_\gamma (\sqrt{d})^{\ell-2} \leq \frac{cm^{\ell-1}}{2}$$

when $t \leq t^-_{\kappa/2\sqrt{d}}$, $C_\gamma \geq C$ and $c_\gamma \leq c/2(\kappa/2)^{\ell-2}$.

Having shown $\Phi_{\mathrm{drift}}(m_t) \geq cm_t^{\ell-1}$, it remains to handle the constant terms in Proposition A.11. We can compute directly $K(t_{\max})$, which yields

$$K(t_{\max}) = \frac{\log(d)^{C - C_{\max}/2}}{C_{\max}\sqrt{d}} + C\gamma \log(d)^{C - C_{\max}} \frac{d}{n_b} \leq \frac{\log(d)^{C - C_{\max}/2}}{C_{\max}\sqrt{d}} + \frac{Cc_\gamma}{\log(d)^{C_{\max}}} d^{-\frac{\ell}{2}} \leq \frac{m_0}{4},$$

by choosing c_γ small enough and $C_{\max}$ large enough.

Finally, we need to show that the events $\mathcal{E}_t$ occur with high probability. This is covered by the following lemma:

Lemma A.12. *Let $T \geq 0$. The following bounds hold:*

- *for the square loss, if $\gamma \leq c_\gamma$ for small enough c_γ,*

$$\mathbb{P}\left(\mathcal{E}_t \text{ holds for all } t \leq t_\eta^+ \wedge T\right) \geq 1 - Te^{-c\log(d)^2};$$

- *for the correlation loss, for small enough ε,*

$$\mathbb{P}\left(\mathcal{E}_t \text{ holds for all } t \leq t_\eta^+ \wedge \tilde{t}_{\gamma,\varepsilon}^+ \wedge T\right) \geq 1 - Te^{-c\log(d)^2}.$$

Proof. We begin with the case of the square loss. From the expression of the gradient expectation in (36), and the estimates (55),

$$\mathbb{E}_t\left[\langle \boldsymbol{w}_t, \boldsymbol{g}_t \rangle\right] = \psi^{\mathrm{sq}}(m_t) - \phi(m_t) m_t = O(1),$$

whenever $t \leq t_\eta^+$. Lemma A.2 applied to $\langle \boldsymbol{w}_t, \boldsymbol{g}_t \rangle$ implies that with probability $1 - ce^{-c\log(n)^2}$

$$|\langle \boldsymbol{w}_t, \boldsymbol{g}_t \rangle| \leq |\mathbb{E}_t\left[\langle \boldsymbol{w}_t, \boldsymbol{g}_t \rangle\right]| + \frac{C\log(d)^C}{\sqrt{n_b}} = O(1)$$

whenever $\mu > 0$. As a result, if $\gamma \leq c_\gamma$ for c_γ small enough, $\mathcal{E}_t$ holds. The proof for the correlation loss proceeds identically, noting this time that

$$\mathbb{E}_t\left[\langle \boldsymbol{w}_t, \boldsymbol{g}_t \rangle\right] = \psi^{\mathrm{corr}}(m_t) - \phi(m_t) m_t = O(1)$$

whenever $t \leq t_\eta^+ \wedge \tilde{t}_{\gamma,\varepsilon}^+$. □

A.4. From the linear regime to one-step recovery

We now prove Theorems 3.4 and 3.5. We focus on the case of the correlation loss; the square loss is identical apart from the additional $\gamma \leq c_\gamma$ requirement of Proposition A.7.

We first assume that $n_b = O(d^{\ell-1})$ and

$$\gamma \leq c_\gamma \log(d)^{-C_\gamma} \min\left(n_b d^{-\left(\frac{\ell}{2} \vee 1\right)},\right) \tag{56}$$

for constants c_γ, C_γ to be chosen later. In particular, the condition $\gamma < c_\gamma n_b d^{-\ell/2}$ of Proposition A.7 is satisfied.

The linear regime The first part of the proof proceeds as in (Ben Arous et al., 2021). Define the function

$$t_\ell(d) = \begin{cases} 1 & \text{if } \ell = 1 \\ \log(d) & \text{if } \ell = 2 \\ d^{\frac{\ell}{2}-1} & \text{if } \ell > 2 \end{cases}, \tag{57}$$

and $t_{\text{conv}} = \max(1, \gamma^{-1} t_\ell(d))$. Proposition A.7 as well as Section 5 from (Ben Arous et al., 2021) implies the following lemma:

Lemma A.13. *There exists a constant $C > 0$ such that if $t_{\max} \geq C t_{\text{conv}}$, then with probability at least $1 - ce^{-c\log(n)^2}$,*

$$\tilde{t}_{\gamma,\varepsilon} \wedge t_\eta^+ \leq C t_{\text{conv}}. \tag{58}$$

We therefore only need to check the condition $t_{\max} \geq C t_{\text{conv}}$. Plugging the expression for γ and $t_{\max}$, we get

$$\begin{aligned} \frac{\gamma t_{\max}}{t_\ell(d)} &\geq \frac{n_b}{\gamma C_{\max} \log(d)^{C_{\max}+1} \gamma d^{1+(\ell/2-1)\vee 0}} \geq \frac{1}{c_\gamma C_{\max}} \log(d)^{C_\gamma - C_{\max} - 1} \\ t_{\max} &\geq \frac{n_b}{C_{\max} \log(d)^{C_{\max}} d\gamma^2} \geq \frac{1}{c_\gamma C_{\max}} \log(d)^{2C_\gamma - C_{\max} - 1} \end{aligned}$$

Whenever $n_b = O(d^{\ell-1})$, by decreasing c_γ and increasing C_γ, for large enough d and any constant C we have

$$\frac{\gamma t_{\max}}{t_\ell(d)} \leq C \quad \text{and} \quad t_{\max} \geq C,$$

as requested in Lemma A.13.

Whenever $\gamma\eta^{\ell-1} \leq \varepsilon$, we have $\tilde{t}_{\gamma,\varepsilon} \geq t_\eta^+$ and hence the proof of Theorem 3.5 is complete. It thus remains to treat the converse case. Note that the latter only happens in the correlation loss case, since we can always choose c_γ such that $c_\gamma \eta^{\ell-1} \leq \varepsilon$; the dynamics of square loss SGD are therefore only in the linearized regime.

One-step recovery above $\tilde{t}_{\gamma,\varepsilon}$ We now treat the case where $\gamma\eta^{\ell-1} \geq \varepsilon$. For simplicity, let $t = \tilde{t}_{\gamma,\varepsilon}$, then Lemmas A.10 and A.9 imply that with probability $1 - ce^{-c\log(n)^2}$

$$\begin{aligned} \|\boldsymbol{g}_t\|^2 &\leq C\left((|\phi(m_t)| + |\psi^{\text{corr}}(m_t)|)^2 + \frac{d}{n_b}\right) \leq C\left(m_t^{2\ell-2} + \frac{d}{n_b}\right) \\ \langle \boldsymbol{g}_t, \boldsymbol{w}^\star \rangle &= \phi(m_t) + m_t \psi^{\text{corr}}(m_t) + O\left(\frac{\text{polylog}(d)}{\sqrt{n_b}}\right) \geq c m_t^{\ell-1} + O\left(\frac{\text{polylog}(d)}{\sqrt{n_b}}\right) \end{aligned}$$

By definition of t, we have $1 \leq \varepsilon^{-1} \gamma m_t^{\ell-1}$, and whenever $\gamma \leq c_\gamma \sqrt{n_b/d}$ one has

$$\frac{\gamma}{\sqrt{n_b}} \leq c_\gamma d^{-1/2} \quad \text{and} \quad \gamma^2 \frac{d}{n_b} \leq c_\gamma^2.$$

Therefore, for large enough d,

$$m_t + \gamma\langle \boldsymbol{g}_t, \boldsymbol{w}^\star\rangle \geq \gamma\langle \boldsymbol{g}_t, \boldsymbol{w}^\star\rangle \geq c\gamma m_t^{\ell-1}$$
$$\|\boldsymbol{w}_t + \gamma \boldsymbol{g}_t\| \leq 1 + \gamma\|\boldsymbol{g}_t\| \leq 1 + C\gamma\left(m_t^{\ell-1} + \frac{d}{n_b}\right) \leq (C + \varepsilon^{-1} + c_\gamma^2\varepsilon^{-1})\gamma m_t^{\ell-1}$$

But by taking the scalar product of Equation (4) with $\boldsymbol{w}^\star$,

$$m_{t+1} = \frac{m_t + \gamma\langle \boldsymbol{g}_t, \boldsymbol{w}^\star\rangle}{\|\boldsymbol{w}_t + \gamma \boldsymbol{g}_t\|} \geq \frac{c}{C + \varepsilon^{-1} + c_\gamma^2\varepsilon^{-1}} =: \eta'.$$

Theorem 3.5 ensues by redefining $\eta = \min(\eta, \eta')$.

B. Informal derivation of Proposition 4.1

In this appendix we provide an informal derivation of the low-dimensional deterministic expressions describing the dynamics of the sufficient statistics (Prop. 4.1). While the formal rigorous characterization should in principle follow directly from (Veiga et al., 2022), it requires a significant amount of work for full mathematical rigor.

Let $\mathcal{D}$ be the set of labeled data $\{\boldsymbol{z}^\nu, y^\nu\}_{\nu\in[n]}$, with label generated by:

$$y^\nu = f^\star(W^\star z^\nu) + \sqrt{\Delta}\xi^\nu, \tag{59}$$

where $W^\star \in \mathbb{R}^{k\times d}$ where we assume $k = O(1)$. We are implying that y^ν depends on $\boldsymbol{z}^\nu \sim \mathcal{N}(0, I_d)$ just throught a low-dimensional representation (linear latent variable). $\xi^\nu \sim \mathcal{N}(0,1)$ is the artificial noise.

We can track the overlap matrix using standard manipulation. We introduce the local fields as:

$$\boldsymbol{\lambda}^\nu := W\boldsymbol{z}^\nu \in \mathbb{R}^p, \quad \boldsymbol{\lambda}^{\star\nu} := W^\star \boldsymbol{z}^\nu \in \mathbb{R}^k \qquad \forall \nu \in [n] \tag{60}$$

We fit these data using a two-layer neural network. Let the first layer weights be $W \in \mathbb{R}^{p\times d}$, the second layer weights $\boldsymbol{a} \in \mathbb{R}^p$; the full expression of the network is given by

$$f(\boldsymbol{z}) = \frac{1}{p}\sum_{j=1}^{p} a_j \sigma(\boldsymbol{w}_j^\top \boldsymbol{z}),$$

where w_j are the rows of W and σ is the activation function.

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

$$\Omega := \begin{pmatrix} Q & M \\ M^\top & P \end{pmatrix} = \begin{pmatrix} WW^\top & WW^{\star\top} \\ W^\star W^\top & W^\star W^{\star T} \end{pmatrix} \in \mathbb{R}^{(p+k)\times(p+k)} \tag{61}$$

These is the low-dimensional matrix (its dimensions stay finite even when $d \to +\infty$) that contains all the information needed for the dynamics.

We are going to train the network with layer-wise SGD without replacement, using at each time step t a fresh new batch of size n_b:

$$\ell_t = \frac{1}{2n_b}\sum_{\nu=1}^{n_b}(y_t^\nu - f(\boldsymbol{z}_t^\nu))^2$$

A stated in the main text, we are interested in the *representation learning* phase. The gradient of the first layer weights

$$\nabla_{\boldsymbol{w}_j}\ell_t = -\frac{1}{pn_b}\sum_{\nu=1}^{n_b} a_j \sigma'(\lambda_{j,t}^\nu)\mathcal{E}_t^\nu \boldsymbol{z}_t^\nu \qquad \forall j \in [p]$$

where we defined for convenience the displacement vector

$$\mathcal{E}_t^\nu := y_t^\nu - f(\boldsymbol{z}_t^\nu). \tag{62}$$

Initially, we focus on plain SGD, without normalizing the weights at every step. Let us take now one gradient step with learning rate γ:

$$\boldsymbol{w}_{j,t+1} = \boldsymbol{w}_{j,t} - \gamma \nabla_{\boldsymbol{w}_j} \ell_t \tag{63}$$

By combining the gradient update equation with the definitions of the matrices (W, W^*) we obtain the following dynamics:

$$\begin{aligned} M_{jr,t+1} - M_{jr,t} =& \frac{\gamma}{pn_b} a_j \sum_{\nu=1}^{n_b} \sigma'(\lambda_{j,t}^\nu) \lambda_r^\star \mathcal{E}_t^\nu \\ Q_{jl,t+1} - Q_{jl,t} =& \frac{\gamma}{pn_b} \sum_{\nu=1}^{n_b} \left(a_j \sigma'(\lambda_{j,t}^\nu) \lambda_{l,t}^\nu + a_l \sigma'(\lambda_{l,t}^\nu) \lambda_{j,t}^\nu \right) \mathcal{E}_t^\nu \\ &+ \frac{\gamma^2}{p^2 n_b^2} a_j a_l \sum_{\nu=1}^{n_b} \sum_{\nu'=1}^{n_b} \sigma'(\lambda_{j,t}^\nu) \sigma'(\lambda_{l,t}^{\nu'}) \mathcal{E}_t^\nu \mathcal{E}_t^{\nu'} \boldsymbol{z}_t^{\nu\top} \boldsymbol{z}_t^{\nu'} \end{aligned} \tag{64}$$

These equations introduce a discrete stochastic process $\{\Omega_t\}_{t\in\mathbb{N}}$ that describes the dynamics in alow-dimensional way. We also introduce the population loss as

$$\mathcal{R}_t = \frac{1}{2} \mathbb{E}_{\Omega_t} \left[\mathcal{E}^2 \right], \tag{65}$$

since it is the quantity telling us the performace of our trained network.

Handling the intra-batch correlations Up to now, we have followed the same derivation as the original Saad&Solla equations (Saad and Solla, 1995a), apart from the effective learning rate scaling. Using larger batches introduces some extra correlations terms that have to be taken into account. Let's split the second term of equation (64) for Q in 2:

$$\sum_{\nu=1}^{n_b} \sum_{\nu'=1}^{n_b} \sigma'(\lambda_{j,t}^\nu) \sigma'(\lambda_{l,t}^{\nu'}) \mathcal{E}_t^\nu \mathcal{E}_t^{\nu'} \boldsymbol{z}_t^{\nu\top} \boldsymbol{z}_t^{\nu'} = \sum_{\nu=1}^{n_b} \sigma'(\lambda_{j,t}^\nu) \sigma'(\lambda_{l,t}^\nu) \mathcal{E}_t^{\nu 2} \boldsymbol{z}_t^{\nu\top} \boldsymbol{z}^\nu + \sum_{\nu=1}^{n_b} \sum_{\nu'=1,\nu'\neq\nu}^{n_b} \sigma'(\lambda_{j,t}^\nu) \sigma'(\lambda_{l,t}^{\nu'}) \mathcal{E}_t^\nu \mathcal{E}_t^{\nu'} \boldsymbol{z}_t^{\nu\top} \boldsymbol{z}_t^{\nu'}$$

The first term is the standard gradient noise term that appears in Sadd&Solla equations, while the second emerge from the large-batch, and has to be treated with new considerations. Let's introduce now the component of the student vectors orthogonal to the teacher space

$$W_t^\perp := W_t - M_t P^{-1} W^\star \text{ and consequently } Q_t^\perp := \left(W_t^\perp \right)^\top W_t^\perp = Q_t - M_t P^{-1} M_t^\top.$$

We can also define the local fields of this subspace, with the interesting property of being independent with the teacher ones

$$\boldsymbol{\lambda}^\perp := W_t^\perp z \qquad \boldsymbol{\lambda}^\perp \sim \mathcal{N}(0, Q_t^\perp) \quad \text{Cov}[\boldsymbol{\lambda}^\star, \boldsymbol{\lambda}^\perp] = 0$$

It is possible to choose a set $\boldsymbol{v}_{\beta,t} \in \left(\text{Span}\,(W_t^\perp) \cup \text{Span}\,(W^\star) \right)^\perp$ of orthonormal vectors, such that $\left\{ \boldsymbol{w}_r^\star, \boldsymbol{w}_{j,t}^\perp, \boldsymbol{v}_{\beta,t} \right\}_{r\in[k], j\in[p], \beta\in[d-p-k]}$ is a basis of $\mathbb{R}^d$. Using the properties of the basis, we can write the identity matrix I_d as

$$I_d = (W^\star)^\top P^{-1} W^\star + (W_t^\perp)^\top (Q_t^\perp)^{-1} W_t^\perp + \sum_{\beta=1}^{d-k-p} \boldsymbol{v}_{\beta,t} \boldsymbol{v}_{\beta,t}^\top$$

We insert the identity matrix $\boldsymbol{z}^\top \boldsymbol{z}$ with $\boldsymbol{z}^\top I_d \boldsymbol{z}$. By recalling that $\boldsymbol{\lambda}^* = W^* \boldsymbol{z}, \boldsymbol{\lambda}^\perp = W_t^\perp \boldsymbol{z}$, we arrive to:

$$\sum_{\nu=1}^{n_b} \sum_{\nu'=1,\nu'\neq\nu}^{n_b} \sigma'(\lambda_{j,t}^\nu) \sigma'(\lambda_{l,t}^{\nu'}) \mathcal{E}_t^\nu \mathcal{E}_t^{\nu'} \left((\boldsymbol{\lambda}^{\nu\star})^\top P^{-1} \boldsymbol{\lambda}^{\nu'\star} + \left(\boldsymbol{\lambda}_t^{\perp\nu} \right)^\top (Q_t^\perp)^{-1} \boldsymbol{\lambda}_t^{\perp\nu'} + \sum_{\beta=1}^{d-k-p} \langle \boldsymbol{v}_{\beta,t}, \boldsymbol{z}^\nu \rangle \langle \boldsymbol{v}_{\beta,t}, \boldsymbol{z}_t^{\nu'} \rangle \right) \tag{66}$$

Now, exploiting the relation:

$$\boldsymbol{\lambda}_t^{\perp} = \boldsymbol{\lambda}_t - M_t P^{-1} \boldsymbol{\lambda}^{\star},$$

and noting that the indeces ν and ν' are independent, all the sum we need to compute are just

$$\sum_{\nu=1}^{n_b} \sigma'(\lambda_{j,t}^{\nu}) \lambda_r^{\star} \mathcal{E}_t^{\nu} \qquad \sum_{\nu=1}^{n_b} \sigma'(\lambda_{j,t}^{\nu}) \lambda_{l,t}^{\nu} \quad \text{and} \quad \sum_{\nu=1}^{n_b} \sum_{\beta=1}^{d-k-p} \langle \boldsymbol{v}_{\beta,t}, \boldsymbol{z}_t^{\nu} \rangle$$

High dimensional limit In our analysis we consider the high-dimensional limit $d \to +\infty$ with the batch size going to infinite as well, with the scaling $n_b = n_0 d^{\mu}$. Note that when $\mu = 0$ the intra-batch correlation disappear and we fall back to standard Saad&Solla setting, given that the learning rate $\gamma = \gamma_0 d^{-\delta}$ is small enough. Indeed, we can informally say that all the sums above converge to their expected value

$$\frac{1}{n_b} \sum_{\nu=1}^{n_b} \sigma'(\lambda_{j,t}^{\nu}) \lambda_r^{\star} \mathcal{E}_t^{\nu} \to \mathbb{E}_{\Omega_t} \left[\sigma'(\lambda_j) \lambda_r^{\star} \mathcal{E} \right] = \psi_{jr,t} \tag{67}$$

$$\frac{1}{n_b} \sum_{\nu=1}^{n_b} \sigma'(\lambda_{j,t}^{\nu}) \lambda_{l,t}^{\nu} \to \mathbb{E}_{\Omega_t} \left[\sigma'(\lambda_j) \lambda_l \mathcal{E} \right] = \phi_{jl,t}^{\text{GF}} \tag{68}$$

$$\frac{1}{n_b} \sum_{\nu=1}^{n_b} \sigma'(\lambda_{j,t}^{\nu}) \sigma'(\lambda_{l,t}^{\nu}) \mathcal{E}_t^{\nu 2} \boldsymbol{z}_t^{\nu \top} \boldsymbol{z}^{\nu} \to d \mathbb{E}_{\Omega_t} \left[\sigma'(\lambda_j) \sigma'(\lambda_l) \mathcal{E}^2 \right] = d \phi_{jl,t}^{\text{GF}} \tag{69}$$

$$\sum_{\nu=1}^{n_b} \sum_{\beta=1}^{d-k-p} \langle \boldsymbol{v}_{\beta,t}, \boldsymbol{z}_t^{\nu} \rangle \to 0 \tag{70}$$

Moreover, using $\boldsymbol{\lambda}^{\perp} = \boldsymbol{\lambda} - M P^{-1} \boldsymbol{\lambda}^{\star}$ we have

$$\mathbb{E}_{\Omega_t} \left[\sigma'(\lambda_j) \mathcal{E} \boldsymbol{\lambda}^{\perp} \right] = \boldsymbol{\phi}_{j,t}^{\text{GF}} - M_t P^{-1} \boldsymbol{\psi}_{j,t}$$

Plugging back in (64), we finally obtain

$$M_{jr,t+1} - M_{jr,t} \approx \frac{\gamma}{p} a_j \mathbb{E}_{\Omega_t} \left[\sigma'(\lambda_j) \lambda_r^{\star} \mathcal{E} \right] \tag{71}$$

$$Q_{jl,t+1} - Q_{jl,t} \approx \frac{\gamma}{p} \mathbb{E}_{\Omega_t} \left[\left(a_j \sigma'(\lambda_j) \lambda_l + a_l \sigma'(\lambda_{l,t}^{\nu}) \lambda_j \right) \mathcal{E} \right] \tag{72}$$

$$+ \frac{\gamma^2 d}{p^2 n_b} a_j a_l \mathbb{E}_{\Omega_t} \left[\sigma'(\lambda_j) \sigma'(\lambda_l) \mathcal{E}^2 \right] \tag{73}$$

$$+ \mathbf{1}_{\{\mu \neq 0\}} \frac{\gamma^2}{p^2} a_j a_l \left(\mathbb{E}_{\Omega_t} \left[\sigma'(\lambda_j) \mathcal{E} \left(\lambda^{\star} \right)^{\top} \right] P^{-1} \mathbb{E}_{\Omega_t} \left[\sigma'(\lambda_l) \mathcal{E} \lambda^{\star} \right] \right) \tag{74}$$

$$+ \mathbf{1}_{\{\mu \neq 0\}} \frac{\gamma^2}{p^2} a_j a_l \left(\mathbb{E}_{\Omega_t} \left[\sigma'(\lambda_j) \mathcal{E} (\lambda^{\perp})^{\top} \right] \left(Q_t^{\perp} \right)^{-1} \mathbb{E}_{\Omega_t} \left[\sigma'(\lambda_l) \mathcal{E} \lambda^{\perp} \right] \right). \tag{75}$$

where the indicator function $\mathbf{1}_{\{\mu \neq 0\}}$ indicates that the last term is only present if the batch is large. If we want to make explicit all the dependencies in d ($\gamma = \gamma_0 d^{-\delta}, n_b = n_0 d^{\mu}$):

$$\begin{aligned}
M_{jr,t+1} - M_{jr,t} &\approx d^{-\delta} \frac{\gamma_0}{p} a_j \psi_{jr,t} = \Psi_{jr,t} \\
Q_{jl,t+1} - Q_{jl,t} &\approx d^{-\delta} \frac{\gamma_0}{p} \left(a_j \phi_{jl,t}^{\text{GF}} + a_l^{\tau} \phi_{lj,t}^{\text{GF}} \right) \\
&\quad + d^{-2\delta+1-\mu} \frac{\gamma_0^2}{p^2 n_0} a_j a_l \phi_{jr}^{\text{HD}} \\
&\quad + d^{-2\delta} \frac{\gamma_0^2}{p^2} a_j a_l \boldsymbol{\phi}_j^{\text{GF}} P^{-1} (\boldsymbol{\phi}_l^{\text{GF}})^{\top} \\
&\quad + d^{-2\delta} \frac{\gamma_0^2}{p^2} a_j a_l \left(\boldsymbol{\phi}_{j,t}^{\text{GF}} - M_t P^{-1} \boldsymbol{\psi}_{j,t} \right) (Q_t - M_t M_t^{\top})^{-1} \left(\boldsymbol{\phi}_l^{\text{GF}} - M_t P^{-1} \boldsymbol{\psi}_l \right)^{\top} = \Phi_{jl,t}
\end{aligned} \tag{76}$$

These equations are the starting point for all our considerations, both when investigating weakly correlation and when characterizing the exact dynamics.

Indeed, when we have a cold start we have to take into account that $\psi_{jr,t}, \phi^{\mathrm{GF}}_{jl,t}, \phi^{\mathrm{HD}}_{jl}$ can also go to 0 when $d \to +\infty$. A careful analysis for the leading terms of these equations around initializations will also give us infromation on the behaviour of the system, and ultimately it will allow to have rules on how to scale δ and μ to have the best performance. An example for these analysis for *generalized linear model* is provided in Appendix B.

On the other hand, when can also assume $\psi_{jr,t}, \phi^{\mathrm{GF}}_{jl,t}, \phi^{\mathrm{HD}}_{jl} = O_d(1)$ for all the dynamics[1] and use the equations for an asymptotic description. Clearly, depending on the values of δ and μ, not all terms are present in the limiting equations. A detailed discussion about this is provided in Section 4.

Spherical projection When we consider the spherical gradient descent, i.e., the modification of eq. (63)

$$\boldsymbol{w}_{j,t+1} = \frac{\boldsymbol{w}_{j,t} - \gamma \nabla_{\boldsymbol{w}_j} \ell_t}{||\boldsymbol{w}_{j,t} - \gamma \nabla_{\boldsymbol{w}_j} \ell_t||}, \tag{77}$$

the overlap dynamics for the spherical large batch SGD can be then approximated as

$$\begin{aligned} M_{jr,t+1} - M_{jr,t} \approx& \Psi_{jr,t}(\Omega) - \frac{M_{jr,t}}{2} \Phi_{jj,t}(\Omega) \\ Q_{jl,t+1} - Q_{jl,t} \approx& \Phi_{jl,t}(\Omega) - \frac{1}{2} Q_{jl,t} \left(\Phi_{jj,t}(\Omega) + \Phi_{ll,t}(\Omega) \right) \end{aligned} \tag{78}$$

This derivation follows from a Taylor expansion of the denominator needed to project the update equations on the sphere. As final note, this aproximation only holds when γ is vanishing when $d \to +\infty$: that's why we need $\gamma = o_d(1)$ in Proppposition 4.1. When γ is too large, all the other orders of Taylor expansion play a role, and we can't have a simple expression for the exact evoluton, even near initialization. Neverthless, the first order expansion is a lower bound of the true dynamic that can provided guarantee of learning in some cases:

$$\begin{aligned} M_{jr,t+1} - M_{jr,t} \geq& \Psi_{jr,t}(\Omega) - \frac{M_{jr,t}}{2} \Phi_{jj,t}(\Omega) \\ Q_{jl,t+1} - Q_{jl,t} \geq& \Phi_{jl,t}(\Omega) - \frac{1}{2} Q_{jl,t} \left(\Phi_{jj,t}(\Omega) + \Phi_{ll,t}(\Omega) \right) \end{aligned} \tag{79}$$

C. Special case: committee machine with matching architecture

We consider a separable teacher, more precisely it is a committee machine with k hidden units, i.e.,

$$f_*(\boldsymbol{z}) = \frac{1}{k} \sum_{r=1}^{k} a^*_r \sigma(\lambda^\star_r)$$

where we additionally consider a matched architecture in which the student and teacher share the same activation function σ.

As we discussed in Section B, the activation and the target appear just in the expected values of Equations (67),(68) and (69),

[1] This happens if $\ell \leq 1$ or when we provide an informed initialization.

that can be further simplified for matching architectures

$$\psi_{jr,t} = \mathbb{E}_{\Omega_t}\left[\sigma'(\lambda_j)\lambda^\star_r\mathcal{E}\right] = \frac{1}{k}\sum_{t=1}^{k} a^*_t \mathbb{E}_{\Omega_t}\left[\sigma'(\lambda_j)\lambda^*_r\sigma(\lambda^*_t)\right] - \frac{1}{p}\sum_{s=1}^{p} a_s \mathbb{E}_{\Omega_t}\left[\sigma'(\lambda_j)\lambda^*_r\sigma(\lambda_s)\right] \tag{80}$$

$$\phi^{\text{GF}}_{jl,t} = \mathbb{E}_{\Omega_t}\left[\sigma'(\lambda_j)\lambda_l\mathcal{E}\right] = \frac{1}{k}\sum_{t=1}^{k} a^*_t \mathbb{E}_{\Omega_t}\left[\sigma'(\lambda_j)\lambda_l\sigma(\lambda^*_t)\right] - \frac{1}{p}\sum_{s=1}^{p} a_s \mathbb{E}_{\Omega_t}\left[\sigma'(\lambda_j)\lambda_l\sigma(\lambda_s)\right] \tag{81}$$

$$\phi^{\text{HD}}_{jl,t} = \mathbb{E}_{\Omega_t}\left[\sigma'(\lambda_j)\sigma'(\lambda_l)\mathcal{E}^2\right] = \frac{1}{k^2}\sum_{r,t=1}^{k} a^*_r a^*_t \mathbb{E}_{\Omega_t}\left[\sigma'\left(\lambda_j\right)\sigma'\left(\lambda_l\right)\sigma\left(\lambda^*_r\right)\sigma\left(\lambda^*_t\right)\right] \tag{82}$$

$$+ \frac{1}{p^2}\sum_{s,u=1}^{p} a_s a_u \mathbb{E}_{\Omega_t}\left[\sigma'\left(\lambda_j\right)\sigma'\left(\lambda_l\right)\sigma\left(\lambda_s\right)\sigma\left(\lambda_u\right)\right] \tag{83}$$

$$- \frac{2}{pk}\sum_{s=1}^{p}\sum_{r=1}^{k} a^*_r a_s \mathbb{E}_{\Omega_t}\left[\sigma'\left(\lambda_j\right)\sigma'\left(\lambda_l\right)\sigma\left(\lambda^*_r\right)\sigma\left(\lambda_s\right)\right] \tag{84}$$

$$+ \Delta\mathbb{E}_{\Omega_t}\left[\sigma'\left(\lambda_j\right)\sigma'\left(\lambda_l\right)\right] \tag{85}$$

In addition, we can also express the population risk as

$$\mathcal{R}_t = \frac{1}{2}\mathbb{E}_{\Omega_t}\left[\mathcal{E}^2\right] = \frac{\Delta}{2} + \frac{1}{p^2}\sum_{s,u}^{p} a_s a_u \mathbb{E}_{\Omega_t}\left[\sigma(\lambda_s)\sigma(\lambda_u)\right] + \frac{1}{k^2}\sum_{r,t}^{k} a^\star_r a^\star_t \mathbb{E}_{\Omega_t}\left[\sigma(\lambda^\star_r)\sigma(\lambda^\star_t)\right] - \frac{2}{pk}\sum_{s,r=1}^{p,k} a_s a^\star_r \mathbb{E}_{\Omega_t}\left[\sigma(\lambda_s)\sigma(\lambda^\star_r)\right]. \tag{86}$$

We introduce auxiliary functions to simplify the mathematical notations:

$$I_2(\omega_{\alpha\alpha},\omega_{\alpha\beta},\omega_{\beta\beta}) = \mathbb{E}\left[\sigma(\lambda_\alpha)\sigma(\lambda_\beta)\right] \tag{87}$$

$$I_3(\omega_{\alpha\alpha},\omega_{\alpha\beta},\omega_{\alpha\gamma},\omega_{\beta\beta},\omega_{\beta\gamma},\omega_{\gamma\gamma}) = \mathbb{E}\left[\sigma'(\lambda_\alpha)\lambda_\beta\sigma(\gamma)\right] \tag{88}$$

$$I_4(\omega_{\alpha\alpha},\omega_{\alpha\beta},\omega_{\alpha\gamma},\omega_{\alpha\delta},\omega_{\beta\beta},\omega_{\beta\gamma},\omega_{\beta\delta},\omega_{\gamma\gamma},\omega_{\gamma\delta},\omega_{\delta\delta}) = \mathbb{E}\left[\sigma'(\lambda_\alpha)\sigma'(\lambda_\beta)\sigma(\lambda_\alpha)\sigma(\lambda_\beta)\right] \tag{89}$$

$$I_2^{\text{noise}}(\omega_{\alpha\alpha},\omega_{\alpha\beta},\omega_{\beta\beta}) = \mathbb{E}\left[\sigma'(\lambda_\alpha)\sigma'(\lambda_\beta)\right]. \tag{90}$$

where we introduced the correlation $\omega_{\alpha\beta} = \mathbb{E}[\lambda_\alpha\lambda_\beta]$, where (α,β) are indices running on either the teacher or the student components. Dropping the time index for clarity, we finally obtain:

$$\psi_{jr} =\frac{1}{k}\sum_{t=1}^{k} a_t^* I_3(Q_{jj}, M_{jr}, M_{jt}, P_{rr}, P_{rt}, P_{tt}) - \frac{1}{p}\sum_{s=1}^{p} a_s I_3(Q_{jj}, M_{jr}, Q_{js}, P_{rr}, M_{sr}, Q_{ss}) \tag{91}$$

$$\phi_{jl}^{\text{GF}} =\frac{1}{k}\sum_{t=1}^{k} a_t^* I_3(Q_{jj}, Q_{jl}, M_{jt}, Q_{ll}, M_{lt}, P_{tt}) - \frac{1}{p}\sum_{s=1}^{p} a_s I_3(Q_{jj}, Q_{jl}, Q_{js}, Q_{ll}, Q_{ls}, Q_{ss}) \tag{92}$$

$$\phi_{jr}^{\text{HD}} =\frac{1}{k^2}\sum_{r,t=1}^{k} a_r^* a_t^* I_4(Q_{jj}, Q_{jl}, M_{jr}, M_{jt}, Q_{ll}, M_{lr}, M_{lt}, P_{rr}, P_{rt}, P_{tt}) \tag{93}$$

$$+ \frac{1}{p^2}\sum_{s,u=1}^{p} a_s a_u I_4(Q_{jj}, Q_{jl}, Q_{js}, Q_{ju}, Q_{ll}, Q_{ls}, Q_{lu}, Q_{ss}, Q_{su}, Q_{uu}) \tag{94}$$

$$- \frac{2}{pk}\sum_{s=1}^{p}\sum_{r=1}^{k} a_r^* a_s I_4(Q_{jj}, Q_{jl}, Q_{js}, M_{jr}, Q_{ll}, Q_{ls}, M_{lr}, Q_{ss}, M_{sr}, P_{rr}) \tag{95}$$

$$+ \Delta I_2^{\text{noise}}(Q_{jj}, Q_{jl}, Q_{ll}) \tag{96}$$

$$\mathcal{R} =\frac{\Delta}{2} + \frac{1}{p^2}\sum_{s,u}^{p} a_s a_u I_2(Q_{ss}, Q_{su}, Q_{uu}) + \frac{1}{k^2}\sum_{r,t}^{k} a_r^\star a_t^\star I_2(P_{rr}, P_{rt}, P_{tt}) \tag{97}$$

$$- \frac{2}{pk}\sum_{s,r=1}^{p,k} a_s a_r^\star I_2(Q_{ss}, M_{sr}, P_{rr}) \tag{98}$$

When analizing a matching architecture setting, we just need to specify I_2, I_3, I_4 and I_2^{noise}. In the following sections we provide the explicit expersion for all the case used in numerical simulation inside this paper.

C.1. Analytic case $\sigma = \operatorname{erf}\left(\cdot/\sqrt{2}\right)$

The expressions can be found in the appendix of (Veiga et al., 2022).

C.2. Analytic case $\sigma = \mathbf{He}_2$

We report here the auxiliary functions:

$$I_2(\omega_{\alpha\alpha}, \omega_{\alpha\beta}, \omega_{\beta\beta}) = \mathbb{E}\left[(\lambda_\alpha^2 - 1)(\lambda_\beta^2 - 1)\right] = \omega_{\alpha\alpha}\omega_{\beta\beta} + 2\omega_{\alpha\beta}^2 - \omega_{\alpha\alpha} - \omega_{\beta\beta} + 1 \tag{99}$$

$$I_3(\omega_{\alpha\alpha}, \omega_{\alpha\beta}, \omega_{\alpha\gamma}, \omega_{\beta\beta}, \omega_{\beta\gamma}, \omega_{\gamma\gamma}) = 2\mathbb{E}\left[\lambda_\alpha\lambda_\beta(\lambda_\gamma^2 - 1)\right] = 2\omega_{\alpha\beta}\omega_{\gamma\gamma} + 4\omega_{\alpha\gamma}\omega_{\beta\gamma} - 2\omega_{\alpha\beta} \tag{100}$$

$$I_4(\omega_{\alpha\alpha}, \omega_{\alpha\beta}, \omega_{\alpha\gamma}, \omega_{\alpha\delta}, \omega_{\beta\beta}, \omega_{\beta\gamma}, \omega_{\beta\delta}, \omega_{\gamma\gamma}, \omega_{\gamma\delta}, \omega_{\delta\delta}) = 4\mathbb{E}\left[\lambda_\alpha\lambda_\beta(\lambda_\gamma^2 - 1)(\lambda_\delta^2 - 1)\right] \tag{101}$$

$$I_2^{\text{noise}}(\omega_{\alpha\beta}) = 4\mathbb{E}\left[\lambda_\alpha\lambda_\beta\right] = 4\omega_{\alpha\beta} \tag{102}$$

We now work on the different terms:

$$4\mathbb{E}\left[\lambda^\alpha\lambda^\beta\left(\lambda^\gamma\right)^2\left(\lambda^\delta\right)^2\right] =4\omega_{\alpha\beta}\omega_{\gamma\gamma}\omega_{\delta\delta} + 8\omega_{\alpha\beta}\omega_{\gamma\delta}^2 + 8\omega_{\alpha\gamma}\omega_{\beta\gamma}\omega_{\delta\delta}+ \tag{103}$$

$$16\omega_{\alpha\gamma}\omega_{\beta\delta}\omega_{\gamma\delta} + 16\omega_{\alpha\delta}\omega_{\beta\gamma}\omega_{\gamma\delta} + 8\omega_{\alpha\delta}\omega_{\beta\delta}\omega_{\gamma\gamma} \tag{104}$$

$$4\mathbb{E}[\lambda_\alpha\lambda_\beta\lambda_\gamma^2] = 4\omega_{\alpha\beta}\omega_{\gamma\gamma} + 8\omega_{\alpha\gamma}\omega_{\beta\gamma} \tag{105}$$

$$4\mathbb{E}[\lambda_\alpha\lambda_\beta\lambda_\delta^2] = 4\omega_{\alpha\beta}\omega_{\delta\delta} + 8\omega_{\alpha\delta}\omega_{\beta\delta} \tag{106}$$

$$4\mathbb{E}[\lambda_\alpha\lambda_\beta] = 4\omega_{\alpha\beta} \tag{107}$$

And then we arrive to:

$$
\begin{aligned}
I_4 = 4\omega_{\alpha\beta}\omega_{\gamma\gamma}\omega_{\delta\delta} + 8\omega_{\alpha\beta}\omega_{\gamma\delta}^2 + 8\omega_{\alpha\gamma}\omega_{\beta\gamma}\omega_{\delta\delta} \quad (108)\\
16\omega_{\alpha\gamma}\omega_{\beta\delta}\omega_{\gamma\delta} + 16\omega_{\alpha\delta}\omega_{\beta\gamma}\omega_{\gamma\delta} + 8\omega_{\alpha\delta}\omega_{\beta\delta}\omega_{\gamma\gamma} \quad (109)\\
-4\omega_{\alpha\beta}\omega_{\gamma\gamma} - 8\omega_{\alpha\gamma}\omega_{\beta\gamma} - 4\omega_{\alpha\beta}\omega_{\delta\delta} - 8\omega_{\alpha\delta}\omega_{\beta\delta} + 4\omega_{\alpha\beta} \quad (110)
\end{aligned}
$$

C.3. Analytic case $\sigma = \mathrm{He}_3$

We report the auxiliary function for this case below.

$$
\begin{aligned}
I_2^{\text{noise}}(\omega_{\alpha\alpha}, \omega_{\alpha\beta}, \omega_{\beta\beta}) :=& \mathbb{E}\left[(3\lambda_\alpha^2 - 3)(3\lambda_\beta^2 - 3)\right]\\
=& 9 - 9\omega_{\alpha\alpha} + 18\omega_{\alpha\beta}^2 - 9\omega_{\beta\beta} + 9\omega_{\alpha\alpha}\omega_{\beta\beta}\\
I_2(\omega_{\alpha\alpha}, \omega_{\alpha\beta}, \omega_{\beta\beta}) :=& \mathbb{E}\left[(\lambda_\alpha^3 - 3\lambda_\alpha)(\lambda_\beta^3 - 3\lambda_\beta)\right]\\
=& 9\omega_{\alpha\beta} - 9\omega_{\alpha\alpha}\omega_{\alpha\beta} + 6\omega_{\alpha\beta}^3 - 9\omega_{\alpha\beta}\omega_{\beta\beta} + 9\omega_{\alpha\alpha}\omega_{\alpha\beta}\omega_{\beta\beta}\\
I_3(\omega_{\alpha\alpha}, \omega_{\alpha\beta}, \omega_{\alpha\gamma}, \omega_{\beta\beta}, \omega_{\beta\gamma}, \omega_{\gamma\gamma}) :=& \mathbb{E}\left[(3\lambda_\alpha^2 - 3)\lambda_\beta(\lambda_\gamma^3 - 3\lambda_\gamma)\right]\\
=& -18\omega_{\alpha\beta}\omega_{\alpha\gamma} + 9\omega_{\beta\gamma} - 9\omega_{\alpha\alpha}\omega_{\beta\gamma} + 18\omega_{\alpha\gamma}^2\omega_{\beta\gamma} + \quad (111)\\
& 18\omega_{\alpha\beta}\omega_{\alpha\gamma}\omega_{\gamma\gamma} - 9\omega_{\beta\gamma}\omega_{\gamma\gamma} + 9\omega_{\alpha\alpha}\omega_{\beta\gamma}\omega_{\gamma\gamma}
\end{aligned}
$$

$$
\begin{aligned}
I_4(\cdots) :=& \mathbb{E}\left[(3\lambda_\alpha^2 - 3)(3\lambda_\beta^2 - 3)(\lambda_\gamma^3 - 3\lambda_\gamma)(\lambda_\delta^3 - 3\lambda_\delta)\right]\\
=& -162\omega_{\alpha\gamma}\omega_{\alpha\delta} + 162\omega_{\alpha\gamma}\omega_{\alpha\delta}\omega_{\beta\beta} + 324\omega_{\alpha\beta}\omega_{\alpha\delta}\omega_{\beta\gamma} - 324\omega_{\alpha\gamma}\omega_{\alpha\delta}\omega_{\beta\gamma}^2 + 324\omega_{\alpha\beta}\omega_{\alpha\gamma}\omega_{\beta\delta} - 162\omega_{\beta\gamma}\omega_{\beta\delta} +\\
& 162\omega_{\alpha\alpha}\omega_{\beta\gamma}\omega_{\beta\delta} - 324\omega_{\alpha\gamma}^2\omega_{\beta\gamma}\omega_{\beta\delta} - 324\omega_{\alpha\delta}^2\omega_{\beta\gamma}\omega_{\beta\delta} - 324\omega_{\alpha\gamma}\omega_{\alpha\delta}\omega_{\beta\delta}^2 + 162\omega_{\alpha\gamma}\omega_{\alpha\delta}\omega_{\gamma\gamma} - 162\omega_{\alpha\gamma}\omega_{\alpha\delta}\omega_{\beta\beta}\omega_{\gamma\gamma} -\\
& 324\omega_{\alpha\beta}\omega_{\alpha\delta}\omega_{\beta\gamma}\omega_{\gamma\gamma} - 324\omega_{\alpha\beta}\omega_{\alpha\gamma}\omega_{\beta\delta}\omega_{\gamma\gamma} + 162\omega_{\beta\gamma}\omega_{\beta\delta}\omega_{\gamma\gamma} - 162\omega_{\alpha\alpha}\omega_{\beta\gamma}\omega_{\beta\delta}\omega_{\gamma\gamma} + 324\omega_{\alpha\delta}^2\omega_{\beta\gamma}\omega_{\beta\delta}\omega_{\gamma\gamma} +\\
& 324\omega_{\alpha\gamma}\omega_{\alpha\delta}\omega_{\beta\delta}^2\omega_{\gamma\gamma} + 81\omega_{\gamma\delta} - 81\omega_{\alpha\alpha}\omega_{\gamma\delta} + 162\omega_{\alpha\beta}^2\omega_{\gamma\delta} + 162\omega_{\alpha\gamma}^2\omega_{\gamma\delta} + 162\omega_{\alpha\delta}^2\omega_{\gamma\delta} - 81\omega_{\beta\beta}\omega_{\gamma\delta} + 81\omega_{\alpha\alpha}\omega_{\beta\beta}\omega_{\gamma\delta} -\\
& 162\omega_{\alpha\gamma}^2\omega_{\beta\beta}\omega_{\gamma\delta} - 162\omega_{\alpha\delta}^2\omega_{\beta\beta}\omega_{\gamma\delta} - 648\omega_{\alpha\beta}\omega_{\alpha\gamma}\omega_{\beta\gamma}\omega_{\gamma\delta} + 162\omega_{\beta\gamma}^2\omega_{\gamma\delta} - 162\omega_{\alpha\alpha}\omega_{\beta\gamma}^2\omega_{\gamma\delta} + 324\omega_{\alpha\delta}^2\omega_{\beta\gamma}^2\omega_{\gamma\delta} -\\
& 648\omega_{\alpha\beta}\omega_{\alpha\delta}\omega_{\beta\delta}\omega_{\gamma\delta} + 1296\omega_{\alpha\gamma}\omega_{\alpha\delta}\omega_{\beta\gamma}\omega_{\beta\delta}\omega_{\gamma\delta} + 162\omega_{\beta\delta}^2\omega_{\gamma\delta} - 162\omega_{\alpha\alpha}\omega_{\beta\delta}^2\omega_{\gamma\delta} + 324\omega_{\alpha\gamma}^2\omega_{\beta\delta}^2\omega_{\gamma\delta} -\\
& 81\omega_{\gamma\gamma}\omega_{\gamma\delta} + 81\omega_{\alpha\alpha}\omega_{\gamma\gamma}\omega_{\gamma\delta} - 162\omega_{\alpha\beta}^2\omega_{\gamma\gamma}\omega_{\gamma\delta} - 162\omega_{\alpha\delta}^2\omega_{\gamma\gamma}\omega_{\gamma\delta} + 81\omega_{\beta\beta}\omega_{\gamma\gamma}\omega_{\gamma\delta} - 81\omega_{\alpha\alpha}\omega_{\beta\beta}\omega_{\gamma\gamma}\omega_{\gamma\delta} +\\
& 162\omega_{\alpha\delta}^2\omega_{\beta\beta}\omega_{\gamma\gamma}\omega_{\gamma\delta} + 648\omega_{\alpha\beta}\omega_{\alpha\delta}\omega_{\beta\delta}\omega_{\gamma\gamma}\omega_{\gamma\delta} - 162\omega_{\beta\delta}^2\omega_{\gamma\gamma}\omega_{\gamma\delta} + 162\omega_{\alpha\alpha}\omega_{\beta\delta}^2\omega_{\gamma\gamma}\omega_{\gamma\delta} -\\
& 324\omega_{\alpha\gamma}\omega_{\alpha\delta}\omega_{\gamma\delta}^2 + 324\omega_{\alpha\gamma}\omega_{\alpha\delta}\omega_{\beta\beta}\omega_{\gamma\delta}^2 + 648\omega_{\alpha\beta}\omega_{\alpha\delta}\omega_{\beta\gamma}\omega_{\gamma\delta}^2 + 648\omega_{\alpha\beta}\omega_{\alpha\gamma}\omega_{\beta\delta}\omega_{\gamma\delta}^2 - 324\omega_{\beta\gamma}\omega_{\beta\delta}\omega_{\gamma\delta}^2 +\\
& 324\omega_{\alpha\alpha}\omega_{\beta\gamma}\omega_{\beta\delta}\omega_{\gamma\delta}^2 + 54\omega_{\gamma\delta}^3 - 54\omega_{\alpha\alpha}\omega_{\gamma\delta}^3 + 108\omega_{\alpha\beta}^2\omega_{\gamma\delta}^3 - 54\omega_{\beta\beta}\omega_{\gamma\delta}^3 + 54\omega_{\alpha\alpha}\omega_{\beta\beta}\omega_{\gamma\delta}^3 +\\
& 162\omega_{\alpha\gamma}\omega_{\alpha\delta}\omega_{\delta\delta} - 162\omega_{\alpha\gamma}\omega_{\alpha\delta}\omega_{\beta\beta}\omega_{\delta\delta} - 324\omega_{\alpha\beta}\omega_{\alpha\delta}\omega_{\beta\gamma}\omega_{\delta\delta} + 324\omega_{\alpha\gamma}\omega_{\alpha\delta}\omega_{\beta\gamma}^2\omega_{\delta\delta} -\\
& 324\omega_{\alpha\beta}\omega_{\alpha\gamma}\omega_{\beta\delta}\omega_{\delta\delta} + 162\omega_{\beta\gamma}\omega_{\beta\delta}\omega_{\delta\delta} - 162\omega_{\alpha\alpha}\omega_{\beta\gamma}\omega_{\beta\delta}\omega_{\delta\delta} + 324\omega_{\alpha\gamma}^2\omega_{\beta\gamma}\omega_{\beta\delta}\omega_{\delta\delta} -\\
& 162\omega_{\alpha\gamma}\omega_{\alpha\delta}\omega_{\gamma\gamma}\omega_{\delta\delta} + 162\omega_{\alpha\gamma}\omega_{\alpha\delta}\omega_{\beta\beta}\omega_{\gamma\gamma}\omega_{\delta\delta} + 324\omega_{\alpha\beta}\omega_{\alpha\delta}\omega_{\beta\gamma}\omega_{\gamma\gamma}\omega_{\delta\delta} +\\
& 324\omega_{\alpha\beta}\omega_{\alpha\gamma}\omega_{\beta\delta}\omega_{\gamma\gamma}\omega_{\delta\delta} - 162\omega_{\beta\gamma}\omega_{\beta\delta}\omega_{\gamma\gamma}\omega_{\delta\delta} + 162\omega_{\alpha\alpha}\omega_{\beta\gamma}\omega_{\beta\delta}\omega_{\gamma\gamma}\omega_{\delta\delta} - 81\omega_{\gamma\delta}\omega_{\delta\delta} +\\
& 81\omega_{\alpha\alpha}\omega_{\gamma\delta}\omega_{\delta\delta} - 162\omega_{\alpha\beta}^2\omega_{\gamma\delta}\omega_{\delta\delta} - 162\omega_{\alpha\gamma}^2\omega_{\gamma\delta}\omega_{\delta\delta} + 81\omega_{\beta\beta}\omega_{\gamma\delta}\omega_{\delta\delta} - 81\omega_{\alpha\alpha}\omega_{\beta\beta}\omega_{\gamma\delta}\omega_{\delta\delta} +\\
& 162\omega_{\alpha\gamma}^2\omega_{\beta\beta}\omega_{\gamma\delta}\omega_{\delta\delta} + 648\omega_{\alpha\beta}\omega_{\alpha\gamma}\omega_{\beta\gamma}\omega_{\gamma\delta}\omega_{\delta\delta} - 162\omega_{\beta\gamma}^2\omega_{\gamma\delta}\omega_{\delta\delta} + 162\omega_{\alpha\alpha}\omega_{\beta\gamma}^2\omega_{\gamma\delta}\omega_{\delta\delta} +\\
& 81\omega_{\gamma\gamma}\omega_{\gamma\delta}\omega_{\delta\delta} - 81\omega_{\alpha\alpha}\omega_{\gamma\gamma}\omega_{\gamma\delta}\omega_{\delta\delta} + 162\omega_{\alpha\beta}^2\omega_{\gamma\gamma}\omega_{\gamma\delta}\omega_{\delta\delta} - 81\omega_{\beta\beta}\omega_{\gamma\gamma}\omega_{\gamma\delta}\omega_{\delta\delta} +\\
& 81\omega_{\alpha\alpha}\omega_{\beta\beta}\omega_{\gamma\gamma}\omega_{\gamma\delta}\omega_{\delta\delta}
\end{aligned}
\quad (112)
$$

D. Weak recovery with Generalized Linear Models

In this section, we restrict our analysis to matching architectures with $p = k = 1$, i.e. *Generalized Linear Models* (GLMs). Moreover, we consider as activation function the Hermite polynomials $\sigma = \mathrm{He}_\ell$, so that we can have control on the information exponent of the problem. Finally, the training algorithm is *projected SGD*, given by Equation (5). We will also assume that $a = a_\star = 1$ throughout all the dynamics.

Let us start by noticing that the set of sufficient statistics reduces to just one single parameter $m = \langle \boldsymbol{w}, \boldsymbol{w}_\star \rangle$. Retracing backward all the steps of the Sections C and B up to (79), we can obtain the lower bound for the update of m. As examples the explicit equation for $\sigma = \mathrm{He}_2$ is

$$\begin{aligned} m_{t+1} - m_t \geq & \gamma_0 d^{-\delta} \Bigg[4m_t - 4m_t^3 - d^{-\delta}\gamma_0 \mathbf{1}_{\{\mu \neq 0\}} \left(8m_t - 8m_t^3\right) + \\ & + d^{-\delta+1-\mu} \frac{\gamma_0}{n_0} \left(24m_t - 24m_t^3 + 2m_t^2\Delta\right) \Bigg], \end{aligned}$$

while for $\sigma = \mathrm{He}_3$ is

$$\begin{aligned} m_{t+1} - m_t \geq & \gamma_0 d^{-\delta} \Bigg[18m_t^2 - 18m_t^4 - d^{-\delta}\gamma_0 \mathbf{1}_{\{\mu \neq 0\}} \left(162m_t + 324m_t^4 - 162m_t^5\right) + \\ & + d^{-\delta+1-\mu} \frac{\gamma_0}{n_0} \left(-1728m_t - 648m_t^3 + 3348m_t^4 - 972m_t^5 - 9\Delta m_t^3\right) \Bigg]. \end{aligned}$$

In general, for an Hermite polynomial activation He_ℓ, the equation of the evolution of m around $m = 0$ is given by

$$m_{t+1} - m_t \geq d^{-\delta}\beta_\ell m^{\ell-1} - d^{-\delta}\left(d^{-\delta+1-\mu}\alpha_\ell + d^{-\delta}\mathbf{1}_{\{\mu\neq 0\}}\phi_\ell\right) m \tag{113}$$

where we fixed $\gamma_0 = n_0 = 1$ for simplicity; $\alpha_\ell, \beta_\ell, \phi_\ell$ are constants. For computing the full equations for any generic ℓ, with the constants values, we refer to the Mathematica notebook published in the repository of this work.

At initialization, $m_0 = {}^1\!/\!\sqrt{d}$. The crucial observation is that a sufficient condition to escape initialization is to have equation (113) being *expansive*, namely $\Delta m > 0$ for m close to zero. This can be true if and only if $\left(d^{-\delta+1-\mu}\alpha_\ell + d^{-\delta}\mathbf{1}_{\{\mu\neq 0\}}\phi_\ell\right) m$ is negligible when compared to $\beta_\ell m^{\ell-1}$, so that the equation for m is lower-bounded by

$$\frac{\Delta m}{\Delta t} \geq \beta_\ell m^{\ell-1} + \text{h.o.t} \quad \text{with} \quad \Delta t = d^{-\delta} \tag{114}$$

Assuming that $\gamma = o_d(1)$, the bound becomes tight and we can also derive some sharp characterization of the escaping time. By simple arguments on differential equations, we can claim that the order of magnitude of steps needed to escape the initial mediocrity is given by

$$T = \begin{cases} O_d\left(\frac{1}{\Delta t}\right) & \ell = 1 \\ O_d\left(\frac{\log m_0}{\Delta t}\right) & \ell = 2 \\ O_d\left(\frac{1}{m_0^{\ell-2}\Delta t}\right) & \ell \geq 3 \end{cases},$$

remembering that $m_0 = {}^1\!/\!\sqrt{d}$ and $\Delta t = d^{-\delta}$, these lead to

$$\log_d T \sim \begin{cases} \max(\delta, 0) & \ell = 1 \\ \log\log d + \delta & \ell = 2 \\ \delta - 1 + {}^\ell\!/2 & \ell \geq 3 \end{cases}. \tag{115}$$

It's clear that to escape as fast as possible, we want δ to be the smallest possible, or in other words, having the learning rate as large as possible. Obviously, δ is constrained by the values that make equation (114) true (or equivalently by the assumptions of the formal proof in Appendix A). The phase diagram of the allowed value of δ and μ is summarized in Figure 4: the green region is where the equations for m is expansive, the red and the yellow region is where the equations in attractive, so there is no escaping, the purple region is where we can't do expansion because the learning rate is too large and the process diverge. Figure 1 in the main text shows the same result in terms of T and n_b, when $\ell \geq 3$.

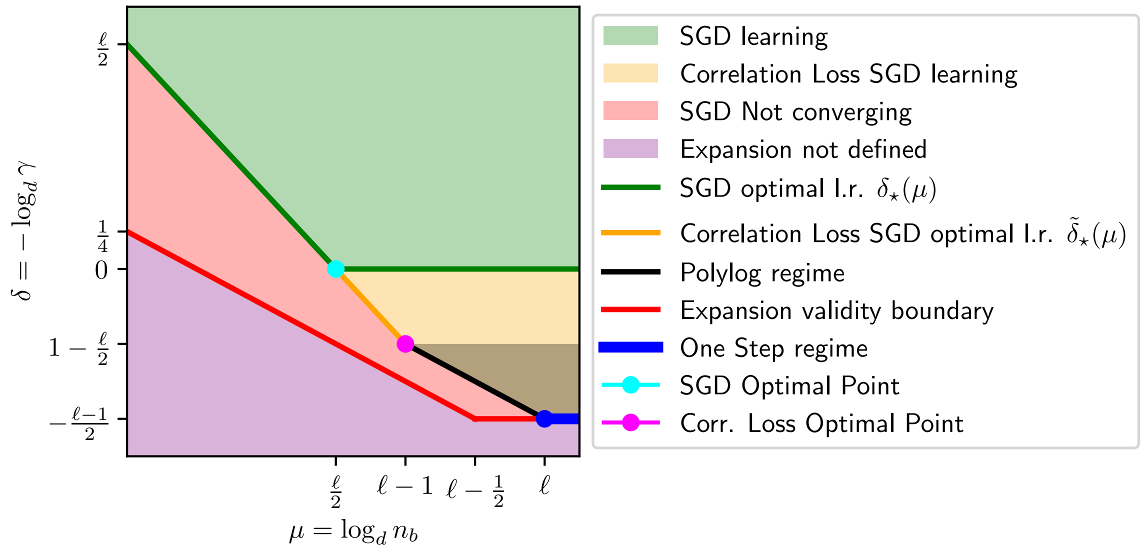

Figure 4. **Phase diagram for the learning rate:** The plot identifies different learning behaviors of standard SGD and *Correlation Loss SGD* for different values of learning rate and batch size when considering randomly initialized networks, i.e. $m_0 = O(1/\sqrt{d})$.

D.1. Correlations loss SGD

When using *Correlation Loss SGD*, Equation (113) rewrite as

$$\frac{\Delta m}{\Delta t} \geq \beta_\ell m^{\ell-1} - d^{-\delta+1-\mu}\alpha_\ell m$$

effectively removing a constraint on the possible values of δ. The modified version of SGD can use smaller values of δ for escaping the initial condition, reaching regions in the phase diagram that are not allowed for SGD: this is colored in yellow in Figures 4 and 1. Of course, if the learning rate becomes too large, all the theory does not work anymore (purple region in the diagram). In Figure 1 we show that the number of steps needed to weakly recover can be pushed down to be smaller than any power scaling with d (black and blue line on the x-axis). The picture becomes clearer if we look at the same diagram in terms of (μ, δ): Correlation Loss SGD can be used with learning very large learning rates ($1 - \ell/2 > \delta > -(\ell-1)/2$) such that the escaping times is $T = O(\text{polylog}(d))$, as proved in Theorem 3.5. We believe that the true number of steps is actually $T = O(\log(d))$, but we could not find any formal proof; in Section E.4, we were able to show that for $\ell = 2$ we have $T = O(\log(d))$, relying the result on numerical integration of our asymptotic theory. Lastly, if the learning rate is of order $\gamma = O(d^{-\delta}) = O(d^{\ell - 1/2})$ we recover the result of (Dandi et al., 2023): the target can be weakly recovered in just one step, when the batch size is $n_b > O(d^\ell)$.

D.2. Simple example: retrieving (Ben Arous et al., 2021)

In this section, we want to show how to find the same result presented in (Ben Arous et al., 2021) starting from our formalism. There, online one-pass SGD is considered, meaning $n_b = 1 \implies \mu = 0$ in our context. Moreover, a vanishing learning rate is assumed, which implies $\delta > 0$, and all the bounds for the evolutions of m are tight. The condition for expansiveness of equation (113) becomes

$$m_0^{\ell-1} \gg d^{-\delta+1} m_0 \implies (\ell - 1)\log_d m_0 \geq -\delta + 1 + \log_d m_0 \implies \delta \geq 1 + (2-\ell)\log_d m_0$$

Plugging in $m_0 = 1/\sqrt{d}$ we finally get $\delta \geq \ell/2$, where the equality is the best possible value of the learning rate in order to make the escaping faster. Note that for $\ell = 1$ we are also bounded from Lemma A.13, so $\delta \geq 1$. Combining with (115),

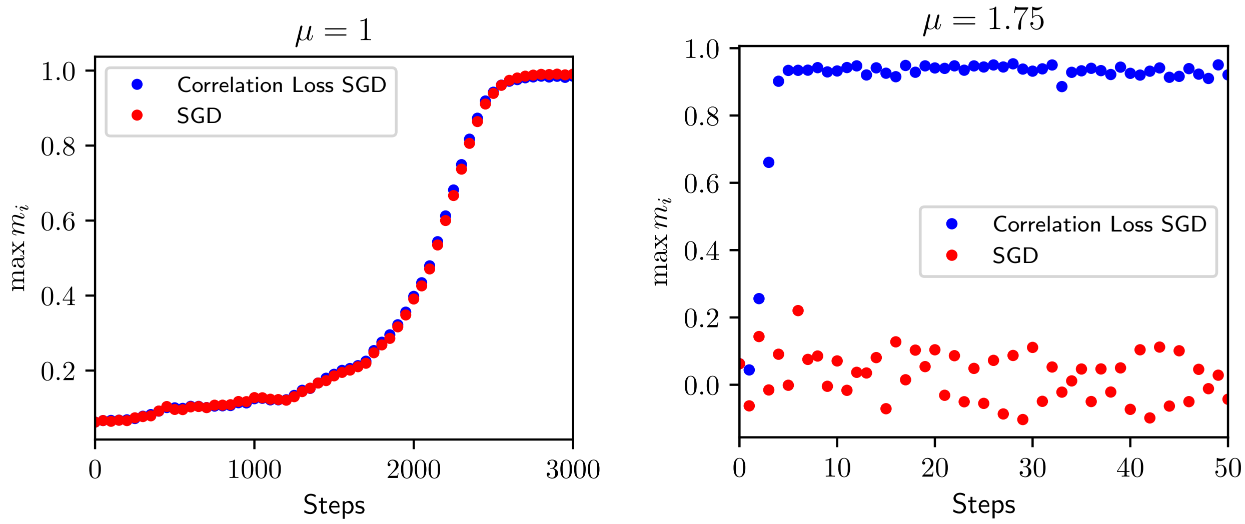

Figure 5. learning single-index teacher with a wide student, when information exponent is $\ell = 3$: $f^\star(\boldsymbol{x}) = \mathrm{He}_3(\boldsymbol{w}^\star \cdot \boldsymbol{x}), f(\boldsymbol{x}) = {}^1/_4 \sum_{i=1}^4 \mathrm{He}_3(\boldsymbol{w}_i \cdot \boldsymbol{x})$. Our theory extends to this case, showing that when $\mu > {}^\ell/_2$ only correlation loss can weakly recover the target. $(d = 256, \gamma = \gamma_0 \cdot p n_b d^{-\ell//2})$

finally gives us the the minimal number of steps needed

$$T \sim \begin{cases} d & \ell = 1 \\ d \log d & \ell = 2 \\ d^{\ell-1} & \ell \geq 3 \end{cases} . \tag{116}$$

This result matches (Ben Arous et al., 2021).

D.3. Extension to $p > 1$

The extension to general two-layer network student functions ($p > 1$), while keeping always the target fixed to be a single-index one, can be readily done by performing an analysis similar to the above. The considerations on the weak recovery trade-offs done in the previous sections are not changed upon re-scaling the learning rate with the hidden layer size $p = O(1)$, i.e. ${}^{\gamma_{\mathrm{2LNN}}}/_p = \gamma_{\mathrm{GLM}}$. Therefore, the scaling laws detailed in the phase diagram (Fig. 1) are not modified, and only prefactors, i.e. quantity not scaling with the input dimension, change with respect to the $p = 1$ case. We illustrate this phenomenon numerically in Fig. 5. We leave the detailed theoretical analysis of the $p > 1$ case for future work, with particular attention to the limit $p \to \infty$ which we believe is an interesting avenue of future research.

E. Additional numerical investigation

In this appendix, we provide additional details on the numerical implementations presented in the main text, along with further explorations. The code to reproduce representative figures is available in https://github.com/IdePHICS/batch-size-time-complexity-tradeoffs.

E.1. Cold start for multi-index models

The theoretical considerations for weak recovery under cold starts presented in Theorems 3.4&3.5 are proven rigorously just for one-hidden neuron network learning single-index targets ($p = k = 1$); this section aims to provide arguments to generalize this to the multi-index case.

Note that for single-index models the initial saddle is the only critical point where the algorithm can get stuck during the dynamics, while this is not true in general for multi-index settings. Indeed, after having weakly recovered a subspace of the

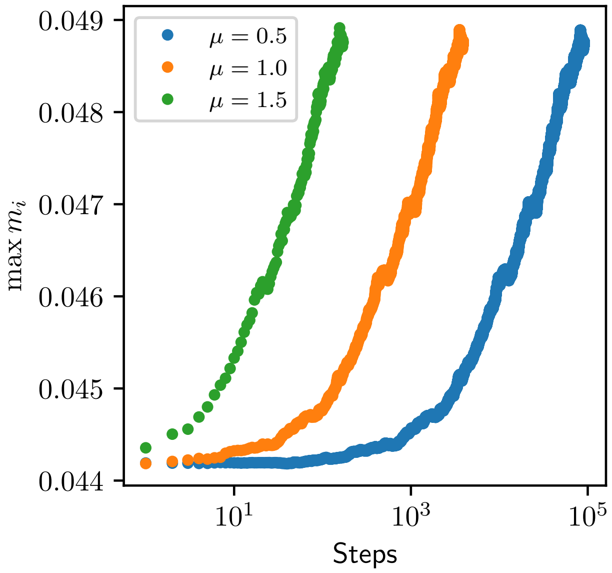

Figure 6. **Multi-index large-batch benefits:** Comparison between the performance of plain SGD learning multi-index model, for different values of μ. The target is $h^\star(z_1, z_2, z_3) = \tanh(z_1 z_2 z_3)$, while the student is a wide 2-layer network $f(\boldsymbol{x}) = \frac{1}{p}\sum_{i=1}^{p} \tanh(\boldsymbol{w}_i \cdot \boldsymbol{x})$ (hence $\ell = 3, p = 30, k = 3, d = 512$). Using a larger batch speeds up the weak-correlation time even when the target is multi-index, and it is learned with a non-matching architecture.

span of the target weights, the learning dynamics can encounter another saddle of the loss function; this behavior is known as saddle-to-saddle dynamics (Jacot et al., 2021). In this manuscript, we focus on escaping from the saddle at initialization, leaving further explorations of the dynamics to future work. We follow (Abbe et al., 2023; Dandi et al., 2023) where the authors generalize the concept of Information Exponent (defined in. (2.1)) to the multi-index setting (See Definition 1 of (Abbe et al., 2023) and Definition 3 of (Dandi et al., 2023)), let us call this quantity the *Leap Index* of the target. We expect that, as long as the dynamics around the saddle at initialization is analyzed, one can substitute the Information Exponent (ℓ) of the teacher in the single-index phase diagram in Fig. 1 with the Leap Index of the target. We explore the Time / Complexity tradeoffs in Figure 6 for a fixed teacher function with Leap Index equal to 3: we observe a relevant decrease in the iterations needed to weakly recover the target subspace as the batch size is increased.

E.2. Behavior of *Spherical SGD*

In many theoretical work (Ben Arous et al., 2021; Abbe et al., 2023), the algorithm used during training uses the *spherical gradient* instead of the simple one. The update rule used instead of Equation (63) is

$$\boldsymbol{w}_{j,t+1} = \frac{\boldsymbol{w}_{j,t} - \gamma\left(I_d - \boldsymbol{w}_{j,t}\boldsymbol{w}_{j,t}^\top\right)\nabla_{\boldsymbol{w}_{j,t}}\ell_t}{\left\|\boldsymbol{w}_{j,t} - \gamma\left(I_d - \boldsymbol{w}_{j,t}\boldsymbol{w}_{j,t}^\top\right)\nabla_{\boldsymbol{w}_{j,t}}\ell_t\right\|} \qquad \forall t \in [T], \forall j \in [p] \tag{117}$$

In practice, only the gradient component orthogonal to the weights is taken into account. This algorithm is particularly convenient for theoretical analysis because it is easier to find a lower bound for the evolution of m_t, since it is always true that

$$\left\|\boldsymbol{w}_{j,t} - \gamma\left(I_d - \boldsymbol{w}_{j,t}\boldsymbol{w}_{j,t}^\top\right)\nabla_{\boldsymbol{w}_{j,t}}\ell_t\right\| \geq 1,$$

while its analogous for Projected SGD does not hold.

In this section, we want to show that *Spherical SGD* is behaving like *Correlation Loss SGD* when γ is not vanishing, namely that is possible to escape mediocrity when the batch size is sufficiently large. For small batch size, Projected SGD and Spherical SGD coincide, while when $\gamma < 0$ their behaviors are drastically different, and only the latter is able to escape

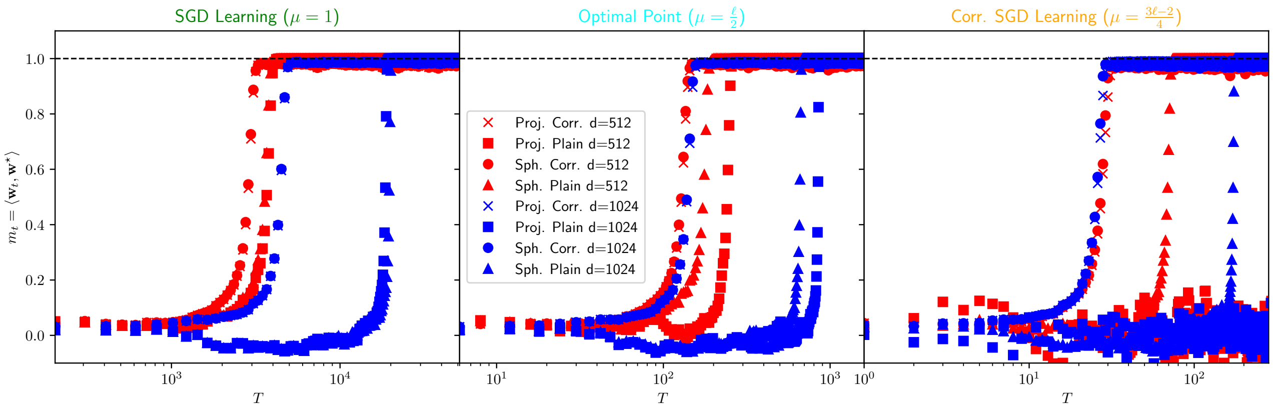

Figure 7. **Correlation Loss SGD weak recovery:** Comparison between the performance of plain SGD, the Correlation Loss SGD and Spherical SGD, in different regions of the phase diagram, and for different sizes d. The plot shows the test error as a function of the optimization steps. Both the teacher and the student activation functions are fixed to $\sigma = h^\star = \mathrm{He}_3$, so the information exponent is $\ell = 3$. In all the three plots we vary the value of μ, while $\delta = \mu - {}^{\ell}\!/_{2}$. Spherical SGD learns even in regions forbidden for plain SGD, as the Correlation Loss does. Note that the Spherical Correlation Loss is equivalent to the Projected Correlation Loss in all the regimes.

mediocrity taking advantage of the large learning rate; a gap between the two is already noticeable at the *Optimal Point*, where they both escape but the spherical is slightly faster. Finally, note that is is possible to introduce a *Correlation Loss Spherical SGD*, by changing the loss in the same way as the usual *Correlation Loss SGD*. There is no practical difference between the two algorithms when working with correlation loss.

E.3. *Adaptive SGD*: combining Correlation Loss SGD with plain SGD

Despite these benefits, the correlation loss is not a good choice to fully learn the target. In this subsection, we explore the idea of combining the two algorithms to escape fast with correlation loss, and then reach the global minimum with the MSE loss. We will call the combination of these two algorithms *Adaptive SGD*.

We are going to test in the simplest case possible: GLM with He_3 as activation function (we remark that there is no benefit in using Correlation Loss SGD over plain SGD when $\ell \leq 2$). If we run the algorithm for multi-index models, it would help to escape the initial saddle, but the algorithm may get stuck in another critical point that is not the global minimum. The study on how to escape fast from a critical point other than the initial one goes beyond the scope of this paper. Our *Adaptive SGD* procedures works as follows:

1. Make a Correlation Loss SGD step;
2. If the Loss is smaller than 60% of the initial loss, jump to Step 3, otherwise go back to Step 1;
3. Reduce the learning rate of a factor 0.995 and do a Standard SGD step;
4. If converged stop, otherwise go back to Step 3.

The learning rate is progressively reduced because the plain SGD requires a lower learning rate compared to the one used by correlation loss to escape fast. Certainly, one can design a much more powerful algorithm than the one we present, but the goal here is just to show that the combination of the two is beneficial, and not to find the possible one. Figure 8 shows how the Adaptive SGD Algorithm is the best one when fully learning the target.

E.4. Polylog regime example

We showed in Section 3 that it is possible to push down the number of steps needed to weakly recover the target until it is growing less than any power law. In order to achieve this, we need to run *Correlation Loss SGD* with $n_b > d^{\ell-1}$ and

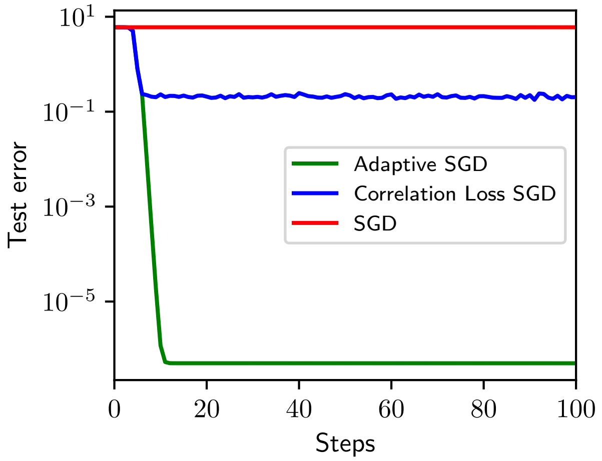

Figure 8. **Adaptive SGD:** The plot compares the performance of SGD and Correlation loss SGD, algorithms with Adaptive SGD; this protocol consists of first using correlation loss SGD to achieve weak recovery, and then switch to adaptive SGD for learning the target. $(\ell = 3, \mu = 1.85, \delta = \mu - \frac{\ell}{2}, \Delta = 10^{-6})$.

$\gamma > O(d^{1-\ell/2})$, as pictured by Figure 4. Proposition 4.1 shows that our theory for cold start, based on expansion of the process (76), is not valid, among other conditions, when $\delta < 0$. Therefore, the only case where we can simultaneously observe the *polylog regime* and have an exact asymptotic description for the full dynamics is when $\ell = 2, \gamma = O(1)$ and $n_b = O(d)$. Let's stick for simplicity with a GLM whose activation is He_2. Note that the total sample complexity is always $N = n_b T = O(d \log d)$.

Figure 9 shows a numerical test of our theory in this particular case. We see that as d grows, also the time needed to escape initial conditions grows. In the right part of the Figure we show that the exact dependence is $T = O(\log d)$, that is indeed a polylog law.

E.5. Large-batch corrections to asymptotic dynamics

Although disappearing when taking the limit, the terms of evolution process (76) coming from intra-batch correlation are useful for providing a better description at large but finite d. Effectively, they are behaving as a first correction to the asymptotic limit.

In this section, we aim to provide numerical arguments about the importance of intra-batch correlations at finite d. We stick with the GLM setting, with erf as activation function. Note that since the information exponent of this target is 1, there is no mediocrity at initialization, we can set $m = \boldsymbol{w}^\top \boldsymbol{w}^\star = 0$ without falling in the *cold start regime*. Figure 10 shows simulations for different values of d (dots), accompanied by the full process dynamic that includes the intra-batch correlation terms (dashed lines); the asymptotic solution of the differential equations (20) is the continuous black line. To enlighten the process even more, we also shows the difference between the asymptotic solution, at the actual finite d one on the right part of the figure. We see that the full process solution always match with the actual project SGD simulation; most importantly, when d grows the simulations are getting closer and closer to the asymptotic solution, confirming that the large batch plays no effect in high-dimension.

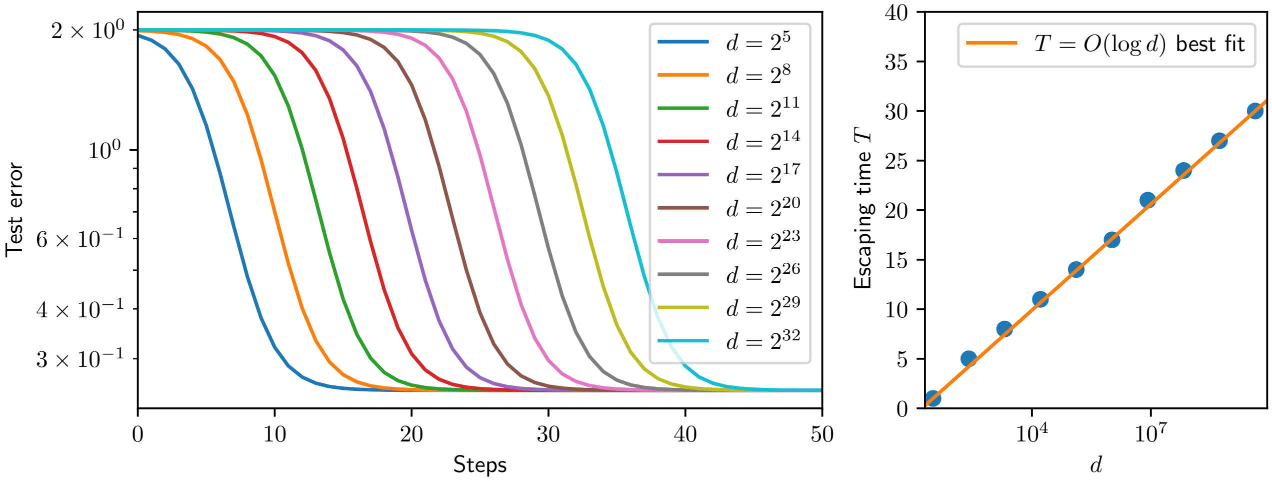

Figure 9. **Phase retrieval with large batch size:** Numerical integration of the process (76), for $f = f^\star = \mathrm{He}_2, \gamma = O(1)$ and $n_b = O(d)$. The escaping time dependence on the number of time steps is a $T = O(\log d)$: we claim this to be valid for all the *polylog regime*.

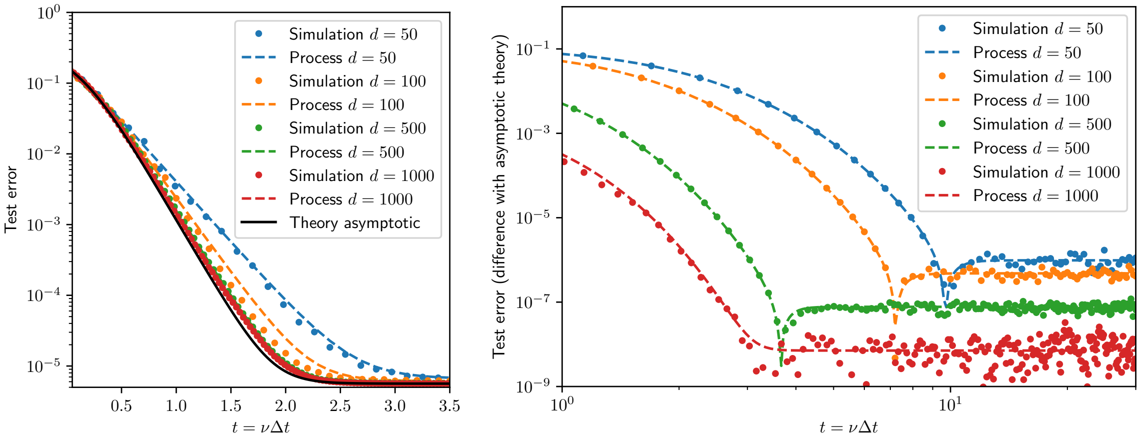

Figure 10. **Non asymptotic corrections:** Comparison between simulations (dots), exact asymptotic solution of ODE (continuous black line), and exact solution including the subleading large-batch corrections (dashed line). As expected, as $d \to +\infty$ the simulations are getting closer and closer to the asymptotic solution; on the other hand, taking into account the batch correlations allows to have a better description of the dynamic even a small d.