The Benefits of Reusing Batches for Gradient Descent in Two-Layer Networks: Breaking the Curse of Information and Leap Exponents

Yatin Dandi [1 2] **Emanuele Troiani** [1] **Luca Arnaboldi** [2] **Luca Pesce** [2] **Lenka Zdeborova** [1] **Florent Krzakala** [2]

Abstract

We investigate the training dynamics of two-layer neural networks when learning multi-index target functions. We focus on multi-pass gradient descent (GD) that reuses the batches multiple times and show that it significantly changes the conclusion about which functions are learnable compared to single-pass gradient descent. In particular, multi-pass GD with finite stepsize is found to overcome the limitations of gradient flow and single-pass GD given by the information exponent (Ben Arous et al., 2021) and leap exponent (Abbe et al., 2023) of the target function. We show that upon re-using batches, the network achieves in just *two* time steps an overlap with the target subspace even for functions not satisfying the staircase property (Abbe et al., 2021). We characterize the (broad) class of functions efficiently learned in finite time. The proof of our results is based on the analysis of the Dynamical Mean-Field Theory (DMFT). We further provide a closed-form description of the dynamical process of the low-dimensional projections of the weights, and numerical experiments illustrating the theory.

1. Introduction

Recent years have witnessed significant theoretical advancements in understanding the dynamics of training neural networks using gradient descent, to unravel the learning mechanisms of these networks, particularly how they adapt to data and identify pivotal features for predicting the target function. Significant progress has been made over the last few years in the case of two-layer networks, in large part thanks to the so-called mean-field analysis (Mei et al., 2018; Chizat & Bach, 2018; Rotskoff & Vanden-Eijnden, 2022; Sirignano & Spiliopoulos, 2020)). Most of the theoretical efforts, in particular, focused either on *one-pass* optimization algorithms, where *each iteration involves a new fresh batch of data*, or to the limit of gradient flow in the population loss. For high-dimensional synthetic Gaussian data, and a low dimensional target function (a multi-index model), the class of functions efficiently learned by these *one-pass* methods has been thoroughly analyzed in a series of recent works, and have been shown to be limited by the so-called information exponent (Ben Arous et al., 2021) and leap exponent (Abbe et al., 2022; 2023) of the target. These analyses have sparked many follow-up theoretical works over the last few months, see, e.g. (Damian et al., 2022; 2023; Dandi et al., 2023; Bietti et al., 2023; Ba et al., 2023; Moniri et al., 2023; Mousavi-Hosseini et al., 2023; Zweig & Bruna, 2023).

However, a common practice in machine learning involves repeatedly traversing the same mini-batch of data. This paper, therefore, aims to go beyond the constraints of single-pass algorithms and to evaluate whether multiple-pass training overcomes these inherent flaws of single-pass methods. We focus on gradient descent, certainly the most straightforward procedure in this family. The theoretical framework we use to prove our main results is based on Dynamical Mean Field Theory (DMFT), which was developed in the statistical physics community (Sompolinsky et al., 1988) to analyze correlated systems, and recently made rigorous in the context of high dimensional machine-learning problems in (Celentano et al., 2021; Gerbelot et al., 2022).

Our findings significantly alter the prevailing narrative in the literature. We demonstrate that gradient descent *surpasses* the limitations imposed by the information and leap exponents, achieving a positive correlation with the target function for a much broader class than the staircase functions (Abbe et al., 2021), even with minimal (that is, two) repetition of data batches. We characterize the (broad) class of functions efficiently learned in finite time. Among the exceptions are *symmetric* functions that remain a challenge due to their extended symmetry-breaking times (a natural feature of physical dynamics (Bouchaud et al., 1998)).

With independent Gaussian datapoints as inputs, both one-

[*]Equal contribution [1]Statistical Physics Of Computation Laboratory, École Polytechnique Fédérale de Lausanne (EPFL) [2]Information Learning and Physics Laboratory, École Polytechnique Fédérale de Lausanne (EPFL). Correspondence to: <yatin.dandi@epfl.ch>, <emanuele.troiani@epfl.ch>.

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

pass SGD and gradient flow on population loss result in pre-activations remaining distributed as Gaussian random variables. This Gaussianity underlies the analysis of such settings, starting from the seminal work of (Saad & Solla, 1995). In contrast, upon re-using batches, the pre-activations develop non-Gaussian components correlated with the targets. This non-Gaussianity is a crucial aspect of the stark contrast in the learning of directions compared to the one-pass setting. While we establish our results for discrete steps with extensive batches of size $\mathcal{O}(d)$ where d denotes the input dimension, we expect our conclusions about the learning of new directions to also hold while performing gradient flow on the empirical loss, since such a setup would lead to the development of similar non-Gaussian components in the pre-activations, in contrast to gradient flow on the population loss where the pre-activations remain Gaussian.

Our results demonstrate that contrary to the common wisdom "the more data the better", gradient descent on the same batch can surpass one-pass SGD on different batches, even when one-pass SGD utilizes a larger number of datapoints. More generally, we believe that our analysis provides insights into incremental learning of features in the presence of correlations between datapoints across batches, which is typical in most high dimensional datasets having a small number of "latent" factors. Our conclusions follow from a rigorous mathematical proof rooted in DMFT, which we also use to provide an analytic description of the dynamic processes of low-dimensional weight projections. This analysis has interest on its own.

2. Setting and main contributions

Let $\mathcal{D}$ be the set of data $\{\mathbf{z}_\nu \in \mathbb{R}^d, y_\nu\}_{\nu\in[n]}$. The input data $\mathbf{z}_\nu$ are taken as a standard i.i.d. Gaussian, while the labels are generated by a teacher, or target, function $y_\nu = f^\star(\mathbf{z}_\nu)$. We consider multi-index target function, dependent on a low-dimensional subspace of the input space:

$$y_\nu = f^\star(\mathbf{z}_\nu) = g^\star\left(\frac{W^\star \mathbf{z}_\nu}{\sqrt{d}}\right), \mathbf{z}_\nu \sim \mathcal{N}(0, \mathbb{1}_d), \quad (1)$$

We assume for convenience that $W^\star = \{\mathbf{w}_r^\star\}_{r\in[k]} \in \mathbb{R}^{k\times d}$ is normalized row-wise on the sphere $\mathcal{S}_{d-1}(\sqrt{d})$, with orthogonal weights i.e $\langle \mathbf{w}_l^\star, \mathbf{w}_m^\star\rangle = 0$ for $l \neq m \in [r]$.

The data are handled to a *two-layer* network (the student) f with first layer weights $W = \{\mathbf{w}_i\}_{i\in[p]} \in \mathbb{R}^{p\times d}$ (p is the number of neurons in the hidden layer) and second layer weights $\mathbf{a} \in \mathbb{R}^p$ with an activation function σ, that is:

$$f\left(\frac{W\mathbf{z}}{\sqrt{d}}\right) = \sum_{j=1}^{p} a_j \sigma\left(\frac{\langle \mathbf{w}_j, \mathbf{z}\rangle}{\sqrt{d}}\right) \quad (2)$$

Our main goal is to analyze the dynamics of gradient descent that minimizes the empirical Mean Squared Error (MSE) loss $\mathcal{L}$ at time $t \in [T]$:

$$\begin{aligned}\mathcal{R}_{\text{emprical}} &= \sum_{\nu=1}^{n} \mathcal{L}\left(\frac{W^{(t)}\mathbf{z}_\nu}{\sqrt{d}}, f^\star(\mathbf{z}_\nu)\right) \quad (3)\\ &= \frac{1}{2}\sum_{\nu=1}^{n}\left[g^\star\left(\frac{W^\star\mathbf{z}_\nu}{\sqrt{d}}\right) - f\left(\frac{W^{(t)}\mathbf{z}_\nu}{\sqrt{d}}\right)\right]^2,\end{aligned}$$

in the high-dimensional limit where $n, d \to \infty$ with $n/d = \alpha = \Theta(1)$. We use a common assumption that is amenable to rigorous theoretical guarantees: we keep the second layer weights $\mathbf{a}$ fixed at initialization. For convenience, we further impose the constraint of symmetric initialization common in such analyses (Dandi et al., 2023; Damian et al., 2022). Concretely, we assume that the number of neurons p is even and the weights satisfy at initialization:

$$a_i = -a_{p-i+1}, \quad \mathbf{w}_i^0 = \mathbf{w}_{p-i+1}^0 \quad \text{for all } i \in [p/2], \quad (4)$$

which ensures that the output $f\left(\frac{W^{(t)}\mathbf{z}_\nu}{\sqrt{d}}\right)$ equals 0 at initialization. For $i, \in [p/2]$, the weights are initialized as $a_i \sim \frac{1}{p}\mathcal{N}(0,1), \mathbf{w}_i^{(0)} \sim \mathcal{N}(0, \mathbb{1}_d)$ Subsequently, with $\mathbf{a}$ fixed, the first layer weights $W = \{\mathbf{w}_i\}_{i\in[p]}$ are learned using gradient descent, producing the following sequence of iterates up to a final time T:

$$\mathbf{w}_i^{(t+1)} = (1-\eta\lambda)\mathbf{w}_i^{(t)} - \eta\sum_{\nu=1}^{n}\nabla_{\mathbf{w}_i^{(t)}}\mathcal{L}\left(\frac{W^{(t)}\mathbf{z}_\nu}{\sqrt{d}}, f^\star(\mathbf{z}_\nu)\right) \quad (5)$$

where $\eta \in \mathbb{R}$ is the learning rate and $\lambda \in \mathbb{R}$ is the explicit regularisation. We may refer to these steps as the *representation learning* steps, in which the first layer weights learn how to adapt to the low dimensional structure identified by the teacher subspace $W^\star$.

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

- We characterize the class of multi-index targets that can be learned efficiently by two-layer networks trained with a finite number of iterations of gradient descent in the high dimensional limit ($d \to \infty$) with large batch sizes ($n = \alpha d, \alpha = O(1)$). We establish a strong separation between what can be learned with one-pass algorithms (that use new fresh batches at every step) and multi-pass gradient approaches that can use the same batch many times (see Figs. 1 and 2 for examples).
- We show that while both gradient flow (Bietti et al., 2023) and single-pass algorithms suffer from the curse of the information exponent (Ben Arous et al., 2021), and are limited to staircase learning (Abbe et al., 2023), requiring a diverging number of iterations for non-staircase functions, some of these problems become *trivial* when allowing reusing samples multiple times, and features can be learned in just $T = 2$ iterations. This disproves, in particular, a recent conjecture by (Abbe et al., 2023).

• The simplest examples of directions that cannot be learned in a finite number of steps relate to symmetries in the target function. This includes phase retrieval (Maillard et al., 2020) or the specialization transition in committees, as discussed in the Bayes optimal approaches of single-index (Barbier et al., 2019) and multi-index (Aubin et al., 2019) models.
• The proof of our results is based on the concept of "hidden progress", and crucially uses the rigorous Dynamical Mean Field Theory (DMFT) (Celentano et al., 2021; Gerbelot et al., 2022). This has an interest on its own as it provides a sharp example of how DMFT can help to understand batch reusing to go beyond the current state-of-the-art results.
• Finally, we use DMFT to provide a closed-form description of the dynamics of gradient descent for two-layer nets. Keeping track of the correlations induced by re-using the same batch leads to a set of integro-differential equations. We provide rigorous theoretical guarantees in the correlated samples regime without assuming the resampling of a fresh new batch for each iteration of the algorithm. We corroborate the theoretical claims with numerical simulations (See https://github.com/IdePHICS/benefit-reusing-batch.).

Other Related works – A major issue in machine learning theory is figuring out how well two-layer neural networks adapt to low-dimensional structures in the data. Different results have tightly characterized the limitations of networks in which the first layer of weights W is kept fixed, i.e. equivalent to kernel approaches (Dietrich et al., 1999; Ghorbani et al., 2019; 2020; Bordelon et al., 2020; Loureiro et al., 2021; Cui et al., 2021). This class of learning algorithms, although amenable to theoretical analysis, is unable to learn features in the data. Therefore, one central avenue of research in this context is to understand the efficiency of the *representation learning* (or feature learning) when training with gradient-based algorithms to overcome the limitations of the kernel regime. Sharp separation results between the performance of neural networks at initialization (random features) and trained with only one step of gradient descent (with a large learning rate) have been offered (Ba et al., 2022; Damian et al., 2022; Dandi et al., 2023).

The class of features efficiently learned with multiple steps of one-pass SGD with one sample per batch is characterized by the *information exponent* (IE) (Ben Arous et al., 2021) of the target function. In the context of single-index learning, denoting ℓ the IE of the target, the algorithm needs $T = O(d^{\ell-1})$ steps to perform weak recovery of the teacher direction, i.e., obtaining an overlap between learned weights and $\mathbf{w}^\star$ better than random guessing (Ben Arous et al., 2021). Recently, these results have been improved up to the Correlational Statistical Query (CSQ) lower bound of $d^{\max(\frac{\ell}{2},1)}$, by considering an appropriate smoothing of the loss (Damian et al., 2023). A generalization to large batch one-pass SGD is in (Dandi et al., 2023).

Similarly, multi-index feature learning presents an unavoidable computational barrier for one-pass algorithms. (Abbe et al., 2021) first characterizes a hierarchical picture of learning in the Boolean data case: informally, the features efficiently learned at each step of the one-pass algorithm need to be linearly connected with the previously learned features. This concept is formalized by the definition of the staircase property (Abbe et al., 2021). This hierarchical picture of learning is extended to large batches in the SGD and non-Boolean data in (Abbe et al., 2022; 2023; Dandi et al., 2023). Moreover, (Abbe et al., 2023) conjecture that re-using the batch can reduce the sample complexity of the target with leap ℓ only up to $O(d^{\max(\frac{\ell}{2},1)})$, corresponding to the lower bound for Correlational Statistical Query (CSQ) algorithms.

We disprove this conjecture and show that the sample complexity for a large class of functions can be reduced to $O(d)$ *independently* of the leap exponent ℓ. More generally, our results show that CSQ lower bounds and the notions of staircase property and information exponent are limited to online-SGD on Gaussian/Boolean data, and do not describe the class of functions inherently easy or hard to learn by gradient-based methods. We also show that learning non-even single-index functions does not require techniques such as spectral warm-start (Chen & Meka, 2020).

Dynamical Mean Field Theory has a long history in statistical physics. Early theories of dynamics in complex systems were pioneered in soft spin glass models (Sompolinsky & Zippelius, 1981) and toy models of random feature deep networks (Sompolinsky et al., 1988). The DMFT approach used in this paper was first proposed as a way to study "hard spins" in spin glass models (Eissfeller & Opper, 1992; 1994), and was later generalized to "soft spins" (Cugliandolo, 2003) and more realistic models in condensed matter (Georges et al., 1996). In the context of learning, DMFT was used for optimization problems (Mannelli et al., 2019a;b; 2020; Mannelli & Urbani, 2021) and for analyzing the behavior and the noise of gradient-based algorithms (Mignacco et al., 2020; 2021; Mignacco & Urbani, 2022). From the mathematics point of view, these DMFT equations were first proven rigorously in the seminal work of (Ben Arous et al., 1997) in the context of spin glasses. Important progress was achieved recently with rigorous proofs of the DMFT equations for multi-index models (Celentano et al., 2021; Gerbelot et al., 2022) that we use to prove our main results.

3. Statement of the results

Here, we introduce the main results covering the theoretical learning guarantees with gradient descent and contrast them with the known one-pass results. We exploit the rigorous DMFT construction to prove the first key result: two-layer networks efficiently learn a large class of multi-index targets in only $T = 2$ iterations, breaking the curse of one-pass algorithms dictated by the information and leap exponents.

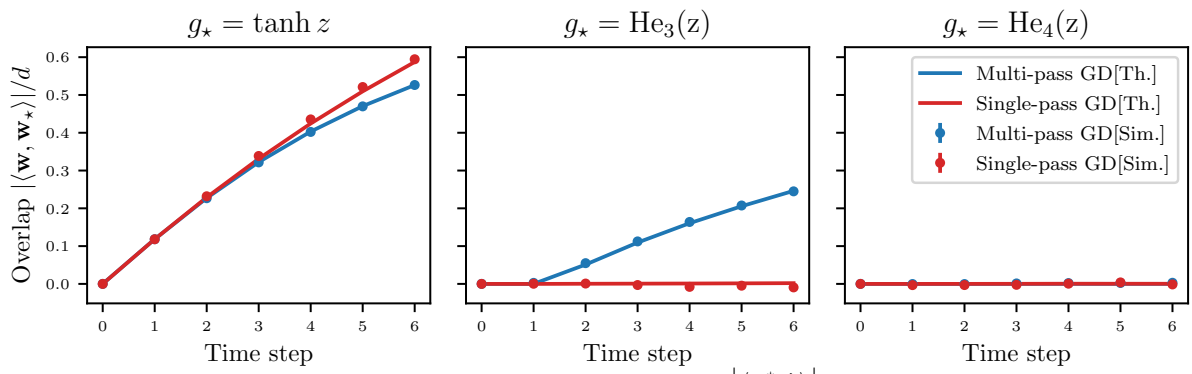

Figure 1. **One-pass and multi-pass GD for single-index models –** The *overlap* $\left|\frac{\langle \mathbf{w}^\star, \hat{\mathbf{w}} \rangle}{d}\right|$ between the learned weight and the target/teacher direction, is plotted as a function of the iteration time of both single-pass (red) and multi-pass (blue) GD. Continuous lines are given theory, dots are simulations. **Left: Easy finite-T learnable single-index target** $g^\star = \tanh$: both one-pass and multi-pass GD obtain positive correlation after a finite number of iterations as the information exponent of the target is $\ell = 1$. **Center: Multi-pass finite-T learnable single-index target**: $g^\star = \mathrm{He}_3$. Multi-pass GD achieves a non-zero correlation in just two steps, but the one-pass algorithm learns nothing. **Right: Finite-time nonlearnable single-index targets** $g^\star = \mathrm{He}_4$; the target function is even and thus, as stated in Thm. 3.2, breaking this symmetry is hard in finite number of steps, resulting in a vanishing correlation with the teacher direction $\mathbf{w}^\star$ for both algorithms in any finite time. (Simulation are averaged over 32 runs, $d = 5000$, with $\sigma = \mathrm{relu}$, $n = 3d$, $p = 1$, $\eta = 0.1$).

3.1. Finite-T Learnable and Non-learnable directions

We first identify which target directions are hard to learn for multi-pass gradient descent. Define $U^\star$ to be the subspace spanned by the rows of the target weights $W^\star$. The "hard" directions are the ones where any transformation of the output $f^\star(\mathbf{z})$ does not lead to a linear correlation along the direction. We now define the subspace of such directions:

Definition 3.1. We define $P^\star$ as the subspace of directions $\mathbf{v}^\star \in U^\star$ such that for any polynomial $F : \mathbb{R} \to \mathbb{R}$ with coefficients in $\mathbb{R}$, the following condition is satisfied:

$$\mathbb{E}_{\mathbf{z}} \left[F(f^\star(\mathbf{z})) \langle \mathbf{v}^\star, \mathbf{z} \rangle \right] = 0, \tag{6}$$

Similarly, we denote by $A^\star$, the subspace of directions where the above condition is satisfied for all real-valued analytic functions F.

One part of our main result shows that directions in $A^\star$ cannot be learned even by re-using batches of size $\mathcal{O}(d)$ in a finite number of gradient steps. Furthermore, under suitable conditions on σ and $\mathbf{a}$ (discussed in Theorem 3.2 and Appendix A.5), we show that after two gradient steps, the first layer learns all directions in the complement $P^\star_\perp$. We are now ready to state our main result:

Theorem 3.2. *Suppose that $n/d = \alpha > 0$. Let $\mathbf{v}^\star \in P^\star_\perp$ denote an arbitrary direction in the orthogonal complement of the subspace $P^\star$ defined in definition 3.1 with norm $\sqrt{d}$ and a fixed representation in the basis $W^\star$. Suppose further that the activation function σ is analytic, with polynomially bounded derivatives satisfying $\mathbb{E}_{z\sim\mathcal{N}(0,1)} [\sigma'(z)] \neq 0$ and $\sigma^k(0) \neq 0 \ \forall k \in \mathbb{N}$. Then for any $g^\star$ with derivatives bounded by polynomials, there exist $\eta > 0, \lambda > 0$ such that almost surely over the choice of $\mathbf{a}$, we have:*

$$\left\| \frac{\mathbf{W}^{(2)} \mathbf{v}^\star}{d} \right\| = \Theta_d(1)\,, \tag{7}$$

with high probability as $n, d \to \infty$. Furthermore, for large enough p, $\mathbf{W}^{(2)}$ asymptotically spans $P^\star_\perp$:

$$\inf_{\mathbf{v}^\star \in P^\star_\perp} \left\| \frac{\mathbf{W}^{(2)} \mathbf{v}^\star}{d} \right\| = \Theta_d(1)\,, \tag{8}$$

with high probability as $n, d \to \infty$. In other words, directions $\mathbf{v}^\star \in P^\star_\perp$ are learned in $T = 2$ gradient steps.

Suppose, however that the teacher subspace $U^\star = A^\star$, then:

$$\sup_{\mathbf{v}^\star \in U^\star} \left\| \frac{\mathbf{W}^{(t)} \mathbf{v}^\star}{d} \right\| = o_d(1)\,, \tag{9}$$

with high probability as $n, d \to \infty$, for any finite time t. Thus, none of the directions are learned in any finite number of GD steps.

The proof is based on the analysis of the DMFT equations discussed in Sec. 4.2, is given in App. A, and we provide an informal heuristic derivation in sec. 4.1. While the above negative result requires all directions in $U^\star$ to be in $A^\star$ and thus in $P^\star$, in App A.6, we discuss the more general setup where learning of certain directions in $P^\star_\perp$ can affect the learning of directions in $U^\star$ in subsequent timesteps.

When the expectation in Equation 6 is non-zero for F being the identity mapping, i.e. $F = \mathrm{id}$, $\mathbf{v}^\star$ is in-fact learned in the first gradient step (Ba et al., 2022; Dandi et al., 2023) or through online SGD (Ben Arous et al., 2022; Abbe et al., 2023). We discuss this further in Section 3.3.

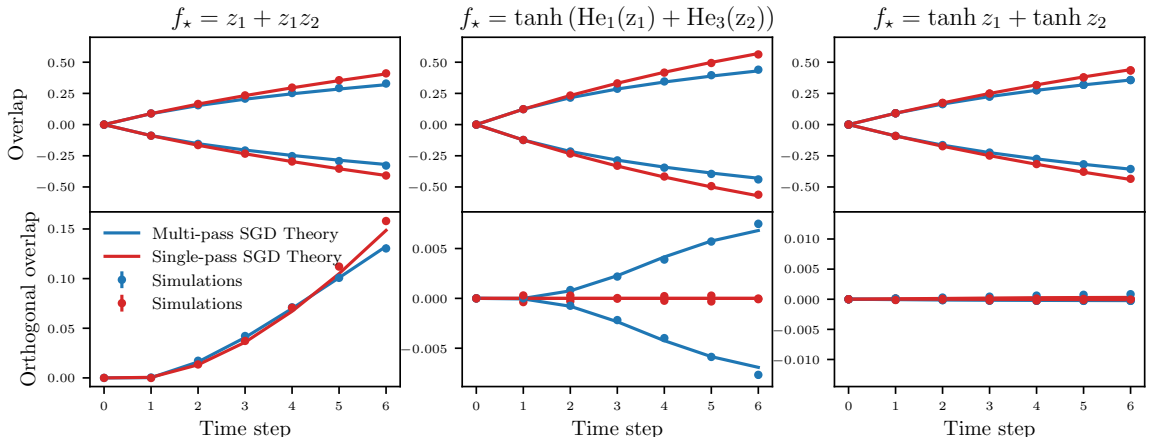

Figure 2. **One-pass and multi-pass GD for multi-index models –** The *overlaps* between the student weights along the first direction learned, namely $\mathbf{C}[f^\star]$, and its orthogonal, is plotted versus the number of iterations for three different classes of functions. **Left: Easy finite-T learnable multi-index target** both the algorithms learn all the relevant directions when an "easy" function is used as a target (here ($p = 8$)). **Center: Multi-pass finite-T learnable multi-index target** both the algorithms learn the first Hermite direction $\mathbf{C}[f^\star]$ but only multi-pass SGD achieve a non-null correlation in the orthogonal. This illustrates how reusing samples allows us to surpass the staircase limitation of single-pass approaches ($p = 2$). **Right: Finite-time non-learnable multi-index target** neither of the two algorithm can learn $\mathbf{C}[f^\star]^\perp$ with this target ($p = 2$). (Simulation are averaged over 32 runs, $d = 5000$, with $\sigma = \mathrm{relu}$, $n = 3d$, $\eta = 0.1$).

Our analysis reveals that the effect of re-using batches is to implicitly transform the output in the subsequent steps, allowing a larger set of directions to be learned. However, for directions in $A^\star$, such transformations are still insufficient.

3.2. Characterization of hard directions through symmetries

While Definition 3.1 characterizes the subspace of hard directions $A^\star$, it requires checking that the equality in Equation 6 holds for any real analytic transformation F. We now show that a sufficient condition for $\mathbf{v}^\star \in A^\star$ is for $f^\star$ to possess certain symmetries along $\mathbf{v}^\star$. This leads us to identify subspaces of hard directions, contained in $A^\star$, linked to symmetries w.r.t certain transformations. We characterize such subspaces below. The simplest such symmetry is defined through reflection along $\mathbf{v}^\star$:

Definition 3.3. For any direction $\mathbf{v}^\star \neq 0 \in \mathbb{R}^d$, let $R_{\mathbf{v}^\star}$ denote the reflection operator along $\mathbf{v}^\star$, i.e. $R_{\mathbf{v}^\star} = \mathbf{I} - 2\frac{1}{\|\mathbf{v}^\star\|^2}\mathbf{v}^\star\mathbf{v}^{\star T}$. We say that a direction $\mathbf{v}^\star \in U^\star$ is even-symmetric w.r.t $f^\star$ if for any $\mathbf{z} \in \mathbb{R}^d$:

$$f^\star(R_{\mathbf{v}^\star}\mathbf{z}) = f^\star(\mathbf{z}) \tag{10}$$

We denote by $E^\star$ the subspace spanned by all even-symmetric directions in $U^\star$.

It is straightforward to see that any $\mathbf{v}^\star \in E^\star$ leads to Equation 6 being satisfied for any transformation F, since $\mathbf{v}^\star$ remains even w.r.t the function $F(f^\star(\cdot))$. Therefore, $E^\star \subseteq A^\star$ However, the set of non-learnable directions can be larger due to the presence of additional symmetries. We now define such a larger subspace of hard directions arising due to a symmetry w.r.t reflections along $\mathbf{v}^\star$ coupled with orthogonal transformations along the orthogonal subspace:

Definition 3.4. For any direction $\mathbf{v}^\star \neq 0$ in $U^\star$, let $R_{\mathbf{v}^\star}$ be as defined in Definition 3.3. Let $O_\perp$ be a matrix in the orthogonal group on the $d-1$ dimensional subspace $\{\mathbf{v}^\star\}_\perp$ i.e the orthogonal complement of the linear subspace spanned by $\mathbf{v}^\star$. We say that a direction $\mathbf{v}^\star$ is orthogonally-even-symmetric w.r.t $f^\star$, if there exists an $O_\perp \in O(\{v^\star\}_\perp)$, such that for any $\mathbf{z} \in \mathbb{R}^d$:

$$f^\star(O_\perp R_{\mathbf{v}^\star}\mathbf{z}) = f^\star(\mathbf{z}) \tag{11}$$

We denote by $OE^\star$ the subspace spanned by all orthogonally-even-symmetric directions in $U^\star$.

By setting $O_\perp$ as the identity mapping in the above definition, we recover the condition for $\mathbf{v}^\star \in E^\star$. Therefore, we have that $E^\star \subseteq OE^\star$. While $OE^\star$ is the largest set of directions we've identified as being hard, the true set of hard directions may be larger still and is given by $A^\star$ in Definition 3.1. We show in Appendix A.8 that the directions in $OE^\star$ are indeed hard as per Definition 3.1:

Proposition 3.5. *Let the subspaces $A^\star$ and $OE^\star$ be as defined in Defs. 3.1 and 3.4 respectively. Then, $OE^\star \subseteq A^\star$ i.e. all directions in $E^\star, OE^\star$ are hard as per Thm. 3.2.*

App.A.7 gives several examples where $P_\perp^\star = E_\perp^\star = U^\star$, such as single-index targets with odd Hermite activations, staircase functions, etc. Interestingly, we show that there exist functions $f^\star$ where the set $OE^\star$ is strictly larger than $E^\star$. Consequently, for such functions, $E^\star$ is strictly contained in

$A^\star, P^\star$. We discuss such target functions in Appendix A.9. For example, we show in Appendix A.9, that for the target function $f^\star(\mathbf{z}) = z_1 z_2 z_3$, the direction $\mathbf{v}^\star = \mathbf{e}_1 + \mathbf{e}_2 + \mathbf{e}_3$ does not lie in $E^\star$ but lies in $OE^\star$ and thus in $A^\star, P^\star$.

3.3. Comparison between one-pass and multi-pass GD

Our results are particularly interesting in the context of a recent line of work on the limitations of one-pass algorithms. (Ben Arous et al., 2021; Abbe et al., 2021; 2022; 2023; Dandi et al., 2023; Bietti et al., 2023; Zweig & Bruna, 2023). We can demonstrate, in particular, a sharp separation performance between one-pass and multiple-pass protocols.

Learning single-index targets – First, we consider single index targets. Targets that are hard to learn for one-pass algorithms starting from uninformed initialization in high dimension are characterized by the *Information Exponent* (IE). Informally, the IE is equivalent to the first non-zero coefficient in the Hermite expansion of the target activation.

Definition 3.6 (Information Exponent)**.** (Ben Arous et al., 2021) Let He_j be the $j-$th Hermite polynomial. Using the definition for the target of eq. (1), reading in the single-index case as $f^\star = g^\star(\langle \mathbf{w}^\star, \mathbf{z}\rangle)$, the IE is defined as:

$$\mathrm{IE} = \min\{j \in \mathbb{N} : \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[g^\star(\xi)\mathrm{He}_j(\xi)\right] \neq 0\} \quad (12)$$

Higher IEs are associated to harder problems for one-pass training protocols. Indeed, (Ben Arous et al., 2021) provably show that one-pass SGD, with one sample per batch, weakly recovers the teacher direction $\mathbf{w}^\star$ only upon iterating the training schedule for $T(\mathrm{IE})$ time iterations:

$$T(\mathrm{IE}) = \begin{cases} \mathcal{O}(d^{\mathrm{IE}-1}) & \text{if IE} > 2 \\ \mathcal{O}(d\log d) & \text{if IE} = 2 \\ \mathcal{O}(d) & \text{if IE} = 1 \end{cases} \quad (13)$$

Recently, the time complexity has been improved up to the Correlational Statistical Query (CSQ) lower bound of $d^{\max(\frac{\mathrm{IE}}{2},1)}$, by considering an appropriate smoothing of the loss (Damian et al., 2023). Definition 3.6 has been extended to larger batch sizes in (Abbe et al., 2022; Dandi et al., 2023), without changing the overall picture; more precisely, even with $n = o(d^{\mathrm{IE}})$ fresh samples per batch, one-pass training procedures are still not able to weakly recover the signal in finite iteration time. The case $\mathrm{IE} = 1$ corresponds to the expectation in Equation 6 being non-zero for $F = \mathrm{id}$. However, since Definition 3.1 allows for general transformations F to the output $f^\star$, $\mathbf{w}^\star$ may not be in $P^\star, A^\star$ even when $\mathrm{IE} > 1$. The presence of general transformations F in definition 3.1 allows our algorithm to bypass CSQ bounds, which are restricted to $F = \mathrm{id}$. Such general transformations are however permitted under the framework of Statistical Query (SQ) algorithms (Kearns, 1998). We thus expect gradient descent with $\mathcal{O}(d)$ sample complexity to inherit the hardness results established for the class of SQ algorithms (Diakonikolas et al., 2020; Goel et al., 2020; Chen et al., 2021; 2022). We emphasize however that unlike explicit SQ algorithms, our analysis shows that gradient descent performs such transformations implicitly, allowing it to reach the optimal complexity of SQ algorithms for certain class of target functions.

We illustrate the sharp contrast between one-pass and multiple-pass protocols with the examples depicted in Figure 1, which shows the scalar product (called *overlap*) between the learned weights and the teacher direction as a function of the time steps and compares simulation (dots) with theoretical predictions (continuous lines). There are 3 cases:

- **Easy finite-T learnable single-index targets** $(\mathrm{IE}=1)$**:** The left panel of Fig. 1 show the learning curve for a problem with $\mathrm{IE}=1$. *Both* single-pass and multiple-pass GD correlates with the target in finite time. The non-symmetric subspace coincides with the teacher one $E^\star_\perp = U^\star$ (Def. 3.3).
- **Multi-pass finite-T learnable single-index targets** $(\mathrm{IE}>1, \mathrm{non-even\ f}^\star)$**:** Fig. 1 (center) depicts the learning curve for a non-even target function, with $\mathrm{IE} = 3$. Here, one-pass GD is not able to achieve any significant correlation with the teacher $\mathbf{w}^\star$ (and it would require a number of iterations $T = \mathcal{O}(d)$ to achieve weak recovery - see eq. (13)). However, multiple-pass GD performs weak recovery in only $T = 2$ steps. As before, the non-symmetric subspace corresponds to the teacher one $E^\star_\perp = U^\star$ (Def. 3.3).
- **Finite-T non-learnable single-index targets** $(\mathrm{IE}>1,\ \mathrm{even\ f}^\star)$**:** Fig. 1 (right) considers an even problem, with $\mathrm{IE} = 4$. Neither of the training procedures achieve weak recovery in finite time. The computational hardness of this problem is associated with the presence of symmetry in the teacher function that requires time to break. Indeed, following Definition 3.3, the even-symmetric subspace $E^\star$ is equivalent to the teacher subspace $U^\star$. These results agree with the emergence of *computational barriers* in symmetric single-index problems like the phase retrieval one (Maillard et al., 2020). In fact, for such problems, regardless of the number of iterations, learnability requires α to be larger than critical values even for the most efficient known algorithms (see (Barbier et al., 2019), Sec. 3.1).

Learning multi-index targets – The hardness of multi-index targets learning has been the subject of numerous recent studies for single-pass algorithms (Abbe et al., 2021; 2022; 2023; Bietti et al., 2023; Zweig & Bruna, 2023; Dandi et al., 2023). The class of multi-index targets efficiently learned by one-pass algorithms has been provably associated with the *Leap Complexity* (LC) of the target to be learned, which generalizes the information exponent:

Remark 3.7. To enhance the clarity of the presentation, we limit the definition of the LC to an informal one. We refer to Section **B.2** (*isoLeap*) in (Abbe et al., 2023) and Definition **3** in (Dandi et al., 2023) (*leap index*) for details.

Informally, the learning dynamics of one-pass routines follow this behavior: initially, the network learns in the first step the first Hermite coefficient of the target $f^\star$. For every time $t \in [T]$ of the one-pass schedule, the network is bound to learn in finite time only features that are *linearly connected* to the previously learned directions; functions possessing only such linearly connected features are leap 1 functions (LC $= 1$), e.g. $f^\star(\mathbf{z}) = z_1 + z_1 z_2$. Similarly, functions that are quadratically connected to the learned features are leap 2 (LC $= 2$), e.g. $f^\star(\mathbf{z}) = z_1 + z_1 z_2 z_3$. Higher LC target functions correspond to harder learning problems for one-pass algorithms: one-pass SGD, with one sample per batch, weakly recovers the teacher subspace $U^\star$ by iterating the training protocol for $T(\mathrm{LC})$ time steps, where the LC substitutes the IE in eq. (13) (Abbe et al., 2023).

We illustrate the behavior of one-pass and multiple-pass algorithms when learning multi-index functions in Fig. 2. Using different two-index teachers ($k = 2$), it shows the scalar product between the learned weights and two reference vectors: a) the first Hermite coefficient of the target $f^\star$, called in the following $\mathbf{C}_1[f^\star]$; b) the vector in the teacher subspace $U^\star$ orthogonal to $\mathbf{C}_1[f^\star]$, referred as $\mathbf{C}_1[f^\star]^\perp$. The figure exemplifies the correlations metrics as a function of time, labeled as *overlap* (resp. *orthogonal overlap*) in the upper (resp. lower) section. There are, again, 3 cases:

- **Finite-T learnable multi-index targets:** Fig. 2 (left) depicts a target with LC $= 1$. The teacher subspace $U^\star$ spanned by the standard basis vectors $\{\mathbf{e}_1, \mathbf{e}_2\}$ is learned by both one-pass and multi-pass GD in finite time. At $T = 1$, $\mathbf{e}_1 = \mathbf{C}_1[f^\star]$ is learned; this enables the recovery of the direction $\mathbf{e}_2$ at $T = 2$ as the target is linear in z_2 once e_1 has been learned. This hierarchical picture of learning is called *staircase* mechanism. Using Def. 3.3 notations, the non-symmetric teacher subspace $E^\star_\perp$ is equivalent to the full teacher subspace $U^\star$.
- **Multi-pass finite-T learnable multi-index targets:** The central panel in Fig. 2 illustrates a teacher with LC $= 3$. Both algorithms are successful in weakly recovering the direction $\mathbf{C}_1[f^\star]$ in the first step. However, as the training continues, one-pass GD never recovers the full teacher subspace in finite time (exemplified by the zero orthogonal overlap in the lower panel). Conversely, multi-pass GD is able to perform weak recovery of the full teacher subspace $U^\star$ by achieving a non-vanishing correlation with $\mathbf{C}_1[f^\star]^\perp$ (non-zero orthogonal overlap in the lower section) in just $T = 2$ steps. Again, the non-symmetric subspace $E^\star_\perp = U^\star$ is equivalent to the full teacher subspace $U^\star$ (Def. 3.3).
- **Finite-T non-learnable multi-index targets:** The right panel of Fig. 2 considers a committee machine teacher with symmetric activation, i.e. $f^\star(\mathbf{z}) = \sum_{r=1}^{2} \sigma^\star(\langle \mathbf{z}, \mathbf{e}_r \rangle)$, here LC $= 2$. Both protocols, in this case, are only able to learn a single-index approximation of the target function in finite time, achieving non-zero correlation only with $\mathbf{C}_1[f^\star] \propto \mathbf{e}_1 + \mathbf{e}_2$ throughout the dynamics. The computational hardness of this problem is associated with the presence of a neuron exchange symmetry. Indeed, using Def. 3.3 notations, we observe that the even-symmetric subspace $E^\star = \{\frac{1}{\sqrt{2}}(\mathbf{e}_2 - \mathbf{e}_1)\}$ is a non-empty subspace of the teacher one $U^\star$. Therefore, as for one-pass routines, multiple-pass ones are bound to learn only $\mathbf{v}^\star = (\mathbf{e}_1 + \mathbf{e}_2)/\sqrt{2}$ in finite time steps. Such difficulties have been described in the analysis of the *specialization* transition in the information-theoretic/Bayes optimal case of symmetric committees (Aubin et al., 2019). As for single index models, breaking the symmetry requires α to be large enough and, even in this case, the best-known algorithms require a diverging number of iterations (see (Aubin et al., 2019), Sec. 3).

3.4. From weak recovery to generalization

While Th. 3.2 provides conditions for the *weak recovery* (a finite overlap with directions in $U^\star$), once this is done, it becomes straightforward to learn the function up to any desired accuracy with only $\mathcal{O}(d)$ additional samples. Indeed, strong generalization guarantees can be proven by utilizing existing results either for subsequent training with online SGD (Ben Arous et al., 2021) (to use their terminology, once you escape *mediocrity*, the ballistic phase is *easy*) or training of the second layer using an independent batch of $\mathcal{O}(d)$ samples as in (Damian et al., 2022; Abbe et al., 2023). See App.A.10 for such generalization sample-complexity results.

4. Main proof ideas

4.1. Learning by hidden progress: heuristic argument

While we give a rigorous proof of Thm. 3.2 in App. A, we provide now an informal description of the *hidden progress* in the first step of gradient descent that allows subsequent development of overlaps in the second step, that is at the root of the difference between single and multi-pass algorithms. For simplicity, we focus on the case of a single hidden neuron ($p = 1$). We denote $h_\mu^{(t)} = \langle \mathbf{w}^{(t)}, \mathbf{z}_\mu \rangle / \sqrt{d}$ the pre-activation for the ν_{th} training point along the neuron with $\mu \in [n]$, and $\mathbf{v}^\star$ a vector in the span of $\mathbf{W}^\star$ with $\|\mathbf{v}^\star\| = \sqrt{d}$.

From the gradient update in Eq. (5), the update lies in the span of the training inputs $\{\mathbf{z}_\nu\}_{\nu=1}^n$, with the gradient of the ν_{th} training example given by $\mathbf{g}_\nu = a\mathcal{L}'\left(h_\nu^{(t)}, f^\star(\mathbf{z}_\nu)\right)\sigma'(h_\nu^{(t)})\mathbf{z}_\nu/\sqrt{d}$. For squared loss, assuming that $f\left(\frac{W^{(t)}\mathbf{z}_\nu}{\sqrt{d}}\right) \approx 0$, the gradient reads:

$$\mathbf{g}_\nu^{(t)} \approx -a f^\star(\mathbf{z}_\nu)\sigma'(h_\nu^{(t)})\mathbf{z}_\nu/\sqrt{d}\,. \tag{14}$$

At initialization $h_\nu^{(0)} = \langle \mathbf{w}^{(0)}, \mathbf{z}_\nu \rangle/\sqrt{d}$ and the projections along the teacher subspace (which we denote $\mathbf{h}_\nu^\star = \frac{1}{\sqrt{d}} W^\star \mathbf{z}_\nu \in \mathbb{R}^k$) are approximately independent since $\mathbf{w}^{(0)}$ is approximately orthogonal to the teacher subspace as well as to the inputs $\{\mathbf{z}_\nu\}_{\nu=1}^n$. The projection of the gradient

along the teacher subspace is given by:

$$W^\star\left(\sum_{\nu=1}^{n}\frac{\mathbf{g}_\nu^{(t)}}{\sqrt{d}}\right)\approx-\sum_{\nu=1}^{n}af^\star(\mathbf{z}_\nu)\sigma'(h_\nu^{(t)})\frac{\mathbf{h}_\nu^\star}{d}\,. \tag{15}$$

We do expect that, due to concentration, the component of the full-batch gradient update along the teacher subspace lies along the direction given by:

$$\mathbb{E}\left[f^\star(\mathbf{z}_\nu)\sigma'(h_\nu^{(0)})\mathbf{h}_\nu^\star\right]\approx\mathbb{E}\left[\sigma'(h_\nu^{(0)})\right]\mathbb{E}\left[f^\star(\mathbf{z}_\nu)\mathbf{h}_\nu^\star\right], \tag{16}$$

where we used the approximate independence of $h_\nu^{(0)}$ and $\mathbf{h}_\nu^\star$ to factorize the expectation. Thus, the neuron parameters $\mathbf{w}^{(1)}$ at the first step are correlated with the teacher subspace only along the direction $\mathbb{E}\left[f^\star(\mathbf{z}_\nu)\mathbf{h}_\nu^\star\right]$.

If $\mathbb{E}\left[f^\star(\mathbf{z}_\nu)\mathbf{h}_\nu^\star\right]=0$, the parameters remain orthogonal to the teacher subspace. This is true whenever the LC of the target function is larger than 1. To make progress, it is thus necessary for the pre-activations $h_\nu^{(t)}=\langle\mathbf{w}^{(t)},\mathbf{z}_\nu\rangle/\sqrt{d}$ to become correlated with the teacher pre-activation $\mathbf{h}_\nu^\star$. This can happen in two different ways:

(i) By $\mathbf{w}^{(t)}$ directly gaining components along the teacher subspace W^*. Under online SGD, the data is used only once for the gradient updates, so only this mechanism is possible. It allows the directions learned by $\mathbf{w}^{(t)}$ at any step to depend on the directions already learned by $\mathbf{w}^{(t)}$. This underlies the "staircase" phenomenon in online SGD (Abbe et al., 2021; 2022; 2023) as well as the notion of information exponent when applied to a single direction (Ben Arous et al., 2021).

(ii) By $\mathbf{w}^{(t)}$ gaining components along $\mathbf{z}_\nu$. Recall that the target is defined as $y_\nu=f^\star(\mathbf{z}_\nu)$, and thus $\mathbf{w}^{(t)}$ can correlate with $\mathbf{h}_\nu^\star$. This is what happens when using gradient descent with multi-pass in our setting. This implies that even when $\mathbf{w}^{(1)}$ does not learn a direction $\mathbf{v}_*$, the pre-activation $h_\nu^{(t)}$ can develop a dependence on $\mathbf{h}_\nu^\star$ through the component of the gradient update along $\mathbf{z}_\nu$.

Let us see how this phenomenon, which we call *hidden progress*, happens in practice. From (5), the update in the pre-activation $h_\nu^{(1)}$ due to the first gradient step reads:

$$\begin{aligned}h_\nu^{(1)}=&(1-\eta\lambda)h_\nu^{(0)}\\&-\eta\frac{a}{d}\sum_{\nu'=1}^{n}\mathcal{L}'(h_{\nu'}^{(0)},f^\star(\mathbf{z}_{\nu'}))\sigma'(h_{\nu'}^{(0)})\langle\mathbf{z}_{\nu'},\mathbf{z}_\nu\rangle\,.\end{aligned} \tag{17}$$

In this sum there is one term of magnitude $\mathcal{O}\left(\|\mathbf{z}_\nu\|^2/d\right)=O(1)$ corresponding to $\nu'=\nu$, and $d-1$ random terms of order $\mathcal{O}\left(\langle\mathbf{z}^\nu,\mathbf{z}^\nu\rangle/d\right)=O\left(1/\sqrt{d}\right)$. This second group of terms contributes to an effective "noise" of order $O(1)$. The first term however, since $\|\mathbf{z}_\nu\|_2^2\approx d$, depends on $f^\star(\mathbf{z}_\nu)$ (and thus on all components of $\mathbf{h}_\nu^\star$):

$$\mathcal{L}'(h_\nu^{(0)},f^\star(\mathbf{z}_\nu))\sigma'(h_\nu^{(0)})\,. \tag{18}$$

Due to this dependence between $h_\nu^{(1)}$ and $f^\star(\mathbf{z}_\nu)$, in the subsequent steps i.e. $T=2$, the term $\sigma'(h_\nu^{(1)})$ in the update (14) can now influence the direction of the gradient along the teacher subspace, leading to $\mathbf{w}^{(2)}$ gaining correlations with new directions in $W^\star$. It can be seen as follow: let $m_{\mathbf{v}_\star}^{(t)}=\langle\mathbf{w}^{(t)},\mathbf{v}^\star\rangle/\sqrt{d}$, it follows from the GD updates that

$$\begin{aligned}m_{\mathbf{v}^\star}^{(2)}=&(1-\eta\lambda)m_{\mathbf{v}^\star}^{(1)}\\&-\eta\frac{a_j}{d}\sum_{\nu=1}^{n}\mathcal{L}'(h_\nu^{(1)},f^\star(\mathbf{z}_\nu))\sigma'(h_\nu^{(1)})h_\nu^{\mathbf{v}^\star}\end{aligned}$$

Now, suppose that $\mathbf{v}^\star$ is not learned in the first step. However, due to the hidden progress, $h_\nu^{(1)}$ is now dependent on $h_\nu^{\mathbf{v}^\star}=\langle\mathbf{v}^\star,\mathbf{z}_\nu\rangle/\sqrt{d}$, thus allowing the new expectation of the projection of the update along $\mathbf{v}^\star$ given by $\mathbb{E}\left[f^\star(\mathbf{z}_\nu)\sigma'(h_\nu^{(1)})h_\nu^{\mathbf{v}^\star}\right]$ to be non-zero. This explains how the dependence of the pre-activations $h_\nu^{(t)}$ on $f^\star(\mathbf{z}^\nu)$ can allow learning of new directions even when the weights have not gained components along the teacher subspace.

This learning mechanism, however, fails when the target function is symmetric along $\mathbf{v}^\star$. Indeed, for such a direction, $h_\nu^{(1)}$ retains an *even* dependence on $h_\nu^{\mathbf{v}^\star}$, which implies that the expectation of the term $\mathbb{E}\left[f^\star(\mathbf{z}_\nu)\sigma'(h_\nu^{(t)})h_\nu^{\mathbf{v}^\star}\right]$ remains 0 for all time steps $t\in[T]$, with $T=\mathcal{O}(1)$. Such directions are therefore not learned with a finite number of time-steps and batch-size $n=O(d)$ even upon re-using the batches.The rigorous control of all these quantities is a difficult task a priori. One cannot, in particular, express the above sum as an expectation w.r.t independent samples $\mathbf{z}_\nu$ since the weights $W^{(1)}$ now depend on all the samples. Fortunately, this is precisely the difficulty solved by the DMFT equations through an effective stochastic process on the pre-activations that are decoupled across training examples. The rigorous analysis is detailed in App. A. The main lines of the DMFT equations are in Sec. 4.2.

Finally, note that while our proof uses the Gaussian data assumption, the heuristic argument hints that this is not crucial. Additionally, in any *real* dataset samples are very correlated, and thus a given sample (or a very similar one) may appear many times. In this case, even *single-pass* algorithms will behave as predicted by our approach. We thus believe it describes a more realistic scenario than the pure single pass theories with fresh *i.i.d.* data.

4.2. Characterization of the dynamics

Re-using batches at each gradient step requires keeping track of the pre-activations of the parameters. Since the number of pre-activations and the dimensions of the parameters grows with d, we need a low-dimensional effective dynamics characterizing the quantities of interests such as the overlaps between the student and target parameters. DMFT provides

such an effective dynamics through a set of coupled stochastic processes $\boldsymbol{\theta}^{(t)} \in \mathbb{R}^p$ and $\mathbf{h}^{(t)} \in \mathbb{R}^p$ representing the joint-distributions of the student, teacher parameters $W^{(t)}, W^\star$. and the student, teacher pre-activations respectively.

We derive the equations and prove their applicability to our setting using existing results in (Celentano et al., 2021; Gerbelot et al., 2022). Asymptotically, for $d \to \infty$ with $n = \alpha d$, the joint distribution of the student and teacher pre-activations (for each sample), $\mathbf{h}_\nu^{(t)} = W^{(t)}\mathbf{z}_\nu/\sqrt{d} \in \mathbb{R}^p$ and $\mathbf{h}_\nu^\star = W^\star \mathbf{z}_\nu/\sqrt{d} \in \mathbb{R}^k$ converge in distribution to samples from the stochastic process $\mathbf{h}^{(t)}$ and the standard normal variable $\mathbf{h}^\star$. Similarly, the joint distribution of each component of the student and teacher weights $W_i^{(t)} \in \mathbb{R}^p, W_i^\star \in \mathbb{R}^k$ with $i \in [d]$ converge in distribution to samples from the stochastic process $\boldsymbol{\theta}^{(t)}$ and the standard normal variable $\boldsymbol{\theta}^\star$.

$$\boldsymbol{\theta}^{(t+1)} = \left(1 - \eta\lambda - \eta\Lambda^{(t)}\right)\boldsymbol{\theta}^{(t)} + \eta \sum_{\tau=0}^{t-1} R_{\mathcal{L}}^{(t,\tau)} \boldsymbol{\theta}^{(\tau)} - \eta g^{(t)} \boldsymbol{\theta}^\star + \eta \sum_{\tau=0}^{t-1} \tilde{R}_{\mathcal{L}}^{(t,\tau)} \boldsymbol{\theta}^\star + \eta \boldsymbol{u}^{(t)} \tag{19}$$

$$\boldsymbol{h}^{(t)} = -\eta \sum_{\tau=0}^{t-1} R_\theta^{(t,\tau)} \nabla_{\boldsymbol{h}} \mathcal{L}(\boldsymbol{h}^{(\tau)}, \boldsymbol{h}^\star) + \boldsymbol{\omega}^{(t)} \tag{20}$$

Notice that the formula above is the high dimensional equivalent of the gradient descent update (5). Here $\mathbf{u}^{(t)}$ and $(\boldsymbol{\omega}^{(t)}, \boldsymbol{\theta}^\star)$ are zero mean Gaussian Process with covariances $C_{\mathcal{L}}^{(t,\tau)}$ and $\Omega^{(t,\tau)}$ respectively, with

$$C_{\mathcal{L}}^{(t,\tau)} = \alpha \mathbb{E}_{\boldsymbol{h}^{(t)}, \boldsymbol{h}^\star}\left[\nabla_{\boldsymbol{h}} \mathcal{L}(\boldsymbol{h}^{(t)}, \boldsymbol{h}^\star) \nabla_{\boldsymbol{h}} \mathcal{L}(\boldsymbol{h}^{(\tau)}, \boldsymbol{h}^\star)^\top\right]$$

$$\Omega^{(t,\tau)} = \begin{bmatrix} C_\theta^{(t,\tau)} & M^{(t)} \\ M^{(\tau)} & 1 \end{bmatrix} = \mathbb{E}_{\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^\star}\left[\begin{pmatrix} \boldsymbol{\theta}^{(t)} \\ \boldsymbol{\theta}^\star \end{pmatrix} \begin{pmatrix} \boldsymbol{\theta}^{(\tau)} \\ \boldsymbol{\theta}^\star \end{pmatrix}^\top\right]$$

the matrix $\Lambda^{(t)}$ can be viewed as an "effective regularization" on the parameters. $\Lambda^{(t)}$ and the projected gradient $g^{(t)}$ converge in probability to:

$$\Lambda^{(t)} = \alpha \mathbb{E}_{\boldsymbol{h}^{(t)}, \boldsymbol{h}^\star}\left[\nabla_{\mathbf{h}}^2 \mathcal{L}\left(\mathbf{h}^{(t)}, \boldsymbol{h}^\star\right)\right], \tag{21}$$

$$g^{(t)} = \alpha \mathbb{E}_{\boldsymbol{h}^{(t)}, \boldsymbol{h}^\star}\left[\nabla_{\mathbf{h}} \mathcal{L}\left(\mathbf{h}^{(t)}, \boldsymbol{h}^\star\right) \mathbf{h}^{\star\top}\right] \tag{22}$$

The memory kernels $R_{\mathcal{L}}^{(t,\tau)}$, $\tilde{R}_{\mathcal{L}}^{(t,\tau)}$, $R_\theta^{(t,\tau)}$ are defined as:

$$R_\theta^{(t,\tau)} = \mathbb{E}_{\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^\star}\left[\frac{\partial \boldsymbol{\theta}^{(t)}}{\partial \mathbf{u}^{(\tau)}}\right],$$

$$R_{\mathcal{L}}^{(t,\tau)} = \alpha \mathbb{E}_{\boldsymbol{h}^{(t)}, \boldsymbol{h}^\star}\left[\frac{\partial \nabla_{\mathbf{h}} \mathcal{L}\left(\mathbf{h}^{(t)}, \boldsymbol{h}^\star\right)}{\partial \boldsymbol{\omega}^{(\tau)}}\right], \tag{23}$$

$$\tilde{R}_{\mathcal{L}}^{(t,\tau)} = \alpha \mathbb{E}_{\boldsymbol{h}^{(t)}, \boldsymbol{h}^\star}\left[\frac{\partial \nabla_{\mathbf{h}} \mathcal{L}\left(\mathbf{h}^{(t)}, \boldsymbol{h}^\star\right)}{\partial (\boldsymbol{\theta}^\star)^{(\tau)}}\right], \tag{24}$$

and $R_{\mathcal{L}}^{(t,t)} = \tilde{R}_{\mathcal{L}}^{(t,t)} = 0$, $R_\theta^{(t,t)} = 1$. Finally, the low dimensional projections of the weights $M^{(t)}$ will obey

$$M^{(t+1)} = (1 - \eta\lambda) M^{(t)} - \eta g^{(t)}. \tag{25}$$

Notice that these definitions are well-posed because of the causal structure of the gradient descent upgrades, and by extension of (19): the distribution of $(\boldsymbol{\theta}^{(t+1)}, \mathbf{h}^{(t+1)})$ is completely determined by $\{(\boldsymbol{\theta}^{(\tau)}, \mathbf{h}^{(\tau)})\}_{\tau \in [t]}$ and the auxiliary quantities in eqs. (21, 23). Iterating backwards we reach the initial condition $\boldsymbol{\theta}^{(0)}$, which is a simple function of the data distribution and the initial conditions of the weights. For additional details we refer to App. A. Notice that it is also possible to write this set of equations as a function of a single stochastic process on $\mathbf{h}$, as in App. C.

Sketch of proof of the hidden progress — Finally, we explain how the DMFT equations relate to the phenomenon in Sec. 4.1 and allow us to prove Th. 3.2. The term $\nabla_{\mathbf{h}^{(1)}} \mathcal{L}\left(\mathbf{h}^{(1)}\right)$ in (20) precisely corresponds to the contribution to pre-activation of a point $\mathbf{z}^\nu$ (App. A.4) from the gradient at the same point $\mathbf{z}^\nu$. As we discussed in Section 4.1, this term induces a dependence between $\mathbf{h}_\nu^{(1)}$ and $\mathbf{h}_\nu^\star$ even when the overlaps $M^{(t)}$ are 0. At time $T = 1$, the response term simplifies to $R_\theta^{(1,0)} = \mathbf{I_d}$ and the pre-activations can be expressed as the random variable $\nabla_{\mathbf{h}^{(t)}} \mathcal{L}\left(\mathbf{h}^{(t)}\right)$ with added Gaussian noise. Analogous to section 4.1, we denote by $M_{\mathbf{v}^*}^{(t)}$ the limiting value of the overlaps $\frac{1}{d} W^{(t)} \mathbf{v}^*$ for some $\mathbf{v}^* = (\mathbf{u}^*)^\top W^*$ with $\mathbf{u}^* \in \mathbb{R}^p$. Propagating the equations over the first two steps, and using Equation (25), we show that $M_{\mathbf{v}^*}^{(2)}$ can be expressed as an expectation w.r.t the pre-activations $\mathbf{h}^\star$ of a function dependent on the target $g^\star$, the second layer $\mathbf{a}$, and the activation function σ:

$$M_{\mathbf{v}^*}^{(2)} = (1 - \eta\lambda) M_{\mathbf{v}^*}^{(1)} + \eta\alpha \mathbb{E}_{\mathbf{h}^\star}\left[F_{\sigma,a}(g^\star(\mathbf{h}^\star)) \mathbf{h}^{\star\top}\right] \mathbf{u}^\star. \tag{26}$$

The function $F_{\sigma,a}$ is described in App.A.5, Eq.(89). Finally, we show an equivalence between the condition $\mathbb{E}_{\mathbf{h}^\star}\left[F_{\sigma,a}(g^\star(\mathbf{h}^\star)) \mathbf{h}^{\star\top}\right] \mathbf{u}^\star = 0$ to the condition $\mathbb{E}_{\mathbf{z}}\left[F(f^\star(\mathbf{z}))\langle \mathbf{v}^\star, \mathbf{z}\rangle\right] = 0$ for general F in definition 3.1.

General multi-pass schemes — While Theorem 3.2 considers finite number of updates with the same batch of data for each step, it can be naturally generalized to other setups involving multiple-passes over a finite-number of minibatches of size $\mathcal{O}(d)$. For instance, one can cycle over distinct minibatches with each cycle constituting one epoch or pass through the dataset. Theorem 3.2 remains valid under such a setup with the onset of weak-recovery shifting to the start of the second epoch instead of the second gradient step. We provide a sketch of this extension in Appendix B. On the other hand, if the minibatches are sampled with replacement from the dataset, the weak recovery still starts at the second gradient step. We illustrate this in Fig.4 (in appendix). Furthermore, we empirically observe that the phenomenon holds even when considering the limit of minibatch size 1 (Figure 5 in appendix). Proving this, however, remains out of the reach of the present technique.

Acknowledgements

We thank Cedric Gerbelot, Bruno Loureiro, and Ludovic Stephan for insightful discussions. We also acknowledge funding from the Swiss National Science Foundation grant SNFS OperaGOST (grant number 200390), and SMArtNet (grant number 212049).

Impact Statement

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

References

Abbe, E., Boix-Adsera, E., Brennan, M. S., Bresler, G., and Nagaraj, D. The staircase property: How hierarchical structure can guide deep learning. *Advances in Neural Information Processing Systems*, 34:26989–27002, 2021.

Abbe, E., Boix-Adsera, E., and Misiakiewicz, T. 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*, pp. 4782–4887. PMLR, 2022.

Abbe, E., Boix-Adsera, E., and Misiakiewicz, T. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics, 2023.

Agoritsas, E., Biroli, G., Urbani, P., and Zamponi, F. Out-of-equilibrium dynamical mean-field equations for the perceptron model. *Journal of Physics A: Mathematical and Theoretical*, 51(8):085002, 2018.

Andrews, G. E. *Special functions*. Cambridge University Press, 2004.

Aubin, B., Maillard, A., Barbier, J., Krzakala, F., Macris, N., and Zdeborová, L. The committee machine: computational to statistical gaps in learning a two-layers neural network. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124023, December 2019. ISSN 1742-5468. doi: 10.1088/1742-5468/ab43d2. URL http://dx.doi.org/10.1088/1742-5468/ab43d2.

Ba, J., Erdogdu, M. A., Suzuki, T., Wang, Z., Wu, D., and Yang, G. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 37932–37946. Curran Associates, Inc., 2022.

Ba, J., Erdogdu, M. A., Suzuki, T., Wang, Z., and Wu, D. Learning in the presence of low-dimensional structure: a spiked random matrix perspective. In *Neurips 2023*, 2023.

Barbier, J., Krzakala, F., Macris, N., Miolane, L., and Zdeborová, L. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

Bayati, M. and Montanari, A. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57 (2):764–785, 2011.

Ben Arous, G., Guionnet, A., et al. Symmetric langevin spin glass dynamics. *The Annals of Probability*, 25(3): 1367–1422, 1997.

Ben Arous, G., Gheissari, R., and Jagannath, A. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22(106):1–51, 2021.

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

Bietti, A., Bruna, J., and Pillaud-Vivien, L. On learning gaussian multi-index models with gradient flow. *arXiv preprint arXiv:2310.19793*, 2023.

Bolthausen, E. An iterative construction of solutions of the tap equations for the sherrington–kirkpatrick model. *Communications in Mathematical Physics*, 325(1):333–366, 2014.

Bordelon, B., Canatar, A., and Pehlevan, C. Spectrum dependent learning curves in kernel regression and wide neural networks. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1024–1034. PMLR, 13–18 Jul 2020.

Bouchaud, J.-P., Cugliandolo, L. F., Kurchan, J., and Mézard, M. Out of equilibrium dynamics in spin-glasses and other glassy systems. *Spin glasses and random fields*, 12:161, 1998.

Celentano, M., Cheng, C., and Montanari, A. The high-dimensional asymptotics of first order methods with random data. *arXiv:2112.07572*, 2021.

Chen, S. and Meka, R. Learning polynomials in few relevant dimensions. In *Conference on Learning Theory*, pp. 1161–1227. PMLR, 2020.

Chen, S., Klivans, A., and Meka, R. Efficiently learning one hidden layer relu networks from queries. *Advances in Neural Information Processing Systems*, 34:24087–24098, 2021.

Chen, S., Gollakota, A., Klivans, A., and Meka, R. Hardness of noise-free learning for two-hidden-layer neural networks. *Advances in Neural Information Processing Systems*, 35:10709–10724, 2022.

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

Cugliandolo, L. F. Dynamics of glassy systems. In *Slow Relaxations and nonequilibrium dynamics in condensed matter*. Springer, 2003.

Cui, H., Loureiro, B., Krzakala, F., and Zdeborová, L. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 10131–10143. Curran Associates, Inc., 2021.

Damian, A., Lee, J., and Soltanolkotabi, M. Neural networks can learn representations with gradient descent. In Loh, P.-L. and Raginsky, M. (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 5413–5452. PMLR, 02–05 Jul 2022.

Damian, A., Nichani, E., Ge, R., and Lee, J. D. Smoothing the Landscape Boosts the Signal for SGD: Optimal Sample Complexity for Learning Single Index Models. Technical report, Princeton, May 2023. arXiv:2305.10633 [cs, math, stat] type: article.

Dandi, Y., Krzakala, F., Loureiro, B., Pesce, L., and Stephan, L. How two-layer neural networks learn, one (giant) step at a time. *arXiv preprint arXiv:2305.18270*, 2023.

Diakonikolas, I., Kane, D. M., Kontonis, V., and Zarifis, N. Algorithms and sq lower bounds for pac learning one-hidden-layer relu networks. In *Conference on Learning Theory*, pp. 1514–1539. PMLR, 2020.

Dietrich, R., Opper, M., and Sompolinsky, H. Statistical mechanics of support vector networks. *Phys. Rev. Lett.*, 82:2975–2978, Apr 1999. doi: 10.1103/PhysRevLett.82.2975.

Eissfeller, H. and Opper, M. New method for studying the dynamics of disordered spin systems without finite-size effects. *Physical review letters*, 68(13):2094, 1992.

Eissfeller, H. and Opper, M. Mean-field Monte Carlo approach to the Sherrington-Kirkpatrick model with asymmetric couplings. *Physical Review E*, 50(2):709, 1994.

Georges, A., Kotliar, G., Krauth, W., and Rozenberg, M. J. Dynamical mean-field theory of strongly correlated fermion systems and the limit of infinite dimensions. *Reviews of Modern Physics*, 68(1):13, 1996.

Gerbelot, C., Troiani, E., Mignacco, F., Krzakala, F., and Zdeborova, L. Rigorous dynamical mean field theory for stochastic gradient descent methods. *arXiv preprint arXiv:2210.06591*, 2022.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Limitations of lazy training of two-layers neural network. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. When do neural networks outperform kernel methods? In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14820–14830. Curran Associates, Inc., 2020.

Goel, S., Gollakota, A., Jin, Z., Karmalkar, S., and Klivans, A. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. In *International Conference on Machine Learning*, pp. 3587–3596. PMLR, 2020.

Kearns, M. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.

Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mezard, M., and Zdeborová, L. Learning curves of generic features maps for realistic datasets with a teacher-student model. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 18137–18151. Curran Associates, Inc., 2021.

Maillard, A., Loureiro, B., Krzakala, F., and Zdeborová, L. Phase retrieval in high dimensions: Statistical and computational phase transitions. *arXiv preprint arXiv:2006.05228*, 2020.

Mannelli, S. S. and Urbani, P. Just a momentum: Analytical study of momentum-based acceleration methods in paradigmatic high-dimensional non-convex problems. *NeurIPS*, 2021.

Mannelli, S. S., Biroli, G., Cammarota, C., Krzakala, F., and Zdeborová, L. Who is afraid of big bad minima? analysis of gradient-flow in spiked matrix-tensor models. In *Advances in Neural Information Processing Systems*, pp. 8676–8686, 2019a.

Mannelli, S. S., Krzakala, F., Urbani, P., and Zdeborova, L. Passed & spurious: Descent algorithms and local minima in spiked matrix-tensor models. In *international conference on machine learning*, pp. 4333–4342, 2019b.

Mannelli, S. S., Biroli, G., Cammarota, C., Krzakala, F., Urbani, P., and Zdeborová, L. Marvels and pitfalls of the langevin algorithm in noisy high-dimensional inference. *Physical Review X*, 10(1):011057, 2020.

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

Mignacco, F. and Urbani, P. The effective noise of stochastic gradient descent. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(8):083405, aug 2022. doi: 10.1088/1742-5468/ac841d. URL https://doi.org/10.1088/1742-5468/ac841d.

Mignacco, F., Krzakala, F., Urbani, P., and Zdeborová, L. Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification. *Advances in Neural Information Processing Systems*, 33:9540–9550, 2020.

Mignacco, F., Urbani, P., and Zdeborová, L. Stochasticity helps to navigate rough landscapes: comparing gradient-descent-based algorithms in the phase retrieval problem. *Machine Learning: Science and Technology*, 2(3):035029, 2021.

Moniri, B., Lee, D., Hassani, H., and Dobriban, E. A theory of non-linear feature learning with one gradient step in two-layer neural networks. *arXiv preprint arXiv:2310.07891*, 2023.

Montanari, A. and Saeed, B. N. Universality of empirical risk minimization. In Loh, P.-L. and Raginsky, M. (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4310–4312. PMLR, 02–05 Jul 2022.

Mousavi-Hosseini, A., Wu, D., Suzuki, T., and Erdogdu, M. A. Gradient-based feature learning under structured data. *arXiv preprint arXiv:2309.03843*, 2023.

Rotskoff, G. and Vanden-Eijnden, E. 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.

Roy, F., Biroli, G., Bunin, G., and Cammarota, C. Numerical implementation of dynamical mean field theory for disordered systems: application to the lotka–volterra model of ecosystems. *Journal of Physics A: Mathematical and Theoretical*, 52(48):484001, November 2019. ISSN 1751-8121. doi: 10.1088/1751-8121/ab1f32. URL http://dx.doi.org/10.1088/1751-8121/ab1f32.

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

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

Sompolinsky, H. and Zippelius, A. Dynamic theory of the spin-glass phase. *Phys. Rev. Lett.*, 47:359–362, Aug 1981.

Sompolinsky, H., Crisanti, A., and Sommers, H. J. Chaos in random neural networks. *Phys. Rev. Lett.*, 61:259–262, Jul 1988.

Zweig, A. and Bruna, J. Symmetric single index learning. *arXiv preprint arXiv:2310.02117*, 2023.

A. Mathematical Proofs

A.1. Notations

We use the asymptotic notation $f_1(d) = \Theta_d(f_1(d))$ to denote $c\,|f_1(d)| \leq |f_2(d)| \leq C\,|f_1(d)|$ for some constants $c, C > 0$ and large enough d. Similarly, $f_1(d) = o_d(f_1(d))$ denotes $f_1(d) \leq c(f_1(d))$ for any constant $c > 0$ and large enough d. We use $\xrightarrow[P]{d,n\to\infty}$, $\xrightarrow[D]{d,n\to\infty}$ to denote convergence in probability and convergence in distribution respectively as $d, n \to \infty$ with $n/d = \alpha > 0$. We denote subspaces and linear operators, matrices on $\mathbb{R}^d$ through uppercase letters $A, B, C, \cdots$. For any subspace $A \in \mathbb{R}^d$, we denote by $A_\perp$, its orthogonal complement, i.e the subspace of vectors orthogonal to all $\mathbf{v} \in A$.

A.2. DMFT and iterative conditioning

Unlike online SGD, the preactivations after multiple steps no longer remain Gaussian since the weights become dependent on the data. This prevents marginalizing over the orthogonal components over the preactivations and relating the learning of new directions to the Hermite decomposition of the target function. Our proof circumvents these issues by utilizing a simpler effective process that decouples the pre-activations for different samples. The effective process is obtained using a rigorous version of the Dynamical Mean Field Theory derived in (Montanari & Saeed, 2022) and (Gerbelot et al., 2022).

The derivation of Dynamical Mean Field Theory in the above works has the following essential elements:

1. Iterative conditioning: The proof in (Gerbelot et al., 2022; Montanari & Saeed, 2022) for obtaining the DMFT equations relies on the observation that the gradient descent algorithm in Equation (27) for a finite-number of iterations can be described completely through projections of the inputs design matrix $\mathbf{Z} \in \mathbb{R}^{n\times d}$ along a finite number of vectors in $\mathbb{R}^n, \mathbb{R}^d$. The iterative conditioning technique (Bolthausen, 2014; Bayati & Montanari, 2011) then involves replacing the components of $\mathbf{Z} \in \mathbb{R}^{n\times d}$ along directions orthogonal to these projections by independent Gaussian random variables. This leads to a non-Markovian structure in the effective processes for the activations, parameters.
2. The concentration of finite-dimensional order parameters such as overlaps of the neuron parameters with the teacher neurons/subspace as well as expectations w.r.t the empirical measure of the pre-activations and parameters.

Using the above elements, DMFT provides a low-dimensional effective dynamics characterizing the limiting joint empirical measure of the student parameters, as well as the pre-activations. We illustrate the proof for the activations after the first gradient step, illustrating the relationship with the "hidden progress" described in section 4.1

Let $\mathbf{H}^{(t)} = \frac{1}{\sqrt{d}}\mathbf{Z}(\mathbf{W}^{(t)})^\top$ denote the $n \times p$ matrix of pre-activations at time t. Similarly, let $\mathbf{H}^* \in \mathbb{R}^{n\times k}$ denote the matrix of input activations in the target function. We denote by $\nabla_{\mathbf{H}}\mathcal{L}(\mathbf{H}^*, \mathbf{H}^{(t)}) \in \mathbb{R}^{n\times p}$, the matrix derivative of $\mathcal{L}$ w.r.t the corresponding entries of the pre-activations matrix $\mathbf{H}^{(t)}$. After each gradient update, $\mathbf{W}^{(t)}$ and the preactivations $\mathbf{H}^{(t)}$ gain a dependence on $\mathbf{Z}$. The Iterative conditioning technique works around this dependence by conditioning on the sigma algebra generated by $\mathbf{H}^{(t)}, \mathbf{H}^*$ and $\mathbf{W}^{(t)}$ instead of on $\mathbf{Z}$. Since $\mathbf{Z}$ interacts with $\mathbf{W}^{(t)}$ and $\mathbf{H}^{(t)}$ only through projections (along right with $\mathbf{W}^{(t)}$ and left with $\nabla_{\mathbf{H}}\mathcal{L}(\mathbf{H}^*, \mathbf{H}^{(t)})$ respectively), the conditioning allows the components of $\mathbf{Z}$ orthogonal to $\nabla_H\mathcal{L}(\mathbf{H}^{(*)}, \mathbf{H}^{(t)})$ and $\mathbf{W}^{(t)}$ to be replaced by independent Gaussian entries.

For the first-gradient step, we only require conditioning on $\mathbf{H}^{(0)}, \mathbf{H}^*, \mathbf{W}^0$.

From ((5)), we obtain the following update for $\mathbf{H}^{(1)}$:

$$\mathbf{H}^{(1)} = \mathbf{H}^{(0)} + \frac{\eta}{d}\mathbf{Z}(\mathbf{Z})^\top\nabla_{\mathbf{H}}\mathcal{L}(\mathbf{H}^*, \mathbf{H}^{(0)})\mathbf{a}^\top, \tag{27}$$

Let $\bar{\mathbf{W}}^{(0)} = \begin{pmatrix}\mathbf{W}^{(0)} \\ \mathbf{W}^\star\end{pmatrix} \in \mathbb{R}^{p+k\times d}$ By the equivalence of projection and conditioning for Gaussian random variables, we have that the following inequality holds in distribution:

$$\mathbf{Z}|_{\mathbf{H}^0,\mathbf{H}^*,\bar{\mathbf{W}}^{(0)}} \stackrel{d}{=} \mathbf{Z}P_w^\top + \tilde{\mathbf{Z}}(P_w^\perp)^\top, \tag{28}$$

where $\tilde{\mathbf{Z}}$ is independent of $\mathbf{Z}$ and P_w denote the projection operator along $\bar{\mathbf{W}}^{(0)}, \bar{\mathbf{W}}^\star$, defined as:

$$P_w = \bar{\mathbf{W}}^{(0)}(\bar{\mathbf{W}}^{(0)}(\bar{\mathbf{W}}^{(0)})^\top)^{-1}(\bar{\mathbf{W}}^{(0)})^\top.$$

Substituting in Equation ((27)), we obtain:

$$\mathbf{H}^{(1)} \overset{d}{=} \mathbf{H}^{(0)} + \eta \frac{1}{d} \tilde{\mathbf{Z}} (P_w^{\perp})^\top P_w^{\perp} (\tilde{\mathbf{Z}})^\top \nabla_{\mathbf{H}} \mathcal{L}(\mathbf{H}^*, \mathbf{H}^{(0)}) \mathbf{a}^\top + \eta \frac{1}{d} \mathbf{H_0} (\bar{\mathbf{W}}^{(0)} (\bar{\mathbf{W}}^{(0)})^\top)^{-1} (\mathbf{H_0})^\top \nabla_{\mathbf{H}} \mathcal{L}(\mathbf{H}^*, \mathbf{H}^{(0)}) \mathbf{a}^\top \quad (29)$$

Since the projection, P_w is along a low-dimensional subspace of dimension at most p, we have $P_w^{\perp} \approx \mathbf{I}_d$. One can therefore show that $\frac{1}{\sqrt{d}} \tilde{\mathbf{Z}} (P_w^{\perp})^\top P_w^{\perp} (\tilde{\mathbf{Z}})^\top \mathbf{u}$ converges in probability to $\tilde{\mathbf{Z}} \tilde{\mathbf{Z}}^\top \mathbf{u}$. for any deterministic $\mathbf{u} \in \mathbb{R}^d$ with $\|u\| = \mathcal{O}(\sqrt{d})$. Applying it to the vector $\mathbf{u} = \mathcal{L}(\mathbf{H}^*, \mathbf{H}^{(0)})$, conditioned on $\mathbf{H}^*, \mathbf{H}^{(0)}$, we obtain that:

$$\frac{1}{\sqrt{d}} \left\| \frac{1}{d} \tilde{\mathbf{Z}} (P_w^{\perp})^\top P_w^{\perp} (\tilde{\mathbf{Z}})^\top \nabla_{\mathbf{H}} \mathcal{L}(\mathbf{H}^*, \mathbf{H}^{(0)}) - \frac{1}{d} \tilde{\mathbf{Z}} (\tilde{\mathbf{Z}})^\top \nabla_{\mathbf{H}} \mathcal{L}(\mathbf{H}^*, \mathbf{H}^{(0)}) \right\|_F \xrightarrow[P]{n,d \to \infty} 0. \quad (30)$$

Now, the diagonal entries of $\frac{1}{d} \tilde{\mathbf{Z}} (\tilde{\mathbf{Z}})^\top$ convergence in probability to 1 due to the concentration of norms of Gaussian random vectors. This results in the term $\mathcal{L}(\mathbf{H}^*, \mathbf{H}^{(0)})$. This term is precisely the one responsible for the "hidden progress" explained in section 4.1, corresponding to the term in Equation 18. Since $\tilde{\mathbf{Z}}$ is independent of $\mathbf{W}^{(0)}$ and $\mathbf{H}^{(0)}, \mathbf{H}^*$, by central limit-theorem for sub-Gaussian random variables, the remaining off-diagonal terms can be shown to converge to Gaussian noise independent of $\mathbf{H}^{(0)}, \mathbf{H}^*$ with variance $\left\| \nabla_{\mathbf{h}} \mathcal{L}(\mathbf{h}^*, \mathbf{h}^{(0)}) \right\|^2$.

Lastly, the third term in Equation (29) can be shown to converge to Gaussian noise correlated with corresponding entries of $\mathbf{H_0}, \mathbf{H}^*$. Specifically, by removing the conditioning on $\mathbf{H_0}$, we have through law of large numbers and Stein's Lemma, we have that the term $\frac{1}{d} (\mathbf{H_0})^\top \nabla_{\mathbf{H}} \mathcal{L}(\mathbf{H}^*, \mathbf{H}^{(0)})$ converges in probability to $(\mathbf{W}^{(0)} (\mathbf{W}^{(0)})^\top) \mathbb{E} \left[\nabla_{\mathbf{h}}^2 \mathcal{L}(\mathbf{h}^*, \mathbf{h}^{(0)}) \right]$. Therefore, we obtain:

$$\frac{1}{\sqrt{d}} \left\| \frac{1}{d} \mathbf{H_0} (\mathbf{W}^{(0)} (\mathbf{W}^{(0)})^\top)^{-1} (\mathbf{H_0})^\top \nabla_{\mathbf{H}} \mathcal{L}(\mathbf{H}^*, \mathbf{h}^{(0)}) \mathbf{a}^\top - \mathbf{H_0} \mathbb{E} \left[\nabla_{\mathbf{h}}^2 \mathcal{L}(\mathbf{h}^*, \mathbf{h}^{(0)}) \right]) \mathbf{a}^\top \right\| \xrightarrow[P]{n,d \to \infty} 0 \quad (31)$$

Proceeding similarly, one obtains low-dimensional effective processes for $\mathbf{W}^{(t)}, \mathbf{H}^{(t)}$ for any time $t \in \mathbb{N}$. In the following section, we derive the resulting DMFT dynamics for the setup considered in Section 1 through a reduction to the result in (Gerbelot et al., 2022). We refer to (Gerbelot et al., 2022; Celentano et al., 2021) for detailed proofs based on the above technique.

A.3. Derivation of the exact asymptotics

We start by stating a general consequence of the main result in (Gerbelot et al., 2022).

Theorem A.1 (Corollary of Theorem 3.2 in (Gerbelot et al., 2022))**.** *Let $W^0 \in \mathbb{R}^{q \times d}$ be a sequence of matrices such that the overlap matrix*

$$\frac{1}{d} W^0 (W^0)^\top, \quad (32)$$

converges almost surely. Consider a dynamics of the form:

$$W^{(t+1)} = W^{(t)} - \eta \lambda W^{(t)} - \eta \frac{1}{\sqrt{d}} \sum_{\nu=1}^{n} F\left(\frac{W^{(t)} \mathbf{z}_\nu}{\sqrt{d}} \right) \mathbf{z}_\nu^\top \quad (33)$$

where $F : \mathbb{R}^q \to \mathbb{R}^q$ is pseudo-Lipshitz of finite-order. Then the empirical measure of the weights $\mathbf{w}_i^{(t)}$ converges in distribution to the weight process $\theta_i^{(t)}$ and the empirical measure of the preactivations $\mathbf{h}_\nu^{(t)}$ converges in distribution to that of the preactivation process $\mathbf{h}^{(t)}$, defined as

$$\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)} = -\eta \left(\lambda + \Lambda^{(t)} \right) \boldsymbol{\theta}^{(t)} + \eta \sum_{\tau=0}^{t-1} R_\ell^{(t,\tau)} \boldsymbol{\theta}^{(\tau)} + \eta \boldsymbol{u}^{(t)} \quad (34)$$

$$\boldsymbol{h}^{(t)} = -\eta \sum_{\tau=0}^{t-1} R_\theta^{(t,\tau)} F(\boldsymbol{h}^{(\tau)}) + \boldsymbol{\omega}^{(t)} \quad (35)$$

were we have

$$\Lambda^{(t)} = \alpha \mathbb{E} \left[\nabla_{\boldsymbol{h}} F(\boldsymbol{h}^{(t)}) \right] \quad (36)$$

$$R_\theta^{(t,\tau)} = \mathbb{E}\left[\frac{\partial\,\theta^{(t)}}{\partial\,\tau^{(\tau)}}\right] \tag{37}$$

$$R_\ell^{(t,\tau)} = \alpha\mathbb{E}\left[\frac{\partial\,F(\mathbf{h}^{(t)})}{\partial\,\omega^{(\tau)}}\right]. \tag{38}$$

Finally, $\boldsymbol{u}^{(t)}$ and $\boldsymbol{\omega}^{(t)}$ are zero-mean Gaussian processes respectively with covariances given by $C_\ell^{(t,\tau)}$ and $C_\theta^{(t,\tau)}$:

$$C_\ell^{(t,\tau)} = \alpha\mathbb{E}\left[\boldsymbol{u}^{(t)}\left(\boldsymbol{u}^{(\tau)}\right)^\top\right] = \alpha\mathbb{E}\left[F(\boldsymbol{r}^{(t)})F(\boldsymbol{r}^{(\tau)})^\top\right] \tag{39}$$

$$C_\theta^{(t,\tau)} = \mathbb{E}\left[\boldsymbol{\omega}^{(t)}\left(\boldsymbol{\omega}^{(\tau)}\right)^\top\right] = \mathbb{E}\left[\boldsymbol{\theta}^{(t)}\left(\boldsymbol{\theta}^{(\tau)}\right)^\top\right] \tag{40}$$

The convergence in distribution of the empirical measures holds in the following sense: For any $t \in \mathbb{N}$, and any pseudo-Lipschitz functions $\psi : \mathbb{R}^{p(t+1)} \to \mathbb{R}$ and $\phi : \mathbb{R}^{pt} \to \mathbb{R}$:

$$\frac{1}{d}\sum_{i=1}^{d}\psi((W_i^{(0)},...,W_i^{(t)})) \xrightarrow[n,d\to\infty]{\text{w.h.p.}} \mathbb{E}\left[\psi(\boldsymbol{\theta}^{(0)},...,\boldsymbol{\theta}^{(t)})\right], \tag{41}$$

$$\frac{1}{n}\sum_{\nu=1}^{n}\phi((\mathbf{h}_\nu^{(0)},...,\mathbf{h}_\nu^{(t-1)})) \xrightarrow[n,d\to\infty]{\text{w.h.p.}} \mathbb{E}\left[\phi(\mathbf{h}^{(0)},...,\mathbf{h}^{(t-1)})\right], \tag{42}$$

where $W_i^{(t)}$ denotes the i_{th} column of $W^{(t)}$

Notice that it's possible to also define R_θ and R_ℓ in the following way. Let $R_\theta^{(t,\tau)}$ is the solution of the difference equation

$$R_\theta^{(t+1,\tau)} - R_\theta^{(t,\tau)} = -\,\eta\,\left(\lambda + \Lambda^{(t)}\right)R_\theta^{(t,\tau)} + \,\eta\sum_{s=\tau}^{t-1}R_\ell^{(t,s)}R_\theta^{(\tau,s)} \tag{43}$$

with boundary conditions

$$R_\theta^{(t,t)} = 1\,, \tag{44}$$

$$R_\theta^{(t+1,t)} = 1 -\,\eta\,\Lambda^{(t)}\,, \tag{45}$$

while

$$R_\ell^{(t,\tau)} = \alpha\mathbb{E}\left[\nabla_{\boldsymbol{h}}F(\boldsymbol{h}^{(t)})\,T_\ell^{(t,\tau)}\right] \tag{46}$$

and $T_\ell^{(t,\tau)}$ is a collection of stochastic processes with distribution

$$T_\ell^{(t,\tau)} = R_\theta^{(t,\tau)}\nabla_{\boldsymbol{h}}F(\boldsymbol{h}^{(\tau)}) + \sum_{s=\tau+1}^{t-1}R_\theta^{(t,s)}T_\ell^{(s,\tau)} \tag{47}$$

and boundary conditions

$$T_\ell^{(t,t)} = 0\,, \tag{48}$$

$$T_\ell^{(t+1,t)} = 1 -\,\eta\,\Lambda^{(t)}\,, \tag{49}$$

The above result follows directly by substituting $\frac{1}{\sqrt{d}}\mathbf{Z}$ as $\mathbf{X}$ in Theorem 3.2 of (Gerbelot et al., 2022).

To obtain the limiting equations under the setting of gradient descent with teacher weights W^* in section 1, we utilize the generality of the update F in theorem A.1, which allows for a portion of the parameters (W^*) to remain unaffected. We obtain the following result, which generalize the former theorem to the setting of our paper:

Theorem A.2. *Consider the distribution over data defined in section 2 and an update rule on the weights of the form (5), i.e:*

$$\mathbf{w}_i^{(t+1)} = \mathbf{w}_i^{(t)} - \eta\lambda\mathbf{w}_i^{(t)} - \eta\sum_{\nu=1}^{n}\nabla_{\mathbf{w}_i^{(t)}}\mathcal{L}\left(\frac{W^{(t)}\mathbf{z}_\nu}{\sqrt{d}}, \frac{W^\star\mathbf{z}_\nu}{\sqrt{d}}\right), \tag{50}$$

Then under the assumptions of Theorem 3.2, as $d\to\infty$ with $n/d=\alpha>0$, the joint empirical measure of the coordinates of the student weights $\mathbf{w}_i^{(t)}$ and the teacher weights $W^\star$ converges in distribution to the stochastic process $\theta^{(t)}$ and the standard normal variable $\theta^\star$, in the sense of Theorem A.1. Similarly, the joint empirical measure of the student and teacher preactivations $\frac{\mathbf{w}_i^{(t)}\mathbf{z}_\nu}{\sqrt{d}}$, $\frac{\mathbf{w}_i^{\star}\mathbf{z}_\nu}{\sqrt{d}}$ converge in distribution to the stochastic process $\mathbf{h}^{(t)}$ and the standard normal variable $\mathbf{h}^\star$. $\boldsymbol{\theta}^{(t)}$ and $\mathbf{h}^{(t)}$ are defined recursively through the following equations:

$$\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)} = -\eta\left(\lambda+\Lambda^{(t)}\right)\boldsymbol{\theta}^{(t)} + \eta\sum_{\tau=0}^{t-1}R_\ell^{(t,\tau)}\boldsymbol{\theta}^{(\tau)} - \eta g^{(t)}\boldsymbol{\theta}^\star + \eta\sum_{\tau=0}^{t}\tilde{R}_\ell^{(t,\tau)}\boldsymbol{\theta}^\star + \eta\boldsymbol{u}^{(t)} \tag{51}$$

$$\boldsymbol{h}^{(t)} = -\eta\sum_{\tau=0}^{t-1}R_\theta^{(t,\tau)}\nabla_{\boldsymbol{h}}\ell(\boldsymbol{h}^{(\tau)},\boldsymbol{h}^\star) + \boldsymbol{\omega}^{(t)} \tag{52}$$

Here $\mathbf{u}^{(t)}, \boldsymbol{\theta}^\star$ and $(\boldsymbol{\omega}^{(t)}, \boldsymbol{h}^\star)$ are zero mean Gaussian Process with covariances $C_\ell^{(t,\tau)}$ and $\Omega^{(t,\tau)}$ respectively, given bY:

$$C_\ell^{(t,\tau)} = \alpha\mathbb{E}\left[\boldsymbol{u}^{(t)}\left(\boldsymbol{u}^{(\tau)}\right)^\top\right] = \alpha\mathbb{E}_{\boldsymbol{h}^{(t)},\boldsymbol{h}^\star}\left[\nabla_{\boldsymbol{h}}\mathcal{L}(\boldsymbol{h}^{(t)},\boldsymbol{h}^\star)\nabla_{\boldsymbol{h}}\mathcal{L}(\boldsymbol{h}^{(\tau)},\boldsymbol{h}^\star)^\top\right] \tag{53}$$

$$\Omega^{(t,\tau)} = \mathbb{E}\left[\begin{pmatrix}\boldsymbol{\omega}^{(t)}\\ \boldsymbol{h}^\star\end{pmatrix}\begin{pmatrix}\boldsymbol{\omega}^{(t)}\\ \boldsymbol{h}^\star\end{pmatrix}^\top\right] = \begin{bmatrix}C_\theta^{(t,\tau)} & M^{(t)}\\ M^{(t)} & 1\end{bmatrix}, \tag{54}$$

where $C_\theta^{(t,\tau)}, M^{(t)}$ are defined as:

$$\begin{bmatrix}C_\theta^{(t,\tau)} & M^{(t)}\\ M^{(t)} & 1\end{bmatrix} = \mathbb{E}_{\boldsymbol{\theta}^{(t)},\boldsymbol{\theta}^\star}\left[\begin{pmatrix}\boldsymbol{\theta}^{(t)}\\ \boldsymbol{\theta}^\star\end{pmatrix}\begin{pmatrix}\boldsymbol{\theta}^{(\tau)}\\ \boldsymbol{\theta}^\star\end{pmatrix}^\top\right] \tag{55}$$

The effective regularisation $\Lambda^{(t)}$ and the projected gradient $g^{(t)}$ concentrate to

$$\Lambda^{(t)} = \alpha\mathbb{E}_{\boldsymbol{h}^{(t)},\boldsymbol{h}^\star}\left[\nabla_{\mathbf{h}}^2\mathcal{L}\left(\mathbf{h}^{(t)},\boldsymbol{h}^\star\right)\right], \tag{56}$$

$$g^{(t)} = \alpha\mathbb{E}_{\boldsymbol{h}^{(t)},\boldsymbol{h}^\star}\left[\nabla_{\mathbf{h}}\mathcal{L}\left(\mathbf{h}^{(t)},\boldsymbol{h}^\star\right)\mathbf{h}^{\star\top}\right] \tag{57}$$

The memory kernels $R_\ell^{(t,\tau)}$, $\tilde{R}_\ell^{(t,\tau)}$, $R_\theta^{(t,\tau)}$ are defined by the partial derivatives with respect to the noise intended as functions, with the following identities valid for $t>\tau$

$$R_\theta^{(t,\tau)} = \mathbb{E}_{\boldsymbol{\theta}^{(t)},\boldsymbol{\theta}^\star}\left[\frac{\partial\,\theta^{(t)}}{\partial\,\tau^{(\tau)}}\right],$$

$$R_\ell^{(t,\tau)} = \alpha\mathbb{E}_{\boldsymbol{h}^{(t)},\boldsymbol{h}^\star}\left[\frac{\partial\,\nabla_{\mathbf{h}}\mathcal{L}\left(\mathbf{h}^{(t)},\boldsymbol{h}^\star\right)}{\partial\,\boldsymbol{\omega}^{(\tau)}}\right], \tag{58}$$

$$\tilde{R}_\ell^{(t,\tau)} = \alpha\mathbb{E}_{\boldsymbol{h}^{(t)},\boldsymbol{h}^\star}\left[\frac{\partial\,\nabla_{\mathbf{h}}\mathcal{L}\left(\mathbf{h}^{(t)},\boldsymbol{h}^\star\right)}{\partial\,(\boldsymbol{\omega}^\star)^{(\tau)}}\right], \tag{59}$$

and $R_\ell^{(t,t)} = \tilde{R}_\ell^{(t,t)} = 0$, $R_\theta^{(t,t)} = 1$. Finally, $M^{(t)}$ obeys the update equation

$$M^{(t+1)} = (1-\eta\lambda)M^{(t)} - \eta g^{(t)}, \tag{60}$$

where $g^{(t)}$ is defined as:

$$\alpha\mathbb{E}\left[\nabla_{\boldsymbol{h}}\ell(\boldsymbol{h}^{(t)})\,(\boldsymbol{h}^*)^\top\right] \tag{61}$$

Proof. Analogous to the embedding of planted vectors in (Celentano et al., 2021), we start by considering an a lifted dynamics defined by concating $W^{(t)}$ and W^*. First define $\tilde{W}^{(t)} \in \mathbb{R}^{(p+k)\times d}$ with update rule:

$$\begin{bmatrix} W^{(t+1)} \\ W^* \end{bmatrix} = \begin{bmatrix} W^{(t)} \\ W^* \end{bmatrix} - \eta \begin{bmatrix} \lambda W^{(t)} \\ 0 \end{bmatrix} - \eta \sum_{\nu=1}^{n} \begin{bmatrix} \nabla_{W^{(t)}} \mathcal{L}\left(\frac{W^{(t)}\boldsymbol{z}^\nu}{\sqrt{d}}\right) \\ 0 \end{bmatrix}, \tag{62}$$

The above form of updates can be seen to be a special case of Theorem A.1 with $q = p + k$ and $F : R^{p+k} \to \mathbb{R}^{p+k}$ given by:

$$F : \begin{pmatrix} \mathbf{h} \\ \mathbf{h}^\star \end{pmatrix} \to \begin{pmatrix} \nabla_{\mathbf{h}} \mathcal{L}(\mathbf{h}, \mathbf{h}^\star) \\ \mathbf{0} \end{pmatrix} \tag{63}$$

The assumptions on g^*, σ and on the distribution of $W^{(0)}, W^*$ imply that F is pseudo-Lipschitz of finite-order and that the overlap matrix at initialization converges to a limit. Therefore, theorem A.1, the effective process for the weights and the pre-activations is described by:

$$\begin{bmatrix} \boldsymbol{\theta}^{(t+1)} \\ \boldsymbol{\theta}^* \end{bmatrix} - \begin{bmatrix} \boldsymbol{\theta}^{(t)} \\ \boldsymbol{\theta}^* \end{bmatrix} = -\eta \begin{bmatrix} \lambda + \Lambda^{(t)} & \tilde{\Lambda}^{(t)} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}^{(t)} \\ \boldsymbol{\theta}^* \end{bmatrix} + \eta \sum_{\tau=0}^{t} \begin{bmatrix} R_\ell^{(t,\tau)} & \tilde{R}_\ell^{(t,\tau)} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}^{(\tau)} \\ \boldsymbol{\theta}^* \end{bmatrix} + \eta \begin{bmatrix} \boldsymbol{u}^{(t)} \\ 0 \end{bmatrix} \tag{64}$$

$$\begin{bmatrix} \boldsymbol{h}^{(t)} \\ \boldsymbol{h}^* \end{bmatrix} = -\eta \sum_{\tau=0}^{t-1} \begin{bmatrix} R_\theta^{(t,\tau)} & \tilde{R}_\theta^{(t,\tau)} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \nabla_{\boldsymbol{h}} \mathcal{L}(\boldsymbol{h}^{(\tau)}, \boldsymbol{h}^\star) \\ 0 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\omega}^{(t)} \\ \boldsymbol{\omega}^* \end{bmatrix} \tag{65}$$

Notice how these equations are simpler due to W^* not being updated in (62). We can simplify (65) and (64), obtaining:

$$\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)} = -\eta \left(\lambda + \Lambda^{(t)}\right) \boldsymbol{\theta}^{(t)} + \eta \sum_{\tau=0}^{t-1} R_\ell^{(t,\tau)} \boldsymbol{\theta}^{(\tau)} - \eta \tilde{\Lambda}^{(t)} \boldsymbol{\theta}^* + \eta \sum_{\tau=0}^{t} \tilde{R}_\ell^{(t,\tau)} \boldsymbol{\theta}^* + \eta \boldsymbol{u}^{(t)} \tag{66}$$

$$\boldsymbol{h}^{(t)} = -\eta \sum_{\tau=0}^{t-1} R_\theta^{(t,\tau)} \nabla_{\boldsymbol{h}} \mathcal{L}(\boldsymbol{h}^{(\tau)}, \boldsymbol{h}^\star) + \boldsymbol{\omega}^{(t)} \tag{67}$$

where we noticed that $\boldsymbol{h}^* \sim \boldsymbol{\omega}^*$. These equations are the same as in A.1, with just two extra terms in (66), $\tilde{\Lambda}^{(t)}$ and $\tilde{R}_\ell^{(t,\tau)}$. An application of the Stein's Lemma further simplifies the term $\tilde{\Lambda}^{(t)}$ to $g^{(t)}$ in the Theorem as follows:

$$\tilde{\Lambda}^{(t)} = \alpha \mathbb{E}\left[\nabla_{\boldsymbol{h}^*} \nabla_{\boldsymbol{h}} \ell(\boldsymbol{h}^{(t)}, \boldsymbol{h}^\star)\right] = \alpha \mathbb{E}\left[\nabla_{\boldsymbol{h}} \ell(\boldsymbol{h}^{(t)}, \boldsymbol{h}^\star) \left(\boldsymbol{h}^*\right)^\top\right] = g^{(t)}, \tag{68}$$

□

The above effective process characterizes the limits of several quantities determined by the weights and pre-activations. In particular, it provides the limits of the overlaps of the student teacher overlaps:

Corollary A.3. *Under the assumptions of Theorem 3.2,*

$$W^{(t)}(W^\star)^\top / d \xrightarrow[P]{n,d\to\infty} M^{(t)}, \tag{69}$$

where $M^{(t)}$ is defined as in Theorem A.2.

Proof. Observe that $\langle \mathbf{w}_i^{(t)}, \mathbf{w}^\star \rangle / d$ can be expressed as an expectation of a pseudo-lipschitz function w.r.t the joint empirical measure over the coordinates of $\mathbf{w}_i^{(t)}, \mathbf{w}^\star$ with the value at the j_{th} coordinate given by $\{\mathbf{w}_i^{(t)}\}_j \{\mathbf{w}^\star\}_j$. Therefore 3.2 implies that $W^{(t)}(W^\star)^\top / d$ converges in probability to the expected overlaps of the effective process $\theta^{(t)}, \theta^\star$ which equal $M^{(t)}$ by definition. □

We also include a useful corollary, describing the evolution of the overlaps of the weights.

Lemma A.4. *Under the assumptions of A.2 the covariance* $C_\theta^{(t,\tau)}$

$$C_\theta^{(t+1,\tau)} - C_\theta^{(t,\tau)} = -\eta\left(\lambda + \Lambda^{(t)}\right) C_\theta^{(t,\tau)} + \eta \sum_{s=0}^{t-1} R_\ell^{(t,s)} C_\theta^{(s,\tau)} + \eta \sum_{s=0}^{\tau-1} R_\theta^{(t,s)} C_\ell^{(s,\tau)} - \tag{70}$$

$$-\eta\left(g^t - \sum_{s=0}^{t} \tilde{R}_\ell^{(t,s)}\right)\left(M^{(\tau)}\right)^\top \tag{71}$$

This is a consequence of linearity of expectation on (51). Concretely, viewing $\theta^{(t)}$ as a function of the Gaussian random variables $\{u^{(\tau)}\}_{\tau=1}^t$, we apply the multi-variate Stein's Lemma to obtain:

$$\mathbb{E}\left[\theta^{(t)}\mathbf{u}^{(\tau)}\right] = \sum_{s=\tau}^{t-1} R_\theta^{(t,s)} C_\ell^{(s,\tau)}\,. \tag{72}$$

In particular, we obtain the following expression for the covariances upto the first time-steps:

Lemma A.5. *The covariances* $C_\theta^{(0,1)}, C_\theta^{(0,0)}$ *satisfy:*

$$C_\theta^{(0,1)} - C_\theta^{(0,0)} = -\eta\left(\lambda + \Lambda^{(0)}\right) C_\theta^{(0,0)} - \eta\left(g^{(0)} - \Lambda^{(0)} M^{(0)}\right)\left(M^{(0)}\right)^\top \tag{73}$$

$$C_\theta^{(1,1)} - C_\theta^{(0,1)} = -\eta\left(\lambda + \Lambda^{(0)}\right) C_\theta^{(0,1)} + \eta C_\ell^{(0,1)} - \eta\left(g^{(0)} - \Lambda^{(0)} M^{(0)}\right)\left(M^{(1)}\right)^\top \tag{74}$$

A.4. Pre-activations at the end of the first gradient update

For $T = 1$, Equation (52) simplifies to:

$$\boldsymbol{h}^{(1)} = -\eta \nabla_{\boldsymbol{h}} \ell(\boldsymbol{h}^{(0)}, \mathbf{h}^\star) + \boldsymbol{\omega}^{(1)} \tag{75}$$

We now show that the first term exactly correspond to the contributions considered in section 4.1.

Lemma A.6. *Under the notation in section 4.1 and assumptions of Theorem 3.2:*

$$\frac{a}{d}\ell'\left(\mathbf{h}_\nu^{(0)}, \mathbf{h}_\nu^\star\right)\sigma'(\mathbf{h}_\nu^{(0)})\langle \mathbf{z}_\nu, \mathbf{z}_\nu\rangle \xrightarrow[D]{n,d\to\infty} \nabla_{\boldsymbol{h}}\ell(\boldsymbol{h}^0, \mathbf{h}^\star) \tag{76}$$

where $\boldsymbol{h}^0 \in \mathbb{R}^p, \boldsymbol{h}^* \in \mathbb{R}^k$ *are independent Gaussian random variables distributed as in Theorem A.2.*

Proof. We simply apply the conditioning by projection technique described in Section A.2 to $\mathbf{z}_\nu$ by expressing it as: $\mathbf{z}_\nu = \mathbf{h}_\nu^{(0)} + \frac{d-1}{d}\mathbf{z}'_\nu$, where $\mathbf{z}'_\nu$ is independent of $\mathbf{h}_\nu^{(0)}$. The result then follows from convergence in probability of $\frac{1}{d}\langle \mathbf{z}'_\nu, \mathbf{z}'_\nu\rangle$ to 1. □

Next, we characterize $\boldsymbol{\omega}^{(1)}$. We consider two cases:

- $M^{(1)} = 0$: In this case, Corollary A.3 implies that the first-layer does not develop any overlap with directions in U^*. Equation (54) then implies that $\boldsymbol{\omega}^{(1)}$ is uncorrelated with ω^*.
- $M^{(1)} \neq 0$: In this case the first-layer develops an overlap along U^*. By initialization, we have that $M^{(0)} = 0$. Equation (61) implies that $M^{(1)}$ is given by:

$$M^{(1)} = -\eta\alpha\mathbb{E}\left[\nabla_{\boldsymbol{h}}\ell(\boldsymbol{h}^{(0)})\,(\boldsymbol{h}^*)^\top\right] \tag{77}$$

  Due to the choice of symmetric initialization (Equation 4), we have $f(\boldsymbol{h}^{(0)}) = 0$. Therefore, $\nabla_{\boldsymbol{h}}\mathcal{L}(\boldsymbol{h}^{(0)}) = -ag^*(\mathbf{h}^*)\sigma'(\boldsymbol{h}^{(0)})$. We thus obtain

$$M^{(1)} = \eta\alpha\mathbf{a}\mathbb{E}\left[g^*(\mathbf{h}^*)\boldsymbol{h}^{(0)}(\boldsymbol{h}^*)^\top\right] = \eta\alpha\mathbf{a}\odot\mathbb{E}\left[\boldsymbol{h}^{(0)}\right]\mathbb{E}\left[g^*(\mathbf{h}^*)(\boldsymbol{h}^*)^\top\right], \tag{78}$$

where $\odot$ denotes element-wise multiplication and we used the independence of $\mathbf{h}^*, \boldsymbol{h}^{(0)}$. Therefore, the rows of $M^{(1)}$ gains a rank-one spike along $\mathbb{E}\left[g^*(\mathbf{h}^*)(\boldsymbol{h}^*)\right]$. This matches the corresponding results for single $\mathcal{O}(d)$ batch gradient steps under the online-setting (Ba et al., 2022; Dandi et al., 2023).

By Equation (54), $\boldsymbol{\omega}^{(1)}$ can be expressed as:

$$\boldsymbol{\omega}^{(1)} = \boldsymbol{\omega}_{\perp}^{(1)} + M^{(1)}\boldsymbol{h}^* \,, \tag{79}$$

where $\omega_{\perp}$ is independent of $\boldsymbol{h}^*$.

A.5. Proof of Theorem 3.2

To illustrate the learning of directions solely due to the hidden progress explained in section 4.1, we first focus on the case where $M^{(1)} = 0$ i.e when the parameters develop no overlap along the target subspace in the first step.

From Corollary A.3, and Slutsky's theorem, we have that:

$$\frac{1}{d}W^2(W^\star)^\top - \frac{1}{d}W^1(W^\star)^\top \xrightarrow[P]{n,d\to\infty} -\eta\alpha\mathbb{E}\left[\nabla_{\boldsymbol{h}}\mathcal{L}(\boldsymbol{h}^{(1)}, \mathbf{h}^\star)\left(\boldsymbol{h}^\star\right)^\top\right], \tag{80}$$

where from Equation (75), $\mathbf{h}^{(1)}$ can be expressed as a combination of $\nabla_{\boldsymbol{h}}\mathcal{L}(\boldsymbol{h}^{(1)}, \mathbf{h}^\star)$ and a Gaussian random variable $\boldsymbol{\omega}^{(1)}$ independent of $\mathbf{h}^\star$. Furthermore, Lemma A.4 implies that the regularization strength λ and step-size η can be set such that the entries of $\boldsymbol{\omega}^{(1)}$ have unit-variance. Now, suppose $\mathbf{v}^\star = (W^\star)^\top\mathbf{u}^\star$ for some fixed vector $\mathbf{u}^\star \in \mathbb{R}^p$. First, consider the case when $\mathbf{v}^\star$ lies in the subspace $P_{\perp}^*$ as defined in definition 3.1.

By projecting Equation (80) along $\mathbf{u}^\star$, we obtain that:

$$\frac{1}{d}W^2\mathbf{v}^\star - \frac{1}{d}W^1\mathbf{v}^\star \xrightarrow[P]{n,d\to\infty} -\eta\alpha\mathbb{E}\left[\nabla_{\boldsymbol{h}}\mathcal{L}(\boldsymbol{h}^{(1)}, \mathbf{h}^\star)\left(\boldsymbol{h}^\star\right)^\top\mathbf{u}^\star\right], \tag{81}$$

For squared loss, we have $-\nabla_{\boldsymbol{h}_j}\mathcal{L}(\boldsymbol{h}^{(t)}, \mathbf{h}^\star) = a_j(g^\star(\mathbf{h}^\star) - f(\boldsymbol{h}^{(t)}))\sigma'(\boldsymbol{h}_j^{(t)})$.

Therefore, the overlap for the j_{th} neuron can be expressed as :

$$\frac{1}{d}\langle\mathbf{w}_i^{(2)}, \mathbf{v}^\star\rangle - \frac{1}{d}\langle\mathbf{w}_i^{(1)}, \mathbf{v}^\star\rangle \xrightarrow[P]{n,d\to\infty} \eta\alpha\mathbb{E}\left[a_j g^\star(\mathbf{h}^\star)\sigma'(\boldsymbol{h}_j^{(1)})\left(\boldsymbol{h}^\star\right)^\top\mathbf{u}^\star\right] - \eta\alpha\mathbb{E}\left[a_j f(\boldsymbol{h}^{(1)})\sigma'(\boldsymbol{h}_j^{(1)})\left(\boldsymbol{h}^\star\right)^\top\mathbf{u}^\star\right]. \tag{82}$$

We focus on the first term in the RHS. By assumption, $\frac{1}{d}\langle\mathbf{w}_i^{(0)}, \mathbf{v}^\star\rangle, \frac{1}{d}\langle\mathbf{w}_i^{(1)}, \mathbf{v}^\star\rangle$ converge in probability to 0. Therefore, using Equation (75), we obtain:

$$\frac{1}{d}\langle\mathbf{w}_i^{(2)}, \mathbf{v}^*\rangle \xrightarrow[P]{n,d\to\infty} \eta\alpha\mathbb{E}\left[a_j g^\star(\mathbf{h}^\star)\sigma'(\boldsymbol{h}_j^{(1)})\left(\boldsymbol{h}^\star\right)^\top\mathbf{u}^\star\right] \tag{83}$$

$$= \eta\alpha\mathbb{E}\left[a_j g^\star(\mathbf{h}^\star)\sigma'(-\nabla_{\boldsymbol{h}}\mathcal{L}(\boldsymbol{h}^{(0)}) + \boldsymbol{\omega}^{(1)})\left(\boldsymbol{h}^\star\right)^\top\mathbf{u}^\star\right]. \tag{84}$$

Recall that by initialization $C_\theta^{(0,0)}$ are diagonal with entries 1 except for the off-diagonal entries corresponding to pairing of neurons through the symmetric initialization. Furthermore, by initialization, $M^{(0)} = \mathbf{0}$ and by assumption $M^{(1)} = \mathbf{0}$.

From Lemma A.4, and the definitions of we have that by setting $\eta\gamma = 1$, the covariance $C_\theta^{(0,1)}, C_\theta^{(1,1)}$ simplify to:

$$C_\theta^{(0,1)} = -\eta\Lambda^{(0)}C_\theta^{(0,0)} \tag{85}$$

$$C_\theta^{(1,1)} = \Lambda^{(0)}C_\theta^{(0,1)} + \eta C_\ell^{(0,1)} \tag{86}$$

By definition, $\Lambda^{(0)}, C_\ell^{(0,1)}$ have diagonal entries proportional to a_j, a_j^2 respectively. Therefore, we can further set $\eta > 0$ such that the j_{th} diagonal entry of $C_\theta^{1,1}$ equals $1 + a_j^2$. By case 1 in section A.4, we further have that $\omega^{(1)}$ is independent of $\mathbf{h}^*$.

Substituting $\nabla_{\boldsymbol{h}}\mathcal{L}(\boldsymbol{h}^{(0)}) = -ag^*(\mathbf{h}^*)\sigma'(\boldsymbol{h}^{(0)})$, we obtain the precise condition on σ for the j_{th} neuron to learn direction $\mathbf{v}^\star$ the second timestep. The condition is given by:

$$\phi(a_j) = \mathbb{E}\left[g^\star(\mathbf{h}^\star)\sigma'(\eta a_j g^\star(\mathbf{h}^\star)\sigma'(h_j^0) + \xi_1 + a_j\xi_2)\langle \mathbf{h}^\star, \mathbf{u}^\star\rangle\right] \neq 0, \tag{87}$$

where $\mathbf{h}^\star$ and ξ_1, ξ_2 are independent Gaussian random variables. Since $\mathbf{h}^\star$ matches in distribution $\frac{1}{\sqrt{d}}\mathbf{W}^\star\mathbf{z}$, the above condition can equivalently be expressed as the following condition on $f^\star$:

$$\phi(a_j) = \mathbb{E}_{\mathbf{z}}\left[F_{\sigma,\mathbf{a}}(f^\star(\mathbf{z}))\langle \mathbf{v}^\star, \mathbf{z}\rangle\right] \neq 0, \tag{88}$$

where:

$$F_{\sigma,\mathbf{a}}(x) = \mathbb{E}_{\xi_1,\xi_2}\left[x\sigma'(\eta a_j\sigma'(u)x + a_j\xi)\right], \tag{89}$$

where u is a standard normal variable, corresponding to $h_j^{(0)}$. The above expectation ϕ is an analytic function of a_j. To show that it is identically non-zero, we consider the derivative w.r.t a_j at $a_j = 0$. We have, using the dominated-convergence theorem:

$$\begin{aligned}\phi'(0) &= \eta\mathbb{E}\left[\sigma'(u)g^\star(\mathbf{h}^\star)^2\sigma''(\xi_1)\langle \mathbf{h}^\star, \mathbf{u}^\star\rangle\right] \\ &= \eta\mathbb{E}\left[g^\star(\mathbf{h}^\star)^2\langle \mathbf{h}^\star, \mathbf{u}^\star\rangle\right]\mathbb{E}\left[\sigma'(u)\right]\sigma''(0) \\ &= \eta\mathbb{E}\left[g^\star(\mathbf{h}^\star)^2\langle \mathbf{h}^\star, \mathbf{u}^\star\rangle\right]\nu_1(\sigma)\sigma''(0),\end{aligned}$$

where $\nu_1(\sigma)$ denotes respectively the 1_{st} Hermite-coefficients of σ and are non-zero by assumption on σ. Similarly, iterating k-times, we have:

$$D_{a_j}^k\phi(\mathbf{a}) = \eta^{k+1}\mathbb{E}\left[(g^\star(\mathbf{h}^*))^{k+1}\langle \mathbf{h}^\star, \mathbf{u}^\star\rangle\right]\nu_1^k(\sigma)\sigma^{k+1}(\sigma), \tag{90}$$

where $\nu_k(\sigma)$ denotes the k_{th} Hermite-coefficient of σ. Note that for $\phi(\mathbf{a})$ to not be identically zero, it is sufficient that $D_{a_j}^{k-1}$ is non-zero for some $k \in \mathbb{N}$. Since by assumption, $\mathbf{v}^* \in P_\perp^*$, and since the monomials $1, x, x^2, \cdots$ span the space of polynomials, we have that there exists a $k \in \mathbb{N}$ such that: $\mathbb{E}\left[(g^\star(\mathbf{h}^*))^k\langle \mathbf{h}^\star, \mathbf{u}^\star\rangle\right] \neq 0$.

Therefore $\phi(a_j)$ is a not identically 0. Since $\phi(a_j)$ is an analytic function non-identically zero, and the law of $\mathbf{a}^{(0)} \sim \mathcal{N}(0, \frac{1}{p}\mathbb{1}_p)$ is absolutely continuous w.r.t the Lebesgue measure, we have that $\phi(a_j)$ is non-zero almost surely over the initialization. Now, the second term in Equation 81 is again an analytic function in $\mathbf{a}$, distinct from $\phi(a_j)$, and can therefore be almost surely absorbed into the non-zero overlap. This proves the first part of Theorem 3.2 for developing an overlap along a fixed direction in $P_\perp^*$ when $M^{(1)} = 0$. We now proceed to show that the weights W^2 span $P_\perp^*$.

Let r denote the dimension of the subspace $P_\perp^*$. Suppose that $v_1^* = (W^\star)^\top\mathbf{u}_1^\star, v_2^*, \cdots = (W^\star)^\top\mathbf{u}_2^\star, \cdots, v_r^* = (W^\star)^\top\mathbf{u}_r^\star$ form an orthonormal basis of $P_\perp^*$. Let $V^* \in \mathbb{R}^{d\times r}$ matrix $M_u^* \in \mathbb{R}^{k\times r}$ denote matrices with columns$\mathbf{v}_1^\star, \cdots, \mathbf{v}_r^\star$ and $\mathbf{u}_1^\star, \cdots, \mathbf{u}_r^\star$ respectively.

Analogous to Equation (82), we obtain:

$$\frac{1}{d}W^2V^\star - \frac{1}{d}W^1V^\star \xrightarrow[P]{n,d\to\infty} -\eta\alpha\mathbb{E}\left[\nabla_{\boldsymbol{h}}\mathcal{L}(\boldsymbol{h}^{(1)}, \mathbf{h}^*)\left(\boldsymbol{h}^\star\right)^\top M_u^*\right], \tag{91}$$

Following the derivation of Equation (88), we obtain that the rows of the matrix $\mathbb{E}\left[\nabla_{\boldsymbol{h}}\mathcal{L}(\boldsymbol{h}^{(1)}, \mathbf{h}^*)\left(\boldsymbol{h}^\star\right)^\top U^*\right]$ are independent for neurons i, j for $j \neq p - i + 1$ (due to the symmetric initialization in Equation (4). Furthermore each row of the matrix is absolutely continuous w.r.t the Lebesgue measure on $\mathbb{R}^r$. This implies that $\frac{1}{d}W^2V^\star - \frac{1}{d}W^1V^\star$ has full row-rank almost surely for large enough p.

Now, suppose that $\mathbf{v}^\star$ instead lies in the even-symmetric subspace $A^\star$. By induction and closure properties of analytic functions, we have that $\boldsymbol{h}^{(t)}$ can be expressed as:

$$\boldsymbol{h}^{(t)} = \mathcal{F}_t(\boldsymbol{h}^{(*)}, \omega_1, \boldsymbol{\omega}^{(1)}, \cdots, \boldsymbol{\omega}^{(t)}), \tag{92}$$

for an analytic mapping $\mathcal{F}_t$. Now, similar to Equation (82), we have that:

$$\frac{1}{d}W^t\mathbf{v}^\star - \frac{1}{d}W^{t-1}\mathbf{v}^\star \xrightarrow[P]{n,d\to\infty} -\eta\alpha\mathbb{E}\left[\nabla_{\boldsymbol{h}}\mathcal{L}(\boldsymbol{h}^{(t)}, \mathbf{h}^*)\left(\boldsymbol{h}^\star\right)^\top \mathbf{u}^\star\right], \tag{93}$$

Using Fubini's theorem, we may take expectation w.r.t to express each entry of $\mathbb{E}\left[\nabla_{\boldsymbol{h}}\mathcal{L}(\boldsymbol{h}^{(t)}, \mathbf{h}^*)\left(\boldsymbol{h}^\star\right)^\top \mathbf{u}^\star\right]$ as:

$$\mathbb{E}_{\mathbf{z}}\left[F_{t,\mathbf{a}}(f^\star(\mathbf{z}))\langle \mathbf{v}^\star, \mathbf{z}\rangle\right] , \tag{94}$$

for some analytic $F_{t,\mathbf{a}}$ This ensures that the expectation in (81) remains 0 for all time t. This proves the second part of Theorem 3.2.

A.6. Effect of previously learned directions

We now consider the case when $M^{(1)} \neq 0$, i.e when the first-layer develops an overlap along U^*. As shown in 4.1, the rows of $M^{(1)}$ lie along the same direction given by $\mathbb{E}\left[g^*(\mathbf{h}^*)(\boldsymbol{h}^*)\right]$. Without loss of generality, we assume that the direction $\mathbb{E}\left[g^*(\mathbf{h}^*)(\boldsymbol{h}^*)\right]$ corresponds to $\mathbf{e}_1$ in the input space $\mathbb{R}^d$ and that W^* has rows along the standard basis $\mathbf{e}_1, \cdots, \mathbf{e}_k$. Note that $\mathbf{e}_1$ itself lies in $P_\perp^*$ by setting $F(x) = x$ in definition 3.1.

From Equation (79), we obtain:

$$\boldsymbol{\omega}_j^{(1)} = \boldsymbol{\omega}_{\perp \boldsymbol{j}}^{(1)} + \eta C a_j h_1^* \tag{95}$$

where C denotes a constant dependent on g^*. Since $\boldsymbol{\omega}^{(1)}$ is now correlated with h_1^*, the condition in Equation (87) is modified to:

$$\phi(a_j) = \mathbb{E}\left[g^\star(\mathbf{h}^\star)\sigma'(\eta a_j \sigma'(u) g^\star(\mathbf{h}^\star) + \xi_1 + a_j\xi_2 + \eta C a_j h_1^*)\langle \mathbf{h}^\star, \mathbf{u}^\star\rangle\right] \neq 0, \tag{96}$$

Again, differentiating w.r.t a_j, we obtain:

$$\phi'(0) = \eta\mathbb{E}\left[g^\star(\mathbf{h}^\star)^2\langle \mathbf{h}^\star, \mathbf{u}^\star\rangle\right]\nu_1(\sigma)\nu_2(\sigma) + \eta\mathbb{E}\left[g^\star(\mathbf{h}^\star)h_1^*\langle \mathbf{h}^\star, \mathbf{u}^\star\rangle\right] \tag{97}$$

Similar to section A.5, we have that $\mathbf{v}^* \in P_\perp^*$ is sufficient for the first term to be non-zero almost surely over a_j. If the second-term is non-zero, we have that $\mathbf{u}^\star$ is learned through the staircase mechanism, since it implies that $g^\star(\mathbf{h}^\star)$ contains terms dependent on h_1^* and linearly coupled with $\langle \mathbf{h}^\star, \mathbf{u}^\star\rangle$. In either case, we obtain that $W^{(2)}$ almost surely obtains an overlap along $\mathbf{v}^*$. This concludes the proof of the first part of Theorem 3.2.

More, generally, suppose that $\mathbf{e}_1, \cdots \mathbf{e}_m$ denote a basis of the directions in U^* learned up to time t. Then, the modified condition for learned a new direction $\mathbf{v}^*$ at time $t+1$ is:

$$\mathbb{E}_{\mathbf{z}}\left[F(f^\star(\mathbf{z}), z_1, \cdots z_m)\langle \mathbf{v}^\star, \mathbf{z}\rangle\right] \neq 0, \tag{98}$$

for a polynomial $F : \mathbb{R}^{m+1} \to \mathbb{R}$. Therefore, new directions can be learned through a combination of the staircase and hidden-progress mechanism.

A.7. Typical examples where $E_\perp^* = P_\perp^* = U^*$

For several target functions of interest, the class $P_\perp^*$ can be shown to cover the entire target space U^*. We list some of them below:

- Single-index odd polynomials with all non-negative/non-positive coefficients. This follows since $\mathbb{E}_{\mathbf{z}}\left[(f^\star(\mathbf{z}))^k\langle \mathbf{v}^\star, \mathbf{z}\rangle\right]$ decomposes into sums of non-negative/non-positive terms.
- Single-index odd Hermite polynomials. We prove this below in Lemma A.7
- Staircase function $f^*(\mathbf{z}) = z_1 + z_1z_2 + z_1z_2z_3$. This follows directly by evaluating $\mathbb{E}_{\mathbf{z}}\left[(f^\star(\mathbf{z}))^2 z_i\right]$ for $i = 2, 3$.

In general, for polynomial $f^\star$, the condition:

$$\mathbb{E}_{\mathbf{z}}\left[(f^\star(\mathbf{z}))^k\langle \mathbf{v}^\star, \mathbf{z}\rangle\right] = 0, \forall k \in \mathbb{N}, \tag{99}$$

specifies an overdetermined system of infinite homogenous polynomial equations on the coefficients of $f^\star$. Therefore we expect the condition to fail almost surely for typical choices of $f^\star$. We leave an investigation of this using algebraic tools to future investigation.

Lemma A.7. *For any odd Hermite-polynomial $H_{2k'+1}$ for $k' \in \mathbb{N}$,:*

$$\mathbb{E}_z \left[(H_{2k'+1}(z))^3 z \right], \neq 0 \tag{100}$$

where $z \sim \mathcal{N}(0,1)$

Proof. Using Stein's Lemma, we have:

$$\mathbb{E}_z \left[(H_{2k'+1}(z))^3 z \right] = 3\mathbb{E}_z \left[(H_{2k'+1}(z))^2 \frac{d}{dz} H_{2k'+1}(z) \right] \tag{101}$$

Next, we recall the following relation between Hermite polynomials and their derivatives:

$$\frac{d}{dz} H_n(z) = n H_{n-1}(z), \forall n \in \mathbb{N}. \tag{102}$$

Substituting in Equation (101), we obtain:

$$\mathbb{E}_z \left[(H_{2k'+1}(z))^3 z \right] = 3(2k'+1)\mathbb{E}_z \left[(H_{2k'+1}(z))^2 H_{2k'}(z)(z) \right]. \tag{103}$$

The above expectation can be obtained analytically using the linearization formulas for Hermite polynomials (Andrews, 2004) to show that $\mathbb{E}_z \left[(H_{2k'+1}(z))^2 H_{2k'}(z)(z) \right] \neq 0$ for all $k' \in \mathbb{N}$. □

A.8. Proof of Proposition 3.5

Suppose that $\mathbf{v}^\star \in OE^\star$ i.e $\mathbf{v}^\star$ is orthogonally even-symmetric w.r.t $f^\star$ for some transformation $O_\perp \in O(\{v^\star\}_\perp)$. Let $z' = O_\perp R_{\mathbf{v}^\star} \mathbf{z}$. Then, by the invariance of the Gaussian measure under orthogonal transformations, we have:

$$\mathbb{E}_{\mathbf{z}} \left[f^\star(\mathbf{z}) \langle \mathbf{v}^\star, \mathbf{z} \rangle \right] = \mathbb{E}_{\mathbf{z}'} \left[f^\star(\mathbf{z}') \langle \mathbf{v}^\star, \mathbf{z}' \rangle \right]. \tag{104}$$

However, the expectation on the right can equivalently be expressed as:

$$\begin{aligned} \mathbb{E}_{\mathbf{z}'} \left[f^\star(\mathbf{z}') \langle \mathbf{v}^\star, \mathbf{z}' \rangle \right] &= -\mathbb{E}_{\mathbf{z}} \left[f^\star(O_\perp R_{\mathbf{v}^\star} \mathbf{z}) \rangle \langle \mathbf{v}^\star, \mathbf{z} \rangle \right] \\ &= -\mathbb{E}_{\mathbf{z}} \left[f^\star(\mathbf{z}) \langle \mathbf{v}^\star, \mathbf{z} \rangle \right] \end{aligned}$$

where in the second equality we used $\langle \mathbf{v}^\star, \mathbf{z}' \rangle = -\langle \mathbf{v}^\star, \mathbf{z} \rangle$ and in the third the definition 3.4. Therefore, for any $\mathbf{v}^\star \in OE^\star$, we have:

$$\mathbb{E}_{\mathbf{z}} \left[f^\star(\mathbf{z}) \langle \mathbf{v}^\star, \mathbf{z} \rangle \right] = 0 \tag{105}$$

Furthermore, it is straightforward to see that $\mathbf{v}^\star$ remains orthogonally even-symmetric w.r.t the composition $F(f^\star(\cdot))$. Therefore, we have that $E^\star \subseteq OE^\star \subseteq A^\star \subseteq P^\star$.

A.9. Illustration of non-even symmetric hard directions

We now show the existence of target functions where $E^* \neq OE^*$. Without loss of generality, we assume that the rows of W^* lie along the standard Euclidean basis $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k$

Lemma A.8. *Suppose that $f^*(\mathbf{z}) = z_1 z_2 z_3$. Let $\mathbf{v}^* = \mathbf{e}_1 + \mathbf{e}_2 + \mathbf{e}_3$. Then $\mathbf{v}^* \notin E^*$ but $OE^* = U^*$.*

Proof. $\mathbf{v}^* \notin E^*$ follows directly by noting that $f^*(\mathbf{z})$ is even-symmetric along $\mathbf{e}_1 - \mathbf{e}_2$ and $\mathbf{e}_1 - \mathbf{e}_3$. We further have that a target f^* satisfying $E^* = U^*$ must satisfy $f^*(-\mathbf{z}) = f^*(\mathbf{z})\ \forall \mathbf{z} \in \mathbb{R}^d$. Therefore, since $f^*(-\mathbf{z}) = -f^*(\mathbf{z})$, f^* cannot be even-symmetric along $\mathbf{e}_1 + \mathbf{e}_2 + \mathbf{e}_3$. Next, we show that $OE^* = U^*$. Since $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ span U^*, and f^* is symmetric w.r.t permutations of $z_1 z_2 z_3$, it suffices to show that condition in definition 3.4 holds for $\mathbf{v}^* = \mathbf{e}_1$. The orthogonal complement $\{\mathbf{v}^*\}_\perp$ is given by $\mathrm{span}(\mathbf{e}_2, \mathbf{e}_3)$. Therefore the transformation O_2 defined by $z_2 \to -z_2$ is a valid orthogonal transformation $O_\perp$ as per definition 3.4. We have:

$$\begin{aligned} f^*(O_2 R_{\mathbf{v}^*} \mathbf{z}) &= (-z_1)(-z_2) z_3 \\ &= z_1 z_2 z_3 = f^*(\mathbf{z}). \end{aligned}$$

This shows that $\mathbf{e}_1$ lies in OE^*. Similarly, we have by symmetry $\mathbf{e}_1 \in OE^*$

We present a numerical illustration of another such example in figure 3. □

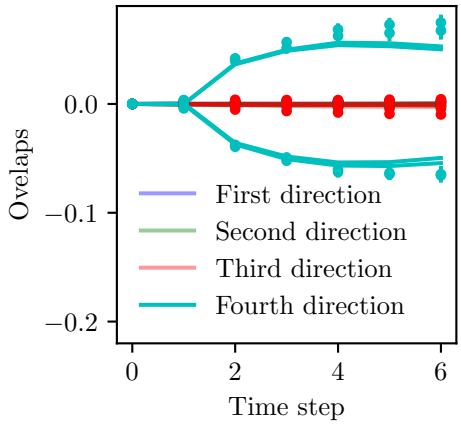

Figure 3. An illustration of a hard, non-even target $f^\star(\mathbf{z}) = z_1 z_2 z_3 + \mathrm{He}_3(z_4)$ being learned by a student with $p = 4$ hidden units. We can see that, *even when reusing the batch*, the teacher can only learn the direction associated with z_4, while keeping a zero overlap otherwise. The continuous lines are from the DMFT numerical integration, the dots are simulations with $d = 10000$. In the legend the overlap with the n-th direction is the projection of the student weights in the subspace associated with z_n. For this figure we have $\sigma = \mathrm{relu}$, $n = 5d$, $\eta = 0.2$.

One can in-fact construct a family of functions with a direction $\mathbf{v}^*$, for instance $\mathbf{v}^* = \mathbf{e_1}$ lying in OE^* but in general not in E^*. To see this, let f_1 be a function $\mathbb{R}^d \to \mathbb{R}$, depending only on projections of $\mathbf{z}$ along $\{\mathbf{e}_1\}_\perp$ and let $O_\perp$ by an involutory orthogonal transformation on $\{\mathbf{e}_1\}_\perp$ i.e an orthogonal transformation satisfying $O_\perp^2 = \mathbf{I}$ or equivalently $O_\perp = (O_\perp)^\top$. Now, let $f_2 : \mathbb{R} \to \mathbb{R}$ be an odd function. Then, consider the function:

$$f^*(\mathbf{z}) = (f_1(O_\perp \mathbf{z}) - f_(\mathbf{z}))f_2(z_1). \tag{106}$$

We observe that:

$$\begin{aligned} f^*(OR_{\mathbf{e}_1}\mathbf{z}) &= (f_1(O_\perp^2 \mathbf{z}) - f_(O_\perp \mathbf{z}))f_2(-z_1) \\ &= (f_1(O_\perp \mathbf{z}) - f_(\mathbf{z}))f_2(z_1) \\ &= f^*(\mathbf{z}), \end{aligned}$$

where we used that $O_\perp^2 = \mathbf{I}$ and $f_2(-z_1) = -f_2(z_1)$. Therefore, for any such function $f^*(\mathbf{z})$, $\mathbf{e}_1 \in OE^*$.

A.10. Implications for generalization

Since the specific guarantees of such results depend on the choice of activation and target functions, we illustrate this for the case of single-index target functions with matching activations:

Corollary A.9. *Consider the setting of a single-index target and student network with matching activations i.e. $\sigma = g^\star$, such that σ is a polynomial with finite degree, satisfying the following assumption, $\exists k \in \mathbb{N}$ such that:*

$$\mathbb{E}\left[\sigma^k(z)z\right]\mathbb{E}\left[D^k\sigma(z)\right] \neq 0, \tag{107}$$

, where D^k denotes the k_{th} derivative. Let $\hat{\mathbf{w}}$ be the parameters obtained after two steps of gradient descent with batch size $\mathcal{O}(d)$ using η as in Theorem 3.2. Then, almost surely over the initialization $a \sim \mathcal{N}(0,1)$, for any $\epsilon > 0$, there exists a step size η' such that online SGD on squared loss reaches generalization error $< epsilon$ in time $\mathcal{O}(d)$.

We verify numerically that the above assumption holds in particular for all odd Hermite polynomials upto order 50. The corollary implies that such target functions can be learned with $\mathcal{O}(d)$ sample complexity using gradient descent alone, without resorting to specialized algorithms and techniques such as spectral initialization.

Proof. Let $\mathbf{w}^*$ denote the single-direction in the teacher subspace with $\|\mathbf{w}^*\| = \sqrt{d}$. We note that Equation (107) is proportional to the $k - 1_{th}$ derivative of $\phi(a_j)$ defined in Equation (88). Therefore, the condition is sufficient to ensure that

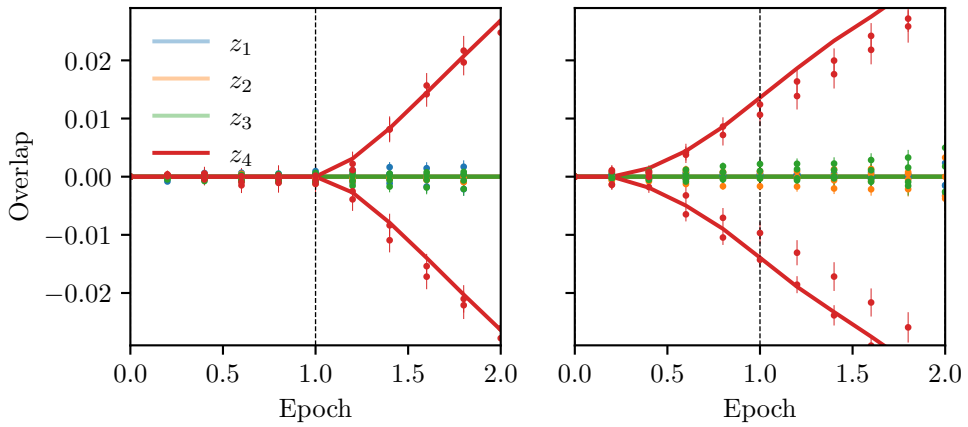

Figure 4. Comparison of theory and experiments for Gradient Descent on the target $z_1 z_2 z_3 + \mathrm{He}_3(z_4)$. Each gradient step uses a mini-batche of $n/5$ samples. On the **left** we use the data sequentially, on the **right** we sample the batch from the dataset with replacement. The continuous lines are from the DMFT numerical integration, the dots are simulations with $d = 10000$ averaged over 32 realisations. In the legend the overlap with the n-th direction is the projection of the student weights in the subspace associated with z_n. For this figure we have $\sigma = \mathrm{relu}$, $n = 5d$, $\eta = 0.2$.

$\phi(a_j)$ is not identically zero and the student neuron almost surely develops an overlap along $\mathbf{w}^*$. The result then follows from Proposition 2.1 in (Ben Arous et al., 2021), which proves that upon weak recovery i.e a non-zero overlap the target direction $\mathbf{v}^*$, online SGD on a differentiable activation with polynomially bounded derivatives converges to strong recovery Concretely, for any starting non-zero overlap $\theta > 0$, for any $\epsilon' > 0$, there exists $C_{\epsilon',\theta}$ and small-enough step-size such that online SGD with time $C_{\epsilon',\theta} d$ achieves overlap $1 - \epsilon'$ along $\mathbf{w}^*$ Due to the matching activations, this suffices to obtain arbitrary generalization error. □

B. General Multi-Pass Schemes

B.1. Sketch of Proof for Extending Theorem 3.2 to Cycling over Epochs

Let $\mathbf{Z}^1, \cdots, \mathbf{Z}_e^n$ denote n_e independent minibatches of size n_b such that $\frac{n_b}{d} = \mathcal{O}(1)$ with n_e being finite. The effective dynamics for a finite number of epochs can be obtained by noting that Theorem 3.2 in (Gerbelot et al., 2022) allows generalizing Theorem A.1 to dynamics of the form:

$$W^{(t+1)} = W^{(t)} - \eta\lambda W^{(t)} - \sum_{i=1}^{n_e} \eta \frac{1}{\sqrt{d}} \sum_{\nu=1}^{n_b} F_i^t \left(\frac{W^{(t)} \mathbf{z}_\nu^i}{\sqrt{d}} \right) (\mathbf{z}_\nu^i)^\top. \tag{108}$$

The above form of the dynamics allows a different update to be utilized for data corresponding to different blocks $\mathbf{Z}^1, \cdots, \mathbf{Z}_e^n$. In particular, setting F_i^t to 0 whenever $t \mod i \neq 0$ and $\nabla_{\boldsymbol{h}} \mathcal{L}$ otherwise, results in a cycling schedule over the mini-batches $\mathbf{Z}^1, \cdots, \mathbf{Z}_e^n$. Subsequently, one can show that the update from $\mathbf{Z}^i$ in the first-epoch leads to the hidden-progress effect on M^t when the model re-uses $\mathbf{Z}^i$ in the second epoch.

We believe a similar result would hold for $n_b = \mathcal{O}(1)$ samples in the minibatch, as displayed in Figure 5

C. Details on the numerics

C.1. DMFT equations with a single stochastic process

In this section, we present a set of exact equations equivalent to the ones in the main text, but that depend on a single stochastic process. It is possible to show that asymptotically in the proportional limit, i.e. for $d \to \infty$ and $n = \alpha d$, the

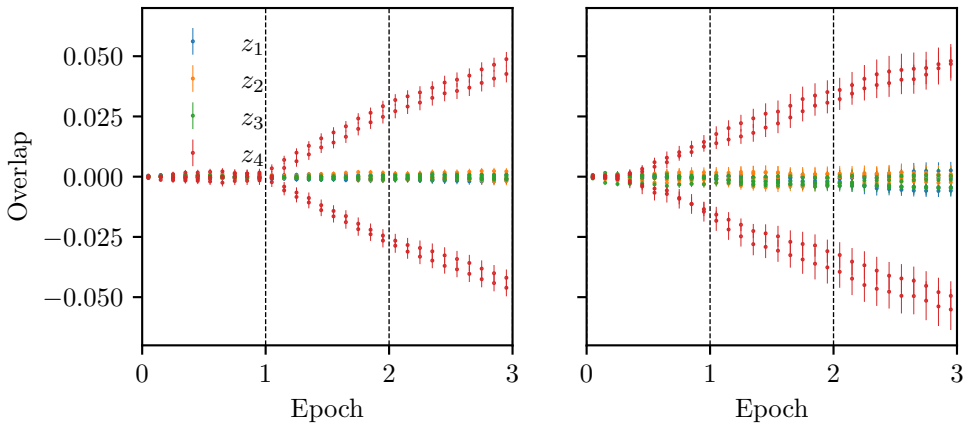

Figure 5. Experiments for Gradient Descent on the target $z_1 z_2 z_3 + \mathrm{He}_3(z_4)$. We use minibatches with 1 sample each. On the **left** we use the data sequentially, on the **right** we sample the data point from the dataset with replacement. The dots are simulations with $d = 10000$ averaged over 32 realisations. In the legend the overlap with the n-th direction denotes the projection of the student weights in the subspace associated with z_n. For this figure we have $\sigma = \mathrm{relu}$, $n = 5d$, $\eta = 0.2$.

pre-activations of the student are distributed as $\mathbf{h}^{(t)} = \mathbf{r}^{(t)} + M^{(t)}\mathbf{h}^*$, with the constraint:

$$\mathbf{r}^{(t+1)} = \mathbf{r}^{(t)} - \eta \left[\left(\lambda + \Lambda^{(t)} \right) \mathbf{r}^{(t)} + \nabla_{\mathbf{h}^{(t)}} \ell \left(\mathbf{h}^{(t)} \right) - \sum_{\tau=0}^{t-1} R_\ell^{(t,\tau)} \mathbf{r}^{(t)} + \zeta^{(t)} \right]$$

Here $\zeta^{(t)}$ is a zero mean Gaussian Process with covariance

$$\mathbb{E}_\zeta \left[\zeta^{(t)} \zeta^{(\tau)\top} \right] = \alpha \mathbb{E}_{\mathbf{r},\mathbf{h}^*} \left[\nabla_{\mathbf{h}^{(t)}} \ell \left(\mathbf{h}^{(t)} \right) \nabla_{\mathbf{h}^{(\tau)}} \ell \left(\mathbf{h}^{(\tau)} \right)^\top \right]$$

and the effective regularisation $\Lambda^{(t)}$ concentrates to

$$\Lambda^{(t)} = \alpha \mathbb{E}_{\mathbf{h}} \left[\nabla^2_{\mathbf{h}^{(t)}} \ell \left(\mathbf{h}^{(t)} \right) \right] . \tag{109}$$

The memory kernel $R_\ell^{(t,\tau)}$ is identically zero for $t \leq \tau$ while for $t > \tau$ it concentrates to

$$R_\ell^{(t,\tau)} = \alpha \mathbb{E}_{\mathbf{h}} \left[\frac{\partial \nabla_{\mathbf{h}^{(t)}} \ell \left(\mathbf{h}^{(t)} \right)}{\partial \zeta^{(\tau)}} \right] \tag{110}$$

Finally, the low dimensionaly projections of the weights $M^{(t)}$ will obey the relation

$$M^{(t+1)} = M^{(t)} - \eta \alpha \mathbb{E}_{\mathbf{h},\mathbf{h}^*} \left[\nabla_{\mathbf{h}^{(t)}} \ell \left(\mathbf{h}^{(t)} \right) \mathbf{h}^{*\top} \right] \tag{111}$$

The procedure is explained in detail in appendix D of (Gerbelot et al., 2022), and can be equivalently derived using non-rigorous field theory techniques (Agoritsas et al., 2018).

C.2. Remark on the numerical integration of the DMFT equations

DMFT is an invaluable tool in itself to probe the behaviour of gradient based algorithms. It trades the update equation over heavily coupled weights in (5) with the ones over completely decoupled preactivations (109) which implies that a Monte Carlo estimation based on (109) is going to be vastly more efficient and it's a trivially parallelisable computation. Furthermore, equation (109) is exact in limit of large d, which removes completely all finite size effects. In practice, an implementation of the DMFT equations is extracting n times using from the initial condition distribution of the practivations and iterating forward. The Gaussian process is sampled by rotating white Gaussian noise by the LU factor of the covariance.

Sampling the gaussian process is by far the costlier operation, as each time step T has a complexity $\mathcal{O}(T^3)$, for a total $\mathcal{O}(T^4)$ complexity considering all the steps up to T. Notice that this is a much more direct implementation than what is done in the literature (Roy et al., 2019; Mignacco et al., 2020), which usually starts with a guess for all the quantities and proceedes with a damped fixed point iteration until convergence, with an overall complexity $\mathcal{O}(mT^3)$, where m is the number of fixed point iterations. While it could appear that simply iterating forward is suboptimal, it is a much more stable and reliable procedure: if you are using n processes and you iterate forward, you are sure that at at each time step you have the best possible Monte Carlo estimate of your samples.

C.3. Details on the numerical simulations

In all the figures the continuous lines are from the numerical integration of the DMFT equations while the dots are from a direct simulation of the gradient descent dynamics. The specific hyperparameters for each setting are near each figure.

For both we fixed the second layer weights to $\pm 1/\sqrt{p}$, as for the cases under consideration this is an equivalent choice to of Gaussian second layer weights $\mathcal{N}(0, \frac{1}{p}\mathbb{1}_p)$. For the DMFT integration we used a minimum of 10^6 Monte Carlo samples in order to have accurate lined. The error bars are too small to be visualised. The direct simulation of the gradient descent dynamics was performed either using PyTorch or a direct implementation in Numpy. In all plots we used a minimum size $d = 5000$ for the input dimension, and averaged over at least 32 independent instances of the dynamics.

In Figure 2 we plot the overlap matrix $M^{(t)}$ projected on two different directions: the parallel to the subspace that is learned in the first step and one direction in the orthogonal of this space. The projection operator is computed by performing explicitly the integrals in (77)

The code is made available through the following Github repository: https://github.com/IdePHICS/benefit-reusing-batch.