GRADIENT DYNAMICS OF LOW-RANK FINE-TUNING BEYOND KERNELS

Anonymous authors
Paper under double-blind review

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 sucess, from a mathematical perspective it remains poorly understood what learning mechanisms ensure that gradient descent converges to useful low-rank perturbations.

In this work we initiate the study of 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, the training dynamics are nonlinear. Nevertheless, in this regime we prove under mild assumptions that a student model which is initialized at the base model and trained with online gradient descent 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.

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 weights of the pre-trained model and only train *low-rank perturbations* to the weight matrices. Remarkably, this works well even with rank 1 perturbations, reducing number of trainable parameters by up to four 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 sufficiently deep and wide pre-trained networks, any sufficiently simple target model can be approximated by a low-rank perturbation of the larger model (Zeng & Lee, 2024), it is largely unknown what mechanisms ensure that gradient-based training converges to these perturbations. Recent works have made initial 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 consider a setting where 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 gradient descent for low-rank fine-tuning converge to a good solution even when the kernel approximation breaks down?

To answer this question, we initiate the study of fine-tuning in a natural student-teacher setting where the training dynamics are inherently non-linear.

1.1 Problem formulation

We consider some family $\mathcal{F} = \{f_\theta\}_{\theta\in\Theta}$ of neural networks, each parametrized by a collection θ of weight matrices. 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,

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

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*. Despite the apparent simplicity of 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. The base model then takes the form

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

where $\theta_0 = (\lambda, W) \in \mathbb{R}^k \times \mathbb{R}^{k\times d}$ and σ is a known scalar activation applied entrywise.

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

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

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 $L(\hat\theta) \leq \varepsilon$ for $\hat\theta \triangleq (\lambda, W + \xi \hat c \hat u^\intercal)$ with high probability over the training data $\{(x_i, y_i)\}_{i=1}^N$.

Connection to GLMs, feature learning, and lazy training. 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 $\xi = \sqrt{k}$.

[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.

[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.

In this case, if the teacher models' parameters are given by $\theta = (\lambda, W^*)$ where W^* is defined in Eq. (3), then the teacher model 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 the question of GLM regression. The behavior of gradient descent for this question is by now very well-understood, shedding light on the training dynamics of neural networks in the *feature learning* regime (sometimes also called the "rich" or "μP" regime) in a stylized but rich model (Bietti et al., 2022).

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, note that 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 (also called the "NTK regime"). For this reason, the scale parameter ξ gives a natural way to interpolate between feature learning and lazy training dynamics.

1.2 Our contributions

1.2.1 Assumptions

Our guarantees will apply to a very wide family of activations σ including all standard ones, e.g. ReLU, sigmoid, polynomial, etc. As the conditions are rather technical, we defer them to Assumption 5 in the supplement and henceforth refer to such activations as *nice*.

More importantly, 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 D.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.

The third assumption 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 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 1).

As for the randomness of c, while we conjecture that fine-tuning should be tractable even in the worst case over c (see Remark 3) 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, in the spirit of smoothed analysis (Spielman & Teng, 2004). One could equivalently formulate our guarantees as holding under a certain set of deterministic nondegeneracy conditions on the rank-1 perturbation.

1.2.2 Training algorithm

In this work, we will focus on learning the factor u in the rank-1 perturbation $\Delta = \xi c u^\intercal$ 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 D.3 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

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

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

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

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.

1.2.3 Statement of results

In this work, we consider two regimes: (1) when $\{w_i\}$ are orthogonal, and (2) when $\{w_i\}$ have very mild angular separation but are otherwise arbitrary.

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.

Our first result is to show that online SGD efficiently converges to the correct rank-1 perturbation.

Theorem 1 (Informal, see Theorem 6)**.** *Let $0 < \varepsilon < 1$, and let $\xi \asymp \sqrt{k}$ for sufficiently small absolute constant factor. Suppose the rows of W are orthogonal. Then under Assumptions 1-3 and for any nice activation σ (see Assumption 5), 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 u_0: online SGD (see Eq.* (4)*) run with step size $\eta = \tilde{\Theta}(\varepsilon^3/dk^{7/2})$ and $T = \tilde{\Theta}(dk^4/\varepsilon^4)$ iterations results in u_T for which $\langle u_T, u\rangle^2 \ge 1 - \varepsilon$.*

Interestingly, the iteration complexity does not depend on fine-grained properties of the activation σ. In contrast, as we discuss in Section 2, the iteration complexity of noisy gradient descent for learning GLMs depends heavily on the decomposition of σ in the Hermite basis. Given that the GLM setting can be recovered from the fine-tuning setting in the $\xi \to \infty$ limit, Theorem 1 implies that the gradient dynamics for fine-tuning exhibit a transition in behavior at some scale $\xi = \Omega(\sqrt{k})$.

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 assumption:

Assumption 4 (Angular separation)**.** *For all $i \neq j$, we have $|\langle w_i, w_j\rangle| \le 1 - \log k/\sqrt{k}$.*

Theorem 2 (Informal, see Theorem 7)**.** *Under the same assumptions as Theorem 1, except with $\xi = 1$ and assuming the rows of W satisfy Assumption 4 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 $\eta = \tilde{\Theta}(\varepsilon^3/dk^{5/2})$ and $T = \tilde{\Theta}(dk^3/\varepsilon^4)$ iterations results in u_T for which $\langle u_T, u\rangle^2 \geq 1-\varepsilon$.*

Given the generality of Assumption 4, 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 $\xi = \Theta(1)$, the linearization of the teacher model around the base model 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.

Finally, we show a rigorous separation between what can be done in the fine-tuning setting and what can be done learning a two-layer network from scratch (see Appendix D.2 for details):

Theorem 3 (Informal, see Theorem 9)**.** *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 10) with a carefully chosen c, u. Even though c is not random, we prove online SGD still converges to the ground truth perturbation in $\tilde{O}(d)$ iterations.

1.3 Related work

Parameter-efficient fine-tuning. Following the popularization of LoRA (Hu et al., 2021), there have been a large number of proposed refinements thereof (Fu et al., 2023; Dettmers et al., 2024; Lialin et al., 2023); a thorough review of the empirical literature is beyond the scope of this work.

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) primarily presented empirical evidence of kernel behavior for prompt-based fine-tuning methods, including LoRA, in the few-shot regime. Their main theoretical result regarding LoRA roughly states that if standard (full-rank) fine-tuning exhibits kernel behavior, then low-rank fine-tuning exhibits kernel behavior, provided the rank of the perturbation is at least $\Omega(1/\varepsilon^2)$. Jang et al. (2024) build upon this as follows. In the kernel regime where the student model is well-approximated by its linearization around the base model throughout training, they consider the resulting linearized empirical loss for an arbitrary dataset. This is still non-convex if one tries jointly training the factors of the low-rank perturbation, but they nevertheless show that this loss has a rank-$O(\sqrt{N})$ global minimizer, where N is the number of training examples. They then show that all local minimizers of this loss are global minimizers, using tools from prior work on low-rank matrix factorization.

These works are incomparable to ours in several regards. Firstly, they operate in the few-shot regime so that the number of training examples N is relatively small, and the perturbation is small enough that one can work with a linear approximation. In contrast, we consider "full" low-rank fine-tuning, for which N must scale at least with the ambient dimension, and we are trying to learn much larger perturbations; as we show in Figure 2, this puts us well outside the regime where the kernel approximation does well. In addition, the aforementioned works do not handle the regime where the rank is extremely small, even though LoRA still works quite well 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. Generalized linear models have received significant attention in learning theory as a stylized model for feature learning, see Dudeja & Hsu (2018) for an overview of older works on this. 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*, defined to be 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).

By a result of Szörényi (2009), the information exponent also dictates the worst-case complexity of learning generalized linear models: for noisy gradient descent (and more generally, correlational statistical query algorithms), $d^{1\vee l^*/2}$ samples are necessary. Various works have focused on deriving algorithms that saturate this lower bound and related lower bounds for 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; 2024; Abbe et al., 2023). A key finding of our work is that quantities like information exponent do not dictate the complexity of fine-tuning.

PAC learning neural networks. Within the theoretical computer science literature on learning neural networks, there has 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 & Narayanan (2024); Chen et al. (2023) 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 & 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 Section 2, even if the base model or teacher model satisfy 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 s-th order tensor power of w, and for two tensors T_1, T_2 we use $\langle T_1, T_2\rangle$ to denote their elementwise dot product and $\|T_1\|_F \triangleq \sqrt{\langle T_1, T_1\rangle}$ for the corresponding Frobenius norm. Note the identity $\sum_{i,j=1}^k \lambda_i \nu_j \langle w_i, v_j\rangle^s = \langle \sum_{i=1}^k \lambda_i w_i^{\otimes s}, \sum_{i=1}^k \nu_i v_i^{\otimes s}\rangle$ 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 $\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}[h_p(\langle u, x\rangle) h_l(\langle v, x\rangle)] = \mathbb{1}\{l = p\}\langle u, v\rangle^p$.

2 Expression for the population gradient

To give intuition for our analysis of online SGD, we first consider the dynamics of gradient descent on the *population loss*, defined as

$$\Phi(\hat{u}) \triangleq \mathbb{E}_{x\sim\mathcal{N}(0,I)}[(f^*(x) - \lambda^\intercal \sigma((W_0 + \xi \hat{c}\hat{u}^\intercal)x))^2], \tag{5}$$

recalling that f^* is the teacher, and $\hat{c}$ is frozen at its random initialization in $\{\pm 1/\sqrt{k}\}^k$.

In this section we derive a closed-form expression for the gradient of this loss and provide high-level discussion on how a key scaling factor term in this expression influences the gradient dynamics.

We begin by calculating the population gradient (see Appendix A.1 for the proof):

Proposition 1. *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

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

This admits a natural interpretation: $-\hat{\nabla}\Phi(\hat{u})$ is a scaling of the ground truth direction u after it has been projected to the orthogonal complement of the current SGD iterate $\hat{u}$. 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.

Nevertheless, it suffices to prove that this scaling $h(\langle u, \hat{u}\rangle)$ is lower bounded throughout the trajectory of gradient descent. To see this, let $\overline{u}_t$ denote the iterates of population gradient descent and define $\overline{m}_t \triangleq \langle \overline{u}_t, u\rangle$. Under one step of population gradient descent, we get the following update:

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

where the approximation is because we are ignoring the projection step in this informal overview, for simplicity. Rearranging, we find that in one step, $1-\overline{m}_t$ contracts by a factor of $1-\eta h(\overline{m}_t)(1+\overline{m}_t)$. In particular, assuming $\overline{m}_t > 0$, this contraction is non-negligible as long as $h(\overline{m}_t)$ is non-negligible.

Lower bounding h will thus be the main focus of our analysis.

Recovering generalized linear model dynamics. Consider taking $\xi \to \infty$. In the definition of h in Eq. (1), for each l we see that all of the summands $s > 0$ are of lower order, so that

$$h(m) \to 2\sum_{l=0}^{\infty}\mu_{l+1}(\sigma)^2 T(l,0) m^l\,. \tag{6}$$

Note that $T(l,0)$ only depends on the parity of l: we have $T(l,0) = (\sum_i \lambda_i)^2$ if l is odd and $T(l,0) = \langle \sum_i \lambda_i c_i, \sum_i \lambda_i \hat{c}_i\rangle$ if l is even, and we can assume these terms are non-negligible. The reason is that they capture the first-order behavior of the degree-l component of the target model after its inputs have been scaled down by a factor of ξ. In particular, if the $T(l,0)$ vanish, then the rank-1 perturbation is information-theoretically not learnable.

In the $\xi \to \infty$ limit, Eq. (6) informally recovers the well-known fact that the complexity of online SGD for generalized linear model regression depends on the *information exponent* of σ: the behavior of h is dictated by the degree of the smallest non-negligible term in its series expansion, i.e. the smallest p for which $|\mu_p(\sigma)| \gg 0$. In particular, the larger this is, the longer it takes for the dynamics to escape from the value of m at initialization, namely $\overline{m}_0 = \langle \overline{u}_0, u\rangle \approx 1/\sqrt{d}$.

In this work, we focus on low-rank fine-tuning rather than generalized linear models and thus consider the finite ξ scaling instead. As we will see, the dynamics under this scaling exhibit very different behavior and are far less sensitive to the particulars of the activation function σ.

3 Lower bounding the population gradient throughout training

In this section we state our main results on lower bounding the scaling factor $h(m)$ from Proposition 1 and provide key intuitions for the proofs, the full details of which are in the supplement. Note that the population gradient can be potentially quite non-linear, and it is not apriori clear whether it would vanish for $m \neq \pm 1$. However, $h(m)$ being non-vanishing across training is crucial, since it is the main term guiding the dynamics. In this section, we argue that under our assumptions, when the sign of m is aligned with $h(0)$, the function $h(m)$ admits a lower bound.

3.1 Orthonormal Features

Here we assume $w_1, \dots, w_k$ are orthonormal, so that the form of $T(l,s)$ in Proposition 1 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

$$h(m) = 2\left[k\sum_{i,j=1}^{k} \lambda_i\lambda_j c_i\hat{c}_j\right] \sum_{\substack{l=0\\ \text{even}}}^{\infty} \left(\frac{\xi^2}{k}\right)^{l+1} (l+1)\mu_{l+1}(\sigma)^2 \left(\frac{k}{k+\xi^2}\right)^{l+1} m^l$$

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

for some coefficients $b_l \geq 0$. Informally, 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 $\Theta(\sqrt{k})$, with high probability over the randomness of $c, \hat{c}$. Then, notice that if $\mu_1(\sigma) \neq 0$, the first term with even l should dominate the second term. In particular, $h(0)$ will dominate the even terms in the second term, and the typical magnitude of h will be $\Theta(\xi^2)$. If $\mu_1(\sigma) = 0$, notice that this is not immediately true since $h(0)$ now could be of a smaller magnitude, but we show that with high probability, the even l, $s = 0$ terms are dominated by the odd l, $s = 1$ terms. Since the odd terms have the same sign as m, as long as the sign of m agrees with that of $h(0)$ we should see relatively monotonic behavior and h should not vanish. In this case, from anti concentration (Proposition 7), we expect a typical magnitude for h to be $\Theta(\xi^2/\sqrt{k})$.

3.2 Separated Features

We now drop the orthonormality assumption and only assume angular separation of the w_i's (Assumption 4). In this case, the population loss does not simplify. However, when $\xi = 1$, we can show that the higher order even terms in the expansion of $h(m)$ are negligible relative to the constant term. First, note that the sums

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

scale with $\Theta(k^l)$, so their contribution could potentially be large. However, we initially show that if we take only the first $s^* = O(\sqrt{k})$ terms, all the low order even terms are small

$$\sum_{\substack{l\geq 2\\ \text{even}}}^{\infty} \sum_{s\geq 0} \left(\frac{\xi^2}{k}\right)^{l+1} \binom{l+s}{l} (l+s+1)\mu_{l+s+1}(\sigma)^2 = O(k^{-\frac{3}{2}-\varepsilon})$$

so that the maximum contribution after adding the factors is $k^{-\frac{1}{2}-\varepsilon}$, for some $\varepsilon > 0$ that depends on the activation. Hence, we separate the factor of the even terms into its diagonal and off-diagonal components:

$$\sum_{i,j=1}^{k} \lambda_i\lambda_j c_i\hat{c}_j \langle w_i, w_j\rangle^s = \sum_{i=1}^{k} \lambda_i^2 c_i\hat{c}_i + \sum_{i\neq j} \lambda_{i=1}^2 c_i\hat{c}_j \langle w_i, w_j\rangle^s$$

Notice that the diagonal components are $\Theta(\lambda_{\min}^2/\sqrt{k})$ with high probability. For these terms and large s, we use the decay of the Hermite coefficients of σ to bound their total contribution by $O(k^{-\frac{1}{2}-\varepsilon})$. For the off-diagonals, we use the angular separation of the weights: Note $(|\langle w_i, w_j\rangle|)^{\gamma\sqrt{k}} \leq (1 - \frac{\log k}{\sqrt{k}})^{\gamma\sqrt{k}} \leq e^{-\gamma \log k} \leq k^{-\gamma}$. Then, we establish a separation between the magnitudes of $h(0)$ and the higher order even terms by showing $h(0)$ has typical magnitude $\Theta(\lambda_{\min}^2/\sqrt{k})$. Then, we argue that the dynamics must be governed by $h(0)$ and the odd terms.

4 FINITE-SAMPLE ANALYSIS AND PUTTING EVERYTHING TOGETHER

Once we know the population gradient is leading m_t in the right direction, we need to show the noise from the stochastic gradients is negligible in training over a long time horizon. Notice that this does not mean SGD noise is entirely negligible: In fact, over short time horizons, it could potentially dominate the dynamics (see Figure 1). Note that we have the stochastic dynamics

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

where E_t is the random error induced due to the sampling of the gradients and Q_t is the distortion error due to projection onto the unit sphere. Then, unrolling the recursion, we have

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

Now, note the population gradient term guides the dynamics in the right direction, whose effect should scale with ηT. Furthermore, the second term forms a margingale, whose effect should scale with $\eta\sqrt{T}$ by Doob's maximal inequality. Over long horizons, we can choose η, T appropriately to make the noise negligible relative to the progress. We use a similar analysis to Arous et al. (2021), but unlike in that work, here we need to explicitly track dependencies on k and ε. In particular, on the finite sample analysis side, we show the following, which we then apply to various settings in fine-tuning:

Theorem 4 (Informal, see Theorem 8)**.** *If $h(m)$ is nice, and lower bounded by S_k throughout training, and the variance of the noise is bounded above by V_k, online SGD with appropriate step size, initialization, and time horizon $T = \widetilde{O}(\frac{dV_k}{S_k^2 \varepsilon^4})$ satisfies $|m_t| \geq 1 - \varepsilon$ with high probability.*

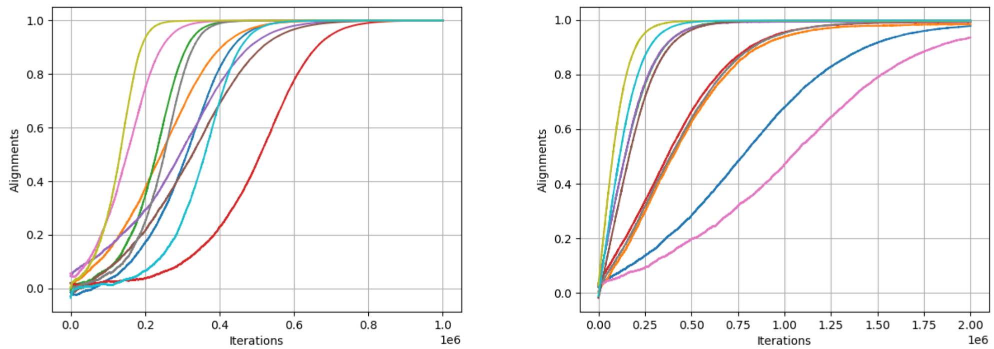
5 NUMERICAL SIMULATIONS

In this section we illustrate (i) the robustness of our theory to small changes in the assumptions (ii) the distinction between our work and kernel methods. In particular, for (i) relax the assumption that c_i are quantized, and we also compare the cases when $\hat{c}$ is frozen and jointly trained with $\hat{u}$. For (ii), we show that linearized networks (kernel approximation) fails at $\xi = \Theta(\sqrt{k})$, and also illustrate some interesting behavior in the joint training of $\hat{u}$ and $\hat{c}$. 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 1.(a) and when $\hat{c}$ is frozen Figure 1.(b). Notably, we see that while freezing $\hat{c}$ leads to longer time scales in training, the qualitative behavior of $\langle u_t, u\rangle$ is similar across the two settings. Next, we test the $\xi = \Theta(\sqrt{k})$ scaling, but we keep the problem setup same otherwise. 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 2.(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 (see Figure 2.(b)). This suggest that the initial phase of joint training might be similar to the training with frozen $\hat{c}$.

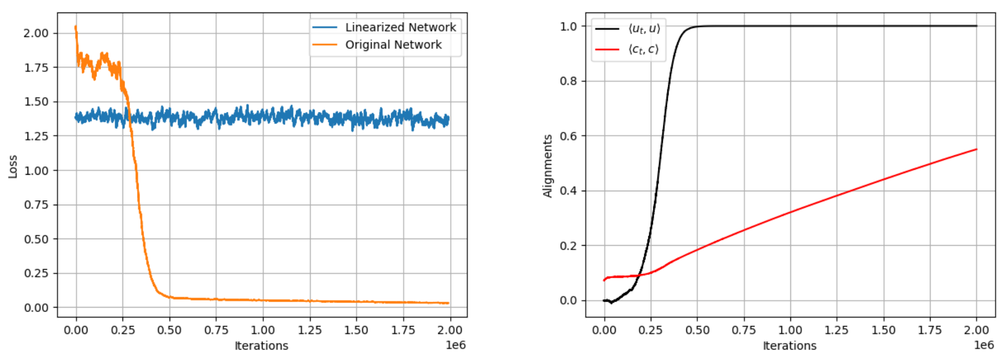
6 OUTLOOK

In this work we took the first steps towards understanding the gradient dynamics 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.



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

Figure 1: 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 2: 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.

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 to have equal-magnitude entries, and the assumption that c is random.

For these questions, a fruitful starting point could be to target a specific, analytically tractable activation function like quadratic activation, 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 – Figure 1 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?

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*, pp. 2552–2623. PMLR, 2023.

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

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, 2024.

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

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 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*, pp. 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*, pp. 1514–1539. PMLR, 2020.

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

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

Zihao Fu, Haoran Yang, Anthony Man-Cho So, Wai Lam, Lidong Bing, and Nigel Collier. On the effectiveness of parameter-efficient fine-tuning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 12799–12807, 2023.

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*, pp. 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.

Vladislav Lialin, Sherin Muckatira, Namrata Shivagunde, and Anna Rumshisky. Relora: High-rank training through low-rank updates. In *The Twelfth International Conference on Learning Representations*, 2023.

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*, pp. 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)*, pp. 21–30. IEEE, 2005.

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

Daniel A Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM (JACM)*, 51(3):385–463, 2004.

Balázs Szörényi. Characterizing statistical query learning: simplified notions and proofs. In *International Conference on Algorithmic Learning Theory*, pp. 186–200. Springer, 2009.

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

Supplement

Table of Contents

A Intuition and Statement of Results

A.1 Assumption on the activation function

We first state the technical assumptions on the activation function σ:

Assumption 5 (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.*
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*

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

Remark 1. *These conditions are satisfied for any reasonable activation used in practice. For the last condition in assumption 5, note that 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 $p(x) = \sum_{n=0}^{d} a_n x^n$, note that

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

Then, applying Cauchy-Schwarz, we have

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

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

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

$$f^*(x) = \sum_{i=1}^{k} \lambda_i \sigma\left(\langle v_i, x \rangle\right) \tag{7}$$

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.

A.2 Computing the population gradient in a general setting

Because σ admits a hermite expansion, for $v, \hat{v} \in \mathbb{S}^{d-1}$ we can evaluate expectations of the form $\mathbb{E}_x[\sigma(\langle v, x \rangle)\sigma(\langle \hat{v}, x \rangle)] = \sum_{p=0}^{\infty} \mu_p(\sigma)^2 \langle v, \hat{v} \rangle^p$. Then, we can compute the population loss and gradient as follows

Proposition 2 (Population Loss and gradient)**.** *We have the population loss*

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

and the population spherical gradient

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

where we define $h : \mathbb{R} \to \mathbb{R}$ to be

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

with

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

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

Remark 2 (Generalizing single index models)**.** *If we fix l^* and let $\xi = \bar{\xi}\sqrt{k}$, and sent $\bar{\xi} \to \infty$ the term*

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

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

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

Then, notice that this is the setting of single index models, where 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 this sense, our fine tuning model is a generalization of single index models.

Remark 3 (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 1 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.*

A.3 Intuition for SGD Dynamics and Sample Complexities

We will initally provide some intuition regarding the gradient dynamics, in terms of the function $h(m)$. Notably, in this setting, the behavior of the function h will determine the behavior of the dynamics. Now, recall the iteration for u_t:

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

We formally analyze the SGD dynamics in Appendix C, so for the sake of intuition, suppose we write the spherical projection error as $\hat{Q}_t$

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

Furthermore, decompose $\hat{\nabla} L(u_t; x_t) = \hat{\nabla}\Phi(u_t) + \hat{\nabla} E(u_t; x_t)$ where E_t is a stochastic error term with mean 0. Then, Let $m_t = \langle u_t, u \rangle$, and we get

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

where Q_t error due to ignoring the spherical projection. Then, unrolling the recursive expression and defining $E_t = \hat{\nabla} E(u_t; x_t)$, we obtain

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

Then, notice that the term $M_t = \eta \sum_{j=0}^{t-1} \langle E_j, u \rangle$ forms a martingale, and $\sum_{j=0}^{t-1} Q_j$ is a stochastic error term. In short time scales, these two error terms could potentially dominate the dynamics. However, in long time scales, the contribution of these terms scale 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, given the population gradient is non-vanishing. Then, notice that we can always keep $\eta T = \Theta(1)$ while letting $\eta\sqrt{T} = o(1)$. The exact choice of η, T depends crucially on the *signal to noise ratio* of the problem. In particular, if we have a lower bound S_k on the population gradient (signal), and an upper bound dV_k on the variance $\mathbb{E}_x[E_t^2]$ (noise), then we show the sample complexity scales with the inverse of the signal to noise ratio, which is $\frac{dV_k}{S_k^2}$. We analyze this precisely in Appendix C.

Nevertheless, even after ignoring the noise and assuming we have a population dynamics, it is not immediately clear from the form of h that the dynamics should converge to the ground truth (or its negation, due to the inherent symmetry in the problem). For the sake of intuition, consider the population dynamics, ignoring the spherical projection

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

If we rearrange, we can write $|1 - \overline{m}_{t+1}| = |1 - \overline{m}_t||1 - \eta h(\overline{m}_t)(1 + \overline{m}_t)|$. Then, if $h(\overline{m}_t)$ is non-vanishing throughout the dynamics, the population dynamics should quickly converge to 0 even if h can potentially be non-linear and exhibit complicated behavior. In particular, suppose $h(\overline{m}_t) \geq s$

throughout training, then we have $|1-\overline{m}_{t+1}| \le |1-\overline{m}_t||1-\eta s|$, in which case the population dynamics is greatly simplified. Furthermore, notice that $\overline{m}_t$ would converge to 1 only if $h(\overline{m}_t)$ is non-vanishing across the trajectory since this would lead to converging to a different stationary point. Hence, the main goal of the subsequent analysis is to prove that h indeed satisfies this lower bound property, and quantitatively determine what the lower bound is.

A.4 Results for Fine Tuning with Online SGD in Different Regimes

Note that we consider two kinds of randomness in our probabilistic bounds. There is the randomness due to the $c, \hat{c}$, and also due to the randomness of the training trajectory due to the data. Furthermore, we consider initializations that satisfy $m_0 \ge \frac{\beta}{\sqrt{d}}\operatorname{sign}(h(0))$. Note that the magnitude condition $|m_0| \ge \frac{\beta}{\sqrt{d}}$ 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. The magnitude assumption is standard in this type of analysis. For the sign condition, empirically, the results are not sensitive: However, handling both sign initializations requires knowing more about the structure of $h(m)$ and we defer it to future work. However, note that the sign condition holds with probability $1/2$, and if not, flipping the sign of u_0 will ensure that the sign condition holds.

A.4.1 Orthogonal Setting

In this section, we assume $\langle w_i, w_j\rangle = 0$ whenever $i \neq j$. Then, notice that h reduces in form to the following:

$$h(m) = 2k\left(\sum_i \lambda_i c_i\right)\left(\sum_i \lambda_i \hat{c}_i\right)\sum_{l \text{ even}}^{\infty} a_l m^l + \sum_{l \text{ odd}} \hat{a}_l m^l + k\left(\sum_i \lambda_i^2 c_i \hat{c}_i\right)\sum_{l \text{ even}} b_l m^l$$

where the $a_l, \hat{a}_l, b_l$ are all positive coefficients. Then, we are interested in the magnitudes of the random quantities in the above sum to characterize the behavior of h. We do this in the next appendix. Essentially, if the first hermite coefficient is non-zero, the term $(\sum_i \lambda_i c_i)(\sum_i \lambda_i \hat{c}_i)$ governs the lower bound for h. In the other case, we show the term $\sum_i \lambda_i^2 c_i \hat{c}_i$ governs the lower bound.

Theorem 5 (Orthogonal setting, $\xi = 1$)**.** *Let Assumption 2 hold, and* $0 < \varepsilon < 1$.

1. *For activations with $\mu_1(\sigma) \neq 0$, for a sufficiently small $C_\delta = \Theta(1)$, let $\delta = \frac{C_\delta \gamma \lambda_{\min}^2 \varepsilon^3}{(\log \lambda_{max}^4 dk^2)^2}$. Furthermore, let $\alpha = \frac{\log(\lambda_{max}^4 dk^2)}{\lambda_{\min}^2 \gamma\varepsilon\delta}$. Then, with probability $1 - o\left(\frac{\lambda_{\max}^2}{\lambda_{\min}^2}\right) - O(\gamma^{1/2})$ randomness of $c, \hat{c}$, for initializations satisfying $\langle u_0, u\rangle \cdot \operatorname{sign}(h(0)) \ge \frac{\beta}{\sqrt{d}}$, online SGD run with step size $\eta = \frac{\delta}{\lambda_{\max}^4 dk^2}$ and time $T = \lceil \alpha\lambda_{\max}^4 dk^2\rceil$ satisfies $\langle u_T, u\rangle^2 \ge 1-\varepsilon$ with high probability over the randomness of the data.*
2. *For activations with $\mu_1(\sigma) = 0$, for a sufficiently small $C_\delta = \Theta(1)$, let $\delta = \frac{C_\delta \gamma \lambda_{\min}^2 \varepsilon^3}{(\log \lambda_{max}^4 dk^2)^2\sqrt{k}}$. Furthermore, let $\alpha = \frac{\log(\lambda_{max}^4 dk^2)\sqrt{k}}{\lambda_{\min}^2 \gamma\varepsilon\delta}$. Then, with probability $1 - o\left(\frac{\lambda_{\max}^2}{\lambda_{\min}^2}\right) - O(\gamma^{1/2})$ randomness of $c, \hat{c}$, for initializations satisfying $\langle u_0, u\rangle \cdot \operatorname{sign}(h(0)) \ge \frac{\beta}{\sqrt{d}}$, online SGD run with step size $\eta = \frac{\lambda_{\max}^4 \delta}{dk^2}$ and time $T = \lceil \alpha\lambda_{\max}^4 dk^2\rceil$ satisfies $\langle u_T, u\rangle^2 \ge 1-\varepsilon$ with high probability over the randomness of the data.*

Proof. For the first point, notice that Lemma 1 and Lemma 2 imply that Assumption 7, Assumption 8 hold with

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

for some small γ with probability $1 - o\left(\frac{\lambda_{\max}^2}{\lambda_{\min}^2}\right) - O(\gamma^{1/2})$. Then, applying theorem 8 with the set S_k, V_k and ε, we get the desired result. The second case follows similarly. $\square$

Remark 4. *In the orthogonal setting with $\xi = 1$, when $\mu_1(\sigma) \neq 0$, we need $T = O(\frac{\lambda_{\max}^4}{\lambda_{\min}^4 \gamma^2} \cdot \frac{dk^3}{\varepsilon^4})$ iterations. Similarly, when $\mu_1(\sigma) = 0$, we need $\mu_1(\sigma) \neq 0$, we need $T = O(\frac{\lambda_{\max}^4}{\lambda_{\min}^4 \gamma^2} \cdot \frac{dk^3}{\varepsilon^4})$ iterations.*

Theorem 6 (Orthogonal setting, $\xi = \bar{\xi}\sqrt{k}$)**.** *Let Assumption 2 hold, and $0 < \varepsilon < 1$.*

1. *For activations with $\mu_1(\sigma) \neq 0$, for a sufficiently small $C_\delta = \Theta(1)$, let $\delta = \min\left\{\frac{C_\delta \bar{\xi}^2 k \gamma \lambda_{\min}^2 \varepsilon^3}{(\log \lambda_{max}^4 dk^2)^2}, 1\right\}$. Furthermore, let $\alpha = \frac{\log(\lambda_{max}^4 dk^2)}{\bar{\xi}^2 \lambda_{\min}^2 k \gamma \varepsilon \delta}$. Then, with probability $1 - o\left(\frac{\lambda_{\max}^2}{\lambda_{\min}^2}\right) - \exp\{-\frac{2}{e\bar{\xi}}\} - O(\gamma^{1/2})$ randomness of $c, \hat{c}$, for initializations satisfying $\langle u_0, u\rangle \cdot \mathrm{sign}(h(0)) \geq \frac{\beta}{\sqrt{d}}$, online SGD run with step size $\eta = \frac{\delta}{\bar{\xi}^2 \lambda_{\max}^4 dk^4}$ and time $T = \lceil \alpha \lambda_{\max}^4 \bar{\xi}^2 dk^4 \rceil$ satisfies $\langle u_T, u\rangle^2 \geq 1 - \varepsilon$ with high probability over the randomness of the data.*
2. *For activations with $\mu_1(\sigma) = 0$, for a sufficiently small $C_\delta = \Theta(1)$, let $\delta = \min\left\{\frac{C_\delta \bar{\xi}^2 \gamma \lambda_{\min}^2 \varepsilon^3 \sqrt{k}}{(\log \lambda_{max}^4 dk^2)^2}, 1\right\}$. Furthermore, let $\alpha = \frac{\log(\lambda_{max}^4 dk^2)}{\bar{\xi}^2 \lambda_{\min}^2 \gamma \varepsilon \delta \sqrt{k}}$. Then, with probability $1 - o\left(\frac{\lambda_{\max}^2}{\lambda_{\min}^2}\right) - \exp\{-\frac{2}{e\bar{\xi}}\} - O(\gamma^{1/2})$ randomness of $c, \hat{c}$, for initializations satisfying $\langle u_0, u\rangle \cdot \mathrm{sign}(h(0)) \geq \frac{\beta}{\sqrt{d}}$, online SGD run with step size $\eta = \frac{\delta}{\lambda_{\max}^4 \bar{\xi}^2 dk^4}$ and time $T = \lceil \alpha \lambda_{\max}^4 \bar{\xi}^2 dk^4 \rceil$ satisfies $\langle u_T, u\rangle^2 \geq 1 - \varepsilon$ with high probability over the randomness of the data..*

Proof. For $\bar{\xi} \leq 1$, the results in Lemma 1 and Lemma 2 imply that Assumption 7, Assumption 8 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 $1 - o(1) - \exp\{-\frac{2k}{e\bar{\xi}^2}\} - O(\gamma^{1/2})$. Then, applying theorem 8 with the set S_k, V_k and ε, we get the desired result. The second case follows similarly. □

Remark 5. *In the orthogonal setting with $\xi = \bar{\xi}\sqrt{k}$, when $\mu_1(\sigma) \neq 0$, we need $T = O(\frac{\lambda_{\max}^4}{\lambda_{\min}^4 \varepsilon^4 \gamma^2 \bar{\xi}^2} \cdot dk^3)$ iterations. Similarly, when $\mu_1(\sigma) = 0$, we need $\mu_1(\sigma) \neq 0$, we need $T = O(\frac{\lambda_{\max}^4}{\lambda_{\min}^4 \varepsilon^4 \bar{\xi}^2 \gamma^2} \cdot dk^4)$ iterations.*

A.4.2 Angularly Separated, Spectral Scaling Setting

Now, we do not necessarily assume the weights are angularly separated. However, we assume the features are not too correlated, so that weight vectors have angular separation $1 - \frac{\log k}{\sqrt{k}}$. Then, we have the following result for $\xi = 1$.

Theorem 7 (Separated setting, $\xi = 1$)**.** *Let Assumption 2 hold, and $0 < \varepsilon < 1$. For a sufficiently small $C_\delta = \Theta(1)$, let $\delta = \frac{C_\delta \gamma \lambda_{\min}^2 \varepsilon^3}{(\log \lambda_{max}^4 dk^2)^2 \sqrt{k}}$. Furthermore, let $\alpha = \frac{\log(\lambda_{max}^4 dk^2)\sqrt{k}}{\lambda_{\min}^2 \gamma \varepsilon \delta}$. Then, with probability $1 - o(1) - O(\gamma^{1/2})$ randomness of $c, \hat{c}$, for initializations satisfying $\langle u_0, u\rangle \cdot \mathrm{sign}(h(0)) \geq \frac{\beta}{\sqrt{d}}$, online SGD run with step size $\eta = \frac{\lambda_{\max}^4 \delta}{dk^2}$ and time $T = \lceil \alpha \lambda_{\max}^4 dk^2 \rceil$ satisfies $\langle u_T, u\rangle^2 \geq 1 - \varepsilon$.*

Proof. Note that Lemma 1 and Lemma 2 imply that Assumption 7, Assumption 8 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 8 with the set S_k, V_k and ε, we get the desired result. The second case follows similarly. □

Remark 6. *In the angularly separated and $\xi = 1$ case, online SGD strongly recovers the true parameter up to a sign with $T = \lceil \frac{\lambda_{\max}^4}{\lambda_{\min}^4 \gamma^4} \cdot \frac{dk^3}{\varepsilon^4} \rceil$ iterations.*

B Bounding Relevant Quantities to the SGD Dynamics

The goal of this appendix is to prove the following statements:

Lemma 1 (General Case Upper Bounds)**.** *Under Assumptions ???, we have the following:*

1. *Variance Upper Bound:* $\max\left\{\left\|\frac{\hat{\nabla} L(\hat{u};x)}{\sqrt{d}}\right\|^{2p}, |\langle \hat{\nabla} L(\hat{u};x), u\rangle|^{2p}\right\}^{1/p} \leq C_{p,\sigma}\lambda_{\max}^4 \frac{k^3\xi^2 \min\{k, 4\xi^2\}}{k+\xi^2}$
2. *Population Gradient Upper Bound:* $\left\|\hat{\nabla}\Phi(\hat{u})\right\| \leq C_\sigma \lambda_{\max}^2 \frac{k\xi^2}{1+\xi^2/k}$

Lemma 2 (Population gradient lower bounds)**.** *Under Assumptions 1 to 5 we have the following:*

1. *Orthonormal case, $\mu_1(\sigma) \neq 0$: With probability $1 - \exp\left\{-\frac{2k}{e\xi^2}\right\} - O(\gamma^{1/2}) - o\left(\frac{\lambda_{\max}^2}{\lambda_{\min}^2}\right)$, for $m \geq 0$, we have*

$$h(\mathrm{sign}(h(0))m)\mathrm{sign}(h(0)) \geq \frac{|h(0)|}{2} \geq \frac{\gamma\xi^2\mu_1(\sigma)^2}{1+\frac{\xi^2}{k}}$$

2. *Orthonormal case, $\mu_1(\sigma) = 0$: With probability $1 - \exp\left\{-\frac{2k}{e\xi^2}\right\} - O(\gamma^{1/2}) - o\left(\frac{\lambda_{\max}^2}{\lambda_{\min}^2}\right)$, for $m \geq 0$ we have*

$$h(\mathrm{sign}(h(0))m)\mathrm{sign}(h(0)) \geq \frac{|h(0)|}{2} \geq \frac{\gamma C_{s^*}\xi^2}{\left(1+\frac{\xi^2}{k}\right)^{s^*}\sqrt{k}}$$

   where s^* is the smallest s for which $\mu_s(\sigma) \neq 0$.

3. *Angularly Separated case, $\xi = 1$: With probability $1 - O(\gamma^{1/2}) - o\left(\frac{\lambda_{\max}^2}{\lambda_{\min}^2}\right)$, for $m \geq 0$ we have*

$$h(\mathrm{sign}(h(0))m)\mathrm{sign}(h(0)) \geq \frac{|h(0)|}{2} \geq \frac{\gamma}{\sqrt{k}}$$

Remark 7. *Our analysis naturally extends to the case when $\xi \neq 1$ but $\xi = \Theta(1)$, but for notational simplicity, we set $\xi = 1$.*

B.1 Upper bounds on the Variances of Gradients and the magnitude of population gradient

We state the following assumption we will use while bounding the variance of the gradients. The following assumption holds for many classes of activations including Lipschitz activations (e.g. ReLU, absolute value, sigmoid, tanh) and finite degree polynomial activations.

Proposition 3 (Moments of squared error)**.** *Let p be given, and Assumption 5 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 4 (Bound on the expected magnitude of $\hat{f}$)**.** *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

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

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,

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

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,

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

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 5 (Population Gradient Bounds)**.** *We have*

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

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

B.2 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 $s \ge 1$, $T(l,s)$ reduces to

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

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.

We initially bound the maximum possible contribution coming from the even l terms with $s = 0$.

Claim 1 (Even l, $s = 0$ contribution)**.** *With probability* $1 - \exp\{-\frac{2k}{e\xi^2}\}$*, the following holds.*

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

Proof. 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

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

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

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

Then, note that it suffices to show that, with high probability, we have

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

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

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

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

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

as desired. □

Proposition 6. *Let $\mu_1(\sigma)\neq 0$, then with probability $1-\exp\{-\frac{2k}{e\xi^2}\}-o(1)-O(\frac{\lambda_{\max}}{\lambda_{\min}}\gamma^{1/2})$, we have $h(\operatorname{sign}(h(0))m)\operatorname{sign}(h(0))\geq\frac{|h(0)|}{2}\geq\frac{\gamma\xi^2\mu_1(\sigma)^2}{1+\frac{\xi^2}{k}}$ for $m\geq 0$.*

Proof. WLOG assume $(\sum_i \lambda_i c_i)(\sum_i \lambda_i \hat{c}_i) > 0$. In this case, using Claim 1, with probability $1 - \exp\{-\frac{2k}{e\xi^2}\}$ we have

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

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 (Proposition 7), we have that the constant term is $\frac{\gamma \lambda_{\max}^2 \mu_1(\sigma)^2 \xi^2}{1+\xi^2/k}$ 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 $(\sum_i \lambda_i c_i)(\sum_i \lambda_i \hat{c}_i) > 0$. Then, we can bound the even terms by half of the constant term, and get the desired result. $\square$

Claim 2. *Let $\mu_1(\sigma) = 0$, then with probability $1 - o(1) - \exp\{-\frac{2k}{e\xi^2}\} - O(\gamma^{1/2})$, for $m \geq 0$ we have $h(\operatorname{sign}(h(0))m)\operatorname{sign}(h(0)) \geq |h(0)| \geq \frac{\gamma C_{s^*}\xi^2}{\left(1+\frac{\xi^2}{k}\right)^{s^*}\sqrt{k}}$ where s^* is the smallest s for which $\mu_s(\sigma) \neq 0$.*

Proof. Again, WLOG assume $\operatorname{sign}(h(0)) > 0$ so that $\langle c_\lambda, \hat{c}_\lambda\rangle > 0$. In this case, with probability $1 - \exp\{-\frac{2k}{e\xi^2}\}$ note that

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

where the b_l are non-negative coefficients. Then, note that $\xi^2\langle c_\lambda, \hat{c}_\lambda\rangle = |m||\langle c_\lambda, \hat{c}_\lambda\rangle|\xi^2$. Then, by anti-concentration (Proposition 7), 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 $h(\operatorname{sign}(h(0)m)\operatorname{sign}(h(0)) \geq |h(0)|$ 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 $\mu_s \neq 0$. $\square$

B.3 Angularly Separated Case: Population Gradient Lower Bounds

B.3.1 Computation of the population gradient

Note that specializing $\xi = 1$, we get

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

B.3.2 Bounding the higher order even terms

Initially, we aim to bound the even terms in the power series (i.e. $l > 1$).

Lemma 3. *Suppose Assumptions 1 to 4 hold. Then, with probability at least $1 - \frac{1}{k^3}$ over the randomization of $c, \hat{c}$, for $\varepsilon = \min\{\frac{\rho}{4}, 1 - \frac{1}{1+2\rho}\}$ we have*

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

Proof. 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 $O(k^{-\varepsilon})$ for some small $\varepsilon > 0$. 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,

$$\mathbb{E}\left[\sum_{i,j=1}^{k} \lambda_i \lambda_j c_i \hat{c}_j \langle w_i, w_j \rangle^s\right] = \sum_{i,j=1}^{k} \lambda_i \lambda_j \langle w_i, w_j \rangle^s \mathbb{E}[c_i \hat{c}_j] = 0$$

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

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 $\|f_s\|_2 \le \lambda_{\max}^2$. Then, by (O'Donnell, 2014, Theorem 9.23) we have

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

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

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

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

$$\begin{aligned}
&\left|\sum_{n=0}^\infty \left(\frac{1}{k}\right)^{2n+2}\sum_{s=0}^{s^*}\binom{2n+2+s}{2n+2}(2n+s+3)\mu_{2n+s+3}(\sigma)^2\left(\frac{k}{k+1}\right)^{2n+s+3}\left\langle \sum_{i=1}^k \lambda_i c_i w_i^{\otimes s}, \sum_{i=1}^k \lambda_i \hat{c}_i w_i^{\otimes s}\right\rangle\right| \\
&\quad \le \gamma\lambda_{\max}^2 \log k \sum_{n=0}^\infty \left(\frac{1}{k}\right)^{2n+2}\sum_{s=0}^{s^*}\binom{2n+2+s}{2n+2}(2n+s+3)\mu_{2n+s+3}(\sigma)^2\left(\frac{k}{k+1}\right)^{2n+s+3}
\end{aligned} \tag{8}$$

Now, it suffices to give a $O(k^{-\frac{1}{2}-c\varepsilon})$ bound for the infinite sum for $c > 1$. We will separate it into cases $s \le (s^*)^{1-\varepsilon}$ and $(s^*)^{1-\varepsilon} \le s \le s^*$. 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 $\binom{n}{k} \le \left(\frac{en}{k}\right)^k$ we have

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

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

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

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

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

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

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

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

Choosing $\varepsilon = 1 - \frac{1}{1+2\rho} > 0$ 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 (8) by $O(\lambda_{\max}^2 k^{-\frac{1}{2}-3\varepsilon})$ where $\varepsilon = 1 - \frac{1}{1+2\rho}$.

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

(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

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

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

$$\begin{aligned}
\sum_{l=0}^{\infty}\left(\frac{1}{k}\right)^l \sum_{s=0}^{\infty}\binom{l+s}{l}(l+s+1)\mu_{l+s+1}(\sigma)^2\left(\frac{k}{k+1}\right)^{l+s+1} &= \sum_{p=1}^{\infty}\left(\frac{k}{k+1}\right)^p p\mu_p(\sigma)^2 \sum_{l=0}^{p-1}\left(\frac{1}{k}\right)^l\binom{p-1}{l} \\
&= \sum_{p=1}^{\infty}\left(\frac{k}{k+1}\right)^p \left(1+\frac{1}{k}\right)^{p-1} p\mu_p(\sigma)^2 \\
&\leq \sum_{p=1}^{\infty} p\mu_p(\sigma)^2 = \tilde{C}_\sigma \qquad (9)
\end{aligned}$$

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

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

[4] $\hat{C}_\sigma$ depends on σ through the definition of ρ in Assumption 5.

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 (9). Combining all the bounds, for $\varepsilon = \min\{\frac{\rho}{4}, 1 - \frac{1}{1+2\rho}\}$, with probability at least $1 - \gamma \frac{1}{k^{\gamma/e - \frac{1}{2}}}$, we have

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

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

B.4 Anti-Concentration Inequalities 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. These functions capture the random behavior of the function h 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 for gaussian polynomial. To that end, we first state some known bounds from literature.

Lemma 4 (Carbery-Wright inequality (Carbery & Wright, 2001))**.** *Let Q be a normalized multilinear polynomial with degree d as in Definition 1. There exists a constant B such that for $g \sim \mathcal{N}(0, I_n)$ we have*

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

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

Now, notice that the random quantities that depend on $c, \hat{c}$ in the function h are all of this form. They are not normalized, but we can always normalize them by factoring out the ℓ_2 norm. Now, consider the following CLT-like result that we will use :

Lemma 5 (Invariance principle, (Mossel et al., 2005, Theorem 2.1))**.** *Let P be as in Definition 1. 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*

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

To be able to leverage these results, we need to quantify the influence of functions $x^\intercal Q y$ with Q being p.s.d. Intuitively, the only way the influence of a term can be non-vanishing is if one of the rows is too large relative to the frobenius norm. For a normalized psd matrix (i.e. $Q_{ii} = 1$), factorizing $Q_{ij} = \langle q_i, q_j \rangle$ we want to state that one q_i cannot be correlated to too many q_j (the row sum is large) if the q_j are not correlated within each other (the other row sums are small). Formally, we have the following:

Claim 3. *Let $\delta > 0$ and $M \triangleq \lceil \frac{2}{\delta^2} \rceil$. Furthermore, let $w_i \in \mathbb{R}^d$ be unit vectors for $i \in [M]$, for arbitrary d. Furthermore, let $\tilde{w} \in \mathbb{R}^d$ be a unit vector such that $|\langle \tilde{w}, w_i \rangle| \ge \delta$ for all $i \in [M]$. Then, for $\varepsilon = \frac{\delta^2}{2}$ we have $|\langle w_i, w_j \rangle| \ge \varepsilon$ for some $i \ne j \in [M]$*

Proof. We will prove by contradiction. Suppose for unit vectors w_i with $|\langle w_i, w_j \rangle| \le \varepsilon$ we have $|\langle \tilde{w}, w_i \rangle| \ge \delta$. Construct the matrix W whose columns are the w_i. Then,

$$\delta^2 M \le \sum_{i=1}^{k} \langle w_i, \tilde{w} \rangle^2 = \|W^\intercal \tilde{w}\|^2 \le \|W^T\|_{\mathrm{op}}^2 \le \lambda_{\max}(W^T W)$$

However, $W^T W$ is the gram matrix with all non-diagonals absolute value less than ε. By Gershgorin, the eigenvalues (and therefore the operator norm) is bounded by $1 + (M-1)\varepsilon$. Set $\varepsilon = \frac{1}{M}$ so that the RHS is strictly bounded by 2. Then, let $M = \lceil \frac{2}{\delta^2} \rceil$. Hence, we get a contradiction $2 \le \delta^2 M < 2$. $\square$

This is essentially saying that if $\tilde{w}$ has non-vanishing correlation with a set of vectors w_i, this set either cannot be too orthogonal or cannot be too large. Specifically, we fix the size of the set and lower bound the correlations. Then, consider the following claim that relates the max ℓ_2 norm of a row of a psd matrix to its frobenius norm.

Claim 4 (Influence of row of PSD matrix)**.** *Let $\delta > 0$ and $k > K(\delta) = O(1/\delta^9)$ be sufficiently large. Then, for any $Q \in \mathbb{R}^{k\times k}$ PSD matrix with $Q_{ii} = 1$. We have*

$$\frac{\max_i \sum_{j\in[k]} Q_{ij}^2}{\sum_{i,j=1}^{k} Q_{ij}^2} \le 2\delta$$

In particular, this implies that

$$\lim_{k\to\infty} \sup_{\substack{Q\in\mathbb{R}^{k\times k},\\ Q \text{ psd},\\ Q_{ii}=1}} \frac{\max_i \sum_{j\in[k]} Q_{ij}^2}{\sum_{i,j=1}^{k} Q_{ij}^2} = 0$$

at a rate of $\frac{1}{\sqrt[9]{k}}$

Proof. Fix some $\delta > 0$. Then, notice that because Q is psd, we can factor it as $Q_{ij} = \langle q_i, q_j \rangle$ where the q_i are unit norm since $\|q_i\|^2 = Q_{ii} = 1$. First, note that the denominator is at least k. Take the maximizing i in the numerator and let it be $\tilde{q} = q_i$, and define $S_k = \{j \in [k] : |\langle q_j, \tilde{q} \rangle| \ge \delta\}$. If we have $|S_k| \le \delta k$, then the contribution from the terms in S_k is at most δk. The contribution from the others is at most $\delta^2 k$ since these terms are less than δ^2. Hence, $\sum_{j\in[k]} Q_{ij}^2 \le \delta(1+\delta)k \le 2\delta k$. Then,

$$\frac{\sum_{j\in[k]} Q_{ij}^2}{\sum_{i,j=1}^k Q_{ij}^2} \le 2\delta = O(\delta)$$

Now, suppose $|S_k| > \delta k$. Then, let $M \triangleq \lceil \frac{2}{\delta^2} + 1 \rceil$ as defined in Claim 3 and let $\varepsilon \triangleq \varepsilon_\delta$ be the constant from the claim. Then, notice that any subset of S_k with size more than M must contain two distinct vectors with correlation at least ε.

Then, consider the following process. For all the remaining vectors, we create a maximal set of vectors that are almost orthogonal (i.e. with correlation at most ε). By definition of maximality, all the remaining vectors should have correlation at least ε with some vector in this subset.

Formally, for $i \ge 1$, initialize a set $S_{k,i}$ (we set $S_{k,0} = \emptyset$) by taking a maximal set of vectors from $S_k \setminus \bigcup_{j<i} S_{k,j}$ such that for all distinct pairs $j \neq j \in S_{k,j}$ we have $|\langle q_j, q_l \rangle| < \varepsilon$. That is, we construct a set such that vectors in the set are almost orthogonal, and we cannot add any more vectors to this subset. Once we cannot add any more vectors, remove these vectors from the set and move to $i+1$.

Continue this process until termination (which must happen since we can add at least 1 element every round) and by Claim 3, we must have $|S_{k,j}| \le M$. This means, there will be at least $\frac{\delta k}{M} = \Omega(k)$ of these subsets. Now, consider $i < j$ and some $v_j \in S_{k,j}$. By construction, v_j was not added to $S_{k,i}$ so it must be the case that $|\langle v_i, v_j \rangle| \ge \varepsilon$ for some $v_i \in S_i$. Furthermore, notice that each set is disjoint. So, if we take all the pairs (i,j) with $i < j$ and pairs of vectors $|\langle v_i, v_j \rangle| \ge \varepsilon$, we have

$$\sum_{i<j} |\langle v_i, v_j \rangle|^2 \ge \varepsilon^2 \frac{\delta^2 k^2}{4M^2}$$

where all pairs (i,j) are disjoint. Then, we have

$$\frac{\sum_{j\in[k]} Q_{ij}^2}{\sum_{i,j=1}^k Q_{ij}^2} \le \frac{k}{\varepsilon^2 \frac{\delta^2 k^2}{4M^2}} \le \frac{64}{\delta^8 k}$$

for $k \ge \frac{\delta^9}{32}$ we have that the above is less than 2δ. The limit statement follows immediately by the definition of limit and the uniformity of all the bounds. $\square$

Corollary 1. *Let $0 < q_{\min}^2 \le q_{\max}^2$ be absolute constants such that for all k, we have $q_{\min}^2 \le Q_{ii} \le q_{\max}^2$. Then, we have*

$$\lim_{k\to\infty} \sup_{\substack{Q\in\mathbb{R}^{k\times k}, \\ Q \text{ psd}, \\ Q_{ii}=1}} \frac{\max_i \sum_{j\in[k]} Q_{ij}^2}{\sum_{i,j=1}^k Q_{ij}^2} = 0$$

Proof. In the proof of the previous claim, we have $q_{\min} \le \|q_i\| \le q_{\max}$. Define normalized vectors $\tilde{q}_i = \frac{q_i}{\|q_i\|}$. Notice that this means we can upper bound $Q_{ij}^2 \le q_{\max}^2 \langle \tilde{q}_i, \tilde{q}_j \rangle^2$ and similarly $Q_{ij}^2 \ge q_{\min}^2 \langle \tilde{q}_j, \tilde{q}_j \rangle^2$. Hence,

$$\frac{\max_{i\in[k]} \sum_{j\in[k]} Q_{ij}^2}{\sum_{i,j=1}^k Q_{ij}^2} \le \frac{q_{\max}^2}{q_{\min}^2} \frac{\max_{i\in[k]} \sum_{j\in[k]} \tilde{Q}_{ij}^2}{\sum_{i,j=1}^k \tilde{Q}_{ij}^2}$$

where now $\tilde{Q}_{ii} = 1$ is a psd matrix. Applying the result of Claim 4, we get the desired result. $\square$

Now, we will use the above results to prove the following fact:

Lemma 6 (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 $Q_{ii} = 1$. Then,*

$$\sup_Q \Pr_{x,y\sim\mathrm{Unif}\{\pm1\}^k} [|x^\intercal Q y| \le \varepsilon \|Q\|_F] \le o(1) + O(\varepsilon^{1/2})$$

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 Claim 4, so that $\tau = o(1)$. Hence, by the invariance principle (Lemma 5), for any Q, we have

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

However, applying Carbery-Wright inequality for the anti-concentration of gaussian polynomials (Lemma 4), we get the desired result. ☐

Corollary 2 (Anti-Concentration of Balanced P.S.D. Quadratics on the Hypercube)**.** *The result above holds when Q_{ii} are not-necessarily equal, but there exists $q_{\min}, q_{\max}$ such that $q_{\min}^2 \le Q_{ii} \le q_{\max}^2$, and we replace $o(1)$ with $o(\frac{q_{\max}^2}{q_{\min}^2})$.*

Proof. Proof follows exactly the same, except by using the influence of a row for balanced psd matrices. ☐

B.4.1 Relating to quantities that arise in h

Claim 5 (Constant term variance, spectral setting)**.** *Let $f : \{-1,1\}^{2k} \to \mathbb{R}$ be such that*

$$\begin{aligned} f(b,\hat{b}) &= \sum_{i,j}^{k} b_i \hat{b}_j \left(\frac{\lambda_i\lambda_j}{k} \sum_{s=0}^{\infty} (s+1)\mu_{s+1}(\sigma)^2 \left(\frac{k}{k+1}\right)^{s+1} \langle w_i, w_j\rangle^s \right) \\ &\triangleq \sum_{i,j=1}^{k} b_i\hat{b}_j Q_{ij} \end{aligned} \tag{10}$$

Then, we have $\Omega(\lambda_{\min}^2) \le \|f\|_2 \le O(\lambda_{\max}^2)$

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

Lemma 7. *Let f be of the form in Equation* (10)*. Then,*

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

where $\tau = o(1)$ and $b, \hat{b}$ are independent uniform draws from $\{-1,1\}^k$.

Proof. Note that entrywise powers of psd matrices are psd, so $(W^T W)^{\odot s}$ is psd. Notice that $Q_{ij} = (\lambda_i\lambda_j)\left(\sum_{s=0}^{\infty}(s+1)\mu_{s+1}(\sigma)^2 \left(\frac{k}{k+1}\right)^{s+1} \langle w_i, w_j\rangle^s\right)$ which is a psd matrix since it is the sum of psd matrices (for s). This is due to the fact

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

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, $q_{\max}/q_{\min} = \frac{\lambda_{\max}}{\lambda_{\min}}$ The proof follows immediately once we normalize as $\frac{f}{\|f\|_2}$ and apply the above results. $\square$

Proposition 7 (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*

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

and

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

Proof. For the first one let $Q = \frac{1}{k}\lambda\lambda^{\intercal}$. and for the second one let $Q = \frac{1}{k}I(\lambda \odot \lambda)$. Both are balanced psd matrices, and the anti concentration result lemma 6 holds. Then, the results follow. $\square$

C Finite Sample Dynamics Analysis

We start with starting the generic assumptions we will work with in this section that are satisfied with the various models we consider.

C.1 Assumptions that capture various regimes in Online SGD

We analyze the finite sample gradient dynamics under the following assumptions:

Assumption 6 (Unbiased Gradient Estimates)**.** *For all $\hat{u}$, the sample gradient is an unbiased estimate of the population gradient. I.e. we have*

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

This assumption is standard in the literature. Note that this assumption holds when σ is almost everywhere differentiable (w.r.t. gaussian measure), and σ' has almost linear polynomial growth. This is because $\nabla_{\hat{u}}L(\hat{u};x)$ has at most linear polynomial growth, so 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 theorem.

Assumption 7 (Magnitudes of variances)**.** *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 $u, \hat{u}$,* $\max\left\{\frac{\mathbb{E}_x\left\|\hat{\nabla}_u L(\hat{u};x)\right\|_2^{2p}}{d^p}, \mathbb{E}_x\langle \hat{\nabla}_{\hat{u}}L(\hat{u};x), u\rangle^{2p}\right\}^{1/p} \leq \mu_p V_k$
2. ***Population gradient bound**: For all $\hat{u}$,* $\left\|\hat{\nabla}_{\hat{u}}\Phi(\hat{u})\right\|^2 \leq V_k$.

where the μ_p may depend on p and the activation, but on nothing else.

We will consider this assumption only for a few p that will be tuned during the proofs, so the moment bounds only have to hold up to a certain p.

Assumption 8 (Population Gradient Lower Bound)**.** *The population gradient is of the form $\hat{\nabla}_{\hat{u}}\Phi(\hat{u}) = -h(\langle \hat{u}, u\rangle)(u - \hat{u}\langle u, \hat{u}\rangle)$. Furthermore, there exists a constant $\max\{S_k, S_k^2\} \leq V_k$ that has at most polynomial decay, such that h satisfies the following:*

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

Theorem 8. *Let Assumptions 6 to 8 hold. Let* $0 < \varepsilon < 1$. *Let* $m_t = \langle u_t, u \rangle$ *and set the learning rate* $\eta = \frac{\delta}{dV_k}$ *with scaling* $\delta = \min\left\{\frac{S_k \varepsilon^3}{4\mu_1 (\log dV_k)^2}, 1\right\}$, *for total time* $T = \lceil \alpha dV_k \rceil$ *with time scaling* $\alpha = \frac{4(\log dV_k)}{\varepsilon \delta S_k}$ *and initialization at* $|m_0| \geq \frac{\beta}{\sqrt{d}}$ *with* $m_0 h(0) > 0$. *Under Assumptions 1-4, with probability at least* $1 - o(1)$ *the following holds for* $T = \lceil \alpha dV_k \rceil$ *and* $T_{weak} = \lceil \frac{4dV_k}{\delta S_k} \rceil = o(T)$.

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

The proof of this theorem is constructed throughout this section, and concluded at the end of the section.

C.2 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.

C.2.1 Bounding spherical projection error

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

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

Then, notice that the total contribution of these terms up to time t can be written as

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

First, notice that η^3 gives a $\frac{\delta^3}{d^3 V_k^3}$ scaling, but $\|L_t\|^2 |\langle L_t, u \rangle|$ 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 $\alpha \delta^2 \leq \varepsilon$.

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

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

Similarly, we have

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

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

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

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

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

Hence, using the moment bounds (Assumption 7) on $\|L_t\|^2$ and $|\langle L_t,u\rangle|^2$, for $p=2,1$ respectively, we have

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

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

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

Setting $\gamma=\frac{\beta}{10\sqrt{d}}$ gives us the first result. For the second, we can use $\alpha\delta^2\leq\varepsilon$ 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 8 (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 Assumption 7. 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 Assumption 7, 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 $\mathbb{E}[Y_t \mathbb{1}\{Y_t \le T^\nu\}] \le \mathbb{E}[Y_t] = O(1)$, and we have the almost sure bound

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

and the conditional variances

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

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 $\alpha\delta^2 \le \frac{\varepsilon^2}{\log d}$. Then*

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

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

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

□

C.3 Controlling the error martingale

Claim 9. *Let $\alpha\delta^2 \le \varepsilon^2(\log d)^{-1}$. 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*

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

Furthermore, we have

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

Proof. The fact that M_t is a martingale follows directly from Assumption 6 and the fact that each x_t is a fresh sample. By Doob's maximal inequality for martingales, we have

$$\begin{aligned}\Pr\left[\sup_{0\le t\le T}|M_t|>\gamma\right] &\le \frac{\mathbb{E}M_T^2}{\gamma^2}\\ &\le \frac{2\mu_1\eta^2 TV_k}{\gamma^2}=\frac{2\mu_1\alpha\delta^2}{d\gamma^2}\end{aligned}$$

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

C.4 Weak Recovery & 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.

C.4.1 Defining an Event for the Error Bounds and initial correlation

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

$$\mathcal{A}=\{m_0\ge\frac{\beta\cdot\operatorname{sign}(h(0))}{\sqrt{d}}\} \tag{11}$$

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

$$\mathcal{B}=\left\{\sup_{0\le t\le T}|M_t|\le\min\left\{\frac{\beta}{10\sqrt{d}},\frac{\varepsilon}{36}\right\}\right\}\cap\left\{\sup_{0\le t\le T}\eta^3\sum_{j=0}^{t-1}\|L_j\|^2|\langle L_j,u\rangle|\le\min\left\{\frac{\beta}{10\sqrt{d}},\frac{\varepsilon}{18}\right\}\right\} \tag{12}$$

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

Proposition 8. *Let* $\delta=\frac{\varepsilon^3 S_k}{4C_\delta\log(dV_k)}$ *where* $C_\delta>\max\{1,\mu_1\}$. *Furthermore suppose that* $\alpha=\frac{4(\log dV_k)}{\varepsilon\delta S_k}$. *Then, for* $T=\lceil\alpha dV_k\rceil$, *we have* $\Pr(\mathcal{B}(\varepsilon,d,\beta,k,T))=1-O\left(\max\left\{\frac{1}{\beta\sqrt{d}},\alpha(dV_k)^{-\frac{\beta^2}{C}(\log dV_k)+1},\frac{\varepsilon^2}{\beta^2\log d},\frac{1}{d\log d}\right\}\right)=1-o(1)$.

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

C.4.2 Defining Stopping Times for the dynamics

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

$$\begin{aligned}\tau_q^+&=\inf\{t\ge0:m_t\ge q\}\\ \tau_q^-&=\inf\{t\ge0:m_t\le q\}\end{aligned}$$

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

$$\begin{aligned}\tau_r^+&=\inf\{t\ge0:m_t>r\}\\ \tau_0^-&=\inf\{t\ge0:m_t<0\}\\ \tau_{1-\varepsilon/6}^+&=\inf\{t\ge0:m_t\ge1-\frac{\varepsilon}{6}\}\end{aligned}$$

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, $\tau_{1-\varepsilon/6}^+$ is used to analyze the stage before we achieve the initial strong correlation, we will show m_t will stay above $1 - \varepsilon$ after $t > \tau_{1-\varepsilon/6}^+$. I.e. the progress made for strong recovery is not eliminated by the noisy gradients.

C.4.3 Analyzing the Dynamics conditioning on $\mathcal{B}$

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

Lemma 9 (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 $\mathrm{sign}(h(0)) = 1$. Then, for all $t \leq \tau_0^-$, we must have $m_t \geq S_k$. Furthermore, for all $t \leq \tau_r^+$, we have $1 - m_t^2 > \frac{1}{2}$. Then, applying the inequalities in $\mathcal{B}$, 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 $\tau_0^- > \tau_r^+ \wedge T$. $\square$

Lemma 10 (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*

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

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

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

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

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

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

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 $\left|1 - \frac{1}{r_j}\right| \le \eta^3 \|L_j\|^2 |\langle L_j, u\rangle| + \eta^2 \|L_j\|^2$ gives us

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

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

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

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

Now, we are in a position to prove Theorem 8.

Proof of Theorem 8. First, notice that due to assumption 8 and the initalization requirement in the theorem, $\mathcal{A}$ holds. Then, per Proposition 8, $\mathcal{B}$ holds with probability $1 - o(1)$. Then, conditioning in $\mathcal{B}$, per Lemma 9 and Lemma 10, we can drop the requirement that $t \le \tau_0^-$. So, let $t \le T \wedge \tau_r^+$. Conditioning on $\mathcal{B}$, per Lemma 9, we have

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

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 $\tau_r^+ \wedge T \le T_{\text{weak}}$. Then, it suffices to show $T_{\text{weak}} \le T$. Notice that $T_{weak} = \lceil \frac{2dV_k}{\delta S_k} \rceil$ and $T = \lceil \alpha d V_k \rceil = \lceil \frac{4(\log dV_k)}{\varepsilon \delta S_k} \rceil > T_{\text{weak}}$ when $\varepsilon < 1, V_k > 1$ and $d > 3$. Then, we conclude $\tau_r^+ \le T_{\text{weak}} \le T$.

Now, conditioning on τ_r^+, for all $t \ge \tau_r^+$, with $t \le T$ per Lemma 10, we have

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

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

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

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

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

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

D Example Constructions Mentioned in the Main Text

D.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 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, $\|c\| = \|c'\|$.

D.2 Example of a base network whose perturbation requires many samples to learn from scratch

We are looking for an example where the target model is hard to learn from scratch but fine tuning is easy. Since the activations are hermite, it suffices to give an example of a target function that has orthonomal weights. Then, 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. Then, consider the following, 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 10. *When $d > 1 + \frac{k(k+1)}{2}$, for $\lambda_i = 1$, there exists unit norm weights $\{w_i\}_{i=1}^k$, a perturbation $u \perp \operatorname{span}(w_i)$, 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 $l \in [d], l \leq k$, 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 $i \neq j$ a coordinate, and we will have at least 1 coordinate left. Then, notice that the inner product $\langle w_i, w_j \rangle$ for $i \neq j$ 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 $i \neq j$, as desired. □

Proposition 9. *Let $\xi = 1$, and consider the example in Claim 10. 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 $h(m)\text{sign}(h(0)) \geq \frac{|h(0)|}{2}$.

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 $|\langle w_i, w_j \rangle| \leq \frac{1}{k}$, we have

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

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 (Lemma 6). Then, with high probability $h(m)\text{sign}h(0) \geq \frac{|h(0)|}{2}$ uniformly. □

Hence, in the construction given in Claim 10, even though the c_i's are non-random, we still have with high probability over the randomization of $\hat{c}$ that h satisfies Assumption 8. Then, we have the following

Theorem 9. *Fine tuning on Claim 10, learns the teacher network perturbation u in $O(\frac{dk^2}{\varepsilon^4})$ samples, whereas training from scratch using any CSQ algorithm requires at least $O(d^{p/2})$ queries or $\tau = O(d^{-d/4})$ tolerance.*

Proof. The first part follows directly from the fact that h satisfies the gradient lower bound in Assumption 8 with a $\Omega(\frac{\lambda_{\min}^2}{\sqrt{k}})$ lower bound, and Theorem 8. 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 succes, one either needs $n = \Theta(d^{p/2})$ queries or tolerance $\tau = \Theta(d^{-p/4})$. $\square$

D.3 Second Layer Training

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

$$\mathcal{L}_\lambda = \left\{ \sum_{i=1}^{k} \lambda_{i,1} \sigma \left(\left\langle \frac{w_i + \frac{\xi}{\sqrt{k}} \hat{u}}{\sqrt{1 + \xi^2/k}}, x \right\rangle \right) + \lambda_{i,2} \sigma \left(\left\langle \frac{w_i - \frac{\xi}{\sqrt{k}} \hat{u}}{\sqrt{1 + \xi^2/k}}, x \right\rangle \right) : \lambda \in \mathbb{R}^k \times \mathbb{R}^k \right\} \quad (13)$$

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 10 (Learning u is sufficient to learn f^*)**.** *Suppose $\hat{u}$ is such that $1 - |\langle u, \hat{u} \rangle| \leq \varepsilon \cdot \frac{k+\xi^2}{2C_\sigma \lambda_{\max}^2 \xi^2 k^2}$ which is $\Theta(\varepsilon/k)$ for $\xi = \Theta(1)$ and $\Theta(\varepsilon/k^2)$ for $\xi = \Theta(\sqrt{k})$ Then, there exists a model $h \in \mathcal{L}_\lambda$ as defined in Equation* (13) *such that $\mathbb{E}_x(f^*(x) - h(x))^2 \leq \varepsilon$. 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. (13)) 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 $\mathbb{E}_x(f^*(x) - \hat{f}(x))^2 \leq \varepsilon$. 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 $\frac{\varepsilon}{\lambda_{\max}^2 k^2}$. 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 $1 - \langle u, \hat{u} \rangle \leq \varepsilon \cdot \frac{k+\xi^2}{2C_\sigma \lambda_{\max}^2 \xi^2 k^2}$ as desired. $\square$

Remark 8. *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).*