Low-rank fine-tuning lies between lazy training and feature learning

Arif Kerem Dayı KEREMDAYI@COLLEGE.HARVARD.EDU
Harvard College

Sitan Chen SITAN@SEAS.HARVARD.EDU
Harvard SEAS

Editors: Nika Haghtalab and Ankur Moitra

Abstract

LoRA has emerged as one of the *de facto* methods for fine-tuning foundation models with low computational cost and memory footprint. The idea is to only train a low-rank perturbation to the weights of a pre-trained model, given supervised data for a downstream task. Despite its empirical success, mathematically it remains poorly understood what learning mechanisms ensure that gradient descent converges to useful low-rank perturbations.

In this work we study low-rank fine-tuning in a student-teacher setting. We are given the weights of a two-layer base model f, as well as i.i.d. samples $(x, f^*(x))$ where x is Gaussian and f^* is the teacher model given by perturbing the weights of f by a rank-1 matrix. This generalizes the setting of generalized linear model (GLM) regression where the weights of f are zero.

When the rank-1 perturbation is comparable in norm to the weight matrix of f, we show that the training dynamics are genuinely distinct from both the lazy linearized dynamics of the kernel regime, and the rich feature learning dynamics captured by GLM regression. We prove under mild assumptions that a student model which is initialized at the base model and trained with online SGD will converge to the teacher in $dk^{O(1)}$ iterations, where k is the number of neurons in f. Importantly, unlike in the GLM setting, the complexity does not depend on fine-grained properties of the activation's Hermite expansion. We also prove that in our setting, learning the teacher model "from scratch" can require significantly more iterations.

Keywords: Fine-tuning, two-layer networks, LoRA, single-index models, feature learning

1. Introduction

Modern deep learning at scale involves two phases: pre-training a foundation model with self-supervised learning, and fine-tuning the model towards various downstream tasks. Given the significant computational cost of the former, effective fine-tuning has been essential to the deployment of these models under hardware constraints and the development of powerful open-source models.

In this space, Low-Rank Adaptation (LoRA) has emerged as one of the most successful and widely adopted methods (Hu et al., 2021). The idea is to freeze the weight matrices of the pre-trained model and only train *low-rank perturbations* to them. Remarkably, this works well even with rank 1 perturbations, reducing the number of trainable parameters by up to 4 orders of magnitude.

Despite the surprising effectiveness of LoRA in practice, it is poorly understood from a theoretical perspective why this method works so well. While it is known that for deep and wide enough pre-trained networks, any sufficiently simple target can be approximated by a low-rank perturbation of the larger model (Zeng and Lee, 2024), it is largely unknown what mechanisms ensure that stochastic gradient descent (SGD) converges to these perturbations. Recent works have made progress towards understanding this question from the perspective of kernel approximations of neural networks in the lazy training regime (Jang et al., 2024; Malladi et al., 2023). These works assume

the perturbation is small enough relative to the weights of the pre-trained model that the fine-tuned model is well-approximated by its linearization around the pre-trained model.

While the kernel picture provides useful first-order intuition for the dynamics of fine-tuning, it only partially explains its success. For one, the kernel approximation is mainly relevant in the few-shot setting where the network is only fine-tuned on a small number of examples (e.g. a few dozen), but the gap between what is possible with few- vs. many-shot fine-tuning is significant. Even within the few-shot setting, Malladi et al. (2023) found that fine-tuning for certain language tasks is not well-explained by kernel behavior, and neither is prompt-based fine-tuning if the prompt is insufficiently aligned with the pre-training task. The gap is even more stark for fine-tuning without prompts. In this work we ask:

Why does SGD for low-rank fine-tuning converge to a good solution even when the kernel approximation breaks down?

To answer this question, we study a natural student-teacher setting. Within this model, our key conceptual finding is that the dynamics of fine-tuning are not only genuinely richer than the lazy, linearized dynamics of the *kernel regime* ("NTK"), but also strictly more tractable than the dynamics in the rich, *feature learning regime* ("μP") where a nonlinear target function is learned from scratch.

1.1. Problem formulation

Let $\mathcal{F} = \{f_\theta\}_{\theta \in \Theta}$ be some family of neural networks, each parametrized by a set of weights θ. Suppose we are given $\theta_0 \in \Theta$, corresponding to a pre-trained *base model*, and then get access to training data $\{(x_i, y_i)\}_{i=1}^N$ for fine-tuning. In this work, we focus on the setting of *realizable Gaussian data* in which the x_i's are i.i.d. Gaussian and there exists a perturbation of the base model, $\theta = \theta_0 + \Delta$ where Δ is low-rank, for which f_θ perfectly fits the training data. That is,

$$x_i \sim \mathcal{N}(0, \mathrm{Id}_n), \quad f_\theta(x_i) = y_i \tag{1}$$

for all $i = 1, \ldots, N$. We call f_θ the *teacher model*.[1]

The goal is to find $\hat{\theta} = \theta_0 + \hat{\Delta}$, where $\hat{\Delta}$ is also low-rank, such that the objective $L(\hat{\theta})$ is small. Here the objective is given by $L(\hat{\theta}) \triangleq \mathbb{E}_x[\ell(f_{\hat{\theta}}(x), f_\theta(x))]$, where $\ell : \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$ is some loss function. In this work we specialize to squared loss.

Algorithms for fine-tuning in practice are based on training the student model, which is initialized to the base model, with gradient descent on L. That is, the parameter $\hat{\Delta}$ is repeatedly updated via stochastic gradient descent on the function $\hat{\Delta} \mapsto L(\theta_0 + \hat{\Delta})$. To ensure that $\hat{\Delta}$ is low-rank throughout the course of training, it is typically parametrized by a low-rank factorization, and the matrices in this factorization are the ones with respect to which one performs gradient descent.

Unfortunately, rigorously analyzing the gradient dynamics at this level of generality is well outside the reach of current theory. Instead, in this work we will focus on a specific instantiation of the above setting, namely *two-layer networks* and *rank-1 perturbations*. Two-layer networks have received significant attention within the learning theory community (see e.g. Chen et al. (2023); Chen and Narayanan (2024); Diakonikolas and Kane (2024) for an overview of this extensive literature) as an important testbed for understanding algorithmic aspects of deep learning, and in this work we use them as a canvas against which to understand fine-tuning. Despite the apparent simplicity of

1. In fact our analysis directly extends to the setting where there is unbiased, moment-bounded label noise, but we focus on the noiseless setting as it is slightly cleaner while exhibiting all the relevant phenomena.

this setting, the dynamics here already exhibit rich behavior beyond the kernel regime, and as we will see, this model strictly generalizes the problem of *generalized linear model (GLM)* regression,[2] a widely studied toy model in the theoretical foundations of deep learning (see Section 1.3).

Concretely, given $k \in \mathbb{N}$, take $\mathcal{F}$ to be the set of all two-layer networks of width k. For $\theta_0 = (\lambda, W) \in \mathbb{R}^k \times \mathbb{R}^{k \times d}$ and σ a known scalar activation, the base model then takes the form

$$f_{\theta_0}(x) \triangleq \lambda^\top \sigma(Wx)\,. \tag{2}$$

The low-rank perturbation defining the teacher model will be given by $\theta \triangleq (\lambda, W^*)$ where

$$W^* = W + \Delta \quad \text{for} \quad \Delta = \xi c u^\top \tag{3}$$

for $\xi > 0$ a known *scale* parameter and for unit vectors $c \in \mathbb{S}^{k-1}$, $u \in \mathbb{S}^{d-1}$. Given a target level of error ϵ, our goal is to find unit vectors $\hat{c}, \hat{u}$ for which $L(\hat{\theta}) \leq \epsilon$ for $\hat{\theta} \triangleq (\lambda, W + \xi \hat{c}\hat{u}^\top)$ with high probability over the training data $\{(x_i, y_i)\}_{i=1}^N$.

1.2. Our contributions

We will consider training a student network $\hat{f}(x) \triangleq \lambda^\top \sigma(W_0 + \xi \hat{c}\hat{u}^\top)x)$ with the same parametrization as the teacher network, using projected online SGD with $\hat{c}$ and $\hat{u}$ initialized randomly. The details of the training algorithm will be given in Section 2. We consider two regimes: (1) when $\{w_i\}$ are orthogonal, and (2) when $\{w_i\}$ have very mild angular separation but are otherwise arbitrary. In both of these regimes, we make the following assumptions to facilitate the analysis.

1.2.1. ASSUMPTIONS

In both of these regimes, we make the following assumptions on the base model and teacher model. Denote the rows of W, i.e. the pre-trained features, by $w_1, \ldots, w_k \in \mathbb{R}^d$. Then we have:

Assumption 1 (Normalization) $\|w_i\|_2 = 1$ *for all* $i = 1, \ldots, k$.

Assumption 2 (Orthogonality of perturbation) *The vector* u *for the teacher model (see Eq. (3)) is orthogonal to the span of* $w_1, \ldots, w_k$.

Assumption 3 (Random quantized c**)** c *is sampled uniformly from* $\{\pm 1/\sqrt{k}\}^k$.

Assumption 1 is without loss of generality when σ is positive homogeneous like in the case of ReLU activation. For general activations, note that one can also handle the case of $\|w_i\|_2 = R$ for all i for arbitrary constant $R > 0$ by redefining σ. This assumption is not essential to our analysis and we assume the scales of the pre-trained features are the same to keep the analysis transparent.

Assumption 2 is crucial to our analysis. To motivate this, in Appendix F.1, we give a simple example where it fails to hold and the low-rank fine-tuning problem ends up having *multiple global optima*, suggesting that the dynamics in the absence of Assumption 2 may be significantly more challenging to characterize. We leave this regime as an interesting area for future study.

Assumption 3 consists of two parts: 1) the entries of c are constrained to lie within $\{\pm 1/\sqrt{k}\}$, and 2) they are random. The former is for technical reasons. First note that the connection to GLMs

2. This is sometimes referred to as *single-index model* regression. While closely related, the latter technically refers to the setting where the activation σ is unknown.

still holds under this assumption. Our main reason to make this is that our proof uses Hermite analysis, and while it is in principle possible to handle neurons with different norms, assuming the c_i's are quantized renders our analysis more transparent without sacrificing descriptive power. As our simulations suggest, the phenomena we elucidate persist without this assumption (see Figure 2).

As for the randomness of c, while we conjecture that fine-tuning should be tractable even in the worst case over c (see Remark 20) albeit with more complicated dynamics, in this work we only show guarantees that hold with *high probability* over c. We primarily use the randomness to ensure that certain quantities that are generically non-vanishing indeed do not vanish. One could equivalently formulate our guarantees as holding under a certain set of deterministic nondegeneracy conditions on the rank-1 perturbation.

Orthonormal features. For this case, we will consider the regime where the scale ξ of the rank-1 perturbation defining the teacher model is large, namely $\xi = \Theta(\sqrt{k})$. Because the norm of the perturbation is comparable to the Frobenius norm of the weight matrix of the base model, the teacher model is not well-approximated by its linearization around the base model. This is therefore a minimal, exactly solvable setting for low-rank fine-tuning where kernel approximation fails and the dynamics fall squarely outside of the lazy training regime. More formally, we make the following assumption:

Assumption 4 (Orthogonality of features) *For all $i \neq j$, we have $\langle w_i, w_j \rangle = 0$.*

Our first result is that online SGD efficiently converges to the correct rank-1 perturbation, under some technical assumptions which we defer to the supplement.

Theorem 1 (Informal, see Theorem 12) *Let $0 < \epsilon < 1$, and let $\xi \asymp \sqrt{k}$ for sufficiently small absolute constant factor. Suppose the rows of W are orthogonal. Let Assumptions 1-3 and Assumption 6 hold. Then, under Orthogonal weights (Assumption 4), the following holds with high probability over the randomness of $c, \hat{c}$ and the examples encountered over the course of training, and with constant probability over the random initialization: online SGD (see Eq. (4)) run with step size $\eta = \tilde{\Theta}(\epsilon^3/dk^{7/2})$ and $T = \tilde{\Theta}(dk^4/\epsilon^4)$ iterations results in $\hat{u}$ for which $\langle \hat{u}, u \rangle^2 \geq 1 - \epsilon$.*

Interestingly, the iteration complexity is linear in d, independent of σ. In contrast, as we discuss in Section 3.1, the iteration complexity of noisy gradient descent, and more generally the complexity of any *statistical query* algorithm, for learning GLMs depends heavily on the Hermite decomposition of σ and transformations thereof. Given that the GLM setting can be recovered from the fine-tuning setting in the $\xi \to \infty$ limit (see end of this subsection), Theorem 1 implies that the gradient dynamics for fine-tuning exhibit a transition in behavior at some scale $\xi = \Omega(\sqrt{k})$.

In other words, even in the orthonormal setting of Theorem 1, we strictly separate the learning dynamics of low-rank fine-tuning from both the linearized dynamics of the kernel regime as well as the feature learning dynamics of GLM regression.

Separated features. While the orthonormal features setting illustrates an important difference between low-rank fine-tuning and GLM regression, the assumption that the features are orthonormal is constraining. We next turn to a more general setting where we only assume that no two pre-trained features are too correlated. Specifically, we make the following very mild assumption:

Assumption 5 (Angular separation) *For all $i \neq j$, we have $|\langle w_i, w_j \rangle| \leq 1 - \log k/\sqrt{k}$.*

Theorem 2 (Informal, see Theorem 13) *Under the same assumptions as Theorem 1, except with $\xi = 1$ and assuming the rows of W satisfy Assumption 5 instead, the following holds with high probability over $c, \hat{c}$ and the examples, and with constant probability over u_0: online SGD run with step size $\eta = \tilde{\Theta}(\epsilon^3/dk^{5/2})$ and $T = \tilde{\Theta}(dk^3/\epsilon^4)$ iterations results in $\hat{u}$ for which $\langle \hat{u}, u \rangle^2 \geq 1 - \epsilon$.*

Given the generality of Assumption 5, we are unable to show a guarantee for learning a rank-1 perturbation at the same scale ξ as Theorem 1. Nevertheless, note that in the regime of $\xi = \Theta(1)$, if one simply considers the linearization of the teacher model around the base model, one is bottlenecked at some fixed level of error. In particular, this means that the kernel approximation to fine-tuning is insufficient to explain why gradient descent converges to the ground truth. One can thus interpret our Theorem 2 as shedding light on the later stages of many-shot fine-tuning whereby the result of the linearized dynamics gets refined to arbitrarily high accuracy.

Separating fine-tuning and learning from scratch. Finally, we show a rigorous lower bound suggesting that fine-tuning is strictly more tractable than learning from scratch in our two-layers setting (see Section C.3 for details):

Theorem 3 (Informal, see Theorem 19) *For any $p > 2$, there exists a base network and a perturbation for which learning the teacher model from scratch using any correlational statistical query algorithm requires either $n = d^{p/2}$ queries or $\tau = d^{-p/4}$ tolerance. However, fine-tuning the base network using Gaussian examples labeled by the teacher only requires $\tilde{O}(d)$ online SGD iterations.*

The proof involves a base model with Hermite activation of degree p whose perturbation has orthonormal weight vectors (see Claim 2) with a carefully chosen c, u. Even though c is not random (i.e. Assumption 3 does not apply to the lower bound instance), we nevertheless prove online SGD still converges to the ground truth perturbation in $\tilde{O}(d)$ iterations. Mainly, this shows in our setting that low-rank fine tuning admits better dimension scaling as opposed to learning from scratch, especially when the information exponent of σ is large.

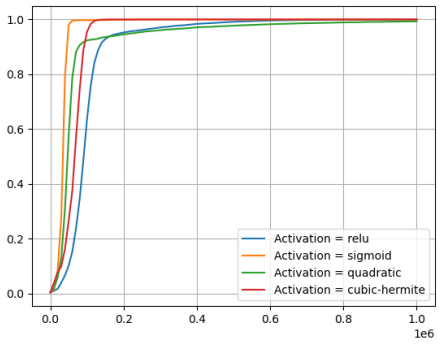
We remark that while there are works showing that gradient-based methods, e.g. multi-pass SGD or online SGD with label transformations, can go beyond correlational statistical query algorithms Damian et al. (2024b); Dandi et al. (2024); Lee et al. (2024); Arnaboldi et al. (2024), these works are restricted to the GLM regression setting, whereas Theorem 3 pertains to the more challenging *multi-index model* setting, for which algorithmic guarantees beyond correlational statistical query are quite limited (Chen and Meka, 2020; Chen et al., 2023).

GLMs, feature learning, lazy training. We conclude this section by elaborating upon the connection between our setting and GLM regression. First note that the special case where the base model is trivial, i.e. when $W = 0_{k \times d}$, recovers the well-studied question of GLM regression. Indeed, consider the case of $c = (1/\sqrt{k}, \ldots, 1/\sqrt{k})$, $\lambda = \frac{1}{k}(1, \ldots, 1)$, and $\xi = \sqrt{k}$. In this case, if the teacher models' parameters are given by $\theta = (\lambda, W^*)$ where W^* is defined in Eq. (3), then the teacher is given by $f_\theta = \sigma(\langle u, x \rangle)$. Learning a direction $\hat{u}$ for which $\mathbb{E}_x[\ell(\sigma(\langle \hat{u}, x \rangle), \sigma(\langle u, x \rangle))]$ is small, given samples $\{(x_i, \sigma(\langle u, x_i \rangle)\}_{i=1}^N$, is precisely GLM regression.

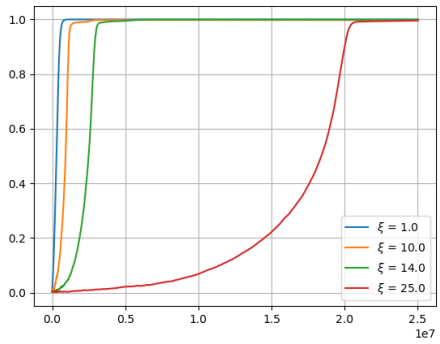
Equivalently, instead of keeping the scale ξ fixed and sending W to zero, we can consider keeping W fixed but nonzero, sending $\xi \to \infty$, and considering ϵ scaling with ξ. This equivalent view is the one we will take in this work as it is more natural for us to regard W as fixed and ξ as a parameter to be varied. Under this view, at the other extreme where $\xi \to 0$, the teacher model

becomes well-approximated by its linearization around the base model, in which case the training dynamics degenerate to the lazy training regime.

The scale parameter ξ thus gives a natural way to interpolate between feature learning ($\xi \to \infty$) and lazy training ($\xi \to 0$) dynamics. Conceptually, our key finding is that the dynamics of low-rank fine-tuning (intermediate ξ) are strictly distinct from either of these two extremes.



(a) Evolution of overlap $\langle u_t, u \rangle$ (y-axis) for $\xi = \Theta(\sqrt{k})$ and different activations



(b) Evolution of overlap $\langle u_t, u \rangle$ (y-axis) while joint training of $\hat{c}$ and $\hat{u}$ for various choices of ξ for $\sigma = \mathrm{He}_3$

Figure 1: (a) Online SGD in the fine tuning-regime exhibits different behavior than the feature learning regime. While the particulars of the activation affect training greatly in feature learning, fine tuning is less sensitive to changes in the activation. (b) Low rank-fine tuning regime interpolates between kernels and feature learning. As $\xi \to \infty$, a plateau starts forming at initialization similar to feature learning.

1.3. Related work

Parameter-efficient fine-tuning. Within the mathematical literature on fine-tuning, the works directly related to ours are the aforementioned results of Malladi et al. (2023); Jang et al. (2024). Malladi et al. (2023) presented empirical evidence of kernel behavior for fine-tuning methods like LoRA in the few-shot regime. Their main theoretical result on LoRA states that if standard (full-rank) fine-tuning exhibits kernel behavior, then low-rank fine-tuning exhibits kernel behavior. Jang et al. (2024) build upon this as follows. In the kernel regime where the student model stays close to its linearization around the base model throughout training, they consider the linearized empirical loss for arbitrary data. They show this loss has a rank-$O(\sqrt{N})$ global minimizer, where N is the number of training examples, and that all local minimizers of this loss are global minimizers.

These works are incomparable to ours as we work well outside the regime where the kernel approximation does well (Fig. 3). In addition, the aforementioned works do not handle the regime where the rank is extremely small, even though LoRA still works quite well in practice in this case. That said, there is no free lunch: our work derives insights in the challenging rank-one, non-linear setting at the cost of working with a specific set of assumptions on the data-generating process.

GLMs and single/multi-index model regression. GLMs have received significant attention in learning theory as a stylized model for feature learning, see Dudeja and Hsu (2018) for an overview

of older works on this. Various works have also considered learning *multi-index models*, i.e. functions that depend on a *bounded-dimension* projection of the input, over Gaussian examples (Bietti et al., 2022; Damian et al., 2022, 2024a; Abbe et al., 2023; Troiani et al., 2024). Most relevant to our work is Arous et al. (2021) which studied the gradient dynamics of learning GLMs models $\sigma(\langle w, \cdot \rangle)$ over Gaussian examples with online SGD. Their main finding was that online SGD achieves high correlation with the ground truth direction in $\tilde{\Theta}(d^{1 \vee l^* - 1})$ iterations/samples, where l^* is the *information exponent*, the lowest degree at which σ has a nonzero Hermite coefficient. We draw upon tools from Arous et al. (2021) to analyze online SGD in our setting, one important distinction being that the population gradient dynamics in our setting are very different and furthermore our finite-sample analysis makes quantitative various bounds that were only proved asymptotically in Arous et al. (2021).

Several works have explored whether gradient descent for feature learning can achieve scaling better than exponential in the information exponent, e.g. if one performs multiple passes over the data (Lee et al., 2024; Arnaboldi et al., 2024; Dandi et al., 2024) or via label transformations (Chen and Meka, 2020; Chen et al., 2022; Damian et al., 2024b). Recently, Damian et al. (2024b) identified an alternative property of the activation called the *generative exponent*, which can be smaller than the information exponent, that dictates the statistical query complexity of learning GLMs. We emphasize however that the generative exponent can still be quite large in general. Yet even for such activations, the runtime bounds we prove in the *fine-tuning* setting still apply and achieve *linear* dependence in d, regardless of the generative exponent.

PAC learning neural networks. Within the theoretical computer science literature on neural networks, there have been numerous works giving algorithms, many of them based on spectral or tensor methods, for learning two-layer networks from scratch over Gaussian examples. The literature is vast, and we refer to Chen and Narayanan (2024); Diakonikolas and Kane (2024) for an overview.

On the hardness side, Diakonikolas et al. (2020) (see also Goel et al. (2020)) proved that for correlational statistical query algorithms, the computational cost of learning such networks from scratch in the worst case must scale with $d^{\Omega(k)}$, which Diakonikolas and Kane (2024) recently showed is tight for this class of algorithms. Additionally, central to these lower bounds for learning two-layer networks is the existence of networks $\sum_i \lambda_i \sigma(\langle w_i, x \rangle)$ for which the tensor $\sum_i \lambda_i w_i^{\otimes s}$ vanishes for all small s. As we discuss at the end of Appendix D.1, even if the base model or teacher model satisfies this in the setting that we consider, it does not appear to pose a barrier for low-rank fine-tuning in the same way that it does for learning from scratch.

1.4. Technical preliminaries

Notation. Let $\mathbb{S}^{d-1} = \{v \in \mathbb{R}^d : \|v\| = 1\}$. For $w \in \mathbb{R}^d$, let $w^{\otimes s}$ denote the order-s tensor power of w, and for two tensors T_1, T_2 we use $\langle T_1, T_2 \rangle$ to denote their element-wise dot product and $\|T_1\|_F \triangleq \sqrt{\langle T_1, T_1 \rangle}$ for the corresponding Frobenius norm. Note the identity $\sum_{i,j=1}^k \lambda_i \mu_j \langle w_i, v_j \rangle^s = \langle \sum_{i=1}^k \lambda_i w_i^{\otimes s}, \sum_{i=1}^k \mu_i v_i^{\otimes s} \rangle$ which arises in our analysis as the interactions between different neurons in the population loss.

Bounds. Our results hold uniformly over the choice of w_i, u, λ under their constraints. We make dependencies on $\lambda_{\min} \triangleq \min_i |\lambda_i|$ and $\lambda_{\max} \triangleq \max_i |\lambda_i|$ explicit, but in our $O(\cdot)$ notation, we ignore constants that only depend on the activation σ. We write $\tilde{O}(\cdot)$ to omit logarithmic factors.

Hermite analysis. We will use Hermite analysis to analytically evaluate expectations of products of functions under the Gaussian measure. We let h_p denote the p-th normalized probabilist's Hermite polynomial, and $\mu_p(\sigma)$ the p-th Hermite coefficient of σ. In particular, Hermite coefficients form an orthonormal basis for functions that are square integrable w.r.t the Gaussian measure. That is, functions σ for which $\|\sigma\|_2^2 \triangleq \mathbb{E}_{g\sim\mathcal{N}(0,1)}[\sigma(g)^2] < \infty$ and we denote $\sigma \in L_2(\mathcal{N}(0,1))$. These functions admit a Hermite expansion $\sigma(a) = \sum_{p=0}^{\infty} \mu_p(\sigma) h_p(a)$, and for two functions $f, g \in L_2(\mathcal{N}(0,1))$, we have $\langle f, g\rangle \triangleq \mathbb{E}_{a\sim\mathcal{N}(0,1)}[f(a)g(a)] = \sum_p \mu_p(f)\mu_p(g)$. Furthermore, for $u, v \in \mathbb{S}^{d-1}$, Hermite polynomials satisfy $\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}[h_p(\langle u, x\rangle)h_q(\langle v, x\rangle)] = \mathbb{1}\{p = q\}\langle u, v\rangle^p$.

2. Training algorithm

In this work, we will focus on learning the factor u in the rank-1 perturbation $\Delta = \xi c u^\top$ from Eq. (3) using gradient descent. As the weight vectors in the teacher model are given by $w_i + \xi c_i u$, the vector u corresponds to the *direction* in which each of the pre-trained features gets perturbed. Learning this direction turns out to be the most challenging part of fine-tuning: once one has converged to a sufficiently good estimate of u, it is straightforward to learn c even using a linear method -- see Appendix F.2 for details. As such, in the student model, we will keep $\hat{c}$ frozen at random initialization and only train $\hat{u}$. Remarkably, as we will see, *the misspecification between $\hat{c}$ and the true c does not significantly affect the learning dynamics*. This robustness to misspecification suggests it may be possible to prove convergence even if c and u were jointly trained, as is done in practice, and we leave this as another important future direction.

We now specify the instantiation of online SGD that we will analyze. Let f^* denote the teacher model and (u_t) the iterates of online SGD with learning rate $\eta > 0$. Let $\hat{c} \in \{\pm 1/\sqrt{k}\}^k$ be sampled uniformly at random at initialization. The algorithm is initialized with

$$u_0 \sim \mathbb{S}_{\Pi^{\perp}_{\mathrm{span}(W)}},$$

i.e. uniformly over the set of unit vectors which are orthogonal to the span of the pre-trained features $w_1, \ldots, w_k$. Given training example $(x, f^*(x))$, define the loss attained by $\hat{u}$ on this example by

$$L(\hat{u}; x) \triangleq (f^*(x) - \lambda^\top \sigma((W_0 + \xi \hat{c}\hat{u}^\top)x))^2 .$$

Denote its *spherical gradient* by $\hat{\nabla} L(\hat{u}; x) \triangleq (I - \hat{u}\hat{u}^\top)\tilde{\nabla} L(\hat{u}; x)$. Note we are working with the gradients restricted to the subspace of training, i.e. $\tilde{\nabla} L(\hat{u}; x) \triangleq \Pi^{\perp}_{\mathrm{span}(W)} \nabla L(\hat{u}; x)$ to keep $\hat{u}$ in this subspace. The update rule is then given by the following: at each step t, defining $\mathrm{proj}(v) \triangleq v/\|v\|$,

$$u_{t+1} = \mathrm{proj}(u_t - \eta \hat{\nabla} L(u_t; x_t)), \qquad x_t \sim \mathcal{N}(0, I). \tag{4}$$

Understanding the gradient dynamics of low-rank fine-tuning in our setting therefore amounts to quantifying the convergence of u_t to the ground truth vector u.

3. Proof overview

While our results are about the convergence of online SGD, the main novelty of our analysis – and the bulk of the technical work in this paper – lies in establishing estimates for the gradients of the *population loss*. Hence, after briefly explaining why it is sufficient to analyze the population

gradient (Section 3.1), we devote the bulk of this overview to the high level ideas of its analysis (Sections 3.2.1 and 3.2.2), deferring the detailed proofs to the Appendix. As we will make clear, the analysis of the population gradient departs significantly from prior works on learning GLMs and shallow neural networks. In Appendix C.2, we describe how to put everything together to give a finite-sample analysis that proves our main results (this part largely draws on existing techniques introduced by Arous et al. (2021)). Finally, in Appendix C.3 we give a self-contained proof of Theorem 19 separating fine-tuning and learning from scratch.

3.1. Reducing convergence analysis to bounding population gradients

We begin by providing some intuition regarding the gradient dynamics. First, recall the update rule from Equation (4). We formally analyze the SGD dynamics in Appendix E; for the sake of intuition, in this overview we will use $\hat{Q}_t$ to denote the error in approximating u_{t+1} by ignoring the normalization, by writing

$$u_{t+1} = u_t - \eta \hat{\nabla} L(u_t; x_t) + \hat{Q}_t \,. \tag{5}$$

Furthermore, decompose $\hat{\nabla} L(u_t; x_t) = \hat{\nabla}\Phi(u_t) + \hat{\nabla} E(u_t; x_t)$ where ∇E is a stochastic error term with mean 0, and $\hat{\nabla}\Phi$ is the population gradient.

Now, we will show that the population gradient term $\hat{\nabla}\Phi(u_t)$ depends non-linearly on the projection $m_t \triangleq \langle u_t, u\rangle$, which will later help us derive a self-governing dynamics for $\langle u_t, u\rangle$. We begin by calculating the population gradient using Hermite analysis (see Appendix D.1 for the derivation):

Proposition 4 *Given $l, s \in \mathbb{Z}_{\geq 0}$, define*

$$T(l,s) = \begin{cases} \left\| \sum_i \lambda_i w_i^{\otimes s} \right\|_F^2 & l \text{ odd} \\ k \left\langle \sum_i \lambda_i c_i w_i^{\otimes s}, \sum_i \lambda_i \hat{c}_i w_i^{\otimes s} \right\rangle & \text{otherwise} \end{cases}$$

Define $h : \mathbb{R} \to \mathbb{R}$ by

$$h(m) = 2 \sum_{l=0}^{\infty} \left(\frac{\xi^2}{k}\right)^{l+1} \left(\sum_{s=0}^{\infty} \binom{l+s}{l} (l+s+1) \mu_{l+s+1}(\sigma)^2 \left(\frac{1}{1+\xi^2/k}\right)^{l+s+1} T(l,s) \right) m^l \,.$$

Then at any $\hat{u} \in \mathbb{S}^{d-1}$, the population spherical gradient is given by

$$\hat{\nabla}\Phi(\hat{u}) \triangleq (I - \hat{u}\hat{u}^\top)\nabla\Phi(\hat{u}) = -h(\langle u, \hat{u}\rangle)(u - \hat{u}\langle \hat{u}, u\rangle) \,.$$

Notice that the population gradient is in the direction of the spherical projection of the ground truth onto $\hat{u}$. Note that the form of the population gradient is significantly more involved than the analogous object in prior works on GLM regression, where the population loss takes the form $2\sum_{p=0}^{\infty} \mu_p(\sigma)^2(1-m^p)$. In the GLM setting, we have $h(m) \propto \sum_{p=1}^{\infty} p \cdot \mu_p(\sigma)^2 m^{p-1}$, and establishing bounds on the population gradient is thus straightforward. In our setting, because the base network involves $k > 1$ neurons and the low-rank perturbation introduces additional degrees of freedom due to the variables c_i, controlling the population gradients becomes significantly more subtle compared to previous work.

Now, let $m_t = \langle u_t, u\rangle$ denote the alignment between the t-th iterate with the ground truth direction u, so that by Eq. (5),

$$m_{t+1} = m_t + \eta h(m_t)(1 - m_t^2) + \eta \langle \hat{\nabla} E_t(u_t; x_t), u\rangle + Q_t \,,$$

where $Q_t = \langle \hat{Q}_t, u \rangle$ is the error from ignoring the normalization. Unrolling the recursive expression and defining $E_t = \hat{\nabla} E(u_t; x_t)$, we obtain

$$m_t = m_0 + \eta \sum_{j=0}^{t-1} h(m_j)(1 - m_j^2) + \eta \sum_{j=0}^{t-1} \langle E_j, u \rangle + \sum_{j=0}^{t-1} Q_j \,. \tag{6}$$

Note that the terms $M_t = \eta \sum_{j=0}^{t-1} \langle E_j, u \rangle$ form a martingale, and $\sum_{j=0}^{t-1} Q_j$ is a stochastic error term. Hence, the dynamics is given by (i) a progress term coming from the population gradient and a (ii) noise term due to the randomness of the data and the error due to spherical projection. The main challenge of this work is to quantify the signal in the population gradient, as opposed to previous work that focused on the finite-sample aspect (e.g. Arous et al. (2021)). In particular, we show that if $h(m)$ is lower bounded by S_k and the gradients have variance V_k, with the right choice of step size we can strongly recover the ground truth u in $O(V_k/S_k^2)$ iterations.

Theorem 5 (Informal, see Theorem 16) *Suppose $h(m)$ is "nice" and lower bounded by S_k, the variance of the gradients are upper bounded by V_k. Let $m_t = \langle u_t, u \rangle$. Then, online SGD run with step size $\eta = \frac{\delta}{dV_k}$ and scaling $\delta = \min\{\frac{S_k \epsilon^3}{4\mu_1 (\log dV_k)^2}, 1\}$ for time $T = \lceil \alpha d V_k \rceil$ with time scaling $\alpha = \frac{4(\log dV_k)}{\epsilon \delta S_k}$ and initialization at $|m_0| \geq \beta/\sqrt{d}$ with $m_0 h(0) > 0$, the correlation satisfies $|m_t| \geq 1 - \epsilon$ at time T.*

See Appendix E for the proof. Henceforth, we will focus on lower bounding the population gradient.

3.2. Population gradient lower bounds

As explained in the previous section, the key step is to show that the population gradient is lower bounded, which involves lower bounding $h(m)$. This is the heart of our technical contribution, and it represents a significant departure from the previous works Arous et al. (2021) as our setup involves significantly more degrees of freedom due to various choices of $w_i, \lambda, c_i, \hat{c}_i, \xi$.

In this section, we overview the key technical claims we prove that subsequently form the core of our proof. Mainly in the two scaling regimes $\xi = \Theta(1)$ and $\xi = \Theta(\sqrt{k})$, we show that with respect to the randomness in the problem initialization, the population gradient will be lower bounded with high probability in d, k. To begin the analysis, we start with the case of orthonormal weights in the large scaling regime and later show a similar result for weights that satisfy the mild separation condition in Assumption 5.

3.2.1. GRADIENT BOUNDS UNDER ORTHONORMAL FEATURES

First we assume $w_1, \ldots, w_k$ are orthonormal, so that the form of $T(l, s)$ in Theorem 4 reduces to:

$$T(l,0) = \begin{cases} k \sum_{i,j=1}^{k} \lambda_i \lambda_j c_i \hat{c}_j & l \text{ even} \\ \left(\sum_{i=1}^{k} \lambda_i\right)^2 & l \text{ odd} \end{cases} \quad \text{and} \quad T(l, s \geq 1) = \begin{cases} k \sum_{i=1}^{k} \lambda_i^2 c_i \hat{c}_i & l \text{ even} \\ \|\lambda\|_2^2 & l \text{ odd} \end{cases}$$

which greatly simplifies our analysis since all the terms where $s \geq 1$ scale with the same expression. Then, notice that we can decompose h into the odd powers of l and even powers of l as

$$h(m) = 2\Big[k\sum_{i,j=1}^{k}\lambda_i\lambda_j c_i\hat{c}_j\Big]\sum_{l\geq 0 \text{ even}}\Big(\frac{\xi^2}{k}\Big)^{l+1}(l+1)\mu_{l+1}(\sigma)^2\Big(\frac{k}{k+\xi^2}\Big)^{l+1}m^l$$
$$+ 2\Big[k\sum_{i=1}^{k}\lambda_i^2 c_i\hat{c}_i\Big]\sum_{\substack{l\geq 0 \text{ even}\\ s\geq 1}}\Big(\frac{\xi^2}{k}\Big)^{l+1}\binom{l+s}{l}(l+s+1)\mu_{l+s+1}(\sigma)^2\Big(\frac{k}{k+\xi^2}\Big)^{l+s+1}m^l + \sum_{l\geq 1 \text{ odd}} b_l m^l\,,$$

for some non-negative coefficients $b_l \geq 0$. Informally, over the randomness of $c, \hat{c}$, the typical magnitude of $k\sum_{i,j=1}^{k}\lambda_i\lambda_j c_i\hat{c}_j$ is $\Theta(k)$, and the typical magnitude of $k\sum_{i=1}^{k}\lambda_i^2 c_i\hat{c}_i$ is $\Theta(\sqrt{k})$. Hence, first, we show that the odd terms with $s = 1$ dominate the even terms with $l > 0, s = 0$. Then, this will mean that the sign of all of these terms will be governed by the sign of m. Hence, as long as m has the same sign as $h(0)$, we expect $h(m)$ to be lower bounded throughout training. To this end, we first state our bound on the even terms with $s = 0$ and then discuss how the $s > 0$, $l = 0$ terms are bounded (see Appendix D.4 for proof).

Claim 1 (Even l, $s = 0$ contribution) *Let Assumption 2 hold. With probability $1 - \exp\{-\frac{2k}{e\xi^2}\}$ over the randomness of $c, \hat{c}$, the following holds.*

$$\text{sign}(m)\Big(2\sum_{\substack{l \text{ odd}\\ s=1}}\Big(\frac{\xi^2}{k}\Big)^{l+1}\binom{l+s}{l}(l+s+1)\mu_{l+s+1}(\sigma)^2\Big(\frac{1}{1+\xi^2/k}\Big)^{l+s+1}T(l,s)$$
$$+ 2\sum_{\substack{l>0\\ \text{even}\\ s=0}}\Big(\frac{\xi^2}{k}\Big)^{l+1}\binom{l+s}{l}(l+s+1)\mu_{l+s+1}(\sigma)^2\Big(\frac{1}{1+\xi^2/k}\Big)^{l+s+1}T(l,s)\Big) \geq 0\,.$$

The proof of this claim relies on using concentration bounds on $T(l, s)$ for even s, and showing that the coefficients of the even terms are bounded by the odd terms up to some constant. This proof is deferred to Appendix D.4. Then, this result should hold with high probability when $\xi = \bar{\xi}\sqrt{k}$, where the probability is exponentially high as $\bar{\xi} \to 0$. Then, we can group these terms together with the odd terms, so that we have the following:

$$h(m) = 2\left[k\sum_{i,j=1}^{k}\lambda_i\lambda_j c_i\hat{c}_j\right]\mu_1(\sigma)^2\Big(\frac{\xi^2}{k+\xi^2}\Big) \tag{7}$$
$$+ 2\Big[k\sum_{i=1}^{k}\lambda_i^2 c_i\hat{c}_i\Big]\sum_{\substack{l\geq 0 \text{ even}\\ s\geq 1}}\Big(\frac{\xi^2}{k}\Big)^{l+1}\binom{l+s}{l}(l+s+1)\mu_{l+s+1}(\sigma)^2\Big(\frac{k}{k+\xi^2}\Big)^{l+s+1}m^l \tag{8}$$
$$+ \tilde{h}(m), \tag{9}$$

where $\tilde{h}(m)m > 0$, with high probability due to Claim 1. In this case, note that we have two cases:

1. $\mu_1(\sigma) \neq 0$: In this case, note the term in eq. (7) is $\Theta(k)$, whereas the term in eq. (8) is $O(\sqrt{k})$ since the sum in eq. (8) is $O(1)$. Hence, we show that the second term is negligible relative to the first, and furthermore that the sign of $h(0)$ is the same as the sign of eq. (7).

2. $\mu_1(\sigma) = 0$: In this case, $h(0)$ has the same sign as all the even terms, so as long as $mh(0) > 0$, we can lower bound $|h(m)|$ with $\Theta(\sqrt{k})$.

Then, the rest of the proof is giving a uniform lower bound on $|h(m)|$ training when $mh(0) > 0$. We do this by showing that the quantities

$$\left(\sum_{i=1}^{k} \lambda_i c_i\right)\left(\sum_{i=1}^{k} \lambda_i \hat{c}_i\right) \qquad \text{and} \qquad \sum_{i=1}^{k} \lambda_i^2 c_i \hat{c}_i$$

are upper and lower bounded with high probability. The upper bounds follow directly since these are quadratic polynomials of Rademacher variables. For the lower bounds, we prove the anti-concentration of these quantities by showing that as functions of c_i, these are "low-influence" quadratic polynomials, so a central limit theorem type result allows us to swap the c_i with Gaussian variables (see Appendix D.3). Here, we show the following:

Lemma 6 (Population gradient lower bounds, orthonormal case) *Suppose Assumptions 1 to 3 hold and $\langle w_i, w_j \rangle = 0$ for all $i \neq j$. Let s^* be the smallest $s \geq 1$ s.t. $\mu_s(\sigma) \neq 0$. With probability $1 - o(1) - \exp\left\{-\frac{2k}{e\xi^2}\right\} - O(\gamma^{1/2})$, for $mh(0) \geq 0$ we have $h(m)\text{sign}(h(0)) \geq |h(0)|/2$ and*

$$|h(0)| \geq \left(\frac{k}{k+\xi^2}\right)^{s^*} C_\sigma \gamma \xi^2 \cdot \begin{Bmatrix} 1 & \mu_1(\sigma) \neq 0 \\ \frac{1}{\sqrt{k}} & \mu_1(\sigma) = 0 \end{Bmatrix}$$

The proof of this lower bound is deferred to Appendix D.4. By the above discussion, this lower bound on the population gradient can then be used to derive our main convergence guarantee for orthonormal features in Theorem 1.

3.2.2. GRADIENT BOUNDS UNDER ANGULARLY SEPARATED FEATURES

Whem $\xi = \Theta(1)$, we can relax the orthogonality assumption, and instead work with angularly separated features w_i, as stated in Assumption 5. For notational simplicity, we consider the case $\xi = 1$, but the analysis is identical when $\xi = \Theta(1)$. Recall the non-linear function h that drives the population dynamics, which simplifies to:

$$h(m) = 2\sum_{l=0}^{\infty}\left(\frac{1}{k}\right)^{l+1}\sum_{s=0}^{\infty}\binom{l+s}{l}(l+s+1)\mu_{l+s+1}(\sigma)^2\left(\frac{k}{k+1}\right)^{l+s+1}T(l,s)m^l$$

Now, note if one could truncate the infinite sum over s at some finite value, the contribution of the higher order terms could be bounded. However, this is not obvious. In fact, one can verify that $\sum_{s=0}^{\infty}\binom{l+s}{l}\left(\frac{k}{k+1}\right)^{s+1} = \Theta(k^l)$ so a naive truncation of the sum over l will not allow to simplify the population gradient. However, note that within $T(l,s)$, the contribution of the non-diagonal ($i \neq j$) terms in the inner products $\langle \sum_i \lambda_i c_i w_i^{\otimes s}, \sum_i \lambda_i \hat{c}_i w_i^{\otimes s} \rangle = \sum_{i,j=1}^{k} \lambda_i \lambda_j c_i \hat{c}_j \langle w_i, w_j \rangle^s$ vanishes as $s \to \infty$, as $\langle w_i, w_j \rangle$ is bounded away from 1 by Assumption 5. Indeed, if $|\langle w_i, w_j \rangle| \leq \nu$ for $i \neq j$, then $|\langle w_i, w_j \rangle|^{\frac{\gamma}{\nu}} \leq \exp\{-\gamma\}$. Furthermore, note that for the diagonal ($i = j$) terms, we can use the decay of the μ_p to bound their contribution to the terms in $h(m)$ corresponding to large s. Finally, for the small s terms (e.g. $s = O(\sqrt{k})$) we simply use the fact that the sum over l in the definition

of $h(m)$ has a factor $(1/k)^{l+1}$ which is decaying as l increases. In summary, we can handle the small s terms because their contribution is not that large, and once we reach s with non-vanishing contribution, we use the decay of $\mu_p(\sigma)$ and the angular separation of the w_i's. We obtain the following bounds on the higher order terms in m with even exponent (see Appendix D.5 for proof):

Proposition 7 *Suppose Assumptions 1 to 6 hold. Then, with probability at least* $1 - \frac{1}{k^3}$ *over the randomness of* $c, \hat{c}$*, for* $\epsilon = \min\{\frac{\rho}{4}, 1 - \frac{1}{1+2\rho}\}$,

$$\sum_{n=0}^{\infty}\Big(\frac{1}{k}\Big)^{2n+2}\sum_{s=0}^{\infty}\binom{2n+2+s}{2n+2}(2n+s+3)\mu_{2n+s+3}(\sigma)^2\Big(\frac{k}{k+1}\Big)^{2n+s+3}\Big\langle\sum_{i=1}^{k}\lambda_i c_i w_i^{\otimes s}, \sum_{i=1}^{k}\lambda_i \hat{c}_i w_i^{\otimes s}\Big\rangle$$

is upper bounded by $O(\lambda_{\max}^2 k^{-\frac{1}{2}-\epsilon})$.

Once we bound the higher order terms in m with even exponent, it remains to show the constant term $h(0)$ dominates the even terms. This follows immediately using the anti-concentration results for $h(0)$, since $|h(0)| = \Omega(\frac{1}{\sqrt{k}})$ with high probability. Then, since the odd terms will have the same sign as m, the rest of the proof is similar to the orthonormal case. In this case, we get the following lower bound on h:

Lemma 8 *With probability* $1 - O(\gamma^{1/2}) - o\left(\frac{\lambda_{\max}^2}{\lambda_{\min}^2}\right)$*, for* $mh(0) \geq 0$ *we have* $h(m)\text{sign}(h(0)) \geq \frac{|h(0)|}{2} \geq \frac{\gamma\lambda_{\min}^2}{2\sqrt{k}}$.

These lower bounds show that the population gradient is lower bounded by some quantity S_k depending on ξ, and combined with the finite sample analysis (see Appendix A.2 and Appendix E) concludes the proof of our main convergence guarantee for separated features in Theorem 2.

4. Outlook

In this work we took the first steps towards understanding the gradient dynamics of low-rank fine-tuning beyond NTK. We identified a rich student-teacher framework, specialized to two-layer networks, and proved in various settings that online SGD efficiently finds the ground truth low-rank perturbation. This student-teacher framework is also appealing because it offers a natural way of interpolating between fine-tuning in the lazy training regime and generalized linear model regression in the feature learning regime. The parameter regime we consider occupies an intriguing middle ground between these extremes where the dynamics are nonlinear yet tractable and not overly sensitive to fine-grained properties like the Hermite coefficients of the activation function.

Our results open up a number of future directions. Firstly, it is important to try to lift our assumptions, in particular the orthogonality of the perturbation relative to the pre-trained features, the assumption that c is quantized, and the assumption that c is random.

For these questions, a fruitful starting point could be to target a specific, analytically tractable activation like quadratic, especially given that based on our findings, the dynamics of low-rank fine-tuning do not depend heavily on particulars of σ. For this special case, we could hope to go beyond Hermite analysis and potentially even obtain an exact characterization of the dynamics.

Other important directions include analyzing the dynamics when $\hat{c}$ and $\hat{u}$ are jointly trained – Fig. 2 suggests that this is roughly twice as efficient as freezing $\hat{c}$ and training $\hat{u}$ in isolation – as well as going beyond two layers and rank-1 perturbations. Finally, it would be interesting to understand the *worst-case complexity* of fine-tuning: are there computational-statistical gaps in this setting?

Acknowledgments

SC is grateful to Adam Klivans, Vasilis Kontonis, and Raghu Meka for enlightening discussions about low-rank fine-tuning, and was supported in part by NSF SLES IIS-2331831. AKD acknowledges support from the Summer Program for Undergraduates in Data Science, during which much of this work was completed.

References

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

Luca Arnaboldi, Yatin Dandi, Florent Krzakala, Luca Pesce, and Ludovic Stephan. Repetita iuvant: Data repetition allows sgd to learn high-dimensional multi-index functions. *arXiv preprint arXiv:2405.15459*, 2024.

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

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. *Advances in Neural Information Processing Systems*, 35:9768–9783, 2022.

Anthony Carbery and James Wright. Distributional and L^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical research letters*, 8(3):233–248, 2001.

Sitan Chen and Raghu Meka. Learning polynomials in few relevant dimensions. In *Conference on Learning Theory*, pages 1161–1227. PMLR, 2020.

Sitan Chen and Shyam Narayanan. A faster and simpler algorithm for learning shallow networks. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 981–994. PMLR, 2024.

Sitan Chen, Adam R Klivans, and Raghu Meka. Learning deep relu networks is fixed-parameter tractable. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 696–707. IEEE, 2022.

Sitan Chen, Zehao Dou, Surbhi Goel, Adam R Klivans, and Raghu Meka. Learning narrow one-hidden-layer relu networks. *arXiv preprint arXiv:2304.10524*, 2023.

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

Alex Damian, Loucas Pillaud-Vivien, Jason D Lee, and Joan Bruna. The computational complexity of learning gaussian single-index models. *arXiv preprint arXiv:2403.05529*, 2024b.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pages 5413–5452. PMLR, 2022.

Yatin Dandi, Emanuele Troiani, Luca Arnaboldi, Luca Pesce, Lenka Zdeborová, and Florent Krzakala. The benefits of reusing batches for gradient descent in two-layer networks: Breaking the curse of information and leap exponents. *arXiv preprint arXiv:2402.03220*, 2024.

Ilias Diakonikolas and Daniel M Kane. Efficiently learning one-hidden-layer relu networks via schurpolynomials. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1364–1378. PMLR, 2024.

Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, and Nikos Zarifis. Algorithms and sq lower bounds for pac learning one-hidden-layer relu networks. In *Conference on Learning Theory*, pages 1514–1539. PMLR, 2020.

Rishabh Dudeja and Daniel Hsu. Learning single-index models in gaussian space. In *Conference On Learning Theory*, pages 1887–1930, 2018.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

Surbhi Goel, Aravind Gollakota, Zhihan Jin, Sushrut Karmalkar, and Adam Klivans. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. In *International Conference on Machine Learning*, pages 3587–3596. PMLR, 2020.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

Uijeong Jang, Jason D Lee, and Ernest K Ryu. Lora training in the ntk regime has no spurious local minima. *arXiv preprint arXiv:2402.11867*, 2024.

Jason D Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with sgd near the information-theoretic limit. *arXiv preprint arXiv:2406.01581*, 2024.

Sadhika Malladi, Alexander Wettig, Dingli Yu, Danqi Chen, and Sanjeev Arora. A kernel-based view of language model fine-tuning. In *International Conference on Machine Learning*, pages 23610–23641. PMLR, 2023.

Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*, pages 21–30. IEEE, 2005.

Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

Emanuele Troiani, Yatin Dandi, Leonardo Defilippis, Lenka Zdeborová, Bruno Loureiro, and Florent Krzakala. Fundamental limits of weak learnability in high-dimensional multi-index models. In *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*, 2024.

Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. In *The Twelfth International Conference on Learning Representations*, 2024.

Appendices

Contents

Appendix A. Main results

Here we provide a formal statement of our results. Before we do so, we formally state the assumption on the activation we use which holds for a large class of relevant activations.

A.1. Assumptions on the activation function

For completeness, we first state the assumptions we make on the activation function σ. These are exceedingly mild and hold for many standard classes of activations including Lipschitz activations (e.g. ReLU, absolute value, sigmoid, tanh) and polynomial activations. As such, this section can be safely skipped upon a first reading. These conditions are only relevant when we need to bound the variance of gradients in Appendix D.2.

Assumption 6 (Activation function) *The activation σ satisfies all of the following:*

1. *σ is almost surely differentiable (with respect to the standard Gaussian measure), with derivative σ' having at most polynomial growth: There exists some $b, c, q > 0$ such that $|\sigma'(a)| \leq b + c|a|^q$ for all $a \in \mathbb{R}$.*

2. *The Hermite coefficients of σ have faster than linear decay: There exists $C_\sigma, \rho > 0$ such that $|\mu_p(\sigma)| \leq C_\sigma p^{-1-\rho}$.*

3. *σ satisfies the following moment condition: For $g_1, g_2 \sim \mathcal{N}(0,1)$ Gaussians (potentially correlated), for some $C_{p,\sigma} > 0$ that only depends the activation and p, we have*

$$(\mathbb{E}|\sigma(g_1) - \sigma(g_2)|^p)^{1/p} \leq C_{p,\sigma} \left(\mathbb{E}|g_1 - g_2|^{2p}\right)^{1/(2p)}$$

Remark 9 *These conditions are satisfied for any reasonable activation used in practice. For the last condition in Assumption 6, note that it holds for any Lipschitz activation (e.g. ReLU, absolute value, sigmoid). Furthermore it is satisfied for any polynomial activation (e.g. finite Hermite expansion). To see why, for a degree s polynomial $p(x) = \sum_{n=0}^{s} a_n x^n$, note that*

$$\left|\sum_{n=1}^{s} a_n g_1^n - \sum_{n=1}^{s} a_n g_2^n\right| \leq s \max\{|g_1|^{s-1}, |g_1|^{s-2}|g_2|, \ldots, |g_2|^{s-1}\}\left(\sum_{n=1}^{s} |a_n|\right)|g_1 - g_2|$$

Then, applying Cauchy-Schwarz, we have

$$\sqrt[p]{\mathbb{E}|p(g_1) - p(g_2)|^p} \leq s\left(\sum_{n=1}^{s} |a_n|\right)\left(\mathbb{E}\max\{|g_1|^{s-1}, \ldots, |g_2|^{s-1}\}^{2p}\right)^{1/(2p)}\left(\mathbb{E}|g_1 - g_2|^{2p}\right)^{1/(2p)}$$

Using the fact that the first expectation can be bounded by a constant that only depends on s concludes the result.

Remark 10 (Generalizing assumption on the activation) *Our analysis can be generalized to activations σ that are absolutely continuous on any inverval $[a, b]$, and the derivative has polynomial growth.*

To see why, note for such functions the derivative is square integrable which implies a $\mu_p(\sigma') = o(1/\sqrt{p})$ rate on the decay of the derivative's coefficients. Using integration by parts, one can show $\mu_p(\sigma) = \frac{1}{\sqrt{p}}\mu_{p-1}(\sigma') = o(1/p)$. Then, the proofs can be modified accordingly to use the rate $o(1/p)$ instead of $O(1/p^{1+\rho})$.

Similarly, for the moment condition, note

$$\begin{aligned}|\sigma(g_2) - \sigma(g_1)|^p = \left|\int_{g_1}^{g_1} \sigma'(s)ds\right|^p &= |g_1 - g_2|^p \left|\frac{1}{g_1 - g_2}\int_{g_1}^{g_2} \sigma'(s)ds\right|^p \\ &\le |g_1 - g_2|^{p-1}\int_{g_1}^{g_2} |\sigma'(x)|^p ds \\ &\le |g_1 - g_2|^p \sup_{s\in[g_1,g_2]} |\sigma'(s)|^p\end{aligned}$$

Then, one can use polynomial growth $|\sigma'(s)| \le a + b|s|^q$ and $|s| \le |g_1| + |g_2|$, and cauchy schwarz to obtain the moment condition.

A.2. Results for fine-tuning with online SGD in different regimes

We consider two different settings: when the weights of the base model are orthogonal and when they have only mild angular separation (Assumption 5). For the former (resp. latter), we prove convergence guarantees when the perturbation has norm as large as the Frobenius (resp. spectral) norm of the base weight matrix.

A.2.1. ORTHOGONAL WEIGHTS

In this section, we assume $\langle w_i, w_j\rangle = 0$ for all $i \neq j$. Then, we can show convergence guarantees for fine-tuning whenever $\xi = O(\sqrt{k})$. In particular, we state the results for $\xi = 1$ and $\xi = \bar{\xi}\sqrt{k}$ in this section.

Theorem 11 (Orthogonal weights, $\xi = 1$) *Let Assumption 2 hold, and $0 < \epsilon < 1$ be given and $\gamma > 0$. For a sufficiently small $C_\delta = \Theta(1)$, set*

$$\delta = \frac{C_\delta \gamma \lambda_{\min}^2 \epsilon^3}{(\log \lambda_{\max}^4 dk^2)^2} \cdot \begin{Bmatrix} 1 & \mu_1(\sigma) \neq 0 \\ \frac{1}{\sqrt{k}} & \text{o.w.} \end{Bmatrix} \quad \text{and} \quad \alpha = \frac{\log(\lambda_{\max}^4 dk^2)}{\lambda_{\min}^2 \gamma \epsilon \delta} \cdot \begin{Bmatrix} 1 & \mu_1(\sigma) \neq 0 \\ \sqrt{k} & \text{o.w.} \end{Bmatrix}$$

Then, with probability $1 - o(\lambda_{\max}^2/\lambda_{\min}^2) - O(\gamma^{1/2})$ over the randomness of $c, \hat{c}$ and for intialization satisfying $\langle u_0, u\rangle \mathrm{sign}(h(0)) \ge \beta/\sqrt{d}$, online SGD run with step size η and time T for

$$\eta = \frac{\delta}{\lambda_{\max}^4 dk^2} \quad \text{and} \quad T = \lceil \alpha \lambda_{\max}^4 dk^2 \rceil = \widetilde{O}\Big(\frac{\lambda_{\max}^4}{\lambda_{\min}^4 \gamma^2} \cdot \frac{dk^3}{\epsilon^4}\Big) \cdot \begin{Bmatrix} 1 & \mu_1(\sigma) \neq 0 \\ k & \text{o.w.} \end{Bmatrix}$$

satisfies $\langle u_T, u\rangle^2 \ge 1 - \epsilon$ with high probability over the randomness of the data.

Theorem 12 (Orthogonal weights, $\xi = \bar{\xi}\sqrt{k}$) *Let Assumption 2 hold, $0 < \epsilon < 1$ and $\xi = \bar{\xi}\sqrt{k}$ for some small $\bar{\xi} > 0$. For a sufficiently small $C_\delta = \Theta(1)$, set*

$$\delta = \min\left\{ \frac{C_\delta \bar{\xi}^2 \sqrt{k} \gamma \lambda_{\min}^2 \epsilon^3}{(\log \lambda_{\max}^4 dk^2)^2} \cdot \begin{Bmatrix} \sqrt{k} & \mu_1(\sigma) \neq 0 \\ 1 & \text{o.w.} \end{Bmatrix}, 1 \right\} \quad \text{and} \quad \alpha = \frac{\log(\lambda_{\max}^4 dk^2)}{\bar{\xi}^2 \lambda_{\min}^2 \sqrt{k} \gamma \epsilon \delta} \cdot \begin{Bmatrix} 1/\sqrt{k} & \mu_1(\sigma) \neq 0 \\ 1 & \text{o.w.} \end{Bmatrix}$$

Then, with probability $1 - o(\lambda_{\max}^2/\lambda_{\min}^2) - \exp(-2/(e\bar{\xi})) - O(\gamma^{1/2})$ *randomness of* $c, \hat{c}$*, for initializations satisfying* $\langle u_0, u\rangle \cdot \operatorname{sign}(h(0)) \geq \beta/\sqrt{d}$*, online SGD run with step size* η *and time* T *for*

$$\eta = \frac{\delta}{\bar{\xi}^2 \lambda_{\max}^4 dk^4} \qquad \text{and} \qquad T = \lceil \alpha \lambda_{\max}^4 \bar{\xi}^2 dk^4 \rceil = \tilde{O}\Big(\frac{\lambda_{\max}^4}{\lambda_{\min}^4 \gamma^2 \bar{\xi}^2} \cdot \frac{dk^4}{\epsilon^4}\Big) \cdot \begin{cases} \frac{1}{k} & \mu_1(\sigma) \neq 0 \\ 1 & \mu_1(\sigma) = 0 \end{cases}$$

satisfies $\langle u_T, u\rangle^2 \geq 1 - \epsilon$ *with high probability over the randomness of the data.*

Note that in the theorems above, we make the very mild assumption that the initialization satisfies $m_0 \geq (\beta/\sqrt{d}) \cdot \operatorname{sign}(h(0))$. Note that the magnitude condition $|m_0| \geq \beta/\sqrt{d}$ is standard in this kind of analysis and will be satisfied with probability $1 - O(\beta)$ since random unit vectors in d dimensions have correlation of order $1/\sqrt{d}$. Hence, we think of β as a small constant, whose value only effects the initialization condition probability and some other tail probabilities. As long as $\beta = \Theta(1)$ (no dimension dependence), we can increase the probability of the initialization condition to be arbitrarily close to 1, and all of our results hold with high probability as $d \to \infty$. For the sign condition $\operatorname{sign}(m_0) = \operatorname{sign}(h(0))$, this holds with probability $1/2$ and, if it doesn't, we can simply re-run the algorithm initialized at $-u_0$.

A.2.2. ANGULARLY SEPARATED WEIGHTS

In this section, we do not assume the pre-trained weights are orthogonal and instead only assume they are not too correlated (Assumption 5). Then, we have the following result for $\xi = 1$.

Theorem 13 (Angularly separated weights, $\xi = 1$**)** *Let Assumption 5 hold, and* $0 < \epsilon < 1$*. For a sufficiently small* $C_\delta = \Theta(1)$*, let*

$$\delta = \frac{C_\delta \gamma \lambda_{\min}^2 \epsilon^3}{(\log \lambda_{\max}^4 dk^2)^2 \sqrt{k}} \qquad \text{and} \qquad \alpha = \frac{\log(\lambda_{\max}^4 dk^2)\sqrt{k}}{\lambda_{\min}^2 \gamma \epsilon \delta}$$

Then, with probability $1 - o(1) - O(\gamma^{1/2})$ *randomness of* $c, \hat{c}$*, for initializations satisfying* $\langle u_0, u\rangle \cdot \operatorname{sign}(h(0)) \geq \frac{\beta}{\sqrt{d}}$*, online SGD run with step size* η *for time* T *where*

$$\eta = \frac{\lambda_{\max}^4 \delta}{dk^2} \qquad \text{and} \qquad T = \lceil \alpha \lambda_{\max}^4 dk^2 \rceil = \tilde{O}\Big(\frac{\lambda_{\max}^4}{\lambda_{\min}^4 \gamma^2} \cdot \frac{dk^3}{\epsilon^4}\Big)$$

satisfies $\langle u_T, u\rangle^2 \geq 1 - \epsilon$ *with high probability over the randomness of the data.*

Remark 14 (Other algorithms for fine-tuning) *Furthermore, the focus of our paper is to understand why* gradient-based *fine-tuning works, motivated by the practical success of LoRA, we would like to note that there are potentially algorithms beyond gradient descent that can solve our proposed learning problem. For example, suppose* $\sigma = \mathrm{ReLU}$ *and let the base model be given by* $f_{\theta_0}(x) = \sum_i \lambda_i \sigma(\langle w_i, x\rangle)$ *and the teacher model be given by* $f_\theta(x) = \sum_i \lambda_i \sigma(\langle w_i + c_i u, x\rangle)$*. Then, note* $\mathbb{E}_x[x(f_{\theta_0}(x) - f_\theta(x))] = \mu_1(\sigma)\left(\sum_{i=1}^k \lambda_i c_i\right) u$*, where* $\mu_1(\sigma)$ *denotes the first normalized Hermite coefficient of the ReLU activation* σ*. This means the ground truth vector* u *could be recovered when* $\sum_i \lambda_i c_i$ *and* $\mu_1(\sigma)$ *do not vanish. One can then recover* c *using our algorithm in Appendix F.2.*

On the other hand, if $\sum_i \lambda_i c_i$ and $\mu_1(\sigma)$ vanish, it is unclear how to proceed. It might be possible to use higher order moment tensors to solve the fine tuning problem, but it is open to understand at what level of generality this would work and whether such an algorithm can achieve sample complexity scaling linearly in d as we show online SGD does. We leave finding general algorithms for the low-rank fine-tuning problem and worst-case lower bounds as an interesting orthogonal direction to be explored.

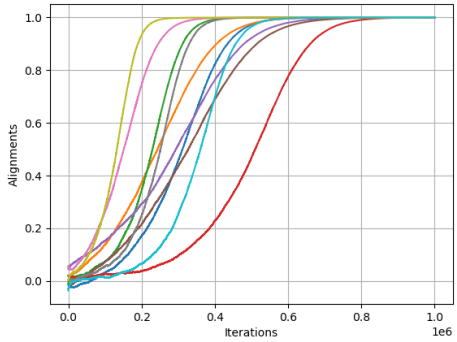
Appendix B. Numerical simulations

In this section we empirically demonstrate (i) the robustness of our theory even as one deviates from the simplifying assumptions we make, (ii) the distinction between the regime we study versus the kernel regime, and (iii) the robustness of the training dynamics to the choice of activation function. In particular, for (i) we provide simulations in which we relax the assumption that c_i are quantized, and we also compare the cases when $\hat{c}$ is frozen versus jointly trained with $\hat{u}$. For (ii), we show that linearized networks (kernel approximation) fail at $\xi = \Theta(\sqrt{k})$, and also illustrate some interesting behavior in the joint training of $\hat{u}$ and $\hat{c}$. For (iii), we corroborate our theoretical finding that the dynamics of fine-tuning are benign essentially regardless of the choice of σ, in sharp contrast to what happens when one trains from scratch.

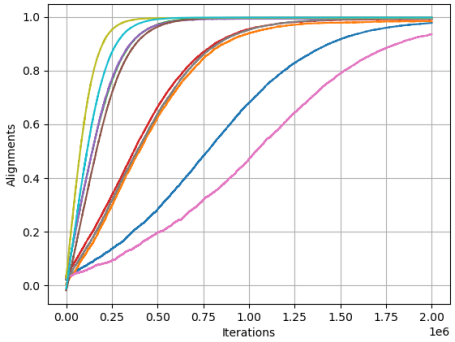
We use the ReLU activation throughout our simulations. We let $f(x) = \frac{1}{\xi}\sum_{i=1}^{k} \lambda_i \sigma(\langle v_i, x\rangle)$ where $v_i = \frac{k}{k+\xi^2}(w_i + \xi c_i u)$ where the $1/\xi$ is to keep the magnitude of gradients consistent. Throughout our simulations, we set $d = 2000$, $k = 50$, and sample the $w_i \in \mathbb{S}^{d-1}$ and $c \in \mathbb{S}^{k-1}$ uniformly at random.

First, in the $\xi = \Theta(1)$ scaling, we plot 10 training curves for random problem instances (see below) for joint training (Figure 2(a)) and when $\hat{c}$ is frozen (Figure 2(b)). Notably, we see that while freezing $\hat{c}$ leads to somewhat longer time scales in training, the qualitative behavior of $\langle u_t, u\rangle$ is similar across the two settings.

Next, we consider the $\xi = \Theta(\sqrt{k})$ scaling, keeping the rest of the setup the same as above. We plot low-rank fine-tuning in orange ($\hat{u}$ and $\hat{c}$ are jointly trained) and linearized training in blue. For the linearization, we Taylor expand around the base model. In Figure 3(a), We demonstrate that linearized dynamics do not explain fine tuning in this regime. Furthermore, when jointly training $\hat{u}$ and $\hat{c}$, we observe there is an initial phase where the loss is high and $\langle u_t, u\rangle$ is increasing but $\langle c_t, c\rangle$ stays at a low level (Figure 3(b)). This suggests that the initial phase of joint training might be similar to the training with frozen $\hat{c}$.

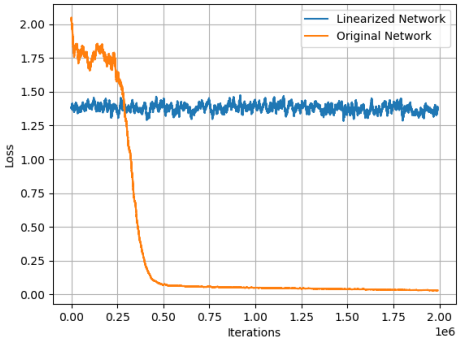


(a) Joint training of $\hat{c}$ and $\hat{u}$, $\xi = \Theta(1)$

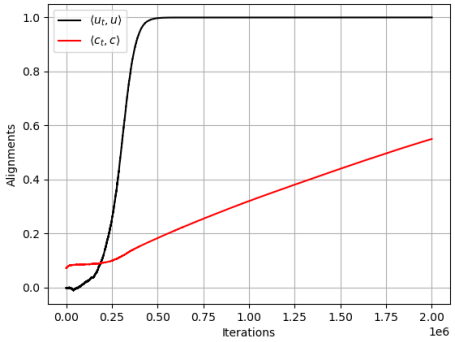


(b) Freezing $\hat{c}$ and training $\hat{u}$, $\xi = \Theta(1)$

Figure 2: Evolution of $\langle u_t, u \rangle$ during online SGD for 10 random instances with joint and frozen-$\hat{c}$ training. Though time scales differ between (a) and (b), trajectories exhibit similar behavior.



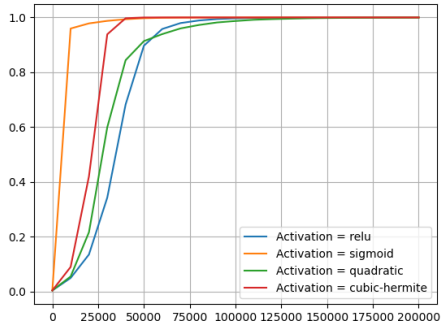
(a) Loss curves for linearized vs. original network



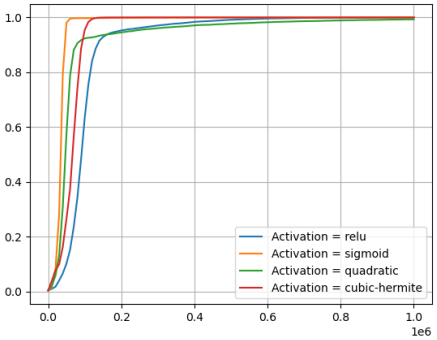
(b) Trajectories of $\langle u_t, u \rangle$ and $\langle c_t, c \rangle$ when $\xi = \Theta(\sqrt{k})$

Figure 3: Linearized networks fail in low-rank fine-tuning, and cannot achieve small loss. When jointly training $\hat{u}$ and $\hat{c}$, we observe incremental behavior in learning, where learning c becomes easier when u is learned to a certain level.

Activation robustness. As our analysis suggests, our findings are not too sensitive to the choice of the activation in the fine-tuning regime ($\xi = O(\sqrt{k})$). To reinforce this point, we overlay training plots for various choices of activations, namely $\sigma \in \{\mathrm{ReLU}, \mathrm{Sigmoid}, \mathrm{Quadratic}, \mathrm{He}_3\}$. In this setting, we jointly train $\hat{c}, \hat{u}$ and do not enforce $\hat{u}$ to be orthogonal to w_i and plot $\langle u_t, u \rangle$ over time in Figure 4. We observe that the qualitative behavior (e.g. no saddle behavior at initialization) is universal across the different choices of activation, which supports our theoretical findings.



(a) Evolution of overlap $\langle u_t, u \rangle$ for $\xi = 1$ and different activations



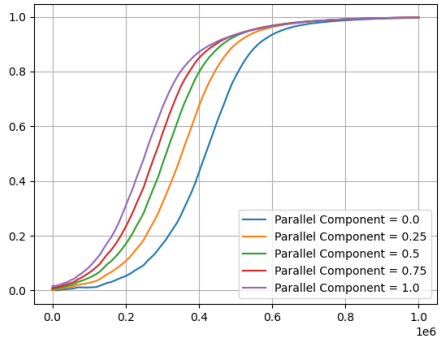
(b) Evolution of overlap $\langle u_t, u \rangle$ (y-axis) for $\xi = \Theta(\sqrt{k})$ and different activations

Figure 4: Evolution of overlap $\langle u_t, u \rangle$ during online SGD, under scaling (a) $\xi = 1$ and (b) $\xi = \Theta(\sqrt{k})$ for different activations. As opposed to learning single-index models, or multi-index models from scratch, the fine-tuning regime is not too sensitive to the choice of activation. Namely, the iteration complexity of fine tuning with SGD does not depend sensitively on information exponent.

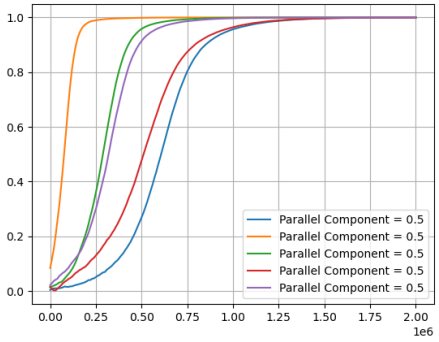
Testing the orthogonality assumption. In addition to the robustness checks w.r.t. to the choice of activation in the fine-tuning regime, we simulate the non-orthogonal case, i.e. when u is not orthogonal to the w_i Figure 5. In particular, define Π_W to be the orthogonal projector onto the span of W. We let $\alpha = \|\Pi_W u\|$, and sample u as

$$u = \alpha u_1 + \sqrt{1 - \alpha^2} u_2$$

where the unit vector u_1 is sampled uniformly from the span of W, and the unit vector u_2 is sampled uniformly from the orthogonal subspace. With the ReLU activation and when $k = 100$, $d = 1000$, we see that the evolution of the overlap $\langle u_t, u \rangle$ over time is qualitatively the same across different levels of parallel component $\|\Pi_w u\|$. This might indicate that under non-pathological activations and initializations, the non-orthogonal setting might be similar to the orthogonal setting. In principle, one can obtain an analytical expression for the population loss with the ReLU activation, and we leave this as an interesting direction for future work.



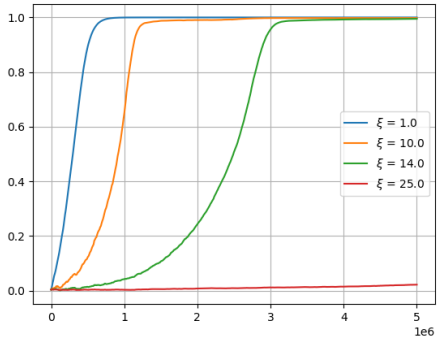
(*a*) Violating the orthogonality assumption and varying $\|\Pi_W u\|$, when $k \ll d$ with $\sigma = \mathrm{ReLU}$.



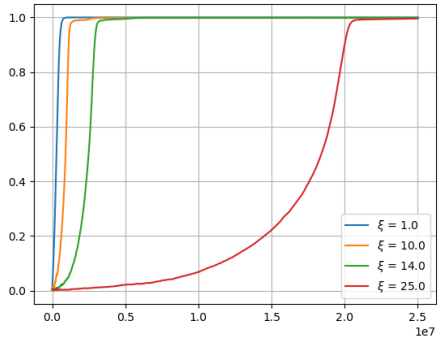
(*b*) Same setup as (a), but different (random) problem instances for $\|\Pi_W u\| = 1/2$

Figure 5: Evolution of $\langle u_t, u \rangle$ during online SGD for (a) varying levels of $\|\Pi_W u\|$ (violating the orthogonality assumption) (b) multiple runs for $\|\Pi_W u\| = \frac{1}{2}$. (a) In certain non-pathological initializations and scales ξ, the orthogonal case might capture behavior related to the non-orthogonal case. (b) Over multiple runs with $\|\Pi_W u\| = \frac{1}{2}$ we see a generic S-curve behavior for ReLU activation.

Interpolating between fine-tuning and feature learning. Recall that our proposed model for low-rank fine tuning allows us to interpolate between feature learning and fine tuning. Now, we demonstrate that training the low-rank perturbation can take significantly many samples when the perturbation is large. To do this, we run various experiments with $\sigma(a) = \mathrm{He}_3(a)$ and vary the scaling ξ. We choose the third hermite activation to clearly illustrate the differences in sample complexity between feature learning and fine tuning. For these simulations, we jointly train c and u, and plot the evolution of the overlap $\langle u_t, u \rangle$ over the course of training. We also do not necessarily enforce u_t to be orthogonal to the w_i, to illustrate the robustness of the results to the choice of algorithm Figure 6.



(*a*) Joint training of $\hat{c}$ and $\hat{u}$ for various choices of ξ (short time-scale)



(*b*) Joint training of $\hat{c}$ and $\hat{u}$ for various choices of ξ (long time-scale)

Figure 6: Evolution of $\langle u_t, u \rangle$ during online SGD for varying scales for ξ (a) short timescale, (b) long timescale. Empirically, we see that while for "small" ξ (e.g. $\xi = O(\sqrt{k})$) online SGD can quickly achieve strong recovery, as $\xi \to \infty$ we see that the time scales for both weak and strong recovery get larger. This illustrates how our low-rank fine-tuning setup allows to interpolate between different regimes (e.g. fine tuning to feature learning in single index models).

For the purpose of illustrating the dimension dependence, we set $d = 1000$ and keep the number of neurons smaller ($k = 20$). While we do not formally analyze the effect of the magnitude of ξ, we expect fine-tuning to get more difficult when the correlations $\frac{k}{k+\xi^2}(\langle w_i + c_i u, w_j + \hat{c}_j \hat{u} \rangle)$ are small (e.g. $O(1/\sqrt{d})$). This happens around $\xi = \Theta(d^{1/4}k^{1/2})$, and we empirically see that the plateau in obtaining the initial weak recovery of u gets larger as $\xi \to \infty$. An interesting future direction is obtaining rigorous results that quantify when this transition happens, and whether one can prove that fine-tuning works up to the feature learning threshold.

Appendix C. Further technical overview

In Section 3.1, we explained how the analysis reduces to proving bounds on the population gradient. Then, in Section 3.2.1 and Section 3.2.2, we overviewed the key elements that go into proving these lower bounds.

Now, in Section C.1, we further elaborate on why the population gradient is crucial to our analysis and the extension to finite sample dynamics. this part largely draws on existing techniques introduced by Arous et al. (2021). Then, in Section C.2, we describe how to put everything together to give a finite-sample analysis that proves our main results; Finally, in Section C.3 we give a self-contained proof of Theorem 19 separating fine-tuning and learning from scratch.

Before proceeding, we pause here to highlight the aforementioned connection to the literature on learning generalized linear models:

Remark 15 (Generalizing GLM regression) *Recall that the fine-tuning setting we study is a generalization of learning generalized models, and Proposition 4 recovers a standard calculation in*

the literature on the latter. Indeed, if we let $\xi = \bar{\xi}\sqrt{k}$ *and send* $\bar{\xi} \to \infty$ *the term*

$$\sum_{s=1}^{\infty} \binom{l+s}{l} (l+s+1)\mu_{l+s+1}(\sigma)^2 \left(\frac{1}{1+\xi^2/k}\right)^s T(l,s)$$

in Eq. (4) will vanish for all l*. Then,* $h(m)$ *around* 0 *reduces to*

$$h(m) \approx \sum_{l=0}^{\infty} l\mu_{l+1}(\sigma)^2 m^l$$

This reproduces a well-known finding for learning generalized linear models, namely that the dynamics at initialization is governed by the information exponent, i.e. the degree of the first non-vanishing Hermite coefficient $\mu_p(\sigma)$*. In that setting, if the information exponent is* l^**, then* $h(m)$ *at initialization scales as* $1/d^{l^*/2}$ *and noisy gradient descent requires iteration complexity scaling as* $d^{\Omega(l^*)}$ *Damian et al. (2022). In contrast, as we will see, the behavior of* $h(m)$ *in our fine-tuning setting when* $\bar{\xi}$ *is bounded is very different, and the dynamics will not be sensitive to the information exponent.*

C.1. Connection of population dynamics to finite sample dynamics

In this section, we give further intuition about why population lower bounds are crucial in our analysis. Going back to the expression in Proposition 4, note that the population gradient can be reduced to a term in the direction of spherical projection of u onto $\hat{u}$, scaled by some non-linear function that only depends on $\langle \hat{u}, u\rangle$. With that, notice that $\langle \hat{\nabla}\Phi(\hat{u}), u\rangle = -h(\langle u, \hat{u}\rangle)(1 - \langle u, \hat{u}\rangle^2)$, so that the population gradient in the ground truth direction only depends on the projection of the estimate $\hat{u}$ onto the ground truth u. The scaling factor $h(\langle u, \hat{u}\rangle)$ thus dictates the rate at which gradient descent moves towards the ground truth, but h depends on the unknown level of correlation $\langle u, \hat{u}\rangle$ in a complicated, highly nonlinear fashion.

Let $m_t = \langle u_t, u\rangle$ denote the alignment between the t-th iterate with the ground truth direction u, so that by Eq. (5),

$$m_{t+1} = m_t + \eta h(m_t)(1 - m_t^2) + \eta\langle \hat{\nabla} E_t(u_t; x_t), u\rangle + Q_t\,,$$

where $Q_t = \langle \hat{Q}_t, u\rangle$ is the error from ignoring the normalization.

Before discussing how to analyze this, let us first consider what happens with the *population dynamics*. Ignoring the error terms E_j and Q_j in the above, we get the recursion

$$\overline{m}_{t+1} = \overline{m}_t + \eta h(\overline{m}_t)(1 - \overline{m}_t^2)\,.$$

Rearranging, we have

$$\frac{|1 - \overline{m}_{t+1}|}{|1 - \overline{m}_t|} = |1 - \eta h(\overline{m}_t)(1 + \overline{m}_t)|\,.$$

So if $h(\overline{m}_t)$ is uniformly bounded from below by some quantity s, then the alignment m_t contracts towards 1 at a geometric rate of $1 - \eta s$ in each step, regardless of whatever complex behavior h exhibits over the course of training. Furthermore, notice that $\overline{m}_t$ would converge to 1 only if $h(\overline{m}_t)$ is non-vanishing across the trajectory since otherwise, the algorithm would converge to a different stationary point.

Let us now turn back to analyzing the finite-sample dynamics. Unrolling the recursive expression and defining $E_t = \hat{\nabla} E(u_t; x_t)$, we obtain

$$m_t = m_0 + \eta \sum_{j=0}^{t-1} h(m_j)(1 - m_j^2) + \eta \sum_{j=0}^{t-1} \langle E_j, u \rangle + \sum_{j=0}^{t-1} Q_j \,. \tag{10}$$

Note that the terms $M_t = \eta \sum_{j=0}^{t-1} \langle E_j, u \rangle$ form a martingale, and $\sum_{j=0}^{t-1} Q_j$ is a stochastic error term. Over short time scales, these terms could potentially dominate the dynamics. However, over the course of training for T iterations, their overall contribution scales with $\eta\sqrt{T}$, whereas the contribution of the population gradient term $\eta \sum_{j=0}^{t-1} h(m_j)(1 - m_j^2)$ scales with ηT, provided that the population gradients are not too small. By choosing η, T appropriately, we can ensure that $\eta T = \Theta(1)$ while $\eta\sqrt{T} = o(1)$. The exact choice depends crucially on the *signal-to-noise ratio* of the problem. In particular, if we have a lower bound S_k on the correlations between the population gradient and u (signal), and an upper bound dV_k on the variances $\mathbb{E}_x[E_t^2]$ of the martingale increments (noise), then we show the sample complexity scales with $\frac{dV_k}{S_k^2}$, the inverse of the signal-to-noise ratio.

Now, we formalize the intuition from the previous paragraphs and state the generic conditions under which we can show convergence of the online SGD dynamics to the ground truth and analyze the number of samples required. First, we require the following mild technical condition which is standard in the literature:

Condition 1 (Unbiased gradient estimates) *For all $\hat{u}$, the sample gradient is an unbiased estimate of the population gradient, i.e.,*

$$\hat{\nabla}_{\hat{u}} \Phi(\hat{u}) \triangleq \hat{\nabla}_{\hat{u}} \mathbb{E}_x[L(\hat{u}; x)] = \mathbb{E}_x[\hat{\nabla}_{\hat{u}} L(\hat{u}; x)] \,.$$

Note this holds when σ is differentiable almost everywhere w.r.t. Gaussian measure, and σ' has almost linear polynomial growth. In particular, it holds under Assumption 6. This is because $\nabla_{\hat{u}} L(\hat{u}; x)$ has at most linear polynomial growth and thus can be bounded by a function $g_k(\langle \hat{u}, x \rangle)$ which has finite expectation under x. Then, the interchange of derivative and expectation follows from dominated convergence.

Next, we need upper bounds on the variance of the empirical gradient and on the norm of the population gradient, in order to bound the martingale term M_t and stochastic error term Q_t in Eq. (6):

Condition 2 (Gradient variance upper bounds) *For each k, and p, there exists some constant $V_k \geq 1$ that has at most polynomial growth in k such that*

1. **Variance bound***: For all $u, \hat{u}$, we have $d^{-p}\, \mathbb{E}_x \|\hat{\nabla}_u L(\hat{u}; x)\|_2^{2p} \vee \mathbb{E}_x \langle \hat{\nabla}_{\hat{u}} L(\hat{u}; x), u \rangle^{2p} \leq (\mu_p V_k)^p$*

2. **Population gradient bound***: For all $\hat{u}$, we have $\|\hat{\nabla}_{\hat{u}} \Phi(\hat{u})\|^2 \leq V_k$*

where the μ_p is a constant that may depend on p and the activation σ, but on nothing else.

We will later prove that these upper bounds hold under the settings we consider, by appealing to appropriate tail bounds in Appendix D.2.

Finally, as described above, a uniform lower bound on the function h ensures that online SGD converges to the correct solution. We formalize this condition as follows. Recall that the population gradient is of the form $\hat{\nabla}_{\hat{u}} \Phi(\hat{u}) = -h(\langle \hat{u}, u \rangle)(u - \hat{u}\langle u, \hat{u} \rangle)$.

Condition 3 (Population gradient lower bound) *There exists a constant* $\max\{S_k, S_k^2\} \leq V_k$ *that has at most polynomial decay, such that* h *satisfies*

$$h(\text{sign}(h(0))m)\text{sign}(h(0)) \geq \frac{|h(0)|}{2} \geq S_k$$

for all $m \geq 0$.

Establishing such a lower bound on the population gradients is the key technical step in our proofs, and in Sections 3.2.1 and 3.2.2 we describe the ideas that go into proving this condition holds in the settings we consider.

The three conditions above are sufficient conditions to conclude that online SGD can recover the ground truth and give an upper bound on the number of samples required. Formally, we have the following generic statement under Conditions 1 to 3, which we will use later to get formal finite-sample guarantees for the different fine-tuning regimes we consider.

Theorem 16 *Let Conditions 1 to 3 hold. Let* $0 < \epsilon < 1$. *Let* $m_t = \langle u_t, u\rangle$ *and set the learning rate* $\eta = \frac{\delta}{dV_k}$ *with scaling* $\delta = \min\{\frac{S_k\epsilon^3}{4\mu_1(\log dV_k)^2}, 1\}$, *for total time* $T = \lceil \alpha d V_k \rceil$ *with time scaling* $\alpha = \frac{4(\log dV_k)}{\epsilon\delta S_k}$ *and initialization at* $|m_0| \geq \beta/\sqrt{d}$ *with* $m_0 h(0) > 0$. *With probability at least* $1 - o(1)$ *the following holds for* $T = \lceil \alpha d V_k \rceil$ *and* $T_{\text{weak}} = \lceil \frac{4dV_k}{\delta S_k} \rceil = o(T)$*:*

- *(Weak recovery):* $\sup_{t \leq T_{\text{weak}}} |m_t| \geq r$
- *(Strong recovery):* $|m_T| \geq 1 - \epsilon$

We prove this in Appendix E, drawing heavily upon the analysis in Arous et al. (2021). In the next two subsections, we show how to establish the conditions needed to apply Theorem 16 in the settings we consider.

C.2. Putting everything together

Now that we have shown that the signal term $h(m)$ is non-vanishing and proven lower bounds on it, it remains to upper bound the variances of the gradients and the population gradient. To that end, we prove the following:

Lemma 17 (General upper bounds) *Under Assumptions 1 to 6, we have the following:*

1. **Variance upper bound***:* $\sup_{u,\hat{u},c,\hat{c}} \left\{ \mathbb{E}_x \left\| \frac{\hat{\nabla} L(\hat{u};x)}{\sqrt{d}} \right\|^{2p} \vee \mathbb{E}_x |\langle \hat{\nabla} L(\hat{u};x), u\rangle|^{2p} \right\}^{1/p} \leq C_{p,\sigma} \lambda_{\max}^4 \frac{k^3\xi^2 \min\{k, 4\xi^2\}}{k+\xi^2}$
2. **Population gradient upper bound***:* $\left\| \hat{\nabla}\Phi(\hat{u}) \right\| \leq C_\sigma \lambda_{\max}^2 \frac{k\xi^2}{1+\xi^2/k}$

We prove this statement in Appendix D.2. Now, we are at a position to combine the lower bounds for the population gradient and the upper bounds for the variance to get a convergence and iteration complexity bound. We begin by proving our main results for the case of orthogonal pretrained features:

Proof [Proof of Theorem 11] For the first point, notice that the variance upper bound in Theorem 17 and the population gradient lower bound in Theorem 6 imply that Conditions 2 and 3 hold with

$$S_k = \frac{\gamma \lambda_{\min}^2 \mu_1(\sigma)^2}{1 + \xi^2/k}$$
$$V_k = C_{p,\sigma} \lambda_{\max}^4 k^2$$

for some small γ with probability $1 - o\left(\frac{\lambda_{\max}^2}{\lambda_{\min}^2}\right) - O(\gamma^{1/2})$. Then, applying Theorem 16 with the set S_k, V_k and ϵ, we get the desired result. The second case follows similarly. ■

Proof [Proof of Theorem 12] For $\bar{\xi} \leq 1$, the results in Theorem 17 and Theorem 6 imply that Condition 2, Condition 3 hold with

$$S_k = \frac{\gamma k \lambda_{\min}^2 \mu_1(\sigma)^2}{2}$$
$$V_k = C_{p,\sigma} \lambda_{\max}^4 \bar{\xi}^2 k^4$$

for some small γ with probability $1 - o(1) - \exp\{-\frac{2k}{\epsilon \bar{\xi}^2}\} - O(\gamma^{1/2})$. Then, applying Theorem 16 with the set S_k, V_k and ϵ, we get the desired result. The second case follows similarly. ■

Finally, we can prove our main result for the case of angularly separated pretrained features:
Proof [Proof of Theorem 13] Note that Theorem 17 and Theorem 8 imply that Condition 2, Condition 3 hold with

$$S_k = \frac{\gamma \lambda_{\min}^2}{\sqrt{k}}$$
$$V_k = C_{p,\sigma} \lambda_{\max}^4 k^2$$

for some small γ with probability $1 - o\left(\frac{\lambda_{\max}^2}{\lambda_{\min}^2}\right) - O(\gamma^{1/2})$. Then, applying Theorem 16 with the set S_k, V_k and ϵ, we get the desired result. The second case follows similarly. ■

C.3. Separations between fine-tuning and feature-learning

Note that if the teacher model is of the form $f(x) = \sum_{i=1}^k \lambda_i \sigma(\langle \tilde{w}_i, x \rangle)$ where the $\tilde{w}_i$ are orthonormal, f will be hard to learn with a CSQ algorithm if the information exponent of σ is large. Hence, in this section we show an example of a base network whose perturbation has orthonormal weights, and show separations between fine-tuning and learning from scratch using this example. We aim to construct $w_i + c_i u \perp w_j + c_j u$ for $i \neq j$. Notice that when $u \perp w_i$, this is equivalent to $\langle w_i, w_j \rangle = -c_i c_j$. Hence, if we can control the pairwise correlations of the w_i as we want, we can construct this example.

For the sake of intuition, consider the following example for $k = 4$, where each row is a w_i, with $c_i = (-\frac{1}{2})^i$.

$$W = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & -\frac{1}{2} & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & -\frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 \end{bmatrix}$$

We aim to generalize this example to general k in the following proposition.

Claim 2 *When $d > 1 + \frac{k(k+1)}{2}$, for $\lambda_i = 1$, there exists unit norm weights $\{w_i\}_{i=1}^k$, a perturbation $u \perp \text{span}(w_i)$, weights $c_i \in \left\{\pm\frac{1}{\sqrt{k}}\right\}$, such that $\frac{\langle w_i + c_i u, w_j + c_j u\rangle}{\|w_i+c_i u\|\|w_j+c_j u\|} = \delta_{ij}$.*

Proof We are looking for a setup where $\langle w_i, w_j\rangle = -c_i c_j$. We will construct k vectors that pairwise only share one non-zero coordinate. For $l \in [d], l \le k$, let $(w_l)_l = \frac{1}{\sqrt{k}}$. Then, for a given coordinate $l \in [d], l > k$, we want exactly two w_i, w_j to have non-zero l'th coordinate. Since $d - k > 1 + \binom{k}{2}$, we can assign every pair (i, j) with $i \ne j$ a coordinate, and we will have at least 1 coordinate left. Then, notice that the inner product $\langle w_i, w_j\rangle$ for $i \ne j$ only depends on 1 coordinate, which is unique for every (i, j). We choose the magnitude of this entry to be $\frac{1}{\sqrt{k}}$. Then, for any $c \in \left\{\pm\frac{1}{\sqrt{k}}\right\}^k$ we can simply choose the signs of these coordinates accordingly to ensure $\langle w_i, w_j\rangle = -c_i c_j$. Notice that each w_i has unit norm, and there is a coordinate, which we can WLOG assume to be the $p \triangleq \frac{k(k+1)}{2}$'th coordinate, that is zero for all w_i. We let $u = e_p$.

Then, notice that $\frac{\langle w_i + c_i u, w_j + c_j\rangle}{\|w_i+c_i u\|\|w_j+c_j u\|} = \frac{\langle w_i, w_j\rangle + c_i c_j}{\|w_i+c_i u\|\|w_j+c_j u\|} = 0$ for $i \ne j$, as desired. ■

Proposition 18 *Let $\xi = 1$, and consider the example in Claim 2. Suppose $\sigma = h_p$ is the p'th Hermite coefficient for some $p > 2$. Then, $h(m) = 2p\left(\frac{k}{k+1}\right)^p \tilde{h}(m)$ where*

$$\tilde{h}(m) = \sum_{i=1}^k \lambda_i^2 c_i \hat{c}_i + O\left(\frac{\lambda_{\max}^2}{k}\right)$$

Moreover, with high probability over the choice of $\hat{c}$, we have $h(m)\text{sign}(h(0)) \ge \frac{|h(0)|}{2}$.

Proof Initially, note

$$h(m) = 2p\left(\frac{k}{k+1}\right)^p \sum_{i,j=1}^k \lambda_i\lambda_j c_i\hat{c}_j(\langle w_i, w_j\rangle + c_i\hat{c}_j m)^{p-1}$$

In this case, notice that because $|\langle w_i, w_j\rangle| \le \frac{1}{k}$ for $i \ne j$, we have

$$\left|\sum_{i,j=1}^k \lambda_i\lambda_j c_i\hat{c}_j(\langle w_i, w_j\rangle + c_i\hat{c}_j\langle u, \hat{u}\rangle)^{p-1} - \sum_{i=1}^k \lambda_i^2 c_i\hat{c}_i(1 + c_i\hat{c}_i\langle u, \hat{u}\rangle)^{p-1}\right| \le \sum_{i\ne j}^k \left|\lambda_i\lambda_j c_i\hat{c}_j\frac{2}{k^{p-1}}\right| \le \frac{\lambda_{\max}^2}{k^{p-2}}$$

Hence, defining $\tilde{h}(m) = 2p\left(\frac{k}{k+1}\right)^p$ to factor out the constant, we have

$$\tilde{h}(m) = \sum_{i=1}^k \lambda_i^2 c_i\hat{c}_i(1 + c_i\hat{c}_i m)^{p-1} + O\left(\frac{\lambda_{\max}^2}{k^{p-2}}\right)$$

Then, expanding the diagonal term, note

$$\sum_{i=1}^k c_i\hat{c}_i\lambda_i^2(1 + c_i\hat{c}_i\langle u, \hat{u}\rangle)^{p-1} = \sum_{s=0}^{p-1}\binom{p-1}{s}\sum_{i=1}^k \lambda_i^2(c_i\hat{c}_i)^{s+1}\langle u, \hat{u}\rangle^s = \sum_{i=1}^k \lambda_i^2 c_i\hat{c}_i + O\left(\frac{\lambda_{\max}^2}{k}\right)$$

Then, for $p \geq 3$, we have

$$\tilde{h}(m) = \sum_{i=1}^{k} \lambda_i^2 c_i \hat{c}_i + O\left(\frac{\lambda_{\max}^2}{k}\right)$$

Then, over the randomization of $\hat{c}$, with high probability, we have $h(0) = \Omega\left(\frac{\lambda_{\min}^2}{\sqrt{k}}\right)$ due to anti-concentration (Theorem 28). Then, with high probability $h(m)\text{sign}h(0) \geq \frac{|h(0)|}{2}$ uniformly. ■

Hence, in the construction given in Claim 2, even though the c_i's are non-random, we still have with high probability over the randomization of $\hat{c}$ that h satisfies Condition 3. Then, we have the following

Theorem 19 *For the teacher and base networks defined in Claim 2, fine-tuning with online SGD learns the teacher network perturbation u in $O(\frac{dk^2}{\epsilon^4})$ samples, whereas training from scratch using any CSQ algorithm requires at least $O(d^{p/2})$ queries or $\tau = O(d^{-p/4})$ tolerance.*

Proof The first part follows directly from the fact that h satisfies the gradient lower bound in Condition 3 with a $\Omega(\frac{\lambda_{\min}^2}{\sqrt{k}})$ lower bound, and Theorem 16. For training from scratch, notice that the target model is of the form

$$f(x) = \sum_{i=1}^{k} \lambda_i h_p(\langle v_i, x\rangle)$$

where the v_i are orthonormal. Fix k. Then, we can embed f into a random k dimensional subspace M by rotating the v_i (since the vectors $w_i + c_i u$ can all be rotated without effecting the construction). The CSQ lower bound in (Abbe et al., 2023, Proposition 6) states that any CSQ algorithm using n queries with tolerance τ cannot achieve less than some small $c > 0$ error with probability $1 - \frac{Cn}{\tau^2} d^{-\frac{p}{2}}$. Hence, to achieve constant probability of success, one either needs $n = \Theta(d^{p/2})$ queries or tolerance $\tau = \Theta(d^{-p/4})$. ■

Appendix D. Deferred proofs for analyzing the online SGD dynamics from Section 3 and Appendix C

D.1. Computing the population gradient in a general setting

Recall that for $\lambda \in \mathbb{R}^k, w_i \in \mathbb{R}^d$ with $\|w_i\| = 1$, $c \in \{\pm\frac{1}{\sqrt{k}}\}^k$, and $u \in \mathbb{S}^{d-1}$ we have the target model

$$f^*(x) = \sum_{i=1}^{k} \lambda_i \sigma\left(\langle v_i, x\rangle\right) \tag{11}$$

where $v_i = \frac{w_i + \xi c_i u}{\|w_i + \xi c_i u\|}$. Furthermore, since $u \perp w_i$, we have $v_i = \frac{w_i + \xi c_i u}{\sqrt{1 + \frac{\xi^2}{k}}}$. Initially, we derive the population loss and gradient without imposing additional assumptions. Because σ admits a hermite expansion, for $v, \hat{v} \in \mathbb{S}^{d-1}$ we can evaluate expectations of the form $\mathbb{E}_x[\sigma(\langle v, x\rangle)\sigma(\langle \hat{v}, x\rangle)] = \sum_{p=0}^{\infty} \mu_p(\sigma)^2 \langle v, \hat{v}\rangle^p$.

Proof [Proof of Theorem 4] Note that $\mathbb{E}_x[(f^*(x)-\hat{f}(x))^2] = \sum_{i,j=1}^k \lambda_i\lambda_j f_i^*(x)f_j^*(x)+\sum_{i,j=1}^k \lambda_i\lambda_j \hat{f}_i(x)\hat{f}_j(x)-2\sum_{i,j=1}^k \lambda_i\lambda_j f_i^*(x)\hat{f}_j(x)$. Then,

$$\langle f_i^*, \hat{f}_j\rangle = \sum_{p=0}^{\infty} \mu_p(\sigma)^2 \langle v_i, \hat{v}_j\rangle^p$$

Working similarly for $\langle f_i^*, f_j^*\rangle$ and $\langle \hat{f}_i, \hat{f}_j\rangle$, we have

$$\mathbb{E}[(f^*(x)-\hat{f}(x))^2] = \left(\sum_{i,j=1}^k \lambda_i\lambda_j \sum_{p=0}^{\infty} \mu_p(\sigma)^2\langle \hat{v}_i, \hat{v}_j\rangle^p\right) + \left(\sum_{i,j=1}^k \lambda_i\lambda_j \sum_{p=0}^{\infty} \mu_p(\sigma)^2\langle v_i, v_j\rangle^p\right) - 2\sum_{i,j=1}^k \lambda_i\lambda_j \sum_{p=0}^{\infty} \mu_p(\sigma)^2\langle v_i, \hat{v}_j\rangle^p$$

Then, under the constraints $u, \hat{u} \perp w_i$ and $\|u\| = \|\hat{u}\| = 1$, notice that $\langle v_i, v_j\rangle = \frac{\langle w_i, w_j\rangle + \xi^2 c_i c_j}{(1+\frac{\xi^2}{k})}$ and similarly $\langle \hat{v}_i, \hat{v}_j\rangle = \frac{\langle w_i, w_j\rangle + \xi^2 \hat{c}_i \hat{c}_j}{(1+\frac{\xi^2}{k})}$. Since we are restricting training and gradients to this constrained space, the gradients of the first two terms with respect to $\hat{u}$ vanish. Then,

$$\hat{\nabla}_{\hat{u}}\mathbb{E}[(f^*(x)-\hat{f}(x))^2] = -2\sum_{i,j=1}^k \lambda_i\lambda_j \frac{\xi^2}{1+\frac{\xi^2}{k}} c_i\hat{c}_j \sum_{p=1}^{\infty} p\mu_p(\sigma)^2\langle v_i, \hat{v}_j\rangle^{p-1}(u-\hat{u}\langle u, \hat{u}\rangle)$$
$$= -2\sum_{i,j=1}^k \lambda_i\lambda_j \xi^2 c_i\hat{c}_j \sum_{p=1}^{\infty} p\mu_p(\sigma)^2 \left(\frac{1}{1+\frac{\xi^2}{k}}\right)^p (\langle w_i, w_j\rangle + \xi^2 c_i\hat{c}_j\langle u, \hat{u}\rangle)^{p-1}(u-\hat{u}\langle u, \hat{u}\rangle)$$

Then, notice that since $\sum_{p=1}^{\infty} p\mu_p(\sigma)^2 < \infty$, the expression above converges absolutely (and uniformly) for any $|\langle u, \hat{u}\rangle| \leq 1$. Let $m = \langle u, \hat{u}\rangle$ and define.

$$h(m) = 2\sum_{i,j=1}^k \lambda_i\lambda_j \xi^2 c_i\hat{c}_j \sum_{p=1}^{\infty} p\mu_p(\sigma)^2 \left(\frac{1}{1+\xi^2}\right)^p (\langle w_i, w_j\rangle + \xi^2 c_i\hat{c}_j m)^{p-1}$$

Because this expression converges absolutely and uniformly for $|m| \leq 1$, we can write its power series expansion around $m = 0$, to get

$$h(m) = 2\sum_{i,j=1}^k \lambda_i\lambda_j \sum_{l=0}^{\infty} (\xi^2)^{l+1}(c_i\hat{c}_j)^{l+1} m^l \sum_{s=0}^{\infty} (l+s+1)\mu_{l+s+1}(\sigma)^2 \binom{l+s}{l} \left(\frac{1}{1+\frac{\xi^2}{k}}\right)^{l+s+1} \langle w_i, w_j\rangle^s$$

Then, notice that for odd l, we have $(c_i\hat{c}_j)^{l+1} = \frac{1}{k^{l+1}}$. For even l, we have $(c_i\hat{c}_j)^{l+1} = \frac{c_i\hat{c}_j}{k^l}$. Then, we can write

$$h(m) = 2\sum_{l=0}^{\infty} \left(\frac{\xi^2}{k}\right)^{l+1} \left(\sum_{s=0}^{\infty} \binom{l+s}{l}(l+s+1)\mu_{l+s+1}(\sigma)^2 \left(\frac{1}{1+\frac{\xi^2}{k}}\right)^{l+s+1} T(l,s)\right) m^l$$

where

$$T(l,s) = \begin{cases} \left\| \sum_i \lambda_i w_i^{\otimes s} \right\|_F^2 & l \text{ odd} \\ k \left\langle \sum_i \lambda_i c_i w_i^{\otimes s}, \sum_i \lambda_i \hat{c}_i w_i^{\otimes s} \right\rangle & \text{otherwise} \end{cases}$$

as claimed. ■

Remark 20 (Role of moment tensors) *The $T(l,s)$ terms in the expression for $h(m)$ involve moment tensors like $\sum_i \lambda_i w_i^{\otimes s}$ and $\sum_i \lambda_i c_i w_i^{\otimes s}$. As mentioned in Section 1.3, there exist networks for which these tensors vanish and for which noisy gradient descent takes a long time to learn them from scratch Diakonikolas et al. (2020); Goel et al. (2020). As such, their appearance in Proposition 4 might seem to suggest that in the worst case over c and (λ_i, w_i), the complexity of fine-tuning could be as bad as the complexity of learning from scratch. While we do not formally address this worst case setting in this work, we expect that the complexity of the former should be dictated by the smallest l for which the sum over s in the definition of $h(m)$ is nonzero. Even if the moment tensors above vanish for many choices of s so that $T(l,s) = 0$ unless s is large, note that such s will still contribute non-negligibly to the aforementioned sum. For this reason, we expect that the worst-case complexity landscape of fine-tuning should be very different (and in general far more benign) than that of learning from scratch.*

D.2. Upper bounds on the variances of gradients and the magnitude of population gradient

Note that the sample gradient is

$$\frac{k}{\xi^2 + k} \nabla_{\hat{u}} L(\hat{u}; x) = 2(f^*(x) - \hat{f}(x)) \sum_{i=1}^{k} \lambda_i \hat{c}_i \sigma' \left(\frac{k}{\xi^2 + k} \langle w_i + c_i u, x \rangle \right) x$$

So, to bound the moments of this quantity, we would like to bound the moments of $(f^*(x) - \hat{f}(x))$ and $\sum_{i=1}^{k} \lambda_i \hat{c}_i \sigma' \left(\frac{k}{\xi^2+k} \langle w_i + c_i u, x \rangle \right) x$ respectively, and then apply Cauchy-Schwarz. To that end, we first prove the following:

Proposition 21 (Moments of squared error) *Let p be given, and Assumption 6 hold. Then, there exists some constant $C_{p,\sigma}$ that only depends on p and σ such that*

$$\mathbb{E}_x[(f^*(x) - \hat{f}(x))^{2p}]^{1/p} \leq C_{p,\sigma} \lambda_{\max}^2 \min\left\{k^2, 4k\xi^2\right\}$$

Proof Let $C_{p,\sigma}$ be a constant that only depends on p and σ, that will change throughout the proof. Note that

$$\begin{aligned} \mathbb{E}_x[(f^*(x) - \hat{f}(x))^{2p}] &\leq k^{2p-1} \sum_{i=1}^{k} \lambda_i^{2p} \mathbb{E}_x(\sigma(\langle v_i, x \rangle) - \sigma(\langle \hat{v}_i, x \rangle))^{2p} \\ &\leq C_{p,\sigma} \lambda_{\max}^{2p} k^{2p-1} \sum_{i=1}^{k} \sqrt{\mathbb{E}_x[|\langle v_i, x \rangle - \langle \hat{v}_i, x \rangle|^{4p}]} \\ &\leq C_{p,\sigma} \lambda_{\max}^{2p} k^{2p} \left\| v_i - \hat{v}_i \right\|^{2p} \end{aligned}$$

Then, note that apriori, $\|v_i - \hat{v}_i\| \leq 2$. Otherwise,

$$\begin{aligned}\|v_i - \hat{v}_i\| &\leq \|\xi c_i u - \xi \hat{c}_i \hat{u}\| + 2\left(1 - \frac{1}{\sqrt{1+\xi^2 c_i^2}}\right) \\ &\leq \frac{2\xi}{\sqrt{k}} + \frac{2\xi^2}{k} = \frac{2\xi}{\sqrt{k}}\left(1 + \frac{\xi}{\sqrt{k}}\right)\end{aligned}$$

However, notice that if $\xi \leq \sqrt{k}$, this is bounded by $\frac{4\xi}{\sqrt{k}}$. Otherwise, we use the bound $\|v_i - \hat{v}_i\| \leq 2$. Then,

$$\|v_i - \hat{v}_i\| \leq \min\left\{2, \frac{4\xi}{\sqrt{k}}\right\}$$

Combining with the above and taking p'th root, we have

$$\begin{aligned}\mathbb{E}_x[(f^*(x) - \hat{f}(x))^{2p}]^{1/p} &\leq C_{p,\sigma}\lambda_{\max}^2 k^2 \min\left\{4, \frac{16\xi^2}{k}\right\} \\ &\leq C_{p,\sigma}\lambda_{\max}^2 \min\left\{k^2, 4k\xi^2\right\}\end{aligned}$$

as desired. ■

Now, we bound the other quantity of interest, which is the moments of squares of the gradient $\hat{\nabla}_{\hat{u}}\hat{f}(x)$. We have the following:

Proposition 22 (Bound on the expected magnitude of $\hat{\nabla}\hat{f}$) *Let p be given. Then, we have*

$$\max\left\{\mathbb{E}_x\left|\frac{\hat{\nabla}_{\hat{u}}\hat{f}(x)}{\sqrt{d}}\right|^{2p}, \mathbb{E}_x\langle\hat{\nabla}_{\hat{u}}\hat{f}(x), u\rangle^{2p}\right\}^{1/p} \leq C_{\sigma,p}\lambda_{\max}^2\frac{k^2\xi^2}{k+\xi^2}$$

Proof Let $C_{\sigma,p}$ be a constant whose value can change throughout the proof. Initially, note that

$$\hat{\nabla}_{\hat{u}}\hat{f}(x) = (I - \hat{u}\hat{u}^\top)x\left[\sum_{i=1}^k \lambda_i \frac{\xi\hat{c}_i}{\sqrt{1+\xi^2\hat{c}_i^2}}\sigma'(\langle v_i, x\rangle)\right]$$

Then, since the spherical projection always leads to a smaller gradient

$$\left\|\hat{\nabla}_{\hat{u}}\hat{f}(x)\right\|^2 \leq \left\|\nabla_{\hat{u}}\hat{f}(x)\right\|^2$$

And furthermore,

$$\begin{aligned}\mathbb{E}_x\left\|\nabla_{\hat{u}}\hat{f}(x)\right\|^{2p} &\leq \sqrt{\mathbb{E}_x\|x\|^{4p}}\sqrt{\mathbb{E}_x\left[\sum_{i=1}^k \lambda_i \frac{\xi\hat{c}_i}{\sqrt{1+\xi^2\hat{c}_i^2}}\sigma'(\langle v_i, x\rangle)\right]^{4p}} \\ &\leq C_{\sigma,p}d^p k^{2p}\lambda_{\max}^{2p}\frac{(\xi^2/k)^p}{(1+\xi^2/k)^p}\max_i \mathbb{E}_x\sqrt{\sigma'(\langle\hat{v}_i, x\rangle)^{4p}}\end{aligned}$$

However, since σ' has at most polynomial growth, so does $(\sigma')^{4p}$ and since $\hat{v}_i$ is unit norm, the last quantity is finite and only depends on σ and p. Then,

$$\left\{\mathbb{E}_x \left\|\nabla_{\hat{u}} \hat{f}(x)\right\|^{2p}\right\}^{1/p} \leq C_{\sigma,p} \lambda_{\max}^2 d \frac{k^2 \xi^2}{k+\xi^2}$$

For the other case, note that the only step that changes is the bound on $\mathbb{E}_x \langle x, u\rangle^{4p}$ does not depend on the dimension, but only on p. So, we lose the dimension dependence. ■

Proposition 23 (Population gradient upper bound) *We have*

$$\left\|\hat{\nabla}_{\hat{u}} \Phi(\hat{u})\right\| \leq C_\sigma \lambda_{\max}^2 \frac{k\xi^2}{1+\xi^2/k}\,.$$

Proof Initially, note the non-expanded form of the population gradient:

$$\hat{\nabla}\Phi = \frac{\xi^2}{1+\xi^2/k} \sum_{i,j=1}^{k} \lambda_i \lambda_j c_i \hat{c}_j \sum_{p=1}^{\infty} p\mu_p(\sigma)^2 \langle v_i, \hat{v}_j\rangle^{p-1} (u - \hat{u}\langle u, \hat{u}\rangle)$$

Then, note $\left|\sum_{i,j=1}^{k} \lambda_i \lambda_j c_i \hat{c}_j\right| \leq k\lambda_{\max}^2$, and $\sum_{p=1}^{\infty} p\mu_p(\sigma)^2 \leq C_\sigma$. Furthermore, $\|u - \hat{u}\langle u, \hat{u}\rangle\| \leq 1$ and $|\langle v_i, \hat{v}_j\rangle| \leq 1$. Then, $\left\|\hat{\nabla}\Phi\right\| \leq C_\sigma \lambda_{\max}^2 \frac{k\xi^2}{1+\xi^2/k}$ as desired. ■

Proof [Proof of Theorem 17] Noting that

$$\mathbb{E}_x \left\|\frac{\hat{\nabla}_{\hat{u}} L(\hat{u}; x)}{\sqrt{d}}\right\|_2^{2p} \leq \sqrt{\mathbb{E}_x (f^*(x) - \hat{f}(x))^{4p} \cdot \mathbb{E}_x \left\|\frac{\hat{\nabla}_{\hat{u}} \hat{f}(x)}{\sqrt{d}}\right\|_2^{4p}}$$

and similarly for $\langle \hat{\nabla}_{\hat{u}} L(\hat{u}; x), u\rangle$ the result immediately follows from Theorem 21, Theorem 22 and Theorem 23. ■

D.3. Anti-concentration tools for lower bounding the population gradient

D.3.1. ANTI-CONCENTRATION FOR QUADRATIC POLYNOMIALS WITH LOW INFLUENCES

In this section, we prove some results related to the anti-concentration of certain quadratic functions on the hypercube. In particular, consider functions $f : \{\pm 1\}^k \times \{\pm 1\}^k \to \mathbb{R}$ of the form $f(x, y) = x^\top Q y$. These functions capture the random behavior of the function h (due to the randomness in $c, \hat{c}$) by determining the magnitudes of the constant term. We will control the magnitudes of functions of boolean variables by relating them to functions of Gaussians, and then applying anti-concentration inequalities for Gaussian polynomials. To that end, we first state some known bounds from literature. Note that these functions fall into the family of multilinear polynomials, which are defined as follows:

Definition 24 (Multilinear polynomial) *We define a normalized degree d multilinear polynomial as*

$$Q(x_1, x_2, \ldots, x_n) = \sum_{S \subset [n], |S| \leq d} a_S \prod_{i \in S} x_i$$

with $\mathrm{Var}(Q) = \sum_{S \subset [n], |S| > 0} a_S^2 = 1$.

Initially, we would have the following anti-concentration result if our randomness was *Gaussian* instead of *rademacher*.

Lemma 25 (Carbery-Wright inequality Carbery and Wright (2001)) *Let Q be a normalized multilinear polynomial with degree d as in Theorem 24. There exists an absolute constant B such that for $g \sim \mathcal{N}(0, I_n)$ we have*

$$\Pr[|Q(g_1, g_2, \ldots, g_n)| \leq \epsilon] \leq B\epsilon^{1/d}$$

For this class of functions, we know the following bound that helps us replace the randomness from Rademacher variables with Gaussian randomness.

Lemma 26 (Invariance principle, (Mossel et al., 2005, Theorem 2.1)) *Let P be as in Theorem 24. Furthermore, define the maximum influence as $\tau = \max_{i \in [n]} \sum_{S \ni i} a_S^2$. Then, for $\xi \sim \mathrm{Unif}\{\pm 1\}^n$ and $g \sim \mathcal{N}(0, I_n)$, we have*

$$\sup_t |\Pr[P(\xi_1, \ldots, \xi_n) \leq t] - \Pr[P(g_1, \ldots, g_n) \leq t]| \leq O(d\tau^{1/8d})$$

To be able to leverage these results, we need to quantify the influence of functions $x^\top Q y$ with Q being p.s.d. Normalizing $\|Q\|_F^2 = 1$ and noting $x^\top Q y = \sum_{i,j=1}^k x_i y_j Q_{ij}$, note that the influence of x_i (or similarly y_i) is $\sum_j Q_{ij}^2$. Hence, bounding the influence of this family of quadratic functions is the same as bounding the ratio of a row norm of Q to the frobenius norm. Hence, we prove the following claim that does exactly this.

Claim 3 (Influence of row of PSD matrix) *Consider all $k \times k$ PSD matrices Q with diagonal entries equal to 1. We have the following uniform bound on the maximum norm of a row to the frobenius norm of Q:*

$$\sup_{\substack{Q \in \mathbb{R}^{k \times k}, \\ Q \text{ psd}, \\ Q_{ii} = 1}} \frac{\max_i \sum_{j \in [k]} Q_{ij}^2}{\sum_{i,j=1}^k Q_{ij}^2} \leq 2k^{-1/2}$$

Proof To bound the given expression, we can write the numerator in an alternative way. Let $N = Q^2$ and note that $N_{jj} = \sum_l Q_{jl} Q_{jl} = \sum_l Q_{jl}^2$ which is the norm of the j'th row of Q. Hence, we are interested in bounding $\max_j N_{jj} / \|Q\|_F^2$. This is equivalently bounding the ratio of a diagonal entry of N to the sum of its eigenvalues.

Concretely, let $v_1, \ldots, v_k \in \mathbb{R}^k$ be the eigenvectors of Q and $s_1, \ldots, s_k$ be the corresponding eigenvalues. Note that N has the same eigenvectors as Q and the eigenvalues $s_1^2, \ldots, s_k^2$. Let $\mathcal{S} = \{i \in [k] : s_i^2 \geq k\}$ and consider $\zeta = \sum_{i \in \mathcal{S}} s_i^2$, which is the sum of 'large' eigenvalues of N.

Apriori, note that $\|Q\|_F^2 \geq \max\{k, \zeta\}$ since the diagonals of Q are 1 and ζ is less than the sum of the squares of eigenvalues of Q.

Now, note $Q = \sum_i s_i v_i v_i^\top$ and $N = \sum_i s_i^2 v_i v_i^\top$. Hence

$$N_{jj} = \sum_{i \in \mathcal{S}} s_i^2 (v_i)_j^2 + \sum_{i \notin \mathcal{S}} s_i^2 (v_i)_j^2$$

To bound the first term in the expression, note that $s_i^2 \geq k$, but $Q_{jj} = \sum_j s_i (v_i)_j^2 = 1$ so that $(v_i)_j^2 \leq k^{-1/2}$. Then, $\sum_{i \in \mathcal{S}} s_i^2 (v_i)_j^2 \leq \sum_{i \in \mathcal{S}} s_i^2 \max_j (v_i)_j^2 \leq k^{-1/2} \zeta \leq k^{-1/2} \|Q\|_F^2$. Similarly, for the second term, note simply that $s_i \leq k^{1/2}$ and $\sum_{i \notin \mathcal{S}} s_i (v_i)_j^2 \leq 1$. Hence, $N_{jj} \leq k^{-1/2} \|Q\|_F^2 + k^{1/2}$. Noting $\|Q\|_F^2 \geq k$, we have $k^{1/2} \leq \|Q\|_F^2 k^{-1/2}$. Hence, $N_{jj} / \|Q\|_F^2 \leq 2k^{-1/2}$. Since none of these bounds depend on the choice of Q, we have proven the claim. ■

Corollary 27

For a given Q, let $0 < q_{\min}^2 \leq q_{\max}^2$ be absolute constants such that for all k, we have $q_{\min}^2 \leq Q_{ii} \leq q_{\max}^2$. Then, we have

$$\sup_{\substack{Q \in \mathbb{R}^{k \times k}, \\ Q \text{ psd}, \\ q_{\min}^2 \leq Q_{ii} \leq q_{\max}^2}} \frac{\max_i \sum_{j \in [k]} Q_{ij}^2}{\sum_{i,j=1}^k Q_{ij}^2} \leq 2 \cdot \frac{q_{\max}^2}{q_{\min}^2} k^{-1/2}$$

Proof The proof follows immediately by modifying the proof of Claim 3 slightly. Note $\sum_i s_i (v_i)_j^2 \leq q_{\max}^2$ and $\|Q\|_F^2 \geq \max\{\zeta, k q_{\min}^2\}$. Then, following the proof similarly, we get

$$N_{jj} \leq k^{-1/2} \zeta + q_{\max}^2 k^{1/2} \leq k^{-1/2} \|Q\|_F^2 + q_{\max}^2 / q_{\min}^2 \|Q\|_F^2 k^{-1/2}$$

Then, noting $1 \leq q_{\max}^2 / q_{\min}^2$ and dividing by $\|Q\|_F^2$, we have that the desired quantity is bounded by $2 \frac{q_{\max}^2}{q_{\min}^2} k^{-1/2}$ as desired. ■

Now, we will use the above results to prove the following, which will help us directly quantify the random behavior of the function h.

Lemma 28 (Anti-Concentration of Normalized P.S.D. Quadratics on the Hypercube) *Let $Q \in \mathbb{R}^{k \times k}$ be positive semi-definite and normalized such that $q_{\min}^2 \leq Q_{ii} \leq q_{\max}^2$ for some $q_{\min}, q_{\max} > 0$. Then,*

$$\sup_{\substack{Q \in \mathbb{R}^{k \times k}, \\ Q \text{ psd}, \\ q_{\min}^2 \leq Q_{ii} \leq q_{\max}^2}} \Pr_{x,y \sim \text{Unif}\{\pm 1\}^k} [|x^\top Q y| \leq \epsilon \|Q\|_F] \leq o(1) + O(\epsilon^{1/2})$$

where the $o(1)$ is in k.

Proof First, note that we have the uniform bound on the influence of a row of Q from Theorem 27, so that $\tau = O(k^{-1/2})$. Hence, by the invariance principle (Theorem 26), for any Q, we have

$$\sup_t \left| \Pr_{x,y \sim \text{Unif}\{\pm\}^k} [x^\top Q y \leq t] - \Pr_{g_1, g_2 \sim \mathcal{N}(0, I_k)} [g_1^\top Q g_2 \leq t] \right| \leq o(1)$$

However, applying Carbery-Wright inequality for the anti-concentration of the Gaussian polynomial $g_1^\top Q g_2$ (Theorem 25), we get the desired result. ■

D.3.2. CONCENTRATION AND ANTI-CONCENTRATION OF $h(0)$

In this section, we analyze the behavior of $h(0)$. In particular, we aim to quantify the magnitude of $h(0)$ with high probability. Recall that when $\xi = 1$, we have

$$h(0) = \sum_{i,j=1}^{k} \lambda_i \lambda_j c_i \hat{c}_j \sum_{s=0}^{\infty} (s+1)\mu_{s+1}(\sigma)^2 \left(\frac{k}{k+1}\right)^{s+1} \langle w_i, w_j \rangle^s$$

Then, notice that writing $c_i = \frac{1}{\sqrt{k}} b_i$ and similarly $\hat{c}_i = \frac{1}{\sqrt{k}} \hat{b}_i$, this is a quadratic function of rademacher variables. In particular, let

$$f(b, \hat{b}) = \sum_{i,j}^{k} b_i \hat{b}_j \left(\frac{\lambda_i \lambda_j}{k} \sum_{s=0}^{\infty} (s+1)\mu_{s+1}(\sigma)^2 \left(\frac{k}{k+1}\right)^{s+1} \langle w_i, w_j \rangle^s \right) \tag{12}$$

so that $f(b, \hat{b}) = b^\top Q \hat{b}$ for some Q. Hence, we can use the anti-concentration results from the previous section to quantify the random behavior of $h(0)$ in the $\xi = 1$ case.

Claim 4 (Variance of $h(0)$, spectral scaling $\xi = 1$) *Let $f : \{\pm 1\}^k \times \{\pm 1\}^k \to \mathbb{R}$ be as in eq. (12). Then, we have $\Omega(\lambda_{\min}^2/k) \le \|f\|_2 \le O(\lambda_{\max}^2)$*

Proof Notice that since each term in the sum is a different basis element of $\{\pm 1\}^{2k}$, we have

$$\|f\|_2^2 = \sum_{i,j=1}^{k} Q_{ij}^2$$

For the first part of the Claim, it suffices to show $\sum Q_{ij}^2 = \Omega(\frac{\lambda_{\min}^4}{k})$ for any choice of λ, w_i. Notice that, for $k \ge 2$,

$$\begin{aligned}
\sum_{i,j=1}^{k} Q_{ij}^2 \ge \sum_{i=1}^{k} Q_{ii}^2 &= \left(\sum_{s=0}^{\infty} (s+1)\mu_{s+1}(\sigma)^2 \left(\frac{k}{k+1}\right)^{s+1} \right)^2 \sum_{i=1}^{k} \frac{\lambda_i^4}{k^2} \\
&\ge \left(\sum_{s=0}^{\infty} \frac{s+1}{2^s} \mu_{s+1}(\sigma)^2 \right)^2 \frac{\lambda_{\min}^4}{k}
\end{aligned}$$

as desired. The other follows directly from

$$\sum_{i,j=1}^{k} Q_{ij}^2 \le \sum_{i,j=1}^{k} \frac{1}{k^2} \lambda_i^2 \lambda_j^2 \left(\sum_{s=0}^{\infty} (s+1)\mu_{s+1}(\sigma)^2 \right)^2 \le \lambda_{\max}^4 \left(\sum_{s=0}^{\infty} (s+1)\mu_{s+1}(\sigma)^2 \right)^2 .$$

■

Hence, if we can show that f anti-concentrates around 0 with high probability, then we will have given a lower bound for $h(0)$ in the $\xi = 1$ scaling regime. This is what we do now.

Proposition 29 (Anti-concentration of $h(0)$, $\xi = 1$) *If $\xi = 1$, we have the following: Let f be of the form in Equation (12). Then,*

$$\sup_{w_i,\lambda_i} \Pr_{b,\hat{b}}[|f(b,\hat{b})| < \epsilon \|f\|_2] = \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \cdot o(1) + O(\epsilon^{1/2})$$

where $b, \hat{b}$ are independent uniform draws from $\{-1,1\}^k$.

Proof Note that entrywise powers of psd matrices are psd, so $(W^TW)^{\odot s}$ is psd. Notice that

$$Q_{ij} = (\lambda_i\lambda_j)\left(\sum_{s=0}^{\infty}(s+1)\mu_{s+1}(\sigma)^2\left(\frac{k}{k+1}\right)^{s+1}\langle w_i, w_j\rangle^s\right),$$

so Q is a psd matrix since it is the sum of psd matrices (for s). This is due to the fact

$$Q = \lambda\lambda^\top * \tilde{Q}$$

where $\tilde{Q}_{ij} = \sum_{s=0}^{\infty}(s+1)\mu_{s+1}(\sigma)^2\left(\frac{k}{k+1}\right)^{s+1}\langle w_i, w_j\rangle^s$ since it is the non-negative sum of psd matrices. Furthermore, $(q_{\max}/q_{\min})^2 = O(\frac{\lambda_{\max}^2}{\lambda_{\min}^2})$. The final result follows immediately once we normalize as $\frac{f}{\|f\|_2}$ and apply Theorem 28. ■

Proposition 30 (Anti-concentration of $(\sum_i \lambda_i c_i)(\sum_i \lambda_i \hat{c}_i)$ and $\sum_i \lambda_i^2 c_i \hat{c}_i$) *We have*

$$\Pr\left[\left|(\sum_i \lambda_i c_i)(\sum_i \lambda_i \hat{c}_i)\right| \le \gamma\lambda_{\min}^2\right] \le o\left(\frac{\lambda_{\max}^2}{\lambda_{\min}^2}\right) + O(\gamma^{1/2})$$

and

$$\Pr\left[\left|\sum_i \lambda_i^2 c_i \hat{c}_i\right| \le \gamma\frac{\lambda_{\min}^2}{\sqrt{k}}\right] \le o\left(\frac{\lambda_{\max}^2}{\lambda_{\min}^2}\right) + O(\gamma^{1/2})$$

Proof For the first one let $Q = \frac{1}{k}\lambda\lambda^\top$. and for the second one let $Q = \frac{1}{k}I(\lambda \odot \lambda)$. Both are balanced psd matrices, and the anti concentration result Theorem 28 holds. Then, the results follow. ■

D.4. Orthonormal case: population gradient lower bounds

Recall the function h.

$$h(m) = 2\sum_{l=0}^{\infty}\left(\frac{\xi^2}{k}\right)^{l+1}\left(\sum_{s=0}^{\infty}\binom{l+s}{l}(l+s+1)\mu_{l+s+1}(\sigma)^2\left(\frac{1}{1+\frac{\xi^2}{k}}\right)^{l+s+1}T(l,s)\right)m^l$$

with $T(l,s)$ being defined as

$$T(l,s) \triangleq \begin{cases} \left\|\sum_i \lambda_i w_i^{\otimes s}\right\|_F^2 & l \text{ odd} \\ k\left\langle\sum_i \lambda_i c_i w_i^{\otimes s}, \sum_i \lambda_i \hat{c}_i w_i^{\otimes s}\right\rangle & \text{otherwise} \end{cases}$$

However, in the orthogonal case, for $s \geq 1$, $T(l,s)$ reduces to

$$T(l, s \geq 1) = \begin{cases} \sum_{i=1}^k \lambda_i^2 & l \text{ odd} \\ k \sum_{i=1}^k \lambda_i^2 c_i \hat{c}_i & \text{otherwise} \end{cases}$$

And for $s = 0$, these reduce to

$$T(l, 0) = \begin{cases} \left(\sum_{i=1}^k \lambda_i\right)^2 & l \text{ odd} \\ k \left(\sum_{i=1}^k \lambda_i c_i\right) \left(\sum_{i=1}^k \lambda_i \hat{c}_i\right) & \text{otherwise} \end{cases}$$

Notice that for all odd l, the power series coefficients are always non-negative. And for all even l, all the power series coefficients have the same sign. Now, we prove Claim 1 as stated in Appendix C which bounds the maximum contribution coming from even l, $s = 0$ terms.

Proof [Proof of Claim 1] Note first that $\mathbb{E}_{c,\hat{c}} \left(\sum_{i=1}^k \lambda_i c_i\right) \left(\sum_{i=1}^k \lambda_i \hat{c}_i\right) = 0$ and moreover,

$$\mathbb{E}_{c,\hat{c}} \left(\sum_{i=1}^k \lambda_i c_i\right)^2 \left(\sum_{i=1}^k \lambda_i \hat{c}_i\right)^2 = \frac{\|\lambda\|_2^4}{k^2}$$

so the standard deviation is $\|\lambda\|_2^2 / k$. Hence, $T(l,0)$ has standard deviation $\|\lambda\|_2^2$ in $c, \hat{c}$. Then, note that

$$\begin{aligned}
&2 \sum_{\substack{l>0 \\ \text{even}}} \left(\frac{\xi^2}{k}\right)^{l+1} (l+1)\mu_{l+1}(\sigma)^2 \left(\frac{1}{1+\frac{\xi^2}{k}}\right)^{l+1} \left(k \sum_{i,j=1}^k \lambda_i \lambda_j c_i \hat{c}_j\right) m^l \\
&= \frac{2m\xi^2}{k} \left(k \sum_{i,j=1}^k \lambda_i \lambda_j c_i \hat{c}_j\right) \sum_{\substack{l>0 \\ \text{even}}} \left(\frac{\xi^2}{k}\right)^{l} (l+1)\mu_{l+1}(\sigma)^2 \left(\frac{1}{1+\frac{\xi^2}{k}}\right)^{l+1} m^{l-1} \\
&= \frac{2m\xi^2}{k} \left(k \sum_{i,j=1}^k \lambda_i \lambda_j c_i \hat{c}_j\right) \sum_{l \text{ odd}} \left(\frac{\xi^2}{k}\right)^{l+1} (l+2)\mu_{l+2}(\sigma)^2 \left(\frac{1}{1+\frac{\xi^2}{k}}\right)^{l+2} m^l \\
&= \frac{2m\xi^2}{k} \left(k \sum_{i,j=1}^k \lambda_i \lambda_j c_i \hat{c}_j\right) \sum_{l \text{ odd}} \frac{1}{l+1}\binom{l+1}{l} \left(\frac{\xi^2}{k}\right)^{l+1} (l+2)\mu_{l+2}(\sigma)^2 \left(\frac{1}{1+\frac{\xi^2}{k}}\right)^{l+2} m^l \\
&= \frac{2m\xi^2}{k} \left(k \sum_{i,j=1}^k \lambda_i \lambda_j c_i \hat{c}_j\right) \sum_{l \text{ odd}, s=1} \frac{1}{l+s}\binom{l+s}{l} \left(\frac{\xi^2}{k}\right)^{l+1} (l+s+1)\mu_{l+s+1}(\sigma)^2 \left(\frac{1}{1+\frac{\xi^2}{k}}\right)^{l+s+1} m^l
\end{aligned}$$

However, notice that the sum precisely corresponds to all odd l with $s = 1$. Then, bounding $l \geq 1$ so that $\frac{1}{l+1} \leq \frac{1}{2}$, we can elementwise compare the odd l terms with $s = 1$ and even l terms with $s = 0$. The odd terms are

$$2 \|\lambda\|_2^2 \sum_{l \text{ odd}} \left(\frac{\xi^2}{k}\right)^{l+1} \binom{l+1}{l} (l+2)\mu_{l+2}(\sigma)^2 \left(\frac{1}{1+\frac{\xi^2}{k}}\right)^{l+2} m^l$$

where all the terms in the sum have the same sign as m. Then, note that it suffices to show that, with high probability, we have

$$\frac{m\xi^2}{k}\left(k\sum_{i,j=1}^{k}\lambda_i\lambda_j c_i\hat{c}_j\right) \leq 2\left\|\lambda\right\|_2^2$$

Then, note that using the standard deviation bound and (O'Donnell, 2014, Theorem 9.23), we have

$$\Pr\left[\left|\frac{m\xi^2}{k}\left(k\sum_{i,j=1}^{k}\lambda_i\lambda_j c_i\hat{c}_j\right)\right| \geq 2\left\|\lambda\right\|_2^2\right] \leq \exp\left\{-\frac{2k}{em\xi^2}\right\} \leq \exp\left\{-\frac{2k}{e\xi^2}\right\}$$

Hence, with probability $1-\exp\left\{-\frac{2k}{e\xi^2}\right\}$, the even $s=0$ terms will not effect the sign of the odd terms. In particular, we have, with probability at least $1-\exp\{-\frac{2k}{e\xi^2}\}$, we have

$$\begin{aligned}\operatorname{sign}(m)\Bigg(&2\sum_{l\text{ odd},s=1}\left(\frac{\xi^2}{k}\right)^{l+1}\binom{l+s}{l}(l+s+1)\mu_{l+s+1}(\sigma)^2\left(\frac{1}{1+\xi^2/k}\right)^{l+s+1}T(l,s)m^l\\ &+2\sum_{l\text{ even},s=0}\left(\frac{\xi^2}{k}\right)^{l+1}\binom{l+s}{l}(l+s+1)\mu_{l+s+1}(\sigma)^2\left(\frac{1}{1+\xi^2/k}\right)^{l+s+1}T(l,s)m^l\Bigg)\geq 0\end{aligned}$$

as desired. ■

Now, we prove the lower bound on the population gradient in the orthonormal case, as stated in Theorem 6, Appendix C.

Proof [Proof of Theorem 6] Initially, suppose $\mu_1(\sigma)\neq 0$. In this case, using Claim 1, with probability $1-\exp\{-\frac{2k}{e\xi^2}\}$ we have

$$\begin{aligned}\operatorname{sign}(m)h(m) \geq{}& \operatorname{sign}(m)\xi^2\left(\sum_{i=1}^{k}\lambda_i c_i\right)\left(\sum_{i=1}^{k}\lambda_i\hat{c}_i\right)\mu_1(\sigma)^2\frac{1}{1+\frac{\xi^2}{k}}+\sum_{l\text{ odd}}b_l|m|^l\\ &+\operatorname{sign}(m)\frac{\xi^2}{1+\xi^2/k}\langle c_\lambda,\hat{c}_\lambda\rangle\sum_{l\text{ even},s\geq 1}\left(\frac{\xi^2}{k}\right)^{l}\binom{l+s}{l}(l+s+1)\mu_{l+s+1}(\sigma)^2\left(\frac{1}{1+\frac{\xi^2}{k}}\right)^{l+s}|m\end{aligned}$$

Now, we investigate the second term. Note that the sum in the second term is bounded by

$$\begin{aligned}&\sum_{l,s\geq 0}\left(\frac{\xi^2}{k}\right)^{l}\binom{l+s}{l}(l+s+1)\mu_{l+s+1}(\sigma)^2\left(\frac{1}{1+\frac{\xi^2}{k}}\right)^{l+s}\\ &=\sum_{p=0}^{\infty}\sum_{s=0}^{p}\left(\frac{\xi^2}{k}\right)^{p-s}\binom{p}{s}(p+1)\mu_{p+1}(\sigma)^2\left(\frac{1}{1+\frac{\xi^2}{k}}\right)^{p}\\ &=\sum_{p=0}^{\infty}(p+1)\mu_{p+1}(\sigma)^2\left(\frac{k}{k+\xi^2}\right)^{p}\left(1+\frac{\xi^2}{k}\right)^{p}\\ &\leq\sum_{p=0}^{\infty}(p+1)\mu_{p+1}(\sigma)^2\leq C_\sigma\end{aligned}$$

Then, notice the the second term is bounded in magnitude by $\frac{C_\sigma \xi^2}{1+\xi^2/k}|\langle c_\lambda, \hat{c}_\lambda\rangle|$. Then, notice that

$$\Pr\left[|\langle c_\lambda, \hat{c}_\lambda\rangle| \geq \frac{\gamma \lambda_{\max}^2}{\sqrt{k}} \log k\right] \leq k^{-\frac{\gamma}{e}}$$

Set $\gamma = 10$. So, with high probability this term is $O\left(\frac{\log k}{\sqrt{k}} \frac{C_\sigma \lambda_{\max}^2 \xi^2}{1+\xi^2/k}\right)$ However, by anti-concentration of the constant term (Theorem 30), we have that the constant term is order $\frac{\gamma \lambda_{\max}^2 \mu_1(\sigma)^2 \xi^2}{1+\xi^2/k}$ with probability $1 - o(1) - O(\frac{\lambda_{\max}}{\lambda_{\min}}\gamma^{1/2})$. Then, the constant term is $O(\sqrt{k}(\log k)^{-1})$ larger than the even terms, and it's sign is dictated by $\text{sign}(m)(\sum_i \lambda_i c_i)(\sum_i \lambda_i \hat{c}_i) > 0$. Then, we can bound the even terms by half of the constant term, and get the desired result.

Now, suppose $\mu_1(\sigma) = 0$. In this case, with probability $1 - \exp\{-\frac{2k}{e\xi^2}\}$ note that

$$\text{sign}(m)h(m) \geq \text{sign}(m)\xi^2\langle c_\lambda, \hat{c}_\lambda\rangle \sum_{\substack{l \text{ even} \\ s\geq 1}} \left(\frac{\xi^2}{k}\right)^l \binom{l+s}{l}(l+s+1)\mu_{l+s+1}(\sigma)^2 \left(\frac{1}{1+\frac{\xi^2}{k}}\right)^{l+s+1} |m|^l + \sum_{l \text{ odd}} b_l |m|^l$$

where the b_l are non-negative coefficients. Then, under the sign assumption on m, note that $\text{sign}(m)\xi^2\langle c_\lambda, \hat{c}_\lambda\rangle = |\langle c_\lambda, \hat{c}_\lambda\rangle|\xi^2$. Then, by anti-concentration (Theorem 30), note that with probability $1 - o(1) - O(\gamma^{1/2})$, $|\langle c_\lambda, \hat{c}_\lambda\rangle| \geq \frac{\gamma\xi^2}{\sqrt{k}}$. Hence, we have $h(\text{sign}(h(0)m)\text{sign}(h(0)) \geq |h(0)|$ for all $m \geq 0$, and $|h(0)| \geq \frac{\gamma C_{s^*}\xi^2}{\left(1+\frac{\xi^2}{k}\right)^{s^*}\sqrt{k}}$ where s^* is the smallest s for which $\mu_s \neq 0$. ■

D.5. Angularly separated case: population gradient lower bounds

D.5.1. COMPUTATION OF THE POPULATION GRADIENT

Note that specializing $\xi = 1$ in Proposition 4, we get

$$h(m) = \sum_{l=0}^{\infty}\left(\frac{1}{k}\right)^{l+1}\sum_{s=0}^{\infty}\binom{l+s}{l}(l+s+1)\mu_{l+s+1}(\sigma)^2\left(\frac{k}{k+1}\right)^{l+s+1} T(l,s)m^l$$

D.5.2. BOUNDING THE HIGHER ORDER EVEN TERMS

Initially, we aim to bound the even terms in the power series (i.e. $l > 1$). We first prove the statement about bounding the higher order even terms as stated in Appendix C, Theorem 7.

Proof [Proof of Theorem 7] Let $s^* = 10\sqrt{k}$. This proof will involve bounding contributions from the following three types of terms:

(i) The contribution from the terms where $s \leq s^*$. These can be bounded naively since there are at most $O(\sqrt{k})$ of them, and the $(1/k)^{2n+2}$ will dominate the growth in k in these terms.

(ii) The contribution for $s \geq s^*$ from diagonal terms: These terms scale with $\sum_{i=1}^k \lambda_i^2 c_i \hat{c}_i$, so it suffices to show the coefficient is $O(k^{-\epsilon})$ for some small $\epsilon > 0$. This is due to the fact that the Hermite coefficients decay at rate $(s^*)^{-1-\rho}$, so the contribution of the large s coefficients have to decay in k at some small rate.

(iii) The contribution for $s \geq s^*$ from non-diagonal terms: Due to the assumption of angular separation between the w_i's, when s is sufficiently large, the decay of the terms $\langle w_i, w_j \rangle^s$ means these terms will be small.

(i) Contribution from terms with $s \leq s^* = O(\sqrt{k})$: Initially, we bound the magnitudes of the randomized terms. Since there are at most $\sqrt{k}$ of them and they concentrate exponentially around their means, we can bound their magnitude by $O(\log k)$ with exponentially high probability. Specifically,

$$\begin{aligned}
\mathbb{E}\left[\sum_{i,j=1}^{k} \lambda_i \lambda_j c_i \hat{c}_j \langle w_i, w_j \rangle^s\right] &= \sum_{i,j=1}^{k} \lambda_i \lambda_j \langle w_i, w_j \rangle^s \mathbb{E}[c_i \hat{c}_j] = 0 \\
\mathbb{E}\left[\left(\sum_{i,j=1}^{k} \lambda_i \lambda_j c_i \hat{c}_j \langle w_i, w_j \rangle^s\right)^2\right] &= \sum_{i,i'=1}^{k} \sum_{j,j'=1}^{k} \lambda_i \lambda_{i'} \lambda_j \lambda_{j'} \langle w_i, w_j \rangle^s \langle w_{i'}, w_{j'} \rangle^s \mathbb{E}[c_i c_{i'} \hat{c}_j \hat{c}_{j'}] \\
&= \sum_{i,i'=1}^{k} \sum_{j,j'=1}^{k} \lambda_i \lambda_{i'} \lambda_j \lambda_{j'} \langle w_i, w_j \rangle^s \langle w_{i'}, w_{j'} \rangle^s \mathbb{E}[c_i c_{i'}] \mathbb{E}[\hat{c}_j \hat{c}_{j'}] \\
&= \frac{1}{k^2} \sum_{i=1}^{k} \sum_{j=1}^{k} \lambda_i^2 \lambda_j^2 \langle w_i, w_j \rangle^{2s} \\
&\leq \frac{\|\lambda\|_2^4}{k^2} \leq \lambda_{\max}^4 .
\end{aligned}$$

Then, define $f_s : \{-1, 1\}^{2k} \to \mathbb{R}$ as $f_s(b, \hat{b}) = \frac{1}{k} \sum_{i,j=1}^{k} \lambda_i \lambda_j b_i \hat{b}_i \langle w_i, w_j \rangle^s$ which is a quadratic polynomial in $b_i, \hat{b}_i$. We have just proved that $\|f_s\|_2 \leq \lambda_{\max}^2$. Then, by (O'Donnell, 2014, Theorem 9.23) we have

$$\Pr_{b,\hat{b}}\left[|f_s(b, \hat{b})| \geq \gamma \log k \, \|f\|_2\right] \leq \exp\{-\frac{\gamma}{e} \log k\} = k^{-\frac{\gamma}{e}}$$

where $\gamma > 0$ is to be chosen later. Then, using the union bound, we have

$$\Pr\left[\max_{s \leq s^*} \left|\sum_{i,j=1}^{k} \lambda_i \lambda_j c_i \hat{c}_j \langle w_i, w_j \rangle^s\right| \geq \gamma \lambda_{\max}^2 \log k\right] \leq s^* k^{-\frac{\gamma}{e}}$$

As $s^* = O(\sqrt{k})$, then with probability at least $1 - k^{-\frac{\gamma}{e} + \frac{1}{2}}$, we have

$$\begin{aligned}
&\left|\sum_{n=0}^{\infty} \left(\frac{1}{k}\right)^{2n+2} \sum_{s=0}^{s^*} \binom{2n+2+s}{2n+2} (2n+s+3) \mu_{2n+s+3}(\sigma)^2 \left(\frac{k}{k+1}\right)^{2n+s+3} \left\langle \sum_{i=1}^{k} \lambda_i c_i w_i^{\otimes s}, \sum_{i=1}^{k} \lambda_i \hat{c}_i w_i^{\otimes s} \right\rangle\right| \\
&\quad \leq \gamma \lambda_{\max}^2 \log k \sum_{n=0}^{\infty} \left(\frac{1}{k}\right)^{2n+2} \sum_{s=0}^{s^*} \binom{2n+2+s}{2n+2} (2n+s+3) \mu_{2n+s+3}(\sigma)^2 \left(\frac{k}{k+1}\right)^{2n+s+3}
\end{aligned} \tag{13}$$

Now, it suffices to give a $O(k^{-\frac{1}{2} - c\epsilon})$ bound for the infinite sum for $c > 1$. We will separate it into cases $s \leq (s^*)^{1-\epsilon}$ and $(s^*)^{1-\epsilon} \leq s \leq s^*$. The reason for this is that we have to use the decay of the

Hermite coefficients as s approaches $\sqrt{k}$, so the two cases need to be handled separately. Hence, for $l \triangleq 2n+2$ using the binomial coefficient bound $\binom{n}{k} \leq \left(\frac{en}{k}\right)^k$ we have

$$\begin{aligned}\sum_{s=0}^{(s^*)^{1-\epsilon}} \binom{l+s}{l}(l+s+1)\mu_{l+s+1}(\sigma)^2\left(\frac{k}{k+1}\right)^{l+s+1} &\leq \sum_{s=0}^{(s^*)^{1-\epsilon}} C_\sigma\left(e\frac{l+s}{l}\right)^l \\ &\leq C_\sigma e^l \sum_{s=0}^{(s^*)^{1-\epsilon}}(1+s)^l \\ &\leq C_\sigma e^l (s^*)^{1-\epsilon}(1+(s^*)^{1-\epsilon})^l \\ &\leq C_\sigma (s^*)^{1-\epsilon}(2e(s^*)^{1-\epsilon})^l\end{aligned}$$

Then, notice that for k larger than some absolute constant, we have

$$C_\sigma (s^*)^{1-\epsilon}\sum_{n=0}^{\infty}\left(\frac{1}{k}\right)^{2n+2}\left(2e(s^*)^{1-\epsilon}\right)^{2n+2} \leq C_\sigma (s^*)^{1-\epsilon}\left(\frac{2e(s^*)^{1-\epsilon}}{k}\right)^2 \frac{1}{1+o(1)} = O(k^{-\frac{1}{2}-\frac{3}{2}\epsilon})$$

since $(s^*)^{3(1-\epsilon)}k^{-2} = O(k^{-\frac{1}{2}-\frac{3}{2}\epsilon})$.

Now, we look at the remaining terms. For $(s^*)^{1-\epsilon} \leq s \leq s^*$, we have

$$\begin{aligned}\left(\frac{1}{k}\right)^l \sum_{(s^*)^{1-\epsilon}\leq s\leq s^*}\binom{l+s}{l}(l+s+1)\mu_{l+s+1}(\sigma)^2\left(\frac{k}{k+1}\right)^{l+s+1} &\leq C_\sigma (s^*)^{-(1-\epsilon)(1+2\rho)} \sum_{(s^*)^{1-\epsilon}\leq s\leq s^*}\left(\frac{2es^*}{k}\right)^l \\ &\leq C_\sigma (s^*)^{1-(1-\epsilon)(1+2\rho)}\left(\frac{2es^*}{k}\right)^l\end{aligned}$$

Taking the sum over all $l \triangleq 2n+2$, we have

$$C_\sigma (s^*)^{1-(1-\epsilon)(1+2\rho)}\sum_{n=0}^{\infty}\left(\frac{2es^*}{k}\right)^{2n+2} \leq C_\sigma (s^*)^{1-(1-\epsilon)(1+2\rho)}\left(\frac{2es^*}{k}\right)^2 \frac{1}{1+o(1)}.$$

Choosing $\epsilon = 1-\frac{1}{1+2\rho} > 0$ for simplicity[3], we have that the sum is bounded by $C_\sigma\left(\frac{2s^*}{k}\right)^2 \frac{1}{1+o(1)} = O(\frac{1}{k})$. Hence, combining with previous steps, we can upper bound the infinite sum in Equation (13) by $O(\lambda_{\max}^2 k^{-\frac{1}{2}-3\epsilon})$ where $\epsilon = 1-\frac{1}{1+2\rho}$.

(ii) The contribution of $s \geq s^*$ for diagonal terms: We first note that

$$\sum_{p=1}^{\infty} p\mu_p(\sigma)^2\left(\frac{k}{k+1}\right)^p \langle w_i + c_i u, w_i + \hat{c}_i\hat{u}\rangle^{p-1} = \sum_{p=1}^{\infty} p\mu_p(\sigma)^2\left(\frac{k}{k+1}\right)^p (\langle w_i, w_j\rangle + c_i\hat{c}_j m)^{p-1}$$

Then, notice that the RHS is maximized in absolute value when $w_i = w_j$, $c_i = \hat{c}_j$ and $m = 1$. In this case, we get

$$\left|\sum_{p=1}^{\infty} p\mu_p(\sigma)^2\left(\frac{k}{k+1}\right)^p \langle w_i + c_i u, w_i + \hat{c}_i\hat{u}\rangle^{p-1}\right| \leq \sum_{p=1}^{\infty} p\mu_p(\sigma)^2 \triangleq \tilde{C}_\sigma$$

3. There are more optimal choices of ϵ that lead to better bounds

In particular, we have absolute convergence of the LHS for all $|m| \leq 1$, so we can freely interchange order of sums. However, notice all steps in this argument works if we replace $\mu_p(\sigma)^2$ with something else that has sufficiently fast decay. In particular, writing $p = l + s + 1$ we have

$$\begin{aligned}\sum_{l=0}^{\infty}\left(\frac{1}{k}\right)^l \sum_{s=0}^{\infty}\binom{l+s}{l}(l+s+1)\mu_{l+s+1}(\sigma)^2\left(\frac{k}{k+1}\right)^{l+s+1} &= \sum_{p=1}^{\infty}\left(\frac{k}{k+1}\right)^p p\mu_p(\sigma)^2 \sum_{l=0}^{p-1}\left(\frac{1}{k}\right)^l\binom{p-1}{l} \\ &= \sum_{p=1}^{\infty}\left(\frac{k}{k+1}\right)^p\left(1+\frac{1}{k}\right)^{p-1} p\mu_p(\sigma)^2 \\ &\leq \sum_{p=1}^{\infty} p\mu_p(\sigma)^2 = \tilde{C}_\sigma \qquad (14)\end{aligned}$$

However, since all the terms in the sum are non-negative, using the same steps, we have

$$\begin{aligned}&\sum_{l=0}^{\infty}\left(\frac{1}{k}\right)^l \sum_{s=s^*}^{\infty}\binom{l+s}{l}(l+s+1)\mu_{l+s+1}(\sigma)^2\left(\frac{k}{k+1}\right)^{l+s+1} \\ &\leq \sum_{l=0}^{\infty}\left(\frac{1}{k}\right)^l \sum_{s=s^*}^{\infty}\binom{l+s}{l}(l+s+1)^{-1-2\rho}\left(\frac{k}{k+1}\right)^{l+s+1} \\ &\leq (s^*)^{-\rho}\sum_{l=0}^{\infty}\left(\frac{1}{k}\right)^l \sum_{s=s^*}^{\infty}\binom{l+s}{l}(l+s+1)^{-1-\rho}\left(\frac{k}{k+1}\right)^{l+s+1} \\ &\leq (s^*)^{-\rho}\sum_{p=1}^{\infty} p^{-1-\rho} = \hat{C}_\sigma (s^*)^{-\rho}\end{aligned}$$

where $\hat{C}_\sigma = \sum_{p=1}^{\infty}\frac{1}{p^{1+\rho}}$.[4] Then,

$$\begin{aligned}&\left|\sum_{n=0}^{\infty}\left(\frac{1}{k}\right)^{2n+2}\sum_{s=s^*}^{\infty}\binom{2n+2+s}{2n+2}(2n+s+3)\mu_{2n+s+3}(\sigma)^2\left(\frac{k}{k+1}\right)^{2n+s+3}\sum_i \lambda_i^2 c_i\hat{c}_i\right| \\ &\leq \hat{C}_\sigma(s^*)^{-\rho}\left|\sum_i \lambda_i^2 c_i\hat{c}_i\right|.\end{aligned}$$

Then, notice that since $\sqrt{\mathbb{E}[(\sum_i \lambda_i^2 c_i\hat{c}_i)^2]} = \sqrt{\frac{1}{k^2}\sum_{i=1}^k \lambda_i^4} \leq \lambda_{\max}^2/\sqrt{k}$, we have

$$\Pr[|\sum_i \lambda_i^2 c_i\hat{c}_i| \geq \gamma\lambda_{\max}^2\frac{\log k}{\sqrt{k}}] \leq k^{-\frac{\gamma}{e}}$$

by another application of (O'Donnell, 2014, Theorem 9.23). Then, with probability at least $1-\frac{1}{k^{\gamma/e}}$, we have

$$\hat{C}_\sigma(s^*)^{-\rho}\left|\sum_i \lambda_i^2 c_i\hat{c}_i\right| \leq \hat{C}_\sigma(s^*)^{-\rho}\gamma\lambda_{\max}^2\frac{\log k}{\sqrt{k}} = O(\lambda_{\max}^2 k^{-\frac{1}{2}-\frac{\rho}{4}})$$

4. $\hat{C}_\sigma$ depends on σ through the definition of ρ in Assumption 6.

as claimed.

(iii) Bounding the non-diagonal terms for $s \geq s^*$: Notice that

$$\left|\sum_{i\neq j}^{k} \lambda_i\lambda_j c_i\hat{c}_j\langle w_i, w_j\rangle^s\right| \leq \sqrt{k^2\sum_{i\neq j}\lambda_i^2\lambda_j^2 c_i^2\hat{c}_j^2\langle w_i, w_j\rangle^{2s}}$$
$$\leq \left(1-\frac{\log k}{\sqrt{k}}\right)^s \|\lambda\|_2^2 .$$

Then, let $s \geq s^* = \gamma\sqrt{k}$. Then,

$$\left(1-\frac{\log k}{\sqrt{k}}\right)^s \|\lambda\|_2^2 \leq e^{-\gamma\log k}\|\lambda\|_2^2 = \frac{\|\lambda\|_2^2}{k^\gamma},$$

so setting $\gamma > \frac{3}{2}$ will suffice. I.e, we have

$$\left|\sum_{n=0}^{\infty}\left(\frac{1}{k}\right)^{2n+2}\sum_{s=s^*}^{\infty}\binom{2n+2+s}{2n+2}(2n+s+3)\mu_{2n+s+3}(\sigma)^2\left(\frac{k}{k+1}\right)^{2n+s+3}\left(\sum_{i\neq j}\lambda_i\lambda_j c_i\hat{c}_j\langle w_i, w_j\rangle^s\right)\right|$$
$$\leq \frac{\|\lambda\|_2^2}{k^\gamma}\sum_{n=0}^{\infty}\left(\frac{1}{k}\right)^{2n+2}\sum_{s=s^*}^{\infty}\binom{2n+2+s}{2n+2}(2n+s+3)\mu_{2n+s+3}(\sigma)^2\left(\frac{k}{k+1}\right)^{2n+s+3}$$
$$\leq \frac{\tilde{C}_\sigma\|\lambda\|_2^2}{k^\gamma},$$

where in the last step we used Equation (14). Combining all the bounds, for $\epsilon = \min\{\frac{\rho}{4}, 1-\frac{1}{1+2\rho}\}$, with probability at least $1-\gamma\frac{1}{k^{\gamma/e-\frac{1}{2}}}$, we have

$$\sum_{n=0}^{\infty}\left(\frac{1}{k}\right)^{2n+2}\sum_{s=0}^{\infty}\binom{2n+2+s}{2n+2}(2n+s+3)\mu_{2n+s+3}(\sigma)^2\left(\frac{k}{k+1}\right)^{2n+s+3}\left\langle\sum_{i=1}^{k}\lambda_i c_i w_i^{\otimes s}, \lambda_i\hat{c}_i w_i^{\otimes s}\right\rangle$$
$$= O(\lambda_{\max}^2\gamma k^{-\frac{1}{2}-\epsilon})$$

Specifically, setting $\gamma = 10$, the result holds with probability at least $1-\frac{1}{k^3}$. ■

D.5.3. PROVING THE LOWER BOUND ON h

Now, we prove Theorem 8 from Appendix C.

Proof [Proof of Theorem 8] Let $\gamma > 0$ be a small constant. Then, notice that using the anti-concentration of $h(0)$ (Theorem 29) and the variance bound $\Omega(\frac{\lambda_{\min}^2}{k})$ for $h(0)$ from Claim 4, with probability $1-O(\gamma^{1/2})-o(\frac{\lambda_{\max}^2}{\lambda_{\min}^2})$, we have

$$|h(0)| \geq \frac{\gamma\lambda_{\min}^2}{\sqrt{k}}.$$

However, note that all the even terms are $O(k^{-\frac{1}{2}-\epsilon})$. Hence, we can bound the even terms by $|h(0)|/2$ with high probability. Then, we get that whenever the sign condition $mh(0) > 0$ is satisfied, we have $\text{sign}(m)h(m) \geq \frac{\gamma\lambda_{\min}^2}{2}$ with high probability, as desired. ■

Appendix E. Finite-sample analysis

The goal of this section is to prove Theorem 16. First, define the following notation for the noisy gradients and gradient error:

$$L_t = \hat{\nabla}_{u_t} L(u_t; x_t)$$
$$E_t = L_t - \hat{\nabla}_{u_t} \Phi(u_t).$$

Now, in the next sections we initially bound the terms contributed by the spherical projection error, and then the error martingale. Finally, we combine our results to show weak recovery and strong recovery under Conditions 1 to 3. Throughout this section, we use $\mathcal{F}_t$ to denote the sigma algebra generated by the iterates u_t.

E.1. Analysis of dynamics under the generic assumptions

Recall the online SGD dynamics

$$u_{t+1} = \frac{u_t - \eta \hat{\nabla}_{u_t} L(u_t; x_t)}{\left\| u_t - \eta \hat{\nabla}_{u_t} L(u_t; x_t) \right\|}$$

where $x_t \sim \mathcal{N}(0, I_d)$ is a fresh Gaussian sample at each time iteration t. Then, define the correlation with ground truth $m_t = \langle u_t, u \rangle$ and the projection magnitude $\Pi_t = \left\| u_t - \eta \hat{\nabla}_{u_t} L(u_t; x_t) \right\|$. Then, notice

$$\begin{aligned} m_{t+1} &= \frac{m_t - \eta \langle \hat{\nabla}_{u_t} L(u_t; x_t), u \rangle}{\Pi_t} \\ &= m_t - \eta \hat{\nabla}_{u_t} \Phi(u_t) - \eta \langle \hat{\nabla}_{u_t} E(u_t; x_t), u \rangle - \left(1 - \frac{1}{\Pi_t}\right) \left(m_t - \eta \langle \hat{\nabla}_{u_t} L(u_t; x_t), u \rangle \right). \end{aligned}$$

Hence, initially, we bound the effect of the spherical projection term.

E.1.1. BOUNDING SPHERICAL PROJECTION ERROR

We initially bound the spherical projection error in terms of L_t and $\langle L_t, u \rangle$. This will later allow us to use the tail bounds on L_t to bound the spherical projection error.

Claim 5 (Relating spherical projection error to L_t)

$$\sum_{j=0}^{t-1} \left| \left(1 - \frac{1}{\Pi_j}\right) (m_j - \eta \langle \hat{\nabla}_{u_j} L(u_j; x_j), u \rangle) \right| \leq \eta^3 \sum_{j=0}^{t-1} \|L_j\|^2 |\langle L_j, u \rangle| + \eta^2 \sum_{j=0}^{t-1} \|L_j\|^2$$

Proof First, notice that because u_t is perpendicular to the spherical gradient $\hat{\nabla}_{u_t} \Phi(u_t)$, we have

$$1 \leq \Pi_t \leq \sqrt{1 + \eta^2 \left\| \hat{\nabla}_{u_t} L(u_t; x_t) \right\|_2^2} \leq 1 + \eta^2 \left\| \hat{\nabla}_{u_t} L(u_t; x_t) \right\|_2^2$$

Then, due to $\left| 1 - \frac{1}{1+x} \right| \leq x$ for $x \geq 0$, we have

$$\left|\left(1-\frac{1}{\Pi_t}\right)\left(m_t-\eta\langle\hat{\nabla}_{u_t}L(u_t;x_t),u\rangle\right)\right|\leq\eta^2\left\|L_t\right\|^2\left(|m_t|+\eta|\langle L_t,u\rangle|\right)$$

Bounding $|m_t|\leq 1$, notice that the total contribution of these terms up to time t can be upper bounded by

$$\eta^3\sum_{j=0}^{t-1}\left\|L_t\right\|^2|\langle L_t,u\rangle|+\eta^2\sum_{j=0}^{t-1}\left\|L_t\right\|^2.$$

■

Then, notice that η^3 gives a $\frac{\delta^3}{d^3V_k^3}$ scaling, but $\left\|L_t\right\|^2|\langle L_t,u\rangle|$ scales only in dV_k^2, and there are $T=\alpha dV_k$ of these. Then, we can use a simple Markov bound to bound these terms when $\alpha\delta^2\leq\epsilon$.

Claim 6 (Bounding cubic terms) *Let α,δ be such that $\alpha\delta^2\leq\epsilon$ and $\delta\leq 1$. Then, we have*

$$\Pr\left[\sup_{0\leq t\leq T}\eta^3\sum_{j=0}^{t}\left\|L_j\right\|^2|\langle L_j,u\rangle|>\frac{\beta}{10\sqrt{d}}\right]\lesssim\frac{1}{\beta\sqrt{d}}$$

Similarly, we have

$$\Pr\left[\sup_{0\leq t\leq T}\eta^3\sum_{j=0}^{t}\left\|L_j\right\|^2|\langle L_j,u\rangle|>\frac{\epsilon}{18}\right]\lesssim\frac{1}{d}$$

Proof Notice that in both cases the maximum is achieved at $t=T$ due to the non-negativity of the terms in the sum. Then, by Markov

$$\Pr\left[\sup_{t\leq T}\eta^3\sum_{j=0}^{t}\left\|L_j\right\|^2|\langle L_j,u\rangle|>\gamma\right]=\Pr\left[\eta^3\sum_{j=0}^{T}\left\|L_j\right\|^2|\langle L_j,u\rangle|>\gamma\right]\leq\frac{\eta^3T\sup_j\mathbb{E}[\left\|L_j\right\|^2|\langle L_j,u\rangle|]}{\gamma}.$$

Now, using Cauchy-Schwarz to bound the expectation, we have

$$\mathbb{E}[\left\|L_j\right\|^2|\langle L_j,u\rangle|]\leq\left\|\left\|L_j\right\|^2\right\|_2\sqrt{\left\||\langle L_j,u\rangle|^2\right\|_1}$$

Hence, using the moment bounds (Condition 2) on $\left\|L_t\right\|^2$ and $|\langle L_t,u\rangle|^2$, for $p=2,1$ respectively, we have

$$\mathbb{E}[\left\|L_j\right\|^2|\langle L_j,u\rangle|]\lesssim dV_k^2$$

Finally, using $\eta=\frac{\delta}{dV_k}$, $T=\alpha dV_k$ and $\alpha\delta^2\leq\epsilon$, $\delta\leq 1$, we have

$$\Pr\left[\sup_{t\leq T}\eta^3\sum_{j=0}^{t}\left\|L_j\right\|^2|\langle L_j,u\rangle|>\gamma\right]\lesssim\frac{\alpha d^2V_k^3\eta^3}{\gamma}=\frac{\alpha\delta^3}{d\gamma}\leq\frac{1}{d\gamma}$$

Setting $\gamma = \frac{\beta}{10\sqrt{d}}$ gives us the first result. For the second, we can use $\alpha\delta^2 \leq \epsilon$ and $\delta \leq 1$ to bound the probability by $\frac{1}{d}$. ■

Now, we turn to the quadratic term. Notice that with the quadratic term, we are not necessarily getting the extra scaling in $1/d$ from η we need, so we need to be more careful while bounding this term. For these terms, we will show that their cumulative effect at any given iteration is smaller than the drift contribution. To do this we need to uniformly bound the cumulative effect up to iteration t. Recall Freedman's inequality Freedman (1975) for submartingales with almost sure bounds:

Lemma 31 (Freedman's inequality) *Let M_t be a submartingale with $\mathbb{E}[(M_{t+1} - M_t)^2|\mathcal{F}_t] \leq V$ and $|M_{t+1} - M_t| \leq K$ almost surely. Then,*

$$\Pr[S_t \leq -\lambda] \leq \exp\left\{\frac{-\lambda^2}{tV + \frac{\lambda}{3}K}\right\}$$

Hence, we will introduce an appropriate clipping of $\|L_t\|$ and separate into cases when it is large and small. When it is large, we will use the fast decay of its tails due to bounded moments the bound the probability of being large. When it is small, we will use the almost sure bound and Freedman's inequality to control the total contribution.

Claim 7 (Bounding the quadratic terms) *Suppose α has at most polynomial growth in d, k. Furthermore suppose, $\alpha\delta^2 \leq 1$, and that V_k has polynomial growth in k. Then, for some constant C, we have*

$$\Pr\left[\inf_{0\leq t\leq T} \eta \sum_{j=0}^{t}\left(\frac{S_k}{4} - \eta\|L_t\|^2\right) < \frac{\beta}{-5\sqrt{d}}\right] \leq \frac{C}{\beta\sqrt{d}} + \alpha(dV_k)^{-\frac{\beta^2}{C}(\log dV_k)+1}$$

Proof Initially, define $Y_t = \frac{\|L_t\|^2}{dV_k}$ and notice that $\|Y_t\|_p \leq \mu_p$ for all $t \geq 0$ where μ_p do not grow in d or k as stated in Condition 2. Then, notice that $\eta\|L_t\|^2 = \delta Y_t$. We write $Y_t = Y_t\mathbb{1}\{Y_t \geq T^\nu\} + Y_t\mathbb{1}\{Y_t < T^\nu\}$. Then, we can decompose the term as

$$\begin{aligned}\eta\sum_{j=0}^{t}\left(\frac{S_k}{2} - \eta\|L_t\|^2\right) &= \eta\sum_{j=0}^{t}\left(\frac{S_k}{2} - \delta\|Y_t\|^2\mathbb{1}\{Y_t \geq T^\nu\}\right) + \eta\sum_{j=0}^{t}\left(\frac{S_k}{2} - \delta\|Y_t\|^2\mathbb{1}\{Y_t < T^\nu\}\right)\\ &\geq -\eta\sum_{j=0}^{t}\delta\|Y_t\|^2\mathbb{1}\{Y_t \geq T^\nu\} + \eta\sum_{j=0}^{t}\left(\frac{S_k}{2} - \delta\|Y_t\|^2\mathbb{1}\{Y_t < T^\nu\}\right)\end{aligned}$$

where we used $\frac{S_k}{2} > 0$ for the last inequality. Then, it suffices to show that the second line is at least $-\frac{\beta}{5\sqrt{d}}$. Hence, we will bound the probability of each term being less than $-\frac{\beta}{10\sqrt{d}}$ and use the union bound.

Then, notice that for fixed choice of $\nu, D > 0$ we have

$$\Pr[Y_t \geq T^\nu] = \Pr[Y_t^{D/\nu} \geq T^D] \leq \frac{\mathbb{E}[Y_t^{D/\nu}]}{T^D}$$

Then, letting $D/\nu = p$ and using the p'th moment bound Condition 2, there exists a constant $C_{\nu,D}$ such that

$$\Pr[Y_t \geq T^\nu] \leq \frac{C_{\nu,D}}{T^D}$$

where we used $V_k \geq 1$. Then, notice that, using Cauchy-Schwarz, we have

$$\mathbb{E}[Y_t \mathbb{1}\{Y_t \geq T^\nu\}] \leq \|Y_t\|_2 \sqrt{\Pr[Y_t \geq T^\nu]} \leq \frac{C_{\nu,D}}{T^{D/2}}$$

where we absorbed the μ_2 constant into the C. Then, we have

$$\Pr\left[\eta \sum_{j=0}^{T-1} Y_t \mathbb{1}\{Y_t \geq T^\nu\} > \gamma\right] \leq \frac{\eta T C_{\nu,D}}{\gamma T^{D/2}}$$

Then, we can choose $D = 1$ (and get rid of the D dependence on the constants), and $\gamma = \frac{\beta}{10\sqrt{d}}$ such that

$$\Pr\left[\eta \sum_{j=0}^{T-1} Y_t \mathbb{1}\{Y_t \geq T^\nu\} > \frac{\beta}{10\sqrt{d}}\right] \lesssim \frac{\sqrt{d}\eta C_\nu}{\beta} \leq \frac{\delta C_\nu}{\sqrt{d} V_k \beta} \leq \frac{\delta C_\nu}{\beta\sqrt{d}}$$

Then, notice that we are left with the term $Y_t \mathbb{1}\{Y_t \leq T^\nu\}$ where ν can be chosen arbitrarily small. Consider setting $\delta \leq \frac{S_k}{4C_\delta \log(dV_k)}$ such that

$$\begin{aligned}
\eta \sum_{j=0}^{t} \left(\frac{S_k}{2} - \delta Y_t \mathbb{1}\{Y_t \leq T^\nu\}\right) &\geq \frac{\eta S_k}{4} \sum_{j=0}^{t} \left(1 - \frac{Y_t \mathbb{1}\{Y_t \leq T^\nu\}}{C_\delta \log(dV_k)}\right) \\
&\geq \frac{\eta S_k}{4 \log(dV_k)} \sum_{j=0}^{t} \left(1 - \frac{Y_t \mathbb{1}\{Y_t \leq T^\nu\}}{C_\delta}\right)
\end{aligned}$$

However, since $\mathbb{E}Y_t$ is bounded by 1, for $C_\delta > \mu_1$, the following forms an $\mathcal{F}_t$ submartingale:

$$Z_t = \frac{\eta S_k}{2 \log(dV_k)} \sum_{j=0}^{t} \left(1 - \frac{Y_t \mathbb{1}\{Y_t \leq T^\nu\}}{C_\delta}\right)$$

Then, it suffices to show

$$\Pr\left[\inf_{0 \leq t \leq T} Z_t < -\frac{\beta}{10\sqrt{d}}\right] = o(1)$$

Then, note $\mathbb{E}[Y_t \mathbb{1}\{Y_t \leq T^\nu\}] \leq \mathbb{E}[Y_t] = O(1)$, and we have the almost sure bound

$$|Z_{t+1} - Z_t| \leq \frac{\eta S_k}{2 \log(dV_k)} \left(1 + \frac{T^\nu}{C_\delta}\right) \leq \frac{\eta S_k}{\log(dV_k)} \frac{T^\nu}{C_\delta}$$

and the conditional variances

$$\mathbb{E}[(Z_{t+1} - Z_t)^2 | F_t] \leq \frac{\eta^2 S_k^2}{4(\log dV_k)^2} \left(1 + \mu_2^2\right) \leq \frac{C\eta^2 S_k^2}{(\log dV_k)^2}$$

where C is a constant that can only depend on μ_2.

Then, using Freedman's inequality for submartingales, for any $0 \le t \le T$ we have

$$\Pr\left[Z_t \le -\frac{\beta}{10\sqrt{d}}\right] \le \exp\left\{\frac{-\frac{\beta^2}{100d}}{\frac{CT\eta^2 S_k^2}{(\log dV_k)^2} + \frac{\beta\eta S_k}{30\sqrt{d}\log(dV_k)}\frac{T^\nu}{C_\delta}}\right\}$$

Let's inspect the expression in the exponent. Note, using $\alpha\delta^2 \le 1$ and equivalently $\delta\alpha^\nu \le 1$, for some updated constant $C = C(\mu_2)$ we have

$$\begin{aligned}
\frac{-\frac{\beta^2}{100d}}{\frac{CT\eta^2 S_k^2}{(\log dV_k)^2} + \frac{\beta\eta S_k}{10\sqrt{d}\log(dV_k)}\frac{T^\nu}{C_\delta}} &= -\frac{\beta^2}{\frac{C\alpha\delta^2 S_k^2}{V_k(\log dV_k)^2} + \frac{10\beta\delta\alpha^\nu S_k}{V_k^{1-\nu}d^{1/2-\nu}\log(dV_k)}} \\
&\le -\beta^2 \min\left\{\frac{V_k(\log dV_k)^2}{CS_k^2}, \frac{V_k^{1-\nu}d^{1/2-\nu}\log(dV_k)}{10\beta S_k}\right\} \\
&\le -\frac{\beta^2}{C}(\log dV_k)^2 V_k^{1/2}
\end{aligned}$$

for sufficiently large d greater than some $O(1)$, where we have $\frac{V_k}{S_k} \ge 1$ and $\frac{V_k}{S_k^2} \ge 1$ when $\nu = 1/4$. Hence, taking the exponent, we have $\exp\{-\frac{\beta^2}{C}(\log dV_k)^2 V_k^{1/2}\} = (dV_k)^{-\frac{\beta^2}{C}(\log dV_k)}$ Then, doing a union bound over all $t \le T$, we have

$$\Pr\left[\inf_{0\le t\le T-1} Z_t \le -\frac{\beta}{10\sqrt{d}}\right] \le T(dV_k)^{-\frac{\beta^2}{C}(\log dV_k)} = \alpha(dV_k)^{-\frac{\beta^2}{C}(\log dV_k)+1}$$

which is $o(1)$ when α has at most polynomial growth and V_k has polynomial growth in k. ■

Claim 8 *Let $\alpha\delta^2 \le \frac{\epsilon^2}{\log d}$. Then*

$$\Pr\left[\sup_{0\le t\le T} \eta^2 \sum_{j=0}^{t} \|L_t\|^2 > \frac{\epsilon}{18}\right] \lesssim \frac{1}{\log d}$$

Proof Note that the maximum is achieved at T since all the summands are non-negative. In that case,

$$\Pr\left[\eta^2 \sum_{j=0}^{T} \|L_t\|^2 > \frac{\epsilon}{18}\right] \lesssim \frac{\eta^2 T\mathbb{E}[\|L_t\|^2]}{\epsilon^2} \le \frac{\mu_1\alpha\delta^2 d^2 V_k^2}{d^2 V_k^2\epsilon^2} = \frac{\mu_1\alpha\delta^2}{\epsilon^2} \le \frac{1}{\log d} = o(1)\,.$$

■

E.2. Controlling the error martingale

Claim 9 *Let $\alpha\delta^2 \le \epsilon^2(\log d)^{-1}$. Furthermore, let $M_t = \eta \sum_{0\le j\le t-1}\langle E_j, u\rangle$. Then, M_t forms a $\mathcal{F}_t$ martingale and*

$$\Pr\left[\sup_{0\le t\le T}|M_t| \ge \frac{\beta}{10\sqrt{d}}\right] \lesssim \frac{\epsilon^2}{\beta^2\log d}$$

Furthermore, we have

$$\Pr\left[\sup_{0\le t\le T_1}|M_t| \ge \frac{\epsilon}{18}\right] \lesssim \frac{1}{d\log d}$$

Proof The fact that M_t is a martingale follows directly from Condition 1 and the fact that each x_t is a fresh sample. By Doob's maximal inequality for martingales, we have

$$\Pr\left[\sup_{0\le t\le T}|M_t| > \gamma\right] \le \frac{\mathbb{E}M_T^2}{\gamma^2} \le \frac{2\mu_1\eta^2 T V_k}{\gamma^2} = \frac{2\mu_1\alpha\delta^2}{d\gamma^2}\,.$$

Setting $\gamma = \frac{\beta}{10\sqrt{d}}$, we get the probability is at most $\frac{\epsilon^2}{\beta^2\log d}$ up to constants. For the second result, set $\gamma = \frac{\epsilon}{18}$ so that the probability is $O(\frac{1}{d\log d})$. ■

E.3. Weak recovery and strong recovery

Before we prove weak and strong recovery, we would like to define events $\mathcal{A}$ and $\mathcal{B}$ that capture the probabilistic bounds on population gradient magnitude and the various error terms in the dynamics.

E.3.1. GOOD EVENT FOR ERROR BOUNDS AND INITIAL CORRELATION

First, define the event $\mathcal{A}$ as

$$\mathcal{A} = \left\{m_0 \ge \frac{\beta\cdot\operatorname{sign}(h(0))}{\sqrt{d}}\right\}. \tag{15}$$

Furthermore, define the event $\mathcal{B} = \mathcal{B}(\epsilon, d, \beta, k, T)$ that corresponds to the error bounds as the following

$$\mathcal{B} = \left\{\sup_{0\le t\le T}|M_t| \le \min\left\{\frac{\beta}{10\sqrt{d}}, \frac{\epsilon}{36}\right\}\right\} \cap \left\{\sup_{0\le t\le T}\eta^3\sum_{j=0}^{t-1}\|L_j\|^2\,|\langle L_j, u\rangle| \le \min\left\{\frac{\beta}{10\sqrt{d}}, \frac{\epsilon}{18}\right\}\right\} \tag{16}$$

$$\cap \left\{\sup_{0\le t\le T}\eta^2\sum_{j=0}^{t}\|L_t\|^2 \le \frac{\epsilon}{18}\right\} \cap \left\{\sup_{0\le t\le T}\eta\sum_{j=0}^{t}\left(\frac{S_k}{4} - \eta\|L_t\|^2\right) \ge -\frac{\beta}{5\sqrt{d}}\right\}$$

Proposition 32 *Let $\delta = \frac{\epsilon^3 S_k}{4C_\delta\log(dV_k)}$ where $C_\delta > \max\{1, \mu_1\}$. Furthermore suppose that $\alpha = \frac{4(\log dV_k)}{\epsilon\delta S_k}$. Then, for $T = \lceil\alpha dV_k\rceil$, we have $\Pr(\mathcal{B}(\epsilon, d, \beta, k, T)) = 1 - O\left(\max\left\{\frac{1}{\beta\sqrt{d}}, \alpha(dV_k)^{-\frac{\beta^2}{C}(\log dV_k)+1}, \frac{\epsilon^2}{\beta^2\log d}, \frac{1}{d\log}\right.\right.$*
$1 - o(1)$.

Proof Notice that the given δ, α satisfy $\alpha\delta^2 \leq \frac{\epsilon^2}{C_\delta \log(dV_k)}$. Hence, all of Claims 6 to 8 hold. Then, combining the results of the claims with a union bound gives the result. ■

E.3.2. STOPPING TIMES FOR THE DYNAMICS

Initially, for a real number $q > 0$, define the stopping times

$$\tau_q^+ = \inf\{t \geq 0 : m_t \geq q\}$$
$$\tau_q^- = \inf\{t \geq 0 : m_t \leq q\}$$

which correspond to the first time m_t is above/below a certain threshold value q. In particular, we will define the following stopping times

$$\tau_r^+ = \inf\{t \geq 0 : m_t > r\}$$
$$\tau_0^- = \inf\{t \geq 0 : m_t < 0\}$$
$$\tau_{1-\epsilon/6}^+ = \inf\{t \geq 0 : m_t \geq 1 - \frac{\epsilon}{6}\}$$

τ_r^+ is defined to analyze the initial stage of training, when m_t is small. This allows us to lower bound the effect of the spherical projection of the gradients $1 - m_t^2$. We will use τ_0^- to be able to lower bound the population gradient, but we will get rid of the requirement with an argument that m_t has to always be non-negative when $\mathcal{B}$ holds. Finally, $\tau_{1-\epsilon/6}^+$ is used to analyze the stage before we achieve the initial strong correlation, we will show m_t will stay above $1 - \epsilon$ after $t > \tau_{1-\epsilon/6}^+$. I.e. the progress made for strong recovery is not eliminated by the noisy gradients.

E.3.3. ANALYZING THE DYNAMICS CONDITIONING ON $\mathcal{B}$

Now, notice that we can WLOG assume $\text{sign}(h(0)) = 1$, since all the proofs will be symmetric as long as the event $\mathcal{A}$ holds. Furthermore, let $r < \frac{1}{\sqrt{2}}$

Lemma 33 (Characterizing dynamics before weak recovery) *Conditioning on $\mathcal{A}, \mathcal{B}$, for $t \leq T \wedge \tau_r^+ \wedge \tau_0^-$, we have*

$$m_t \geq \frac{\beta}{2\sqrt{d}} + \frac{t\eta S_k}{2}$$

Furthermore, we have $\tau_0 > T \wedge \tau_r^+$.

Proof Condition on $\mathcal{A}, \mathcal{B}$. Then, as explained before, WLOG assume $\text{sign}(h(0)) = 1$. Then, for all $t \leq \tau_0^-$, we must have $h(m_t) \geq S_k$. Furthermore, for all $t \leq \tau_r^+$, we have $1 - m_t^2 > \frac{1}{2}$. Then, rearranging using Claim 5 and applying the inequalities that hold with the event $\mathcal{B}$, for $t \leq \tau_r^+ \wedge \tau_0^- \wedge T$, we have

$$\begin{aligned} m_t &\geq m_0 + \eta \sum_{j=0}^{t-1} h(m_j)(1 - m_j^2) - \eta \sum_{j=0}^{t-1} \langle E_j, u\rangle - \eta^2 \sum_{j=0}^{t-1} \|L_j\|^2 - \eta^3 \sum_{j=0}^{t-1} \|L_j\|^2 |\langle L_j, u\rangle| \\ &\geq m_0 + \frac{\eta t S_k}{4} + \eta \sum_{j=0}^{t-1} \left(\frac{S_k}{4} - \eta \|L_j\| \right) - \frac{\beta}{5\sqrt{d}} \end{aligned}$$

Now, using the uniform lower bound on the summation term and $m_0 \geq \frac{\beta}{\sqrt{d}}$, we have

$$m_t \geq \frac{\beta}{2\sqrt{d}} + \frac{\eta t S_k}{4}$$

which concludes the first part. For the second part, suppose for $j \leq \tau_r^+ \wedge T$, we have $j \leq \tau_0^-$. Then, for all $l \in [0, 1, \ldots, j-1]$ we have $m_l \geq 0$, meaning $h(m_l) \geq S_k$. Hence, the above inequality holds for j, meaning $m_j > 0$. Hence, this implies $j < \tau_0^-$. Then, we conclude that it must be the case that $\tau_0^- > \tau_r^+ \wedge T$. ■

Lemma 34 (Dynamics after weak recovery is well approximated by drift term) *Conditioning on $\mathcal{A}, \mathcal{B}, \tau_r^+$, the following holds: For $t \geq \tau_r^+$ with $t \leq T \wedge \tau_0^-$, we have*

$$\left| m_t - m_{\tau_r}^+ - \eta \sum_{j=\tau_r^+}^{t-1} h(m_j)(1 - m_j^2) \right| < \frac{\epsilon}{6}$$

Furthermore, $\tau_0^- > T$.

Proof Notice that under the event $\mathcal{B}$, due to non-negativity of each of the summands, we have the following upper bounds

$$\eta^3 \sum_{j=\tau_r^+}^{t-1} \|L_j\|^2 |\langle L_j, u \rangle| \leq \sup_{0 \leq t \leq T} \eta^3 \sum_{j=0}^{t-1} |\langle L_j, u \rangle| < \frac{\epsilon}{18}$$

$$\eta^2 \sum_{j=\tau_r^+}^{t-1} \|L_j\|^2 \leq \sup_{0 \leq t \leq T} \eta^2 \sum_{j=0}^{t-1} \|L_j\|^2 < \frac{\epsilon}{18}$$

For the martingale term, since the terms are not necessarily non-negative we decompose it as

$$\begin{aligned}
\left| \eta \sum_{j=\tau_r^+}^{t-1} \langle E_j, u \rangle \right| &= \left| \eta \sum_{j=0}^{t-1} \langle E_j, u \rangle - \eta \sum_{j=0}^{\tau_r^+ - 1} \langle E_j, u \rangle \right| \\
&\leq \left| \eta \sum_{j=0}^{t-1} \langle E_j, u \rangle \right| + \left| \eta \sum_{j=0}^{\tau_r^+ - 1} \langle E_j, u \rangle \right| \\
&\leq 2 \sup_{0 \leq t \leq T} \left| \eta \sum_{j=0}^{t-1} \langle E_j, u \rangle \right| < \frac{\epsilon}{18}
\end{aligned}$$

Then, notice that the following holds exactly

$$m_t = m_{\tau_r^+} + \eta \sum_{j=\tau_r^+}^{t-1} h(m_j)(1 - m_j^2) + \eta \sum_{j=\tau_r^+}^{t-1} \langle E_t, u \rangle + \sum_{j=\tau_r^+}^{t-1} \left(1 - \frac{1}{r_j}\right) (m_j - \eta \langle L_j, u \rangle)$$

which after rearranging, using $\left|1-\frac{1}{r_j}\right| \leq \eta^3 \|L_j\|^2 |\langle L_j, u\rangle| + \eta^2 \|L_j\|^2$ gives us

$$\begin{aligned}\left|m_t - m_{\tau_r^+} - \eta \sum_{j=\tau_r^+}^{t-1} h(m_j)(1-m_j^2)\right| &= \left|\eta \sum_{j=\tau_r^+}^{t-1} \langle E_t, u\rangle + \sum_{j=\tau_r^+}^{t-1} \left(1-\frac{1}{r_j}\right)(m_j - \eta\langle L_j, u\rangle)\right| \\ &\leq \left|\eta \sum_{j=\tau_r^+}^{t-1} \langle E_t, u\rangle\right| + \eta^3 \sum_{j=\tau_r^+}^{t-1} \|L_j\|^2 |\langle L_j, u\rangle| + \eta^2 \sum_{j=\tau_r^+}^{t-1} \|L_j\|^2\end{aligned}$$

using the $\epsilon/18$ bound for each of the terms, we get a total bound of $\epsilon/6$. Then, to get rid of the requirement $t \leq \tau_0^-$, notice that

$$m_t - m_{\tau_r^+} \geq -\frac{\epsilon}{3} + \sum_{j=\tau_r^+}^{t-1} h(m_j)(1-m_j^2)$$

Then, notice that if $t \leq \tau_0^-$, we have $m_j \geq 0$ for all $j \leq t-1$, so the sum is non-negative, which gives us $m_t \geq m_{\tau_r^+} - \frac{\epsilon}{3} \geq r - \frac{\epsilon}{3}$. However, notice that choosing $r = \frac{1}{2}$, we always have $\epsilon/3 < r$ so $m_t \geq 0$ as well. Hence, $\tau_0^- > t$, so we must have $\tau_0^- > T$. ■

Now, we are in a position to prove Theorem 16.

Proof [Proof of Theorem 16] First, due to the initalization requirement in the theorem, $\mathcal{A}$ holds. Then, per Theorem 32, $\mathcal{B}$ holds with probability $1-o(1)$. Then, conditioning in $\mathcal{B}$, per Theorem 33 and Theorem 34, we can drop the requirement that $t \leq \tau_0^-$. So, let $t \leq T \wedge \tau_r^+$. Conditioning on $\mathcal{B}$, per Theorem 33, we have

$$m_t \geq \frac{\beta}{2\sqrt{d}} + \frac{t\eta S_k}{2}$$

Then, notice that at time $T_{\text{weak}} = \lceil \frac{2}{\eta S_k} \rceil$, the RHS is larger than 1. Then, it must be the case that $\tau_r^+ \wedge T \leq T_{\text{weak}}$. Then, it suffices to show $T_{\text{weak}} \leq T$. Notice that $T_{weak} = \lceil \frac{2dV_k}{\delta S_k} \rceil$ and $T = \lceil \alpha d V_k \rceil = \lceil \frac{4(\log dV_k)}{\epsilon \delta S_k} \rceil > T_{\text{weak}}$ when $\epsilon < 1, V_k > 1$ and $d > 3$. Then, we conclude $\tau_r^+ \leq T_{\text{weak}} \leq T$.

Now, conditioning on τ_r^+, for all $t \geq \tau_r^+$, with $t \leq T$ per Theorem 34, we have

$$m_t \geq m_{\tau_r^+} + \sum_{j=\tau_r^+}^{t-1} h(m_j)(1-m_j^2) - \frac{\epsilon}{6}$$

Now, consider $t \leq \tau_{1-\epsilon/6}^+ \wedge T$, so that $h(m_j)(1-m_j^2) > S_k \frac{\epsilon}{6}$ for all $j \leq \tau_{1-\epsilon/6}^+$. Hence,

$$m_t \geq r + \frac{\eta(t-\tau_r^+)S_k\epsilon}{6} - \frac{\epsilon}{6} > \frac{\eta(t-\tau_r^+)S_k\epsilon}{6}$$

Hence, notice that the RHS of the inequality is greater than 1 at time $t = \tau_r^+ + \lceil \frac{6}{\eta S_k \epsilon} \rceil \leq T_{\text{weak}} + \lceil \frac{6}{\eta S_k \epsilon} \rceil$. Hence, it must be the case that $\tau_{1-\epsilon/6}^+ \wedge T \leq T_{\text{weak}} + \lceil \frac{6}{\eta S_k \epsilon} \rceil$. However, notice that

$T = \lceil \frac{dV_k(\log dV_k)}{\delta S_k \epsilon} \rceil$ which is larger than $T_{\text{weak}} + \lceil \frac{6}{\eta S_k \epsilon} \rceil$ so it must be the case that $\tau^+_{1-\epsilon/6} \leq T$. Finally, we need to show that m_t stays above $1-\epsilon$ after it crosses $1-\epsilon/6$. However, notice that for $t' \geq t \geq \tau_r^+$, we have

$$\begin{aligned} m_{t'} - m_t &\geq \left| m_t - m_{\tau_r^+} - \eta \sum_{j=0}^{t-1} h(m_j)(1-m_j^2) \right| + \left| m_{t'} - m_{\tau_r^+} - \eta \sum_{j=0}^{t'-1} h(m_j)(1-m_j^2) \right| + \sum_{j=t}^{t'-1} h(m_j)(1-m_j^2) \\ &\geq -\frac{\epsilon}{3} \end{aligned}$$

so that $m_t \geq 1 - \frac{\epsilon}{2}$ for $t \geq \tau^+_{1-\epsilon/6}$. Hence, we conclude that $m_T \geq 1 - \frac{\epsilon}{2}$. Since this result holds for any τ_r^+, we can conclude the proof. ■

Appendix F. Additional details

In this section we provide details for two remaining points mentioned in the main text, namely the existence of multiple global optima when Assumption 2 does not hold, and the fact that one can learn a good approximation to the teacher model once u is learned.

F.1. Multiple global optima when Assumption 2 does not hold

The following example shows that if the direction u of the perturbation lies in the span of the base model weight vectors, then there can exist multiple global optima.

Example 1 *Let $\lambda_1, \lambda = 1$, let $w_1 = (1,0)$, $w_2 = (0,1)$, and consider the activation $\sigma(z) = z^2$. If the base model $f : \mathbb{R}^2 \to \mathbb{R}$ is given by $f(x) = \sum_{i=1}^2 \lambda_i \sigma(\langle w_i, x \rangle)$, then observe that the following two rank-1 perturbations of equal scale are equal.*

First, take $u = (1/\sqrt{2}, 1/\sqrt{2})$ and $u' = (1/\sqrt{3}, \sqrt{6}/3)$. Then define $c = (-(1+\sqrt{2})(2+\sqrt{3}), (1+\sqrt{2})(\sqrt{2}+\sqrt{3}))$ and $c' = -c$. Then one can verify that the teacher models $\sum_{i=1}^2 \lambda_i \sigma(\langle w_i + c_i u, x \rangle)$ and $\sum_{i=1}^2 \lambda_i \sigma(\langle w_i + c_i' u', x \rangle)$ are functionally equivalent, even though $\{w_1 + c_1 u, w_2 + c_2 u\} \neq \{w_1 + c_1' u', w_2 + c_2' u'\}$, regarded as unordered pairs of vectors in $\mathbb{R}^2$. Furthermore, $\|c\| = \|c'\|$.

F.2. Learning the teacher model once u is learned

In this section, we show that learning u is sufficient to learning the teacher model by adding additional features to the model and training the second layer.

Definition 35 (Linear Model Family From Learned Features) *Let $\hat{u}$ be given. Then, define the model family*

$$\mathcal{L}_\lambda = \left\{ \sum_{i=1}^k \lambda_{i,1} \sigma\left(\left\langle \frac{w_i + \frac{\xi}{\sqrt{k}} \hat{u}}{\sqrt{1+\xi^2/k}}, x \right\rangle \right) + \lambda_{i,2} \sigma\left(\left\langle \frac{w_i - \frac{\xi}{\sqrt{k}} \hat{u}}{\sqrt{1+\xi^2/k}}, x \right\rangle \right) : \lambda \in \mathbb{R}^k \times \mathbb{R}^k \right\} \quad (17)$$

Then, we will show that once we learn $\hat{u}$ to a sufficient accuracy, there exist a choice of λ that allows the linear model to closely approximate the teacher model.

Theorem 36 (Learning u is sufficient to learn f^*) *Suppose $\hat{u}$ is such that $1-|\langle u, \hat{u}\rangle| \leq \epsilon \cdot \frac{k+\xi^2}{2C_\sigma \lambda_{\max}^2 \xi^2 k^2}$ which is $\Theta(\epsilon/k)$ for $\xi = \Theta(1)$ and $\Theta(\epsilon/k^2)$ for $\xi = \Theta(\sqrt{k})$ Then, there exists a model $h \in \mathcal{L}_\lambda$ as defined in Equation (17) such that $\mathbb{E}_x(f^*(x) - h(x))^2 \leq \epsilon$. In particular, second layer training on the family of neural networks defined as $\mathcal{L}_\lambda$, we*

Proof WLOG suppose $\langle u, \hat{u}\rangle > 0$, otherwise we flip all the signs of the c_i in the later part of the proof. Consider the candidate model $h \in \mathcal{L}_\lambda$ (given in eq. (17)) given by

$$h(x) = \sum_{i=1}^{k} \lambda_i \sigma\left(\left\langle \frac{w_i + \xi c_i \hat{u}}{\sqrt{1+\xi^2/k}}, x \right\rangle\right)$$

We aim to show $\mathbb{E}_x(f^*(x) - \hat{f}(x))^2 \leq \epsilon$. Notice

$$\mathbb{E}_x(f^*(x) - \hat{f}(x))^2 \leq k \sum_{i=1}^{k} \lambda_i^2 \mathbb{E}_x \left(\sigma(\langle v_i, x\rangle) - \sigma(\langle \tilde{v}_i, x\rangle)\right)^2$$

where v_i is as before and $\tilde{v}_i = \frac{w_i + \xi c_i \hat{u}}{\sqrt{1+\xi^2/k}}$. Then, it suffices to show that the expectation is less than $\frac{\epsilon}{\lambda_{\max}^2 k^2}$. Note

$$\mathbb{E}_x(\sigma(\langle v_i, x\rangle) - \sigma(\langle v_i, x\rangle))^2 \leq C_\sigma \left\|v_i - \hat{v}_i\right\|^2$$

Furthermore, we have

$$\|v_i - \hat{v}_i\| = \frac{\xi/\sqrt{k}\,\|u - \hat{u}\|}{\sqrt{1+\xi^2/k}}$$

So that

$$k \sum_{i=1}^{k} \lambda_i^2 \mathbb{E}_x(\sigma(\langle v_i, x\rangle) - \sigma(\langle v_i, x\rangle))^2 \leq C_\sigma \lambda_{\max}^2 k \frac{2\xi^2(1 - \langle u, \hat{u}\rangle)}{1+\xi^2/k}$$

Then, it suffices to get $1 - \langle u, \hat{u}\rangle \leq \epsilon \cdot \frac{k+\xi^2}{2C_\sigma \lambda_{\max}^2 \xi^2 k^2}$ as desired. ■

Remark 37 *The above result can be extended to the case when the c_i are not necessarily quantized, by quantizing the interval $[-1, 1]$ into a sufficiently granular discrete set of elements. Then, the algorithm follows similarly by adding these features into the model and training the second layer (e.g. via linear regression or SGD).*