A Numerical investigation

In this section, we provide many plots corroborating the theoretical results presented in the main manuscript. We also give information on the numerical procedures needed to reproduce the different figures, along with the details behind the numerical experiments.

The code will be available on GitHub after de-anonymization.

Description of the training algorithm: First, we describe the training protocol reported in Alg. 1: we separately update the first layer with $T-$GD steps of learning rate η, followed by training with standard ridge regression for the second layer with fixed regularization strength λ. We vary adaptively the learning rate to satisfy the hypothesis of Thm. 2, i.e. $\eta = \mathcal{O}(p\sqrt{\frac{n}{d}})$, the regularization parameter is fixed to $\lambda = 1$, and we take noiseless labels. We average over different seeds to get the mean performance, and we use standard deviation for giving confidence intervals.

A.1 Learning with a single giant step

A sizable part of our results concerns the feature learning efficiency of two layer neural networks after one giant step of GD. We provide a toy illustration of the phenomenology in Fig. 1, and rigorously characterize this in Theorems 1 and 2. Moreover, in section 2.3 we provide a plethora of results analyzing the consequences of the theorems above in the actual learning performance of the network. Here, we perform a detailed numerical investigation of the different claims in these results.

Investigating the generalization performance: We start by analyzing the generalization performance of different networks after one giant GD step in Fig. 3. We compare the generalization performances of linear and quadratic kernel methods - horizontal lines marked by different colors computed at $n = n_{max} \sim d^{2.25}$ - with three networks: a) *random features* (red points) random network with fixed weight matrix W_0 at initialization; b) *1 GD step* (green points) two-layer network trained using one step in the protocol of Alg. 1; c) *1 GD step with preprocessing* (blue points) two-layer network trained using a preprocessing step in Alg. 1. The introduction of a preprocessed algorithm is linked with the theoretical results of Theorem 2 and we provide a detailed analysis in the next paragraph.

The importance of preprocessing: As Theorem. 2 provably states, it is not possible to get fully specialized hidden units with one giant step of GD in the $n = \mathcal{O}(d^k)$ regime (with $k > 1$), unless the directions associated to teacher Hermite coefficients lower than k are suppressed, or equivalently, if the leap index ℓ (Def. 1) of the target is equal to k - see Fig. 1 for an illustration. We can circumvent this issue by using a preprocessing step. Given a batch of size $n = \mathcal{O}(d^k)$, we preprocess the labels in Alg. 1 using a method introduced in Damian et al. (2022) for the case $k = 1$:

$$\hat{c}_{j_1,\dots,j_d} \leftarrow \frac{1}{n}\sum_{\nu=1}^{n} y_\nu \,\mathrm{He}_{j_1}(\langle \boldsymbol{e}_1, \boldsymbol{z}_\nu\rangle)\cdots \mathrm{He}_{j_d}(\langle \boldsymbol{e}_d, \boldsymbol{z}_\nu\rangle) \tag{17}$$

$$y_\nu \leftarrow y_\nu - \sum_{j_1,\dots,j_d : j_1+\dots+j_d<k} \frac{\hat{c}_{j_1,\dots,j_d}}{j_1!\cdots j_d!}\mathrm{He}_{j_1}(\langle \boldsymbol{e}_1, \boldsymbol{z}_\nu\rangle)\cdots \mathrm{He}_{j_d}(\langle \boldsymbol{e}_d, \boldsymbol{z}_\nu\rangle) \tag{18}$$

where we denoted with $\hat{c}_{j_1,\cdots,j_d}$ the plug-in estimates from data of the teacher Hermite coefficients, and with $\{\boldsymbol{e}_i\}_{i\in[d]}$ the canonical basis in $\mathbb{R}^d$. By standard concentration arguments Gotze et al. (2019) the plug-in estimation of the coefficients is accurate only in the $n = \omega(d\,\mathrm{polylog}(d))$ regime. Indeed, in Fig. 3 the inefficient estimation of eq. (17) in the $n = o(d)$ sample regime generates a noisy learning curve for the preprocessed algorithm (blue points). The ridge estimator $\hat{\boldsymbol{a}}$ is consequently found by training on the processed labels defined in eq. (18) and the suppressed part is injected back in the predictor only at test

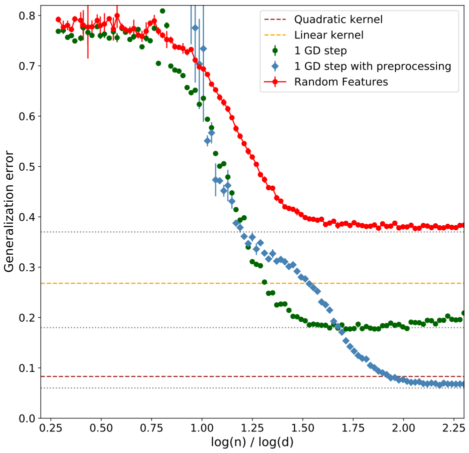

Figure 3: **Learning with training of the second layer.** Simulation illustrating the different regimes in Fig. 1, using $d=512, p=1024$, a symmetric two-index target function $f^\star(\boldsymbol{z}) = \sigma^\star(\langle \boldsymbol{w}_1^\star, \boldsymbol{z}\rangle) + \sigma^\star(\langle \boldsymbol{w}_2^\star, \boldsymbol{z}\rangle)$ with activation $\sigma^\star(z) = \mathrm{He}_1(x) + \mathrm{He}_2(x)/2! + \mathrm{He}_4(x)/4!$, and a relu student. (a) The first algorithm (green) applies a giant step and then learns the second layer. When $n \gg d$, its performance goes beyond the linear predictor that would be obtained with a kernel method and reach the "linear subspace learning" regime in Fig. 1. (b) To go beyond this regime, the second algorithm (blue) preprocesses the data to remove a plug-in estimate of the first Hermite coefficient. It reaches a lower plateau as $n \approx d^2$, now beating the quadratic kernel. We contrast this behavior with the one of the random feature model (red).

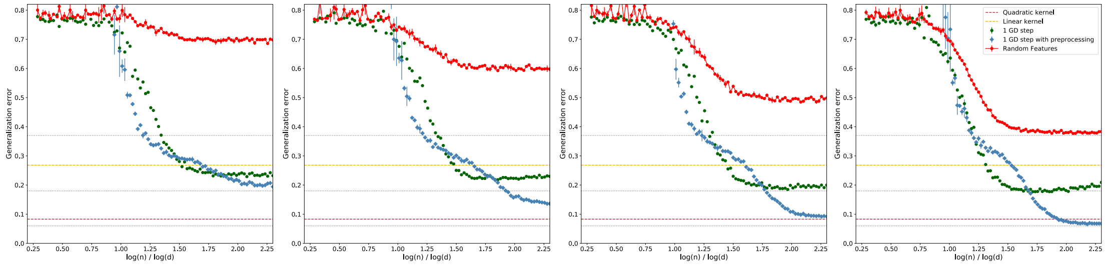

Figure 4: **Learning as a function of the number of hidden neurons.** Simulations illustrating the influence of the number of hidden neurons in Fig. 3. We change the value of $p \in (128, 256, 512, 1024)$ from left to right.

time:

$$\hat{f}(\boldsymbol{z}_\nu) = \frac{1}{\sqrt{p}}\hat{\boldsymbol{a}}^\top \sigma(W\boldsymbol{z}_\nu) + \sum_{j_1,\cdots,j_d:j_1+\cdots+j_d<k} \frac{\hat{c}_{j_1,\ldots,j_d}}{j_1!\cdots j_d!} \mathrm{He}_{j_1}(\langle \boldsymbol{e}_1, \boldsymbol{z}_\nu\rangle)\cdots \mathrm{He}_{j_d}(\langle \boldsymbol{e}_d, \boldsymbol{z}_\nu\rangle) \tag{19}$$

Comparison of different methods: The results presented in Fig. 3 clearly illustrate the theoretical predictions of Thm. 2: in the $n = \mathcal{O}(d)$ regime vanilla Alg. 1 attains the "linear subspace learning" regime (see Fig. 1) and beats the linear kernel, while the preprocessed version cannot. However, implementing preprocessing turns out definitely beneficial in the $n = \mathcal{O}(d^2)$ region. Indeed, while the vanilla Alg. 1 remains stuck on the linear subspace learning plateau, the preprocessed Alg. 1 reaches a lower test error than the quadratic kernel. This is achieved by effectively raising the leap index of the target function. More precisely, given a target with leap index $\ell = 1$ as in Fig. 3, the manipulation in eq. (18) aims exactly at the removal of the first Hermite coefficient of the target by estimating it from the data, allowing feature learning in the $n = \mathcal{O}(d^2)$ regime in accordance with Thm. 2. We complement the above picture by analyzing the influence of the number of hidden neurons p on the generalization performance in Fig. 4: by increasing the expressive power of the network, we attain the single-index regime by using a single giant step of Alg. 1 (in accordance with Conj. 1). Moreover, we note that it is necessary to use $p = 2d$ in order to be able to beat the performance of the quadratic kernel in this learning task (rightmost section).

Investigating representation learning efficiency: We move to an additional numerical investigation of feature learning efficiency, as characterized by Theorems 1 and 2. In Fig. 5 we again consider a single

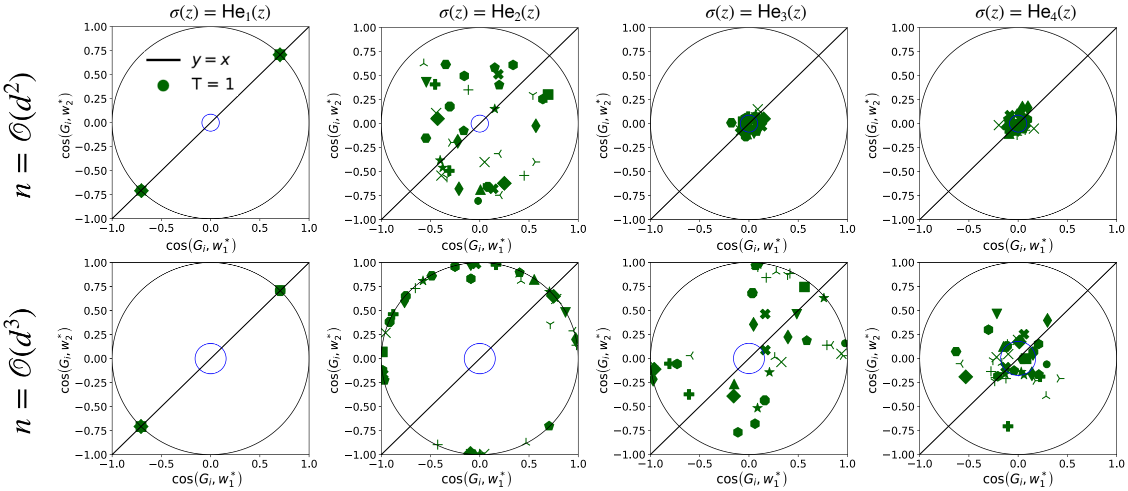

Figure 5: **Feature learning after a single step.** Specialization of hidden units in the $n = \mathcal{O}(d^k)$ regime ($k = 2, 3$). The plots show the cosine similarity of the gradient with respect to the target vectors $(\boldsymbol{w}_1^\star, \boldsymbol{w}_2^\star)$ for $p = 40$ different neurons, identified by different markers. The bisectrix of the first quadrant is shown as a continuous black line, the circle of unitary radius in black, and the circle of radius $2/\sqrt{d}$ in blue. In the upper panel, $(n, d) = (2^{18}, 2^9)$, and $(n, d) = (2^{21}, 2^7)$ in the lower one.
We use a 2-index target $f^\star(\boldsymbol{z}) = \sigma^\star(\langle \boldsymbol{w}_1^\star, \boldsymbol{z}\rangle) + \sigma^\star(\langle \boldsymbol{w}_2^\star, \boldsymbol{z}\rangle)$, with matching student: $\sigma(z) = \sigma^\star(z)$. **Left:** $\sigma(z) = \mathrm{He}_1(z)$. **Center-Left:** $\sigma(z) = \mathrm{He}_2(z)$. **Center-Right:** $\sigma(z) = \mathrm{He}_3(z)$. **Right:** $\sigma(z) = \mathrm{He}_4(z)$. We observe that if the leap index $\ell = 1$, we only learn a single direction, no matter the data quantity, while for $\ell > 1$ we learn every direction as soon as we reach $n = \mathcal{O}(d^\ell)$. The small spread observed for $\sigma(z) = \mathrm{He}_4(z)$ and $n = \mathcal{O}(d^3)$ is due to the small value of d used for the experiments.

step of Alg. 1, focusing now on the analysis of the gradient. We compute the gradient matrix $G \in \mathbb{R}^{p\times d}$ and plot the cosine similarities of all the rows $\{\boldsymbol{G}_i \in \mathbb{R}^d\}_{i=1}^p$ with the teacher vectors $(\boldsymbol{w}_1^\star, \boldsymbol{w}_2^\star)$. The figure clearly illustrates the claims of Thm. 2: in the $n = \mathcal{O}(d^k)$ (with $k > 1$) regime is necessary to analyze targets with leap index k in order to obtain specialized hidden units. Moreover, the leftmost section of Fig. 5 completes the picture offered by Figs. 3&4 about the lack of specialization in presence of teacher functions with non-zero first Hermite coefficient ($\ell = 1$): the gradient is stuck in the linear subspace learning regime theoretically predicted by Thm. 2, regardless of the sample regime considered, preventing feature learning.

A.2 Learning with multiple steps

We move now the numerical investigation of the learning behavior after multiple gradient steps. The general picture of the phenomenology is offered in Fig. 2, following the theoretical characterization of Theorem 3.

Investigating the generalization performance: First, we investigate the generalization behavior in the upper panel of Fig. 6. We modify slightly the training procedure in Alg. 1 to perform the numerical experiments: at every gradient step on the first layer weights we train the second layer sequentially with ridge regression. The analysis of the test error behavior in the upper panel of Fig. 6 sheds light on the consequences of Thm. 3 on the generalization performance of two-layer networks. Indeed, we observe a clear benefit in performing multiple gradient steps if the teacher function has a direction linearly connected to the rank-one spike in the gradient identified by $C_1(f^\star)$ (right panel), while if such linearly connected direction does not exist (left panel) the generalization performance does not improve relevantly over time,

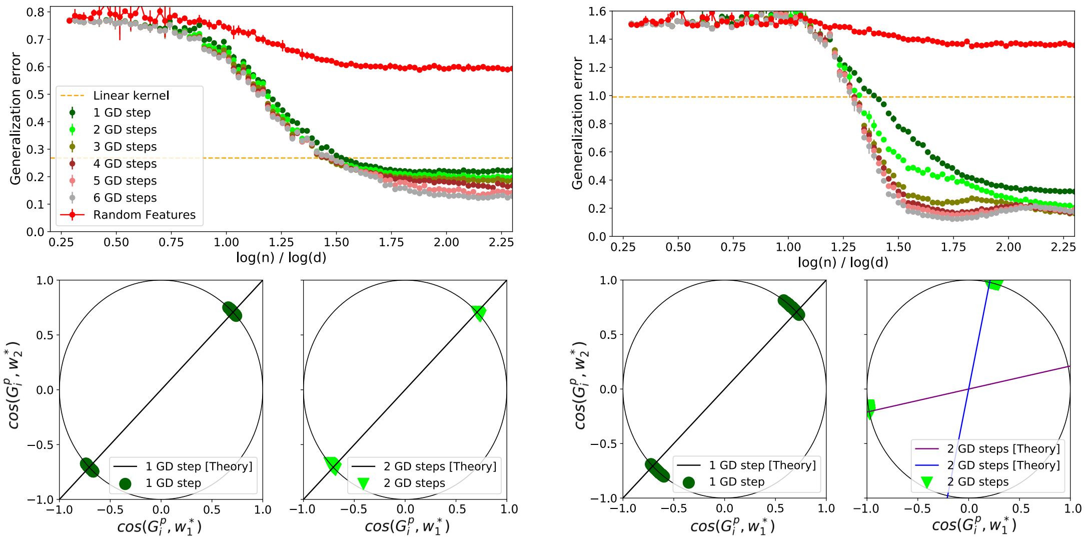

Figure 6: **Feature learning with multiple gradient steps. Top:** Generalization error as a function of n ($d = 512, p = 256$) after iterating the training procedure for six steps. **Bottom:** Cosine similarity of the projected gradient matrix G^p inside the target subspace for all the p neurons at a fixed ratio $^n/_d = 4$, plotted at different stages of the training. The blue and purple lines are the theoretical predictions for the orientation of the gradient in the second step.
We fix a relu student and consider two different 2-index target functions $f^\star(\boldsymbol{z}) = \sigma_1^\star(\langle \boldsymbol{w}_1^\star, \boldsymbol{z}\rangle) + \sigma_2^\star(\langle \boldsymbol{w}_2^\star, \boldsymbol{z}\rangle)$. **Left:** $\sigma_1^\star(z) = \sigma_2^\star(z) = \mathrm{He}_1(z) + {}^{\mathrm{He}_2(z)}/_2 + {}^{\mathrm{He}_4(z)}/_{4!}$ **Right:** $\sigma_1^\star(z) = z - z^2$ and $\sigma_2^\star(z) = z + z^2$. In accordance with Theorem 3, the difference between the two cases is clear already after the first GD step: while on the left the gradient is stuck around the predicted rank-one spike after the first step (black line), on the right the gradient changes orientation in the second step, allowing to learn multiple features.

and the network is stuck on the "linear subspace learning" (see the upper right plot of Fig. 2). These results are in perfect agreement with Thm. 3.

Investigating representation learning efficiency: In this paragraph we further analyze the claims of Thm. 3 in the context of feature learning. The experiments done in the lower panel of Fig. 6 are closely related to the ones of Fig. 5. However, contrary to the previous setting, we study the cosine similarity of the *projected gradient* $G^p = G\Pi^\star$ in the teacher subspace. This quantity differs from the cosine similarity of the full gradient, plotted in Fig. 5, as we lose completely the information about the share of the gradient lying in the subspace orthogonal to the teacher one. This divergence in the choices is due to the different illustrative goals of the figures: while in Fig. 6 we highlight the change in orientation of the gradient inside the teacher subspace after a few steps, hence not caring about the relative magnitude, in Fig. 5 we contrast the magnitude of the true gradient with the one of a random object (blue circles) in order to claim the presence (or lack) of feature learning after a single step. The results in the lower panel of Fig. 6 are obtained iterating 2 steps of the training procedure in Alg. 1: in accordance with Thm. 3 we observe delocalization of the projected gradient only if there are linearly connected directions that can be exploited to escape the spike given by the first Hermite coefficient $C_1(f^\star)$ (right panel). Moreover, we are able to theoretically predict the orientation of the gradient at the second step as well (see Appendix. C). On the contrary, when such linearly connected directions do not exist, the gradient is stuck on the spike $C_1(f^\star)$ (Left panel). We elaborate on this last observation by checking that the lack of specialization persists iterating for more than two GD steps. We present the results in Fig. 7: the gradient is stuck in the linear subspace learning regime even as the training proceeds, again in agreement with Thm. 3. Moreover, we illustrate by changing the

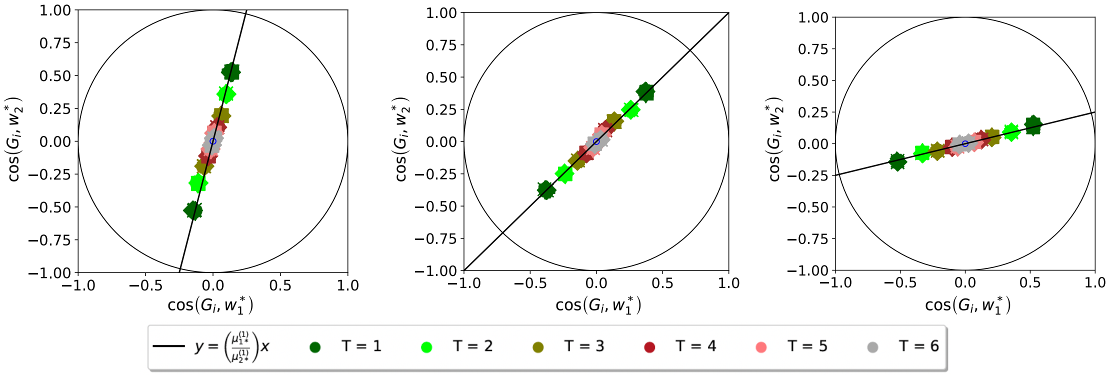

Figure 7: **Lack of feature learning after few GD steps.** The plots show the cosine similarity with respect to the teacher vectors $(\boldsymbol{w}_1^\star, \boldsymbol{w}_2^\star)$ for the gradient at different stages of the training. The predicted orientation (Thm. 2) of the gradient is shown as a continuous black line, the circle of unitary radius in black, and the circle of radius $2/\sqrt{d}$ in blue. We fix $n = d = 2^{13}$, the learning rate $\eta = p$, and we use a relu student. We vary the teacher functions: **Left:** $\sigma_1^\star(z) = 4z^2 + z$, $\sigma_2^\star(z) = z$ **Center:** $\sigma_1^\star(z) = \sigma_2^\star(z) = 4z^2 + z$, **Right:** $\sigma_2^\star(z) = 4z^2 + z$, $\sigma_1^\star(z) = z$. The orientation of the gradient does not change after $T = 6$ steps preventing specialization, in agreement with Thm. 3.

teacher functions, that the theoretical prediction of Thm. 2 on the gradient orientation, are valid beyond the symmetric teachers.

Multiple stairs: We complement the picture offered by Fig. 6 studying functions that have multiple linearly connected directions to the previously learned one, or informally, "multiple-stairs function". The results are presented in Fig. 8 by considering the function $f_\star(\boldsymbol{z}) = z_1/3 + 2z_1z_2/3 + z_2z_3$; we consider 3 steps in the training of Alg. 1, the network is able to learn respectively $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ after the first three steps of training in the proportional sample regime. This is clearly appreciable by studying the cosine similarity of the projected gradient on the teacher subspace: after the first step it is localized around $\boldsymbol{e}_1$, proceeding with training it has projections along $\boldsymbol{e}_2$ while $\boldsymbol{e}_3$ remains hidden, and only at the third step we obtain delocalization of the gradient along $\boldsymbol{e}_3$. These results are in perfect agreement with Thm. 3. Note that the hierarchical learning framework of Thm. 3 allows neurons to simultaneously specialize along different directions, as exemplified in Fig. 2 (see the bottom right plot). We observe one instance of this multidirectional staircase learning in Fig. 9 by considering the target $f_\star(\boldsymbol{z}) = z_1/3 + 2\,\mathrm{He}_2(z_1)z_2 + z_1z_3$: while the results are unchanged in the first step with respect to Fig. 8 (with only the $\boldsymbol{e}_1$ direction being learned), we observe that both directions $\boldsymbol{e}_2$ & $\boldsymbol{e}_3$ are learned at the second step.

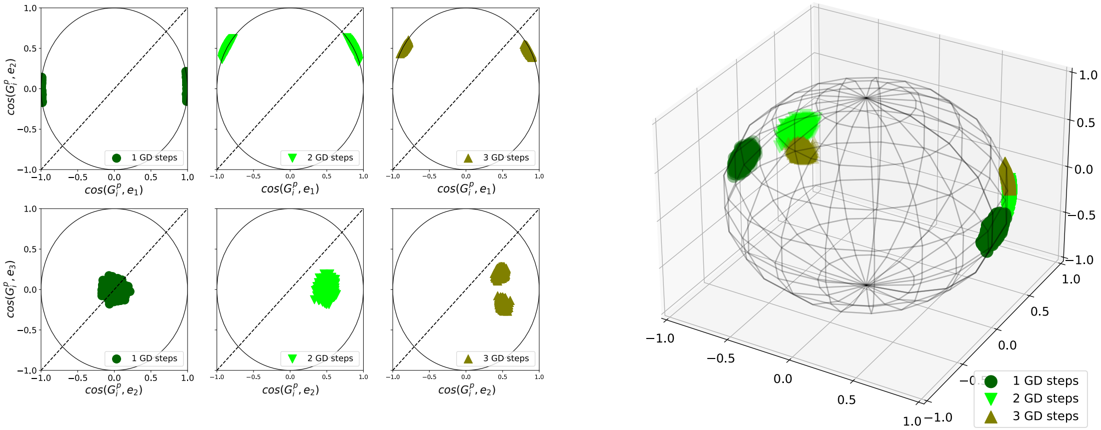

Figure 8: **Climibing multiple stairs.** Fix the teacher function $f_\star(\boldsymbol{z}) = z_1/3 + 2z_1z_2/3 + z_2z_3$ and a relu student. The plots show the cosine similarity of the projected gradient matrix G^p inside the teacher subspace for all the p neurons at a fixed ratio $n/d = 4$, plotted at different stages of the training following Alg. 1. The plot shows the similarity in the $3D$ teacher subspace on the right, and two sections of it on the left: **Up:** $(\boldsymbol{e}_1, \boldsymbol{e}_2)$ plane. **Bottom:**$(\boldsymbol{e}_2, \boldsymbol{e}_3)$ plane. In accordance with Thm. 3, the gradient is first localized around $\boldsymbol{e}_1$, then sequentially learns $\boldsymbol{e}_2$, and only at the third step has components along $\boldsymbol{e}_3$.

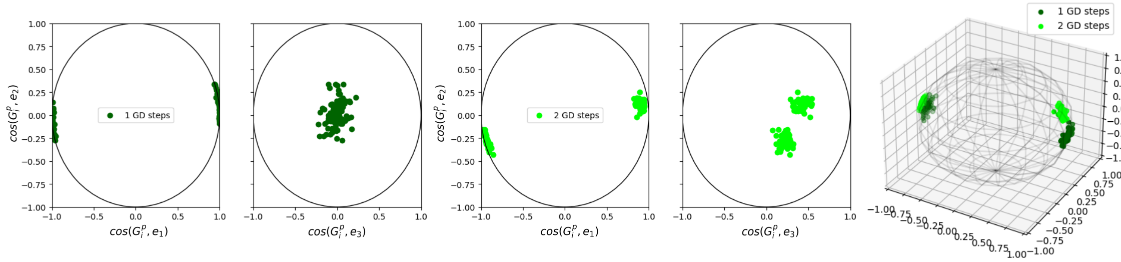

Figure 9: **Learning multiple directions at a time.** Fix the teacher function $f^\star(\boldsymbol{z}) = z_1/3 + 2\,\mathrm{He}_2(z_1)z_2 + z_1z_3$ and a relu student. The plots show the cosine similarity of the projected gradient matrix G^p inside the teacher subspace for all the p neurons at a fixed ratio $n/d = 4$, plotted at different stages of the training following Alg. 1. The plot shows the similarity measure in different cases. **Left:** $(\boldsymbol{e}_1, \boldsymbol{e}_2)$ cross section. **Center:** $(\boldsymbol{e}_3, \boldsymbol{e}_2)$ cross section. **Right:** $3D$ teacher subspace $(\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3)$. In accordance with Thm. 3, the gradient is first localized around the direction $\boldsymbol{e}_1$, and then learns both directions $(\boldsymbol{e}_3, \boldsymbol{e}_2)$ at the second gradient step.

Algorithm 1 Training procedure

```
Choice of parameters Fix the data dimension and the width of the second layer $(d, p)$ and sample $(W_0, \boldsymbol{a}_0)$ obeying eq. (26). Fix a regularization parameter $\lambda$, and a number of GD steps $T_{max}$.
for $n$ in a given range do
  Learning rate tuning Fix the learning rate $\eta = \mathcal{O}(p\sqrt{\frac{n}{d}})$.
  for $t < T_{max}$ do
    Data generation Sample the data matrix $Z \sim \mathcal{N}(0, I_{n\times d})$ and get the labels $Y = f_\star(Z) \in \mathbb{R}^n$
    Update first layer Compute the gradient matrix $G_t = \{\boldsymbol{G}_i^{(t)}\}_{i\in[p]} \in \mathbb{R}^{p\times d}$ and update $W$:
        $$\boldsymbol{G}_i^{(t)} \leftarrow \frac{a_{0,i}}{\sqrt{p}} \cdot \frac{1}{n}\sum_{\nu=1}^{n} \boldsymbol{x}^\nu \sigma'(\langle \boldsymbol{w}_i^{(t)}, \boldsymbol{z}^\nu\rangle) \left(\hat{f}(\boldsymbol{z}^\nu, W_t, \boldsymbol{a}_0) - f^\star(\boldsymbol{z}^\nu)\right) \tag{20}$$
        $$W_{t+1} \leftarrow W_t - \eta G_t \tag{21}$$
    if $t == T_{max}$ then
      Train second layer Get the feature matrix $X_t \leftarrow \sigma(W_t Z)$, and compute estimator:
        $$\hat{\boldsymbol{a}} \leftarrow \begin{cases} X_t^\top \left(X_t X_t^\top + \lambda I_n\right)^{-1} Y & n<p \\ \left(X_t^\top X_t + \lambda I_p\right)^{-1} X_t^\top Y & n>p \end{cases}$$
    end if
  end for
end for
```

B Preliminaries and Assumptions for the proofs

B.1 Preliminaries

Before going to the mathematical detail of the proof, we recall a few definition and useful facts.

Hermite expansion — Given the Gaussian measure γ_m on $\mathbb{R}^m$, we can build a scalar product on $\ell^2(\mathbb{R}^m, \gamma_m)$ as

$$\langle f, g\rangle_\gamma = \int_{\mathbb{R}^m} fg \, \mathrm{d}\gamma_m = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0}, I_m)}[f(\boldsymbol{z})g(\boldsymbol{z})]. \tag{22}$$

It turns out that there is a specific orthonormal basis of interest for this scalar product, that we present in tensor form:

Definition 4 (Hermite decomposition). *Let $f : \mathbb{R}^m \to \mathbb{R}$ be a function that is square integrable w.r.t the Gaussian measure. There exists a family of tensors $(C_j(f))_{k\in\mathbb{N}}$ such that $C_j(f)$ is of order j and for all $\boldsymbol{x} \in \mathbb{R}^m$,*

$$f(\boldsymbol{x}) = \sum_{j\in\mathbb{N}} \langle C_j(f), \mathcal{H}_j(\boldsymbol{x})\rangle \tag{23}$$

where $\mathcal{H}_j(\boldsymbol{x})$ is the j-th order Hermite tensor (Grad, 1949).

Higher-order singular value decomposition — The higher-order singular value decomposition (HOSVD) of a tensor is defined as follows:

Definition 5 (Higher-order SVD). *Let $C \in \mathbb{R}^{m^k}$ be a symmetric tensor of order k. A higher-order SVD of C is an orthonormal set $(\boldsymbol{u}_1, \dots, \boldsymbol{u}_r)$ of $r \leq k$ vectors, as well as a tensor $S \in \mathbb{R}^{r^k}$ such that*

$$C = \sum_{j_1,\dots,j_k=1}^{r} S_{j_1,\dots,j_k} \boldsymbol{u}_{j_1} \otimes \cdots \otimes \boldsymbol{u}_{j_k} \tag{24}$$

The singular values tensor S, as well as the rank r, are unique, but just as the regular SVD, the vectors $(\boldsymbol{u}_1, \dots, \boldsymbol{u}_r)$ are only unique up to rotations.

B.2 Setting and assumptions

We discuss the setting and main assumptions required. The first concerns the class of target functions we consider.

Assumption 2 (Data model). *The training inputs $\boldsymbol{z}^\nu \in \mathbb{R}^d$ are independently drawn from the Gaussian distribution $\mathcal{N}(0, I_d)$. Further, we assume that the target function $y^\nu = f^\star(\boldsymbol{z})$ depends only on a few relevant directions. In other words, there exists a low-dimensional subspace $V^\star \subset \mathbb{R}^d$ of fixed dimension r and a function $g^\star : V^\star \to \mathbb{R}$ such that:*

$$y = f^\star(\boldsymbol{z}) \coloneqq g^\star(\Pi^\star \boldsymbol{z}), \tag{25}$$

where $\Pi^\star$ is the orthogonal projection on $V^\star$.

As we will show later, learning with GD can be seen as a hierarchical process, where depending on the batch size different directions of the target are progressively learned.

Given a batch of training data $(\boldsymbol{z}^\nu, y^\nu)_{\nu=1}^n \in \mathbb{R}^{d+1}$ drawn from the model (2) defined above, we now define how the network weights (W, a) are initialized and updated.

Assumption 3 (Training procedure). *Consider the following random initialization for the weights:*

$$\sqrt{p} \cdot a_i^0 \overset{i.i.d}{\sim} \mathrm{Unif}([-1,1]) \quad \textit{and} \quad \boldsymbol{w}_i^0 \overset{i.i.d}{\sim} \mathrm{Unif}(\mathbb{S}^{d-1}). \tag{26}$$

The distribution of the a_i can be replaced by any other continuous distribution with positive variance. Note that for $p = \mathcal{O}(1)$, we have $\hat{f}(\boldsymbol{z}; W^0, \boldsymbol{a}^0) \neq 0$. To further simplify the analysis, we assume p to be even and further impose the following symmetrization at initialization:

$$a_i^0 = -a_{p-i+1}^0 \quad \textit{and} \quad \boldsymbol{w}_i^0 = \boldsymbol{w}_{p-i+1}^0 \quad \textit{for all } i \in [p/2], \tag{27}$$

which ensures $\hat{f}(\boldsymbol{z}; W^0, \boldsymbol{a}^0) = 0$. Note that this simplification is common in the related literature, e.g. Chizat et al. (2019); Damian et al. (2022), and is mainly necessary when p is small. Given the initial conditions, the weights are trained with the following two-step full-batch gradient descent:

(i) First layer training*: for every gradient step $t \leq T$, a fresh batch of training data $\{(\boldsymbol{z}^\nu, y^\nu)\}_{\nu=1}^n$ is drawn from the model in Assumption 2, and the first layer weights are updated according to:*

$$\boldsymbol{w}_i^{t+1} = \boldsymbol{w}_i^t - \frac{\eta}{2n} \sum_{\nu=1}^n \nabla_{\boldsymbol{w}_i} \left(y^\nu - \hat{f}(\boldsymbol{z}^\nu; W^t, \boldsymbol{a}^0) \right)^2, \tag{28}$$

Hence, the total sample complexity for this step is Tn.
(ii) Second layer training*: once the first layer is trained for T steps, the second layer weights a are trained to optimality by performing ridge regression with the features learned in the first step:*

$$\hat{\boldsymbol{a}} = \underset{\boldsymbol{a} \in \mathbb{R}^p}{\arg\min} \frac{1}{2n} \sum_{\nu=1}^n \left(y^\nu - \hat{f}(\boldsymbol{z}^\nu; W^T, \boldsymbol{a}) \right)^2 + \lambda \|\boldsymbol{a}\|^2. \tag{29}$$

Such a separation of the training between the first and second layer is a common setup for the theoretical study of training (Damian et al., 2022; Abbe et al., 2023; Berthier et al., 2023), and allows for a more tractable study of convergence.

C Gradient descent on the first layer

C.1 Technical assumptions

We shall show our results under the following assumptions. First, since we assume that the leap index of $f^\star$ is at least one, and we setup the network to zero output the following assumption is unrestrictive:

Assumption 4. *The teacher function $f^\star$ and the student activation σ both have their zero-th Hermite coefficient equal to 0.*

We shall also need a smoothness assumption:

Assumption 5. *Both the student activation σ and g^* are continuous, and differentiable except possibly on a finite set of points. Further, the first two derivatives of $g^\star$ and the first three derivatives of σ are uniformly bounded in $\mathbb{R}$.*

C.2 Preliminaries

More on Hermite expansion We recall a few properties of the Hermite tensors of Definition 4. Up to symmetry, the tensors $\mathcal{H}_k(\boldsymbol{x})$ are an orthonormal basis of $\ell^2(\mathbb{R}^m, \gamma)$, in the sense that for any $\boldsymbol{i}, \boldsymbol{j} \in \mathbb{R}^k$,

$$\langle \mathcal{H}_{k,\boldsymbol{i}}(\boldsymbol{x}), \mathcal{H}_{k,\boldsymbol{j}}(\boldsymbol{x}) \rangle_\gamma = \frac{1}{|\mathfrak{o}(\boldsymbol{i})|} \mathbf{1}_{\boldsymbol{i} \text{ is a permutation of } \boldsymbol{j}} \tag{30}$$

where $|\mathfrak{o}(\boldsymbol{i})|$ is the number of distinct permutations of $\boldsymbol{i}$. It can be checked from the definition in Grad (1949) that the $\mathcal{H}_k$, and hence the $C_k(f)$, are basis-invariant, and hence represent an actual k-linear form on $\mathbb{R}^m$. Further, the property (30) yields an immediate expression for the scalar product in $\ell^2(\mathbb{R}^m, \gamma_m)$:

$$\langle f, g \rangle_\gamma = \sum_{k \in \mathbb{N}} \langle C_k(f), C_k(g) \rangle. \tag{31}$$

Further, the Hermite coefficients of low-rank functions are straightforward to compute:

Lemma 1. *Let $g : \mathbb{R}^r \to \mathbb{R}$, and a linear map $A \in \mathbb{R}^{r \times d}$. Then the Hermite coefficients of $f(\boldsymbol{x}) = g(A\boldsymbol{x})$ are*

$$C_k(f) = C_k(g) \cdot (A, \dots, A), \tag{32}$$

where $\cdot$ is the multilinear multiplication operator (Greub, 2012).

In particular, this implies that the singular vectors of $C_k^\star$ all belong to $V^\star$.

Concentration in Orlicz spaces We recall the classical definition of Orlicz spaces:

Definition 6. *For any $\alpha \in \mathbb{R}$, let $\psi_\alpha(x) = e^{x^\alpha} - 1$. Let X be a real random variable; the* Orlicz norm *$\|X\|_{\psi_\alpha}$ is defined as*

$$\|X\|_{\psi_\alpha} = \inf\left\{ t > 0 \ : \ \mathbb{E}\left[\psi_\alpha\left(\frac{|X|}{t} \right) \right] \leq 1 \right\} \tag{33}$$

We refer to the monographs Ledoux and Talagrand (1991); van der Vaart and Wellner (1996) for more information. We say that a random variable is sub-gaussian (resp. sub-exponential) if its ψ_2 (resp. ψ_1) norm is finite. The main use of this definition is the following concentration inequality: for a variable X with finite Orlicz norm,

$$\mathbb{P}(|X - \mathbb{E}X| > t\|X\|_{\psi_\alpha}) \leq 2e^{-t^\alpha}. \tag{34}$$

The Orlicz norms are sub-multiplicative, in the following sense:

Lemma 2. *Let X and Y be two random variables. Then, for any $\alpha > 0$, there exists a constant K_α such that*

$$\|XY\|_{\psi_{\alpha/2}} \leq K_\alpha \|X\|_{\psi_\alpha} \|Y\|_{\psi_\alpha} \tag{35}$$

Finally, we shall use the following theorem:

Theorem 5 (Theorem 6.2.3 in Ledoux and Talagrand (1991) and Lemma 2.2.2 in van der Vaart and Wellner (1996))**.** *Let $X_1, \dots, X_n$ be n independent random variables with zero mean and second moment $\mathbb{E}X_i^2 = \sigma_i^2$. Then,*

$$\left\| \sum_{i=1}^n X_i \right\|_{\psi_\alpha} \leq K_\alpha \log(n)^{1/\alpha} \left(\sqrt{\sum_{i=1}^n \sigma_i^2} + \max_i \|X_i\|_{\psi_\alpha} \right) \tag{36}$$

Preliminary computations We begin with a few useful preliminary computations. First, since $\boldsymbol{w}_i^0 \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, the following lemma holds:

Lemma 3. *With probability at least $1 - cpe^{-c\log(d)^2}$, we have for any $i \in [p]$ and $k \in [r]$:*

$$\|\boldsymbol{\pi}_i^0\| \leq \frac{\sqrt{r}\log(d)}{\sqrt{d}} \tag{37}$$

Let $\boldsymbol{g}_i$ be the negative gradient for the i-th neuron at initialization:

$$\boldsymbol{g}_i = -\nabla_{\boldsymbol{w}_j} \mathcal{L}\left(\hat{f}(\boldsymbol{z}^\nu; W^0, \boldsymbol{a}), f^\star(\boldsymbol{z}^\nu) \right). \tag{38}$$

Since at initialization the output of the network is exactly zero, we have

$$\boldsymbol{g}_i = \frac{a_i}{\sqrt{p}} \cdot \frac{1}{n} \sum_{\nu=1}^n \boldsymbol{z}^\nu \sigma'(\langle \boldsymbol{w}_i^0, \boldsymbol{z}^\nu \rangle) f^\star(\boldsymbol{z}^\nu) \tag{39}$$

Finally, the update equation for $\|\boldsymbol{w}_i\|$ reads

$$\|\boldsymbol{w}_i^1\|^2 = 1 + 2\eta \langle \boldsymbol{w}_i^0, \boldsymbol{g_i} \rangle + \eta^2 \|\boldsymbol{g}_i\|^2 \tag{40}$$

C.3 Computing expectations

We begin by a simple computation of the expectation of $\boldsymbol{g}_i$:

Lemma 4. *For any $i \in [p]$, we have*

$$\mathbb{E}[\boldsymbol{g}_i] = \frac{a_i}{\sqrt{p}} \left(\sum_{k=0}^\infty c_{k+2} \left\langle (\boldsymbol{w}_i^0)^{\otimes k}, C_k^\star \right\rangle \boldsymbol{w}_i + \sum_{k=0}^\infty c_{k+1}\, C_{k+1}^\star \times_{1\dots k} (\boldsymbol{w}_i^0)^{\otimes k} \right) \tag{41}$$

where the last multiplication is a product over the first k axes of C_{k+1} (and thus results in a vector).

Proof. By Stein's lemma, for any $\boldsymbol{w}$, we have

$$\begin{aligned} \mathbb{E}\left[\boldsymbol{z}\sigma'(\langle \boldsymbol{w}, \boldsymbol{z}\rangle) f^\star(\boldsymbol{z})\right] &= \mathbb{E}\left[\nabla_{\boldsymbol{z}} \sigma'(\langle \boldsymbol{w}, \boldsymbol{z}\rangle) f^\star(\boldsymbol{z})\right] + \mathbb{E}\left[\sigma'(\langle \boldsymbol{w}, \boldsymbol{z}\rangle) \nabla_{\boldsymbol{z}} f^\star(\boldsymbol{z})\right] \\ &= \boldsymbol{w}\mathbb{E}\left[\sigma''(\langle \boldsymbol{w}, \boldsymbol{z}\rangle) f^\star(\boldsymbol{z})\right] + \mathbb{E}\left[\sigma'(\langle \boldsymbol{w}, \boldsymbol{z}\rangle) \nabla_{\boldsymbol{z}} f^\star(\boldsymbol{z})\right] \end{aligned}$$

From Lemma 1, the k-th Hermite coefficient of $\boldsymbol{z} \mapsto \sigma''(\langle \boldsymbol{w}, \boldsymbol{z}\rangle)$ is $c_{k+2}\, \boldsymbol{w}^{\otimes k}$, where the $(c_k)_{k\geq 0}$ are the Hermite coefficients of σ. By two applications of the scalar product formula (31), we find

$$\mathbb{E}\left[\boldsymbol{z}\sigma'(\langle \boldsymbol{w}, \boldsymbol{z}\rangle) f^\star(\boldsymbol{z})\right] = \sum_{k=0}^\infty c_{k+2} \langle \boldsymbol{w}^{\otimes k}, C_k^\star \rangle \boldsymbol{w} + \sum_{k=0}^\infty c_{k+1}\, C_{k+1}^\star \times_{1\dots k} \boldsymbol{w}^{\otimes k}. \tag{42}$$

□

Truncating the expansions Now, we show that the expectations in Lemma 4 can be truncated at the leap index term.

Lemma 5. *With probability at least $1 - cpe^{-c\log(d)^2}$, for every $k \geq 0$ and $i \in [p]$, we have*

$$\left|\langle C_k^\star, (\boldsymbol{w}_i^0)^{\otimes k}\rangle\right| \leq c\left(\frac{\sqrt{r}\log(d)}{\sqrt{d}}\right)^k \quad \text{and} \quad \left\|C_{k+1}^\star \times_{1\dots k} (\boldsymbol{w}_i^0)^{\otimes k}\right\| \leq c\left(\frac{\sqrt{r}\log(d)}{\sqrt{d}}\right)^k \tag{43}$$

As a result, if ℓ is the leap index of $f^\star$,

$$\left\|\mathbb{E}[\boldsymbol{g}_i] - C_\ell^\star \times_{1\dots(\ell-1)} (\boldsymbol{w}_i^0)^{\otimes(\ell-1)}\right\| = \mathcal{O}\left(\frac{r^{\ell/2}\operatorname{polylog}(d)}{d^{\ell/2}}\right) \tag{44}$$

Proof. First, we have by Lemma 1,

$$\left|\langle C_k^\star, (\boldsymbol{w}_i^0)^{\otimes k}\rangle\right| = \left|\langle C_k(g^\star), (W^\star \boldsymbol{w}_i^0)^{\otimes k}\rangle\right| \leq \|C_k(g^\star)\|_2 \cdot \|\boldsymbol{\pi}_i^0\|^k,$$

where $\|C_k(g^\star)\|_2$ is the operator norm of $C_k(g^\star)$. Since

$$\|C_k(g^\star)\|_2 \leq \|C_k(g^\star)\|_F \leq \|g^\star\|_\gamma,$$

the first inequality ensues by Lemma 3. Now, let A_{k+1} be the $(k+1)$-th mode unfolding of $C_{k+1}(g^\star)$; then

$$\left\|C_{k+1}^\star \times_{1\dots k} (\boldsymbol{w}_i^0)^{\otimes k}\right\| = \left\|A_{k+1}(W^\star \boldsymbol{w}_i^0)^{\otimes k}\right\| \leq \|A_{k+1}\|_2 \|\boldsymbol{\pi}_i^0\|^k$$

The norm of A_{k+1} is then bounded by the same argument as above. The final equality is obtained by using the above bounds on every term above $k = \ell$ in the first sum, and above $k = \ell - 1$ in the second. □

Student norms We now move on to controlling (40), in expectation. We begin with the cross-term:

Lemma 6. *With probability at least $1 - cpe^{-c\log(d)^2}$, we have for any $i \in [p]$,*

$$\mathbb{E}\left[\langle \boldsymbol{w}_i^0, \boldsymbol{g}_i\rangle\right] = \mathcal{O}\left(\frac{r^{\ell/2}\operatorname{polylog}(d)}{pd^{\ell/2}}\right) \tag{45}$$

Proof. From eq. (44), we have

$$\mathbb{E}\left[\langle \boldsymbol{w}_i^0, \boldsymbol{g}_i\rangle\right] = \frac{a_i}{\sqrt{p}}\left(\langle C_\ell^\star, (\boldsymbol{w}_i^0)^{\otimes \ell}\rangle + \mathcal{O}\left(\frac{r^{\ell/2}\operatorname{polylog}(d)}{d^{\ell/2}}\right)\right)$$

The first part of Lemma 5 gives

$$\langle C_\ell^\star, (\boldsymbol{w}_i^0)^{\otimes \ell}\rangle = \mathcal{O}\left(\frac{r^{\ell/2}\operatorname{polylog}(d)}{pd^{\ell/2}}\right),$$

and the lemma follows since $|a_i| \leq 1/\sqrt{p}$. □

The main object of study is therefore $\|\boldsymbol{g}_i\|^2$. We can write it as

$$\begin{aligned}\|\boldsymbol{g}_i\|^2 &= \frac{a_i^2}{n^2p^2}\sum_{\nu,\nu'=1}^n \langle \boldsymbol{z}^\nu, \boldsymbol{z}^{\nu'}\rangle \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}^\nu\rangle) f^\star(\boldsymbol{z}^\nu)\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}^{\nu'}\rangle) f^\star(\boldsymbol{z}^{\nu'}) \\ &= \frac{a_i^2}{n^2p^2}\left(\sum_{\nu\neq\nu'} \langle \boldsymbol{z}^\nu, \boldsymbol{z}^{\nu'}\rangle \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}^\nu\rangle) f^\star(\boldsymbol{z}^\nu)\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}^{\nu'}\rangle) f^\star(\boldsymbol{z}^{\nu'})\right. \\ &\qquad \left. + \sum_{\nu=1}^n \|\boldsymbol{z}^\nu\|^2 \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}^\nu\rangle)^2 f^\star(\boldsymbol{z}^\nu)^2\right)\end{aligned} \tag{46}$$

Since $\boldsymbol{z}^\nu, \boldsymbol{z}^{\nu'}$ are independent for $\nu \neq \nu'$, this leaves

$$\mathbb{E}\left[\|\boldsymbol{g}_i\|^2\right] = \frac{n(n-1)}{n^2}\|\mathbb{E}[\boldsymbol{g}_i]\|^2 + \frac{a_i^2}{np^2}\mathbb{E}\left[\|\boldsymbol{z}^\nu\|^2\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}^\nu\rangle)^2 f^\star(\boldsymbol{z}^\nu)^2\right] \tag{47}$$

We shall only need orders of magnitude for those terms. These are taken care of in the following lemma:

Lemma 7. *There exists a bounded random variable X independent from d such that, with probability at least $1 - cpe^{-\log(d)^2}$,*

$$\|\mathbb{E}[\boldsymbol{g}_i]\|^2 = a_i^2 X_i \cdot \frac{\left\|\boldsymbol{\pi}_i^0\right\|^{2(\ell-1)}}{p^2} + \mathcal{O}\left(\frac{r^{\ell/2}\,\mathrm{polylog}(d)}{d^{\ell/2}}\right) \tag{48}$$

where $(X_i)_{i\in[p]}$ are i.i.d copies of X. Additionally, there exist two constants c, C such that

$$c \cdot d \leq \mathbb{E}\left[\|\boldsymbol{z}\|^2\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}\rangle)^2 f^\star(\boldsymbol{z})^2\right] \leq C \cdot d \tag{49}$$

Proof. We begin with (48). Define the unit norm vectors

$$\boldsymbol{r}_i = \frac{W^\star \boldsymbol{w}_i^0}{\|\boldsymbol{\pi_i^0}\|},$$

since the $\boldsymbol{w}_i$ are isotropic, the $\boldsymbol{r}_i$ are uniform on $\mathbb{S}^{r-1}$. Then,

$$\left\|C_\ell^\star \times_{1\ldots(\ell-1)} (\boldsymbol{w}_i^0)^{\otimes(\ell-1)}\right\|^2 = \underbrace{\left\|C_\ell(g^\star) \times_{1\ldots(\ell-1)} \boldsymbol{r_i}^{\otimes(\ell-1)}\right\|^2}_{=:X_i} \cdot \left\|\boldsymbol{\pi}_i^0\right\|^{2(\ell-1)}.$$

The random variables X_i thus defined are i.i.d, independent from d, and have positive expectation since $C_\ell(g^\star)$ is nonzero. Equation (48) then results from the expansion in (44).

We now move on to the second part; first, by Hölder's inequality,

$$\mathbb{E}\left[\|\boldsymbol{z}\|^2\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}\rangle)^2 f^\star(\boldsymbol{z})^2\right] \leq \sqrt{\mathbb{E}\left[\|\boldsymbol{z}\|^4\right]}\sqrt[4]{\mathbb{E}\left[\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}\rangle)^8\right]\mathbb{E}\left[f^\star(\boldsymbol{z})^8\right]} \leq C \cdot d, \tag{50}$$

since the last two expectations are independent from d. On the other hand, using the same inequality with $\|\boldsymbol{z}\|^2 - d$, we can write

$$\mathbb{E}\left[\|\boldsymbol{z}\|^2\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}\rangle)^2 f^\star(\boldsymbol{z})^2\right] = d\mathbb{E}\left[\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}\rangle)^2 f^\star(\boldsymbol{z})^2\right] + \mathcal{O}(\sqrt{d}).$$

Since $\mu_\ell \neq 0$ and $f^\star$ has leap index ℓ, there exists $\varepsilon > 0$ two subsets $\mathcal{A} \subseteq \mathbb{R}, \mathcal{B} \subseteq V^\star$ of positive measure such that $\sigma'(x)^2 \geq \varepsilon$ if $x \in \mathcal{A}$ and $f^\star(\boldsymbol{z}^\star) > \varepsilon$ if $\boldsymbol{z}^\star \in \mathcal{B}$. From the fact that $\pi_i \leq 1/2$ with high probability, we conclude that the set

$$\mathcal{C} := \{\boldsymbol{z} \in \mathbb{R}^p \;:\; \langle \boldsymbol{w}_i, \boldsymbol{z}\rangle \in \mathcal{A}, P_{V^\star}\boldsymbol{z} \in \mathcal{B}\}$$

has positive (Gaussian) measure. It follows that

$$\mathbb{E}\left[\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}\rangle)^2 f^\star(\boldsymbol{z})^2\right] \geq \gamma(\mathcal{C})\varepsilon^2, \tag{51}$$

which concludes the proof of Eq. (49). □

C.4 Concentration

We now move on to concentrating the quantities of interest of the previous section. Our aim will be to show the following proposition:

Proposition 2. *With probability at least* $1 - Cpe^{-c\log(n)^2} - Cpe^{-c\log(d)^2}$*, for any* $i \in [p], k \in [r]$,

$$\left\|\boldsymbol{\pi}_i^1 - \mathbb{E}\left[\boldsymbol{\pi}_i^1\right]\right\| = \mathcal{O}\left(\frac{\eta\sqrt{r}\log(n)}{p\sqrt{n}}\right) \tag{52}$$

$$\left|\left\|\boldsymbol{w}_i^1\right\|^2 - \mathbb{E}\left[\left\|\boldsymbol{w}_i^1\right\|^2\right]\right| = \mathcal{O}\left(\frac{\eta\log(n)}{p\sqrt{n}} + \frac{\eta^2 d\log(n)^6}{p^2 n\sqrt{n}} + \frac{\eta^2\log(d)}{p^2 n\sqrt{d}} + \frac{\eta^2 r\log(n)^2}{p^2 n} + \frac{\eta^2\log(n)^\ell r^{(\ell-1)/2}}{p^2 d^{(\ell-1)/2}\sqrt{n}}\right) \tag{53}$$

Importantly, we do not claim that the whole vector $\boldsymbol{w}_i^1$ concentrates; only its norm and its projection on a low-dimensional subspace do. Throughout this section, we define the random vectors

$$\boldsymbol{X}^\nu = \boldsymbol{z}^\nu \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}^\nu\rangle) f^\star(\boldsymbol{z}^\nu). \tag{54}$$

These vectors are i.i.d, with the same distribution as a random vector that we will call $\boldsymbol{X}$.

Concentration of linear functionals We begin with a simple bound, that implies both Eq. (52) and the first term of Eq. (53).

Lemma 8. *Let* $\boldsymbol{w}$ *be a unit vector in* $\mathbb{R}^d$*. There exists a universal constant* c *such that with probability* $1 - 2pe^{-c\log(n)^2}$*, for any* $i \in [p]$ *and* $k \in [r]$,

$$|\langle \boldsymbol{w}, \boldsymbol{g}_i\rangle - \mathbb{E}[\langle \boldsymbol{w}, \boldsymbol{g}_i\rangle]| \leq \frac{\log(n)}{p\sqrt{n}} \tag{55}$$

Proof. By Assumption 5, the function $f^\star$ is Lipschitz, so $f^\star(\boldsymbol{z})$ is a sub-gaussian random variable. The same is obviously true for $\langle \boldsymbol{w}, \boldsymbol{z}\rangle$, and since σ' is bounded the random variable $\langle \boldsymbol{w}, \boldsymbol{X}\rangle = \langle \boldsymbol{w}, \boldsymbol{z}\rangle\sigma'(\langle \boldsymbol{w}_i^0, \boldsymbol{z}\rangle)f^\star(\boldsymbol{z})$ is sub-exponential with bounded sub-exponential norm. We can thus apply Bernstein's inequality (Vershynin, 2018, Corollary 2.8.3) with $t = \log(n)/\sqrt{n}$ to get

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{\nu=1}^n \langle \boldsymbol{w}, \boldsymbol{X}^\nu\rangle - \mathbb{E}\langle \boldsymbol{w}, \boldsymbol{X}\rangle\right| \geq \frac{\log(n)}{\sqrt{n}}\right) \leq 2e^{-c\log(n)^2}. \tag{56}$$

The result ensues upon noticing that $\frac{1}{n}\sum \boldsymbol{X}^\nu$ differs from $\boldsymbol{g}_i$ by a factor of at most $1/p$. □

Decomposing the gradient norm We now move on to the concentration of the q_i. This allows us to write

$$\|\boldsymbol{g}_i\|^2 - \mathbb{E}[\|\boldsymbol{g}_i\|^2] = \frac{1}{n^2p^2}\left(\underbrace{\sum_{\nu=1}^n \|\boldsymbol{X}^{\boldsymbol{\nu}}\|^2 - n\mathbb{E}[\|\boldsymbol{X}\|^2]}_{S_1} + \underbrace{\sum_{\nu\neq\nu'} \langle \boldsymbol{X}^\nu, \boldsymbol{X}^{\nu'}\rangle - n(n-1)\|\mathbb{E}\boldsymbol{X}\|^2}_{S_2}\right) \tag{57}$$

We shall show the concentration of those two terms sequentially.

Concentrating the norms We first focus on S_1:

Lemma 9. *Let $i \in [p]$. There exists a constant $c > 0$ such that with probability $1 - e^{-c\log(n)^2}$,*

$$\mathbb{P}\big(|S_1| \geq \log(n)^6 d\sqrt{n}\big) \leq e^{-c\log(n)^2}. \tag{58}$$

Proof. The random variable $\|\boldsymbol{z}\|/\sqrt{d}$ is sub-gaussian, and so is $f^\star(\boldsymbol{z}^\nu)$. By Lemma 2 and the Hölder inequality, the random variable $\|\boldsymbol{X}^\nu\|^2$ satisfies

$$\|\|\boldsymbol{X}^\nu\|^2\|_{\psi_{1/2}} \leq C\cdot d \quad \text{and} \quad \mathrm{Var}\left(\|\boldsymbol{X}^\nu\|^2\right) \leq C\cdot d^2$$

As a result, we can apply Theorem 5 to the random variables $\|\boldsymbol{X}^\nu\|^2 - \mathbb{E}[\|\boldsymbol{X}\|^2]$, which yields

$$\left\|\sum_{\nu=1}^n \|\boldsymbol{X}^\nu\|^2 - n\mathbb{E}\left[\|\boldsymbol{X}\|^2\right]\right\|_{\psi_{1/2}} \leq c\log(n)^2 d\sqrt{n}. \tag{59}$$

Equation (58) is then a consequence of the Orlicz concentration bound (34). □

Decomposing the cross-term We now move on to S_2. To handle this sum, we use the following decoupling result from Pena and Montgomery-Smith (1995):

Theorem 6. *Let $(f_{ij})_{i,j\in[n]}$ be a set of measurable functions from $\mathcal{S}^2$ to a Banach space $(B, \|\cdot\|)$, and $(X_1, \dots, X_n), (Y_1, \dots, Y_n)$ two sets of independent random variables such that the laws of X_i and Y_i are the same. Then there exists a constant $C > 0$ such that*

$$\mathbb{P}\left(\left\|\sum_{i\neq j} f_{ij}(X_i, X_j)\right\| \geq t\right) \leq C\mathbb{P}\left(\left\|\sum_{i\neq j} f_{ij}(X_i, Y_j)\right\| \geq \frac{t}{C}\right) \tag{60}$$

We apply this theorem to the functions $f_{\nu,\nu'}(\boldsymbol{X}^\nu, \boldsymbol{X}^{\nu'}) = \langle \boldsymbol{X}^\nu, \boldsymbol{X}^{\nu'}\rangle - \|\mathbb{E}\boldsymbol{X}\|^2$. Let $\boldsymbol{Y}^\nu$ be an independent copy of the $\boldsymbol{X}^\nu$ for $\nu \in [n]$, we then have to estimate

$$\mathbb{P}\left(\left|\sum_{\nu\neq\nu'} \langle \boldsymbol{X}^\nu, \boldsymbol{Y}^{\nu'}\rangle - n(n-1)\|\mathbb{E}\boldsymbol{X}\|^2\right| \geq t\right).$$

For convenience, let $\bar{x} = \|\mathbb{E}\boldsymbol{X}\|^2$. Since the $\boldsymbol{X}^\nu$ are sub-exponential vectors, the scalar product $\langle \boldsymbol{X}^\nu, \boldsymbol{Y}^\nu\rangle$ has finite $\psi_{1/2}$-norm. The same bound as Lemma 9 then gives that

$$\mathbb{P}\left(\left|\sum_{\nu=1}^n \langle \boldsymbol{X}^\nu, \boldsymbol{Y}^\nu\rangle - n\bar{x}^2\right| \geq \sqrt{n}d\log(n)^6\right) \leq e^{-c\log(n)^2} \tag{61}$$

Hence, to show Proposition 2, we only need to study the overall sum

$$\tilde{S}_2 := \sum_{\nu,\nu'=1}^n \langle \boldsymbol{X}^\nu, \boldsymbol{Y}^{\nu'}\rangle - n^2\bar{x}^2 \tag{62}$$

Recall that, as in the proof of Lemma 7, the vector $\mathbb{E}\boldsymbol{X}$ belongs to the space $V_i = V^\star + \mathrm{span}(\boldsymbol{w}_i^0)$. We thus make the decomposition

$$\boldsymbol{X}^\nu = \boldsymbol{X}_i^\nu + \boldsymbol{X}_\perp^\nu \quad \text{and} \quad \boldsymbol{Y}^\nu = \boldsymbol{Y}_i^\nu + \boldsymbol{Y}_\perp^\nu \tag{63}$$

where $\boldsymbol{X}_i^\nu, \boldsymbol{Y}_i^\nu \in V_i$. Hence,

$$\sum_{\nu,\nu'=1}^n \langle \boldsymbol{X}^\nu, \boldsymbol{Y}^{\nu'}\rangle - n^2\bar{x}^2 = \underbrace{\langle \sum_{\nu=1}^n \boldsymbol{X}_i^\nu, \sum_{\nu=1}^n \boldsymbol{Y}_i^\nu\rangle - n^2\bar{x}^2}_{S_2'} + \underbrace{\langle \sum_{\nu=1}^n \boldsymbol{X}_\perp^\nu, \sum_{\nu=1}^n \boldsymbol{Y}_\perp^\nu\rangle}_{S_2''} \tag{64}$$

Bounding the last two terms The main step in bounding S_2' is the following lemma:

Lemma 10. *With probability at least* $1 - Ce^{-c\log(n)^2}$,

$$\left\| \sum_{\nu=1}^{n} \boldsymbol{X}_i^\nu - n\mathbb{E}\boldsymbol{X} \right\| \leq C\sqrt{r}\log(n)\sqrt{n} \tag{65}$$

Proof. Let $(\boldsymbol{u}_1, \dots, \boldsymbol{u}_{r+1})$ be an orthonormal basis of V_i. Since we have for any vector $\boldsymbol{x} \in V_i$

$$\|\boldsymbol{x}\|^2 = \sum_{k=1}^{r+1} \langle \boldsymbol{x}, \boldsymbol{u}_k \rangle^2,$$

it suffices to bound such a scalar product with high probability. Each term of the form $\langle \boldsymbol{X}_i^\nu - \mathbb{E}\boldsymbol{X}, \boldsymbol{u_k} \rangle$ is a sub-exponential random variable with zero mean and bounded variance, and hence by another application of Bernstein's inequality

$$\mathbb{P}\left(\left| \langle \sum_{\nu=1}^{n} \boldsymbol{X}_i^\nu - n\mathbb{E}\boldsymbol{X}, \boldsymbol{u}_k \rangle \right| \geq \log(n)\sqrt{n} \right) \leq e^{-c\log(n)^2} \tag{66}$$

The result ensues from a union bound, and the equivalence of norms in finite-dimensional spaces. □

As an easy corollary of this lemma, we get the following bound on S_2':

Corollary 2. *With probability at least* $1 - Ce^{-c\log(n)^2}$,

$$S_2' = \mathcal{O}\left(rn\log(n)^2 + \frac{\log(n)^\ell r^{(\ell-1)/2} n\sqrt{n}}{d^{(\ell-1)/2}} \right) \tag{67}$$

Proof. We use the following decomposition:

$$\begin{aligned} S_2' &= n\langle \sum_{\nu=1}^{n} \boldsymbol{X}_i^\nu - n\mathbb{E}\boldsymbol{X}, \mathbb{E}\boldsymbol{X} \rangle + n\langle \mathbb{E}\boldsymbol{X}, \sum_{\nu=1}^{n} \boldsymbol{Y}_i^\nu - n\mathbb{E}\boldsymbol{X} \rangle + \langle \sum_{\nu=1}^{n} \boldsymbol{X}_i^\nu - n\mathbb{E}\boldsymbol{X}, \sum_{\nu=1}^{n} \boldsymbol{Y}_i^\nu - n\mathbb{E}\boldsymbol{X} \rangle \\ &\leq n\|\mathbb{E}\boldsymbol{X}\| \left(\left\| \sum_{\nu=1}^{n} \boldsymbol{X}_i^\nu - n\mathbb{E}\boldsymbol{X} \right\| + \left\| \sum_{\nu=1}^{n} \boldsymbol{Y}_i^\nu - n\mathbb{E}\boldsymbol{X} \right\| \right) + \left\| \sum_{\nu=1}^{n} \boldsymbol{X}_i^\nu - n\mathbb{E}\boldsymbol{X} \right\| \cdot \left\| \sum_{\nu=1}^{n} \boldsymbol{Y}_i^\nu - n\mathbb{E}\boldsymbol{X} \right\| \end{aligned}$$

by the Cauchy-Schwarz inequality. The result ensues from the high probability bounds of Lemma 10, as well as the bound on $\|\mathbb{E}\boldsymbol{X}\|$ from Lemma 7. □

We finally bound the last term, which closes the proof of Proposition 2.

Lemma 11. *Let* $i \in [p]$. *With probability at least* $1 - 2e^{-c\log(n)^2} - e^{-c\log(d)^2}$, *we have*

$$|S_2''| \leq 2\log(d)n\sqrt{d} \tag{68}$$

Proof. Define $\alpha^\nu = \sigma'(\langle \boldsymbol{w}_i^0, \boldsymbol{z}^\nu \rangle) f^\star(\boldsymbol{z}^\nu)$, and β^ν its equivalent for $\boldsymbol{Y}^\nu$. Since α^ν only depends on $\boldsymbol{z}_i$, the distribution of $\sum \boldsymbol{X}_\perp^\nu$ is the same as $\|\alpha\|\boldsymbol{X}_\perp$, where $\boldsymbol{X}_\perp$ is a normal random vector in $V_i^\perp$. Therefore, we have

$$S_2'' \overset{d}{=} \|\alpha\| \cdot \|\beta\| \cdot \langle \boldsymbol{X}_\perp, \boldsymbol{Y}_\perp \rangle$$

for two independent Gaussian vectors $\boldsymbol{X}_\perp, \boldsymbol{Y}_\perp$. Now, both $\|\alpha\|^2$ and $\|\beta\|^2$ are the sum of n sub-exponential random variables with bounded variance, and $\langle \boldsymbol{X}_\perp, \boldsymbol{Y}_\perp \rangle$ is the sum of d such variables. Hence, by Bernstein's inequality, with probability $1 - 2e^{-c \log(n)^2}$,

$$\|\alpha\|^2 \leq n + \log(n)\sqrt{n} \leq 2n \quad \text{and} \quad \|\beta\|^2 \leq 2n,$$

and with probability at least $1 - e^{-c \log(d)^2}$

$$\langle \boldsymbol{X}_\perp, \boldsymbol{Y}_\perp \rangle \leq \log(d)\sqrt{d},$$

which ends the proof. □

C.5 Proof of Theorems 1 and 2

We begin with a proposition that summarizes everything from the two previous sections.

Proposition 3. *Let ℓ be the leap index of $f^\star$, and assume that $n = \Omega(d^{\ell - \delta})$ for some $\delta > 0$. There is an event with probability at least $1 - cpe^{-\log(d)^2}$ such that for $i \in [p]$:*

$$\left\| \boldsymbol{\pi}_i^1 - \left(\boldsymbol{\pi}_i^0 + \frac{\eta a_i}{\sqrt{p}} \, C_\ell^\star \times_{1\ldots(\ell-1)} (\boldsymbol{w}_i^0)^{\otimes(\ell-1)} \right) \right\| = \mathcal{O}\left(\frac{r^{\ell/2} \operatorname{polylog}(d)}{d^{\ell/2}} + \frac{\sqrt{r}\eta \log(d)}{p\sqrt{n}} \right) \tag{69}$$

$$\left\|\boldsymbol{w}_i^1\right\|^2 = \Theta\left(1 + \frac{\eta^2 X_i \left\|\boldsymbol{\pi_i^0}\right\|^2}{p^2} + \frac{\eta^2 d}{np^2} \right) \tag{70}$$

where the X_i are i.i.d random variables as in Lemma 7.

Proof. The proof amounts to checking that all the bounds proven so far are of the right order. The first equality is simply a combination of Lemma 5 and Proposition 2. For the second part, notice that Lemma 7 implies that

$$\mathbb{E}\left[\left\|\boldsymbol{w}_i^1\right\|^2 \right] = \Theta\left(1 + \frac{\eta^2 X_i \left\|\boldsymbol{\pi_i^0}\right\|^2}{p^2} + \frac{\eta^2 d}{np^2} \right),$$

and it is straightforward (albeit tedious) to check that all bounds in Proposition 2 are negligible with respect to the above expectation. □

Proof of Theorem 1 We first consider the case where $n = \Theta(d^{\ell-\delta})$. A simple triangular inequality yields

$$\|\boldsymbol{\pi}_i^1\| = \mathcal{O}\left(\|\boldsymbol{\pi}_i^0\| + \frac{\eta \|\boldsymbol{\pi}_i^0\|^{\ell-1}}{p} \right)$$

where the second part is due to Lemma 7. On the other hand, the middle term in (70) becomes negligible w.r.t the rightmost one, so we get

$$\left\|\boldsymbol{w}_i^1\right\| = \Omega\left(1 + \frac{\eta \, d^{\delta/2}}{p} \right)$$

This implies

$$\frac{\|\boldsymbol{\pi}_i^1\|}{\|\boldsymbol{w}_i^1\|} = \mathcal{O}\left(\max\left(\|\boldsymbol{\pi}_i^0\|, \frac{\left\|\boldsymbol{\pi}_i^0\right\|^{\ell-1}}{d^{\delta/2}} \right) \right) = \mathcal{O}\left(\frac{\operatorname{polylog}(d)}{d^{(1\wedge\delta)/2}} \right) \tag{71}$$

where the last inequality is due to Lemma 3.

Proof of Theorem 2 Now, we take $n = \Omega(d^\ell)$, and $\eta = d^{(\ell-1)/2}$. Then, the bounds of Proposition 3 become

$$\left\| \boldsymbol{\pi}_i^1 - a_i d^{(\ell-1)/2} C_\ell^\star \times_{1\dots(\ell-1)} (\boldsymbol{w}_i^0)^{\otimes(\ell-1)} \right\| = \mathcal{O}\left(\frac{\sqrt{r}\,\mathrm{polylog}(d)}{\sqrt{d}} \right) \quad \text{and} \quad \left\|\boldsymbol{w}_i^1\right\|^2 = \mathcal{O}(1)$$

Hence, the first part of Theorem 2 is straightforward: from Lemma 7,

$$\frac{\|\boldsymbol{\pi}_i^1\|}{\|\boldsymbol{w}_i^1\|} = \Omega\left(a_i^2 X_i \cdot (\sqrt{d}\|\pi_i^0\|)^{\ell-1} \right), \tag{72}$$

which is a random variable with positive expectation. The latter part is not independent from d, but it dominates e.g. a variable of the form $\|\boldsymbol{z}_r\|$ where $\boldsymbol{z}_r \sim \mathcal{N}(0, I_r/2)$ with probability $1 - ce^{-\log(d)^2}$.

For the second part, we write using the higher-order SVD of $C_\ell^\star$

$$C_\ell^\star \times_{1\dots(\ell-1)} (\boldsymbol{w}_i^0)^{\otimes(\ell-1)} = \sum_{j_1,\dots,j_\ell=1}^{r_\ell} S_{j_1,\dots,j_\ell} \langle \boldsymbol{w}_i^0, \boldsymbol{u}_{j_1}^\star \rangle \dots \langle \boldsymbol{w}_i^0, \boldsymbol{u}_{j_{\ell-1}}^\star \rangle \, \boldsymbol{u}_{j_\ell}^\star$$

which belongs to $V_\ell^\star$. Finally, since S is full-rank, each vector $C_\ell^\star \times_{1\dots(\ell-1)} (\boldsymbol{w}_i^0)^{\otimes(\ell-1)}$ is an i.i.d random variable which is absolutely continuous w.r.t the Lebesgue measure in $V_\ell^\star$. This implies that the collection of such vectors is full-rank with probability one, and ends the proof of Theorem 2.

C.6 Spike+Bulk decomposition

Having proven Theorems 1 and 2, we move to investigate the behavior after multiple gradient steps. First, we relate the discussion above to a "spike+noise" decomposition of the gradient. We start from Equation (39):

$$\boldsymbol{g}_i = \frac{a_i}{\sqrt{p}} \cdot \frac{1}{n} \sum_{\nu=1}^{n} \boldsymbol{z}^\nu \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}^\nu \rangle) f^\star(\boldsymbol{z}^\nu) \tag{73}$$

Define $\sigma'_{>1}(u) : \mathbb{R} \to \mathbb{R}$ as the following function:

$$\sigma'_{>1}(u) = \sigma'(u) - \mu_1, \tag{74}$$

so that $\mathbb{E}\left[\sigma'_{>1}(u)\right] = 0$. We have the following decomposition of the gradient:

$$\boldsymbol{g}_i = \frac{a_j}{\sqrt{p}} \frac{1}{n} \mu_1 \sum_{i=1}^{n} y_i \boldsymbol{x}_i + \underbrace{\frac{1}{n} \frac{a_j}{\sqrt{p}} \mu_1 \sum_{i=1}^{n} \sigma'_{>1}(\boldsymbol{x}_i^\top \boldsymbol{w}^0) \boldsymbol{x}_i y_i}_{\Delta_j}, \tag{75}$$

or in matrix form:

$$\boldsymbol{G} = \boldsymbol{u}\boldsymbol{v}^\top + \Delta, \tag{76}$$

where $\boldsymbol{u} = \frac{\mu_1}{\sqrt{p}} \boldsymbol{a}, \boldsymbol{v} = \frac{1}{n} \sum_{i=1}^{n} y_i \boldsymbol{x}_i$. A similar decomposition was utilized in Ba et al. (2022) to provide an asymptotic characterization of the training and generalization errors in the regime $n = \Theta(d)$ and step-size $\eta = \mathcal{O}(\sqrt{p})$. In particular, they show that the presence of this spike for $\eta = \mathcal{O}(\sqrt{p})$ is not enough to go beyond the linear kernel regime.

However, as we see below, it is possible to obtain a precise characterization in the feature learning regime $\eta = \Theta(p)$ and generalizing to multiple steps, with stronger concentration over the structure of Δ. In particular, we prove that Δ effectively acts as uniform noise that can be incorporated into the initialization $\boldsymbol{W}^{(0)}$.

This is expressed through the following Lemma:

Lemma 12. *With high probability over the initialization W^0, as $n, d \to \infty$ with $n = \Omega(\max(p, d))$, the matrix Δ satisfies the following:*

(i) For any $\boldsymbol{v} \in V^\star$, with $\|\boldsymbol{v}\| = 1$, $\langle \Delta_j, \boldsymbol{v} \rangle = \mathcal{O}\left(\frac{\operatorname{polylog}(d)}{p\sqrt{d}}\right)$.

(ii) $\|\Delta\| = \mathcal{O}(\operatorname{polylog} d/\sqrt{d})$.

(iii) For any $i \neq j, i, j \in [p/2]$, $\Delta_j^\top \Delta_i = \mathcal{O}\left(\frac{\operatorname{polylog}(d)}{p^2\sqrt{d}}\right)$,

where we only consider the first half neurons due to the choice of the symmetric initialization in Equation (26).

Proof. Without loss of generality, we assume that $\mu_1 = 0$ and hence that $\Delta_i = \boldsymbol{g}_i$. By Lemma 4, since $\mu_1 = 0$; we have $\mathbb{E}\left[\Delta_j^\top \boldsymbol{v}\right] = \mathcal{O}\left(\frac{\operatorname{polylog}(d)}{p\sqrt{d}}\right)$. Furthermore, from Lemma 8, we obtain that, with high probability:

$$|\Delta_j^\top \boldsymbol{v} - \mathbb{E}\left[\Delta_j^\top \boldsymbol{v}\right]| = \mathcal{O}\left(\frac{\log(n)}{p\sqrt{d}}\right). \tag{77}$$

This proves Part (i). Part (ii) follows from Lemma 14 in Ba et al. (2022).

It remains to show Part (iii). The same proof as in Proposition 2 (Eq. (53)) implies that, with high probability,

$$\langle \boldsymbol{g_i}, \boldsymbol{g_j} \rangle = \mathbb{E}[\langle \boldsymbol{g_i}, \boldsymbol{g_j} \rangle] + \mathcal{O}\left(\frac{\operatorname{polylog}(d)}{p^2\sqrt{d}}\right), \tag{78}$$

and hence we only need to bound the expectation $\mathbb{E}[\langle \boldsymbol{g_i}, \boldsymbol{g_j} \rangle]$. In turn, the decomposition of Equation (47) still holds, and we get

$$\mathbb{E}[\langle \boldsymbol{g_i}, \boldsymbol{g_j} \rangle] \leq \|\mathbb{E}[\boldsymbol{g_i}]\|\|\mathbb{E}[\boldsymbol{g_j}]\| + \frac{1}{np^2}\mathbb{E}\left[\|\boldsymbol{z}\|^2 \sigma'(\langle \boldsymbol{w}_i^0, \boldsymbol{z} \rangle)\sigma'(\langle \boldsymbol{w}_j^0, \boldsymbol{z} \rangle) f^\star(\boldsymbol{z})^2\right] \tag{79}$$

Since $\mu_1 = 0$, the bound of Lemma 7 becomes

$$\|\mathbb{E}[\boldsymbol{g_i}]\| \leq \frac{\pi_i}{p} = \mathcal{O}\left(\frac{\log(d)}{p\sqrt{d}}\right),$$

and it remains to bound the cross term. The main argument is the following lemma, which is the generalization (with an identical proof) of Lemma D.4 in Arnaboldi et al. (2023):

Lemma 13. *Let $N \geq 0$ be fixed, and $f_1, \ldots, f_N$ be a sequence of functions with bounded first and second derivatives. Consider the function on $N \times N$ matrices*

$$F(\Sigma) = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(0,\Sigma)}[f_1(z_1) \ldots f_N(z_N)] \tag{80}$$

Then, for Σ, Σ' two semidefinite positive matrices with unit diagonal, we have

$$|F(\Sigma) - F(\Sigma')| \leq C\|\Sigma - \Sigma'\|_\infty. \tag{81}$$

Now, we first have by the same arguments as in Lemma 7

$$\mathbb{E}\left[\|\boldsymbol{z}\|^2 \sigma'(\langle \boldsymbol{w}_i^0, \boldsymbol{z} \rangle)\sigma'(\langle \boldsymbol{w}_j^0, \boldsymbol{z} \rangle) f^\star(\boldsymbol{z})^2\right] = d\mathbb{E}\left[\sigma'(\langle \boldsymbol{w}_i^0, \boldsymbol{z} \rangle)\sigma'(\langle \boldsymbol{w}_j^0, \boldsymbol{z} \rangle) f^\star(\boldsymbol{z})^2\right] + \mathcal{O}\left(\sqrt{d}\right),$$

so we only to bound the first term of the RHS. Expanding the definition of $f^\star$, the latter is a sum of k^2 terms of the form

$$\mathbb{E}[\sigma'(\langle \boldsymbol{w}_i^0, \boldsymbol{z} \rangle)\sigma'(\langle \boldsymbol{w}_j^0, \boldsymbol{z} \rangle)\sigma_k^\star(\langle \boldsymbol{w}_k^\star, \boldsymbol{z} \rangle)\sigma_{k'}^\star(\langle \boldsymbol{w}_k^\star, \boldsymbol{z} \rangle)],$$

which falls under the framework Lemma 13 for $N = 4$. In particular, since $\mu_1 = 0$, $F(\Sigma) = 0$ whenever we have $\Sigma_{1i} = \Sigma_{2j} = 0$ for $i \neq 1, j \neq 2$. Hence, by an application of Lemma 13, we have

$$\mathbb{E}[\sigma'(\langle \boldsymbol{w}_i^0, \boldsymbol{z}\rangle)\sigma'(\langle \boldsymbol{w}_j^0, \boldsymbol{z}\rangle)\sigma_k^\star(\langle \boldsymbol{w}_k^\star, \boldsymbol{z}\rangle)\sigma_{k'}^\star(\langle \boldsymbol{w}_k^\star, \boldsymbol{z}\rangle)] \leq C \max(\langle \boldsymbol{w}_i^0, \boldsymbol{w}_j^0\rangle, \pi_i, \pi_j) \leq C\frac{\log(d)}{\sqrt{d}}$$

with high probability, which ends the proof. □

We next prove that the norm of $\boldsymbol{w}_i^1$ after the first gradient step posseses a simplified dimension-independent limit:

Lemma 14. *Suppose $n = \Theta(d)$. Then, there exists a constant C, such that for any neuron i, with high-probability as $d \to \infty$, with step-size η:*

$$\|\boldsymbol{w}_i^1\|^2 = 1 + \eta C a_i^2 + \mathcal{O}(\frac{\operatorname{polylog} d}{\sqrt{d}}) \tag{82}$$

Proof. Recall Equation 47:

$$\mathbb{E}\left[\|\boldsymbol{g}_i\|^2\right] = \frac{n(n-1)}{n^2}\|\mathbb{E}[\boldsymbol{g}_i]\|^2 + \frac{a_i^2}{np^2}\mathbb{E}\left[\|\boldsymbol{z}^\nu\|^2\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}^\nu\rangle)^2 f^\star(\boldsymbol{z}^\nu)^2\right] \tag{83}$$

Lemma 7 implies that $\|\mathbb{E}[\boldsymbol{g}_i]\|^2$ is approximately $a_i^2 X_i \cdot \frac{\|\boldsymbol{\pi}_i^0\|^{2(\ell-1)}}{p^2}$ for a random variable X_i. When $\ell = 1$, X_i simply reduces to a constan depending only on g^*. The second term can be decomposed as:

$$\frac{a_i^2}{np^2}\mathbb{E}\left[\|\boldsymbol{z}^\nu\|^2\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}^\nu\rangle)^2 f^\star(\boldsymbol{z}^\nu)^2\right] = \frac{a_i^2}{np^2}\mathbb{E}\left[d\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}^\nu\rangle)^2 f^\star(\boldsymbol{z}^\nu)^2\right] + \mathbb{E}\left[(d - \|\boldsymbol{z}^\nu\|^2)\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}^\nu\rangle)^2 f^\star(\boldsymbol{z}^\nu)^2\right] + \tag{84}$$

Let $m_0^i \in R^r$ denote the vector of overlaps $\langle \boldsymbol{w}_i^0, \boldsymbol{w}_1^*\rangle, \cdots, \langle \boldsymbol{w}_i^0, \boldsymbol{w}_k^*\rangle$ By Holder's inequality, the second term is of order $\mathcal{O}(\frac{1}{\sqrt{d}})$ while through a change of variables, the first term can be expressed as a function of the overlaps $\langle \boldsymbol{w}_i^0, \boldsymbol{w}_j^*\rangle$ for $j \in [r]$:

$$\begin{aligned}\frac{a_i^2}{np^2}\mathbb{E}\left[d\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}^\nu\rangle)^2 f^\star(\boldsymbol{z}^\nu)^2\right] = \frac{a_i^2 d}{np^2}\mathbb{E}\left[\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}^\nu\rangle)^2 f^\star(\boldsymbol{z}^\nu)^2\right] &= \frac{a_i^2 d}{np^2}F_{\sigma,g^*}(M_0) \\ &= \frac{a_i^2 d}{np^2}F_{\sigma,g^*}(0) + \frac{1}{\sqrt{d}}\end{aligned}$$

□

C.7 Second step: Proof Sketch for Theorem 3

Before providing detailed proof of Theorem 3 for general polynomial activation functions, and a general number of steps, we illustrate the essential idea by analyzing the second gradient step. Let $\boldsymbol{Z}^0$ denote the batch of inputs used for the first gradient step. We condition on $\boldsymbol{Z}^0$ and assume that the high-probability events in Lemma 12 hold. We independently sample another batch of n training inputs $\boldsymbol{Z}$ and perform the gradient update:

$$\boldsymbol{g}_j^1 = -\nabla_{\boldsymbol{w}_j}\mathcal{L}\left(\hat{f}(\boldsymbol{z}^\nu; W^1, \boldsymbol{a}), f^\star(\boldsymbol{z}^\nu)\right) \tag{85}$$

However, unlike the first gradient step, the weights $\boldsymbol{w}^1$ are no-longer approximately orthonormal across neurons and contain significant correlation along the teacher subspace.

We have:

$$\boldsymbol{w}_j^1 = \eta \frac{a_j \mu_1}{\sqrt{p}} \boldsymbol{v} + \boldsymbol{w}_j^0 + \Delta_j, \tag{86}$$

where $\boldsymbol{v} = \frac{1}{n} \sum_{i=1}^n y_i \boldsymbol{z}_i$. By theorem 2, we have that the projection of $\boldsymbol{v}$ along the target subspace V^* converges in probability to $C_1(f)$. Let $\boldsymbol{v}^* = C_1(f)$ We show that the alignment of $\boldsymbol{v}$ along $\boldsymbol{v}^*$ affects the components of the second gradient step along the teacher subspace, allowing the gradient to be sensitive to directions linearly coupled with $\boldsymbol{v}^*$ in the target function.

We proceed by analyzing the projection of the above update along a direction in the teacher subspace. Let $\boldsymbol{v}_j = P_{V^\star}(\boldsymbol{w}_j^1)$ and consider the decomposition $\boldsymbol{w}_j^1 = \boldsymbol{v}_j + P_{V^\star}^{\perp}(\boldsymbol{w}_j^1)$. We further have from Lemma 14 that $\|P_{V^\star}^{\perp}(\boldsymbol{w}_j^1)\|_2^2$ concentrates to a positive bounded value c_j depending only on a_j. For each sample, $\boldsymbol{z}_i$ let $\kappa_i = \langle \boldsymbol{v}_j, \boldsymbol{z}_i \rangle$ denote the projection along the "signal" $\boldsymbol{v}_j$.

We now introduce the following function for $z \in \mathbb{R}$:

$$\sigma_{\kappa,j}(z) = \sigma(c_j z + \eta \kappa). \tag{87}$$

Define $\mu_{1,\kappa,j} = \mathbb{E}_{z \sim \mathcal{N}(0,1)} \left[\sigma_{\kappa,j}(z) z \right]$. Further, let:

$$\sigma'_{>1,\kappa}(u) = \sigma'_{\kappa,j}(u) - \mu_{1,\kappa,j}. \tag{88}$$

From Lemma 5, we have that $P_{V^\star}(\boldsymbol{v}) \xrightarrow{\mathbb{P}} \boldsymbol{v}^\star$. Now, let $\boldsymbol{u} \in V^\star$ be a direction in the teacher subspace orthogonal to $\boldsymbol{v}^\star$. Using, equation (42), we have:

$$\begin{aligned} \mathbb{E}\left[\langle \boldsymbol{u}, \boldsymbol{g}_j^1 \rangle\right] &= \mathbb{E}\left[(f^\star(\boldsymbol{z}) - \hat{f}(\boldsymbol{z}, \boldsymbol{W}^1, \boldsymbol{a}))\sigma'(\langle \boldsymbol{z}, \boldsymbol{w}_j^1 \rangle)_j(\langle \boldsymbol{z}, \boldsymbol{u} \rangle) \right] \\ &= \mathbb{E}\left[(f^\star(\boldsymbol{z}) \mu_{1,\langle \boldsymbol{z}, \boldsymbol{v}_j \rangle}(\langle \boldsymbol{z}_i, \boldsymbol{u} \rangle) \right] - \mathbb{E}\left[\hat{f}(\boldsymbol{z}, \boldsymbol{W}^1, \boldsymbol{a})\sigma'(\langle \boldsymbol{z}_i, \boldsymbol{w}^1 \rangle)_j(\langle \boldsymbol{z}_i, \boldsymbol{u} \rangle) \right] \end{aligned} \tag{89}$$

Where in the first term we took the expectation over $P_{V^\star}^{\perp}(\boldsymbol{w}^1)$ since it is orthogonal to the teacher-subspace.

The second term can be expressed as:

$$\langle (\mathbb{E}\left[\hat{f}(\boldsymbol{z}, \boldsymbol{W}^1, \boldsymbol{a})\sigma'(\langle \boldsymbol{z}, \boldsymbol{w}^1 \rangle_j \boldsymbol{z}_i \right]), \boldsymbol{u} \rangle \tag{90}$$

We have that $\hat{f}(\boldsymbol{z}, \boldsymbol{W}^1, \boldsymbol{a})$ depends only on the directions $\boldsymbol{w}_1^1, \cdots, \boldsymbol{w}_p^1$. By Lemma 12, each of the directions, satisfies $\langle \boldsymbol{w}_i, \boldsymbol{u} \rangle = \mathcal{O}(\frac{\text{polylog}(d)}{p\sqrt{d}})$. Furthermore, one can show that $(\mathbb{E}\left[\hat{f}(\boldsymbol{z}, \boldsymbol{W}^1, \boldsymbol{a})\sigma'(\langle \boldsymbol{z}, \boldsymbol{w}^1 \rangle)_j \boldsymbol{z}_i \right]$ lies in the span of $\boldsymbol{w}_1^1, \cdots, \boldsymbol{w}_p^1$. Therefore, $\mathbb{E}\left[\hat{f}(\boldsymbol{z}, \boldsymbol{W}^1, \boldsymbol{a})\sigma'(\langle \boldsymbol{z}_i, \boldsymbol{w}^1 \rangle)_j \boldsymbol{z}_i \right])^\top \boldsymbol{u} \xrightarrow{d \to \infty} 0$.

Now, consider the first term i.e $\mathbb{E}\left[f^\star(\boldsymbol{z}) \mu_{1,\langle \boldsymbol{z}, \boldsymbol{v}_j \rangle}(\langle \boldsymbol{z}, \boldsymbol{u} \rangle) \right]$. Let $\boldsymbol{v}^\star, \boldsymbol{u}, \boldsymbol{u}'_1 \cdots, \boldsymbol{u}'_{d-2}$ be an orthonormal basis of $\mathbb{R}^d$. Without loss of generality, assume that $\boldsymbol{v}^\star, \boldsymbol{u}, \cdots, \boldsymbol{u}'_{r-2}$. span the teacher subspace $V^\star$. We express y using the product Hermite decomposition under the above basis:

$$y = f^\star(\boldsymbol{z}) = \sum_{j_1, \cdots, j_r = 1}^{\infty} \frac{c^\star_{j_1, \cdots, j_r}}{j_1! j_2! \cdots j_r!} \operatorname{He}_{j_1}(\langle \boldsymbol{v}^\star, \boldsymbol{z} \rangle) \operatorname{He}_{j_2}(\langle \boldsymbol{u}, \boldsymbol{z} \rangle) \cdots \operatorname{He}_{j_r}(\langle \boldsymbol{u}_{r-2}, \boldsymbol{z} \rangle). \tag{91}$$

Since $\boldsymbol{v}_j \xrightarrow{\mathbb{P}} \boldsymbol{v}^\star$ and $\boldsymbol{u} \perp \boldsymbol{u}'_1, \cdots, \boldsymbol{u}'_{r-2}$, only the terms of the form $\operatorname{He}_{j_1}(\langle (\boldsymbol{v}^\star), \boldsymbol{z} \rangle) \operatorname{He}_{j_2}(\langle \boldsymbol{u}, \boldsymbol{z} \rangle)$ contribute to the expectation $\mathbb{E}\left[y \mu_{1,\langle \boldsymbol{z}, \boldsymbol{v}_j \rangle}(\langle \boldsymbol{z}, \boldsymbol{u} \rangle) \right]$ in the limit $d \to \infty$. Consider the contribution of one such term:

$$\mathbb{E}\left[\operatorname{He}_{j_1}(\langle \boldsymbol{v}^\star, \boldsymbol{z} \rangle) \operatorname{He}_{j_2}(\langle \boldsymbol{u}, \boldsymbol{z} \rangle) \mu_{1,\langle \boldsymbol{z}, \boldsymbol{v}_j \rangle, j} \langle \boldsymbol{z}, \boldsymbol{u} \rangle \right] \to \mathbb{E}\left[\operatorname{He}_{j_1}(\langle \boldsymbol{v}^\star, \boldsymbol{z} \rangle) \operatorname{He}_{j_2}(\langle \boldsymbol{u}, \boldsymbol{z} \rangle) \mu_{1,\langle \boldsymbol{z}, \boldsymbol{v}^\star \rangle, j} \langle \boldsymbol{z}, \boldsymbol{u} \rangle \right] \tag{92}$$

Suppose $j_2 \neq 1$, then $\mathbb{E}\left[\mathrm{He}_{j_2}(\langle \boldsymbol{u}, \boldsymbol{z}\rangle)\langle \boldsymbol{z}_i, \boldsymbol{u}\rangle\right] = 0$. Therefore, the non-zero contributions arise from terms of the form $\mathrm{He}_{j_1}(\langle \boldsymbol{v}^\star, \boldsymbol{z}\rangle)\langle \boldsymbol{u}, \boldsymbol{z}\rangle$. It can be checked that directions $\boldsymbol{u}$ having non-zero terms of this form span $U_2^\star$ as defined in Theorem 3. However, in general, the RHS of equation 92 might be 0 for some choices of σ and a_j. Moreover, such non-zero contributions might cancel each other for a chosen direction in $U_2^\star$. Furthermore, to obtain high-probability result on the alignment along $U_t^\star$ for a general number of t steps, one needs to quantitatively propagate the expectations and concentration bounds on the projections and norms of W, and show that the magnitude of the projections can be bounded independent of the dimension. We tackle these issues in the next section and provide a full proof of Theorem 3.

C.8 Proof of Theorem 3

The proof proceeds by induction on the number of time-steps t. To avoid certain degeneracy conditions in the proof, we restrict ourselves to polynomial activations. Let $U_t^\star$ be the learned subspace at time-step t according to the definition 2.

Let $Q_t \in \mathbb{R}^{p\times p}$ denote the overlap matrix for weights of the first-layer neurons at time t, i.e. $Q^t_{i,j} = \langle \boldsymbol{w}_i^t, \boldsymbol{w}_j^t\rangle \; \forall i, j \in [p]$. Let $M_t \in \mathbb{R}^{r\times p}$ denote the target-network overlap matrix i.e .$M^t_{i,j} = \langle \boldsymbol{w}_i^\star, \boldsymbol{w}_j^t\rangle \; \forall i \in [p], j \in [r]$. Let $\boldsymbol{W}^* \in \mathbb{R}^{r\times d}$ denote the matrix with rows $\boldsymbol{w}_{\boldsymbol{1}}^\star, \cdots, \boldsymbol{w}_{\boldsymbol{r}}^\star$.

We denote by $\boldsymbol{Z}_t$, the batch of input sampled at time $t \in [T]$. By assumption $\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_T$ are independent. Let $\mathcal{F}_t$ denote the natural filtration associated to $\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_T$, i.e $\mathcal{F}_t$ is the σ-algebra generated by $\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_t$, and let μ_t denote the corresponding joint-measure of $\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_t$. We let $\boldsymbol{g}_i^t$ denote the gradient for the i_{th} neuron at time t obtained using the batch $\boldsymbol{Z}^{t+1}$.

For any time t, let r_t denote the dimension of U_t^* and let $\boldsymbol{W}_t^* \in \mathbb{R}^{r_t\times d}$ denote a matrix with rows forming a basis of U_t^*, such that $(\boldsymbol{W}^*)\top \boldsymbol{W}_t^*$ is independent of d, n. Thus, $\boldsymbol{W}_t^*$ represents a dimension independent basis of U_t^*. Let $\boldsymbol{v}_{j,\boldsymbol{a}} \in \mathbb{R}^{r_t}$ denote the projections of $\boldsymbol{w}_j^t$ along $\boldsymbol{W}_t^*$ i.e $\boldsymbol{v}_{j,\boldsymbol{a}} = \boldsymbol{W}_t^* \boldsymbol{w}_j^t$. Similarly, for an input $\boldsymbol{z} \in \mathbb{R}^d$, we denote the projection of $\boldsymbol{z}$ along $\boldsymbol{W}_t^*$ by $\kappa = \boldsymbol{W}_t^* \boldsymbol{z}$. In what follows, we shall say that a sequence of events $\mathcal{E}_n$ occurs with high-probability as $n, d \to \infty$ if there exist constants $c, C > 0$ such that $\mathbb{P}(\mathcal{E}_n) \geq 1 - Cpe^{-c\log(n)^2} + Cpe^{-c\log(d)^2}$

At any timestep $t \geq 1$, we prove that the following statements hold with high probability w.r.t μ_t:

(i) $Q^t = \tilde{Q}^t_{\boldsymbol{a}} + \mathcal{O}(\frac{\text{polylog}d}{\sqrt{d}})$, $M^t = \tilde{M}^t_{\boldsymbol{a}} + \mathcal{O}(\frac{\text{polylog}d}{\sqrt{d}})$, where $\tilde{Q}^t_{\boldsymbol{a}}, \tilde{M}^t_{\boldsymbol{a}}$ denote dimension-independent matrices with each entry being a polynomial dependent only on $\boldsymbol{a}, t$ of $\boldsymbol{w}_i^t$, dependent on the second layer $i, \boldsymbol{a}$.

(ii) Let $\boldsymbol{v} \in U_t^\star$, with $\|\boldsymbol{v}\| = 1$ be arbitrary. Denote by $\boldsymbol{v}^m \in \mathbb{R}^k$, the components of $\boldsymbol{v}$ along $\boldsymbol{w}_{\boldsymbol{1}}^\star, \cdots, \boldsymbol{w}_{\boldsymbol{r}}^\star$ i.e $\boldsymbol{v}^m = \boldsymbol{W}^\star \boldsymbol{v}$. Then there exists an almost surely positive random variable $q_{t,\boldsymbol{v}^m,\boldsymbol{a}}$, independent of d, n such that $\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle) = q_{t,\boldsymbol{v},\boldsymbol{a}} + O(\frac{\text{polylog}}{\sqrt{d}})$. Furthermore, $q_{t,\boldsymbol{v},\boldsymbol{a}}$ is a non-constant polynomial in $\boldsymbol{a}, v_1, \cdots, v_k$.

(iii) For any $\boldsymbol{v} \in {U^\perp}_t^\star \cap V^\star$, $|\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle| = O(\frac{\text{polylog}}{\sqrt{d}})$, with high probability, for all $i \in [p]$.

Proof. We proceed by induction over t. Suppose that the statements hold at some timestep t. We start by proving that (i) holds at time $t+1$ in expectation:

Lemma 15. *$\mathbb{E}\left[Q_t\right] = \tilde{Q}^t_{\boldsymbol{a}} + \mathcal{O}(\frac{\text{polylog}d}{\sqrt{d}})$ and $\mathbb{E}\left[M_t\right] = \tilde{M}^t_{\boldsymbol{a}} + \mathcal{O}(\frac{\text{polylog}d}{\sqrt{d}})$ where each entry of $\tilde{Q}^t_{\boldsymbol{a}}, \tilde{M}^t_{\boldsymbol{a}}$ is a polynomial of $\boldsymbol{a}$ with degree independent of d, n.*

Proof. Recall that:

$$\begin{aligned}
Q^{t+1}_{i,j} &= \langle \boldsymbol{w}^{t+1}_i, \boldsymbol{w}^{t+1}_j \rangle \\
&= Q^t_{i,j} + \eta \langle \boldsymbol{g}^t_i, \boldsymbol{w}^t_j \rangle + \eta \langle \boldsymbol{w}^t_i, \boldsymbol{g}^t_j \rangle + \eta^2 \langle \boldsymbol{g}^t_i, \boldsymbol{g}^t_j \rangle. \\
M^{t+1}_{i,j} &= \langle \boldsymbol{w}^\star_i, \boldsymbol{w}^{t+1}_j \rangle \\
&= M^t_{i,j} + \eta \langle \boldsymbol{w}^\star_i, \boldsymbol{g}^t_j \rangle
\end{aligned} \tag{93}$$

By the induction hypothesis, the entries of $\mathbb{E}\left[Q^t\right], \mathbb{E}\left[M^t\right]$ converge with high-probability to polynomial limits with error $\mathcal{O}\left(\frac{\operatorname{polylog} d}{d}\right)$. Therefore, it suffices to show that $\mathbb{E}\left[\langle \boldsymbol{g}^t_i, \boldsymbol{w}^t_j \rangle\right], \mathbb{E}\left[\langle \boldsymbol{w}^t_i, \boldsymbol{w}^t_j \rangle\right], \mathbb{E}\left[\langle \boldsymbol{g}^t_i, \boldsymbol{g}^t_j \rangle\right]$ converge to dimension-inpedendent polynomial limits. First, consider the case $i = j$. We have, analogous to Equation 47

$$\mathbb{E}\left[\|\boldsymbol{g}^t_i\|^2\right] = \underbrace{\frac{1}{np^2}\mathbb{E}\left[\|\boldsymbol{z}^\nu\|^2 \sigma'(\langle \boldsymbol{w}^t_i, \boldsymbol{z}^\nu \rangle)^2 (f^\star(\boldsymbol{z}^\nu) - \hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a}))^2\right]}_{T_1} + \underbrace{\frac{n(n-1)}{n^2}\|\mathbb{E}[\boldsymbol{g}_i]\|^2}_{T_2} \tag{94}$$

The first term T_1 can be decomposed as follows:

$$\begin{aligned}
\frac{1}{np^2}\mathbb{E}\left[\|\boldsymbol{z}^\nu\|^2 \sigma'(\langle \boldsymbol{w}^t_i, \boldsymbol{z}^\nu \rangle)^2 (f^\star(\boldsymbol{z}^\nu) - \hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a}))^2\right] &= \frac{d}{np^2}\mathbb{E}\left[\sigma'(\langle \boldsymbol{w}^t_i, \boldsymbol{z}^\nu \rangle)^2 (f^\star(\boldsymbol{z}^\nu) - \hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a}))^2\right] \\
&+ \frac{1}{np^2}\mathbb{E}\left[(d - \|\boldsymbol{z}^\nu\|^2)\sigma'(\langle \boldsymbol{w}^t_i, \boldsymbol{z}^\nu \rangle)^2 (f^\star(\boldsymbol{z}^\nu) - \hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a}))^2\right].
\end{aligned}$$

Similar to equation 50, Holder's inequality implies that conditioned on the event in $\mathcal{F}_t$ of Q_t, M_t being bounded independent of d, n, the second term is of order $\mathcal{O}(\frac{\sqrt{d}}{n}) = \mathcal{O}(\frac{1}{\sqrt{d}})$. Consider the first term, conditioned on $\mathcal{F}_t$.

$$\frac{d}{np^2}\mathbb{E}\left[\sigma'(\langle \boldsymbol{w}^t_i, \boldsymbol{z}^\nu \rangle)^2 (f^\star(\boldsymbol{z}^\nu) - \hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a}))^2 | \mathcal{F}_t\right] \tag{95}$$

By assumption, $\frac{d}{np^2} = \frac{\alpha}{p^2}$ for some constant α. Therefore, by definition of $f^\star(\boldsymbol{z}^\nu)$ and $\hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a})$, the term inside the expectation only depends on the overlaps of $\boldsymbol{z}^\nu$ with the neurons and teacher subspace i.e $\langle \boldsymbol{w}^t_1, \boldsymbol{z}^\nu \rangle, \cdots, \langle \boldsymbol{w}^t_p, \boldsymbol{z}^\nu \rangle$ $\langle \boldsymbol{w}^\star_1, \boldsymbol{z}^\nu, \rangle, \cdots, \langle \boldsymbol{w}^\star_k, \boldsymbol{z}^\nu \rangle$. By a change of variables the above term can therefore be expressed as an expectation w.r.t the $k + j$ correlated variables corresponding to the above overlaps.

Concretely, we have:

$$\frac{d}{np^2}\mathbb{E}\left[\sigma'(\langle \boldsymbol{w}^t_i, \boldsymbol{z}^\nu \rangle)^2 (f^\star(\boldsymbol{z}^\nu) - \hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a}))^2 | \mathcal{F}_t\right] = F_g(Q_t, M_t), \tag{96}$$

for some function $F : \mathbb{R} \to \mathbb{R}$

Lemma 16. *F_g is a polynomial in Q_t, M_t independent of d, n.*

Proof. By assumption, σ' and $f^\star$ are polynomials in $\langle \boldsymbol{w}^t_1, \boldsymbol{z}^\nu \rangle, \cdots, \langle \boldsymbol{w}^t_p, \boldsymbol{z}^\nu \rangle$ and $\langle \boldsymbol{w}^\star_1, \boldsymbol{z}^\nu \rangle, \cdots, \langle \boldsymbol{w}^\star_k, \boldsymbol{z}^\nu \rangle$ respectively. Therefore, $\sigma'(\langle \boldsymbol{w}^t_i, \boldsymbol{z}^\nu \rangle)^2 (f^\star(\boldsymbol{z}^\nu) - \hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a}))^2$ is a polynomial in the zero mean correlated Gaussian variables $\langle \boldsymbol{w}^t_1, \boldsymbol{z}^\nu \rangle, \cdots, \langle \boldsymbol{w}^t_p, \boldsymbol{z}^\nu \rangle, \langle \boldsymbol{w}^\star_1, \boldsymbol{z}^\nu \rangle, \cdots, \langle \boldsymbol{w}^\star_k, \boldsymbol{z}^\nu \rangle$. Therefore, by Wick's theorem, F_g is a polynomial in Q_t, M_t. □

By the induction hypothesis, with high-probability, $Q_t = \tilde{Q}_{t,\boldsymbol{a}} + \mathcal{O}(\frac{\text{polylog}\, d}{\sqrt{d}})$ and $\tilde{M}_{t,\boldsymbol{a}} + \mathcal{O}(\frac{\text{polylog}\, d}{\sqrt{d}})$, where $\tilde{Q}_{t,\boldsymbol{a}}, \tilde{M}_{t,\boldsymbol{a}}$ denote deterministic matriceswith entries being polynomial functions of $\boldsymbol{a}$. By propagating the errors through the polynomial F_g, we obtain that $F_g(Q_t, M_t) = F_g(\tilde{Q}_{t,\boldsymbol{a}}, \tilde{M}_{t,\boldsymbol{a}}) + \mathcal{O}(\frac{\text{polylog}\, d}{\sqrt{d}})$.

Next, consider the term T_2 in Equation 94. By repeatedly applying Stein's Lemma w.r.t terms $\langle \boldsymbol{w}_i^t, \boldsymbol{z} \rangle$ for $i \in [p]$ and $\langle \boldsymbol{w}^*, \boldsymbol{z} \rangle$ for $j \in [r]$, analogous to Lemma 4, $\mathbb{E}[\boldsymbol{g}_i]$, can be expressed as a linear combination of $\boldsymbol{w}_1^t, \cdots, \boldsymbol{w}_p^t$ and $\boldsymbol{w}_1^*, \cdots, \boldsymbol{w}_r^*$. Furthermore, by Wick's theorem, the coefficients are independent of d and polynomial in Q_t, M_t. Therefore, propagating errors from time t, T_2 can be approximated by polynomials in Q_t, M_t with error $\mathcal{O}(\frac{\text{polylog}\, d}{\sqrt{d}})$

Similary, the terms $\mathbb{E}\left[\langle \boldsymbol{g}_i^t, \boldsymbol{g}_j^t \rangle\right], \mathbb{E}\left[\langle \boldsymbol{w}_i^*, \boldsymbol{g}_j^t \rangle\right]$ converge with high-probability to dimension-independent polynomials in Q_t, M_t □

Next, we prove that (ii) and (iii) holds in expectation:

Lemma 17. *Let $\boldsymbol{v} \in V^\star$, with $\|\boldsymbol{v}\| = 1$ be arbitrary with components $\boldsymbol{v}^m \in \mathbb{R}^r$ along $\boldsymbol{w}_1^\star, \cdots, \boldsymbol{w}_r^\star$, then $\mathbb{E}\left[\langle \boldsymbol{v}, \boldsymbol{g}_j^t \rangle\right] = h(\boldsymbol{v}^m, \boldsymbol{a}, Q_t, M_t) + \mathcal{O}(\frac{\text{polylog}\, d}{p\sqrt{d}})$, where $h(\boldsymbol{v}^m, \boldsymbol{a}, Q_t, M_t)$ satisfies:*

(i) $h(\boldsymbol{v}^m, \boldsymbol{a}, Q_t, M_t)$ is non-zero, almost surely over $\boldsymbol{a}$ if $\boldsymbol{v} \in U_{t+1}^\star$.

(ii) $h(\boldsymbol{v}^m, \boldsymbol{a}, Q_t, M_t) = 0$ otherwise.

Consider the gradient w.r.t the j_{th} neuron's parameters:

$$\boldsymbol{g}_j^t = -\nabla_{\boldsymbol{w}_j} \mathcal{L}\left(\hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a}), f^\star(\boldsymbol{z}^\nu)\right) = \frac{1}{n} a_j \sum_{\nu=1}^{n} \boldsymbol{z}^\nu (f^\star(\boldsymbol{z}^\nu) - \hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a})) \sigma'(\langle \boldsymbol{z}^\nu, \boldsymbol{w}_j^t \rangle) \tag{97}$$

□

Suppose that $\boldsymbol{v} \in U_{t+1}^\star \cap (U_t^\star)^\perp$ i.e when $\boldsymbol{v}$ is a new direction not yet learned upto time t.Using, equation (97), the expectation $\mathbb{E}\left[\langle \boldsymbol{v}, \boldsymbol{g}_j^t \rangle\right]$ can be expressed as:

$$\mathbb{E}\left[\langle \boldsymbol{v}, \boldsymbol{g}_j^t \rangle\right] = \mathbb{E}\left[a_j (f^\star(\boldsymbol{z}) - \hat{f}(\boldsymbol{z}; \boldsymbol{W}^t, \boldsymbol{a})) \sigma'(\langle \boldsymbol{z}, \boldsymbol{w}_j^t \rangle) \langle \boldsymbol{z}, \boldsymbol{v} \rangle\right]. \tag{98}$$

We first consider the term $\mathbb{E}\left[\hat{f}(\boldsymbol{z}; \boldsymbol{W}^t, \boldsymbol{a})) \sigma'(\langle \boldsymbol{z}, \boldsymbol{w}_j^t \rangle) \langle \boldsymbol{z}, \boldsymbol{v} \rangle\right]$. Through a change of variables, and Wick's theorem, one obtains that $\mathbb{E}\left[\hat{f}(\boldsymbol{z}; \boldsymbol{W}^t, \boldsymbol{a}) \sigma'(\langle \boldsymbol{z}, \boldsymbol{w}^1 \rangle)_j \langle \boldsymbol{z}, \boldsymbol{v} \rangle\right]$ is a polynomial in Q and the overlaps $\langle \boldsymbol{w}^i, \boldsymbol{v} \rangle$ for $i \in [p]$ having value 0 when $\langle \boldsymbol{w}^i, \boldsymbol{v} \rangle = 0$ for all $i \in [p]$. By the induction hypothesis, $\langle \boldsymbol{w}^i, \boldsymbol{v} \rangle = \mathcal{O}(\frac{\text{polylog}\, d}{\sqrt{d}})$ with high probability. Therefore $\mathbb{E}\left[\hat{f}(\boldsymbol{z}; \boldsymbol{W}^t, \boldsymbol{a})) \sigma'(\langle \boldsymbol{z}, \boldsymbol{w}_j^t \rangle) \langle \boldsymbol{z}, \boldsymbol{v} \rangle | \mathcal{F}_t\right] = \mathcal{O}(\frac{\text{polylog}\, d}{\sqrt{d}})$ with high probability. Similarly, $\mathbb{E}\left[\hat{f}(\boldsymbol{z}; \boldsymbol{W}^t, \boldsymbol{a})) \sigma'(\langle \boldsymbol{z}, \boldsymbol{w}_j^t \rangle) \langle \boldsymbol{z}, \boldsymbol{v} \rangle | \mathcal{F}_t\right] = \mathcal{O}(\frac{\text{polylog}\, d}{\sqrt{d}})$ holds when $\boldsymbol{v} \notin U_{t+1}^\star$.

Now, consider the term $\mathbb{E}\left[a_j f^\star(\boldsymbol{z}) \sigma'(\langle \boldsymbol{z} \cdot \boldsymbol{w}_j^t \rangle) \langle \boldsymbol{z}, \boldsymbol{v} \rangle\right]$. First, using Fubini's theorem, we take the expectation w.r.t the component $\boldsymbol{z}^\perp$ of z in $V^{\star\perp}$.

Recall that $\boldsymbol{v}_{j,\boldsymbol{a}} = \boldsymbol{W}_t^* \boldsymbol{w}_j^t$ and $\kappa = \boldsymbol{W}_t^* \boldsymbol{z}$. The resulting expectation converges in probability to a function of κ:

$$\mathbb{E}_{\boldsymbol{z}^\perp}\left[a_j f^\star(\boldsymbol{z}) \sigma'(\langle \boldsymbol{z}, \boldsymbol{w}_j^t \rangle) \langle \boldsymbol{z}, \boldsymbol{v} \rangle\right] = \mathbb{E}_\kappa\left[a_j f^\star(\boldsymbol{z}) f_1(\boldsymbol{a}, \kappa) \langle \boldsymbol{z}, \boldsymbol{v} \rangle\right], \tag{99}$$

where $f_1(\boldsymbol{a}, \kappa)$ is defined as follows:

$$\begin{aligned} f_1(\boldsymbol{a}, \kappa) &= \mathbb{E}_{\boldsymbol{z}^\perp}\left[\sigma'(\boldsymbol{z}^\top \boldsymbol{w}_j^t)\right] \\ &= \mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma'(c_{j,\boldsymbol{a}}u + \langle \kappa, \boldsymbol{v}_{j,\boldsymbol{a}}\rangle)\right] \end{aligned}$$

, where $c_{j,\boldsymbol{a}}$ denotes the norm of $\boldsymbol{w}_j^t$ along the orthogonal complement of V^*. $f_1(\boldsymbol{a}, \kappa)$ generalizes the "shifted-hermite" $\mu_{1,\kappa,j}$ defined in the section C.7. By assumption on σ, $\sigma'(c_{j,\boldsymbol{a}}u + \langle \kappa, \boldsymbol{v}_{j,\boldsymbol{a}}\rangle)$ is a polynomial in $c_{j,\boldsymbol{a}}u$, $\langle \kappa, \boldsymbol{v}_{j,\boldsymbol{a}}\rangle$. Furthermore, only the odd terms in $c_{j,\boldsymbol{a}}u$ are zero in expectation $u \sim \mathcal{N}(0,1)$. Therefore, $f_1(\boldsymbol{a}, \kappa)$ is a polynomial in $\langle \kappa, \boldsymbol{v}_{j,\boldsymbol{a}}\rangle$ and $c_{j,\boldsymbol{a}}$ with only even degree terms in $c_{j,\boldsymbol{a}}$. By the induction hypothesis, $c_{j,\boldsymbol{a}}^2$ converges in probability to a polynomial in $\boldsymbol{a}$. Therefore $f_1(\boldsymbol{a}, \kappa)$ is a polynomial in $\boldsymbol{a}, \kappa$.

Subsequently, we consider the expectation w.r.t $\langle \boldsymbol{z}, \boldsymbol{v}\rangle$, at a fixed value of κ. Define the following function of κ:

$$f_2(\kappa) = \mathbb{E}_{\langle \boldsymbol{z}, \boldsymbol{v}\rangle}\left[y\langle \boldsymbol{z}, \boldsymbol{v}\rangle | \kappa\right]. \tag{100}$$

Using the tower law of expectation, we obtain:

$$\mathbb{E}\left[a_j f^\star(\boldsymbol{z})\sigma'(\boldsymbol{z}^\top \boldsymbol{w}_j^t)\langle \boldsymbol{z}, \boldsymbol{v}\rangle\right] = \mathbb{E}_\kappa\left[f_1(a_j, \kappa) f_2(\kappa)\right], \tag{101}$$

When $\boldsymbol{v} \notin U_{t+1}^\star$, $f_2(\kappa)$ is identically 0 and the above expectation vanishes.

We aim to show that the above expectation does not vanish except for a_i belonging to a zero-measure set. By the definition of subspace conditioning (definition 2), $\exists \kappa > 0$ such that $\mathbb{E}_{\langle \boldsymbol{z}, \boldsymbol{v}\rangle}\left[f^\star(\boldsymbol{z})\boldsymbol{z}|\kappa\right]$ has non-zero overlap with $\boldsymbol{v}$.

Therefore, $f_2(\kappa)$ is not identically zero. Furthermore, since $f^\star$ is a polynomial by assumption, and $\boldsymbol{v} \perp V^\star$, a rotation of basis implies that f_2 is a polynomial in κ. Let $\mathcal{S}_{y,t}$ be the set of degrees $s \in \mathbb{N}_0$ such that $\mathbb{E}_{f_2(\kappa)\kappa^s}\left[\kappa\right] \neq 0$. Since f_2 is not identically 0, we have that $\mathcal{S}_{y,t} \neq \phi$.

Now, recall that:

$$\begin{aligned} f_1(\boldsymbol{a}, \kappa) &= \mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma'(c_{j,\boldsymbol{a}}u + \langle \kappa, \boldsymbol{v}_{j,\boldsymbol{a}}\rangle)\right] \\ &= \mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sum_{r=0}^{\deg(\sigma)-1}(r+1)b_{r+1}(c_{j,\boldsymbol{a}}u + \langle \kappa, \boldsymbol{v}_{j,\boldsymbol{a}}\rangle)^r\right] \\ &= \sum_{r=0}^{\deg(\sigma)-1}(r+1)b_{r+1}\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[(c_{j,\boldsymbol{a}}u + \langle \kappa, \boldsymbol{v}_{j,\boldsymbol{a}}\rangle)^r\right]. \end{aligned}$$

Now, let $s \in \mathcal{S}_{y,t}$ be arbitrary. By assumption, $\deg(\sigma) - 1 \geq s$. Let $p_s(\boldsymbol{a})$ denote the coefficient of κ^s in $f_1(\boldsymbol{a}, \kappa)$. Since $c_{j,\boldsymbol{a}}, \boldsymbol{v}_{j,\boldsymbol{a}}$ are non-constant polynomials in $\boldsymbol{a}$, the coefficient of κ^s in $(c_{j,\boldsymbol{a}}u + \langle \kappa, \boldsymbol{v}_{j,\boldsymbol{a}}\rangle)^r$ is a non-constant polynomial in $\boldsymbol{a}$ for any r such that $r - q$ is even. Furthermore, the degree of the coefficient of κ^s in $(c_{j,\boldsymbol{a}}u + \langle \kappa, \boldsymbol{v}_{j,\boldsymbol{a}}\rangle)^r$ is strictly increasing in r. Therefore, for any $s \in \mathcal{S}_{y,t}$, $p_s(\boldsymbol{a})$ is a non-constant polynomial in $\boldsymbol{a}$. Now, consider the term in $p_s(\boldsymbol{a})$ with the least degree in a_j. From the definition of $\boldsymbol{v}_{j,\boldsymbol{a}}$, we have that $\boldsymbol{v}_{j,\boldsymbol{a}} = 0$ whenever $a_j = 0$. Let d_j denote the least $s \in \mathbb{N}_0$ such that the coefficient of a_j^s in $\langle \boldsymbol{v}_{j,\boldsymbol{a}}, \kappa\rangle$ is non-zero. We have that $d_j > 0$. Consequently, the minimum degree of a_j in $(c_{j,\boldsymbol{a}})^q(\langle \kappa, \boldsymbol{v}_{j,\boldsymbol{a}}\rangle)^s$, is $(d_j)^s$ for any q. Therefore, the minimum degree of $p_s(\boldsymbol{a})$ is strictly increasing in s. This implies that $p_s(\boldsymbol{a})$ are linearly independent for $s = 1, \cdots, \deg(\sigma) - 1$.

Now, consider the function defined above in Equation 101:

$$h(t, \boldsymbol{a}) = \mathbb{E}_\kappa\left[f_1(\boldsymbol{a}, \kappa) f_2(\kappa)\right]. \tag{102}$$

By expanding f_1, f_2 along κ, the coefficient of κ^s for each $s \in \mathcal{S}_{y,t}$ results in a non-constant polynomial in $\boldsymbol{a}$. We obtain:

$$h(t, \boldsymbol{a}) = \sum_{s \in \mathcal{S}_{y,t}} c_s p_s(\boldsymbol{a}), \tag{103}$$

where c_s denote constants independent of d, n. Therefore, we have that $h(\boldsymbol{a})$ is a non-constant polynomial in $\boldsymbol{a}$. Using Fubini's theorem, we have that the set of zeros of non-zero multivariate polynomials has 0 measure w.r.t the Lebesque measure (for a generalization, see (Mityagin, 2020)), we obtain that $q(\boldsymbol{a}) \neq 0$ almost surely.

Now, suppose that $\boldsymbol{v} \in (U_t^\star)$, i.e when $\boldsymbol{v}$ is an already learned direction. By the induction hypothesis, $\langle \boldsymbol{w}_i^t, \boldsymbol{v} \rangle$ converges to a non-constant polynomial in $\boldsymbol{a}$. Consider the term $\frac{a_j}{\sqrt{p}} \mathbb{E}\left[\hat{f}(\boldsymbol{z}; \boldsymbol{W}^t, \boldsymbol{a}))\sigma'(\langle \boldsymbol{z}, \boldsymbol{w}_j^t \rangle)\langle \boldsymbol{z}, \boldsymbol{v} \rangle\right]$ in $\langle \boldsymbol{g}_i^t, \boldsymbol{v} \rangle$. By expanding $\hat{f}(\boldsymbol{z}; \boldsymbol{W}^t, \boldsymbol{a})$ we obtain:

$$\frac{a_j}{\sqrt{p}} \mathbb{E}\left[\hat{f}(\boldsymbol{z}; \boldsymbol{W}^t, \boldsymbol{a}))\sigma'(\langle \boldsymbol{z}, \boldsymbol{w}_j^t \rangle)\langle \boldsymbol{z}, \boldsymbol{v} \rangle\right] = \frac{a_j}{p} \sum_{i=1}^{p} a_i \mathbb{E}\left[\sigma(\langle \boldsymbol{w}_i^t, \boldsymbol{z} \rangle)\sigma'(\langle \boldsymbol{z}, \boldsymbol{w}_j^t \rangle)\langle \boldsymbol{z}, \boldsymbol{v} \rangle\right] \tag{104}$$

The term correspondign to the j_{th} neuron has the form:

$$\frac{a_j^2}{p} \mathbb{E}\left[\sigma(\langle \boldsymbol{w}_j^t, \boldsymbol{z} \rangle)\sigma'(\langle \boldsymbol{z}, \boldsymbol{w}_j^t \rangle)\langle \boldsymbol{z}, \boldsymbol{v} \rangle\right] \tag{105}$$

By Wick's theorem and assumption on σ, $\mathbb{E}\left[\sigma(\langle \boldsymbol{w}_j^t, \boldsymbol{z} \rangle)\sigma'(\langle \boldsymbol{z}, \boldsymbol{w}_j^t \rangle)\langle \boldsymbol{z}, \boldsymbol{v} \rangle\right]$ is a non-zero polynomial in $\langle \boldsymbol{w}_j^t, \boldsymbol{v} \rangle$. Let d_j be the degree of a_j in $\langle \boldsymbol{w}_j^t, \boldsymbol{v} \rangle$. Then, the degree of a_j in $\frac{a_j^2}{p} \mathbb{E}\left[\sigma(\langle \boldsymbol{w}_j^t, \boldsymbol{z} \rangle)\sigma'(\langle \boldsymbol{z}, \boldsymbol{w}_j^t \rangle)\langle \boldsymbol{z}, \boldsymbol{v} \rangle\right]$ is at-least $d_j + 2$. Proceeding similarly for the other terms, one can show that the degree of a_j in $\langle \boldsymbol{g}_i^t, \boldsymbol{v} \rangle$ is strictly larger than in $\langle \boldsymbol{w}_i^t, \boldsymbol{u} \rangle$. This ensures that $\langle \boldsymbol{w}_i^{t+1}, \boldsymbol{v} \rangle = \langle \boldsymbol{g}_i^t, \boldsymbol{v} \rangle + \eta \langle \boldsymbol{g}_i^t, \boldsymbol{v} \rangle$ remains a non-constant polynomial upto error $\mathcal{O}(\frac{\operatorname{polylog} d}{\sqrt{d}})$. Therefore, almost surely over $\boldsymbol{a}$, a direction is not "un-learned". Finally, by decomposing along a general $\boldsymbol{v} \in U_{t+1}^\star$, along $U_t^\star$ and $U_{t+1}^\star \cap (U_t^\star)^\perp$, one obtains that points (ii) and (iii) of the induction statements hold in expectation.

Next, we prove that the events $(i), (ii), (iii)$ hold with high probability. By the induction hypothesis, we have that and the above analysis, we have that:

Lemma 18. *Suppose that the induction hypothesis holds at time t. Then, the following events occur with high-probability for all $i, j \in [p]$*

(i) $|\|\boldsymbol{g}_i^{t+1}\|^2 - \mathbb{E}\left[\|\boldsymbol{g}_i^{t+1}\|^2\right]| = \mathcal{O}\left(\frac{\operatorname{polylog} d}{\sqrt{d}}\right)$

(ii) $\|\langle \boldsymbol{g}_i, \boldsymbol{g}_j \rangle - \mathbb{E}\left[\langle \boldsymbol{g}_i, \boldsymbol{g}_j \rangle\right]\| = \mathcal{O}\left(\frac{\operatorname{polylog} d}{\sqrt{d}}\right)$

(iii) For any $k \in [r]$, and any unit vector $\boldsymbol{w}$

$$|\langle \boldsymbol{w}, \boldsymbol{g}_i \rangle - \mathbb{E}[\langle \boldsymbol{w}, \boldsymbol{g}_i \rangle]| = \mathcal{O}\left(\frac{\operatorname{polylog} d}{\sqrt{d}}\right) \tag{106}$$

Proof. We condition on the event in $\mathcal{F}_t$ that Q_t, M_t are bounded by some constants independent of d, n. Subsequently, the proof proceeds similar to Proposition 2, with the additional incorporation of the term due to $\hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a}))$ in the gradient.

We have:

$$\boldsymbol{g}_j^t = \frac{1}{n} a_j \sum_{\nu=1}^{n} \boldsymbol{z}_i (f^\star(\boldsymbol{z}^\nu) - \hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a}))\sigma'(\boldsymbol{z}_i^\top \boldsymbol{w}_j^t) \tag{107}$$

Define:

$$\boldsymbol{X}_i^\nu = \boldsymbol{z}^\nu \sigma'(\langle \boldsymbol{w}_i^t, \boldsymbol{z}^\nu \rangle)(f^\star(\boldsymbol{z}^\nu) - \hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a})). \tag{108}$$

Analogous to the proof of Proposition 2, we have:

$$\|\boldsymbol{g}_i\|^2 - \mathbb{E}[\|\boldsymbol{g}_i\|^2] = \frac{1}{n^2 p^2}\left(\underbrace{\sum_{\nu=1}^{n} \|\boldsymbol{X}_i^{\boldsymbol{\nu}}\|^2 - n\mathbb{E}[\|\boldsymbol{X}^{\boldsymbol{\nu}}\|^2]}_{S_1} + \underbrace{\sum_{\nu \neq \nu'} \langle \boldsymbol{X}_i^\nu, \boldsymbol{X}_i^{\nu'} \rangle - n(n-1)\|\mathbb{E}\langle \boldsymbol{X}_i^\nu, \boldsymbol{X}_i^{\nu'} \rangle\|^2}_{S_2}\right) \tag{109}$$

Similarly, we have:

$$\langle \boldsymbol{g}_i, \boldsymbol{g}_j \rangle - \mathbb{E}[\boldsymbol{g}_i, \boldsymbol{g}_j] = \frac{1}{n^2 p^2}\left(\underbrace{\sum_{\nu=1}^{n} \langle \boldsymbol{X}_i^{\boldsymbol{\nu}}, \boldsymbol{X}_j^{\boldsymbol{\nu}} \rangle - n\mathbb{E}[\langle \boldsymbol{X}_i^{\boldsymbol{\nu}}, \boldsymbol{X}_j^{\boldsymbol{\nu}} \rangle]}_{S_1'} + \underbrace{\sum_{\nu \neq \nu'} \langle \boldsymbol{X}_i^\nu, \boldsymbol{X}_j^{\nu'} \rangle - n(n-1)(\mathbb{E}\langle \boldsymbol{X}_i^\nu, \boldsymbol{X}_j^{\nu'} \rangle)^2}_{S_2'}\right) \tag{110}$$

Note that $f^\star(\boldsymbol{z}^\nu)$ and $\hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a})$ are polynomials in finite-number of correlated Gaussians $\langle \boldsymbol{w}_1^t, \boldsymbol{z}^\nu \rangle, \cdots, \langle \boldsymbol{w}_p^t, \boldsymbol{z}^\nu \rangle$, $\langle \boldsymbol{w}_1^\star, \boldsymbol{z}^\nu \rangle, \cdots, \langle \boldsymbol{w}_r^\star, \boldsymbol{z}^\nu \rangle$. Therefore, by repeatedly applying Lemma 2 and Theorem 5, we obtain that $\sigma'(\langle \boldsymbol{w}_i^t, \boldsymbol{z} \rangle)$, $f^\star(\boldsymbol{z}^\nu)$ and $\hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a}))$ have bounded Orlicz norms of some finite order α_t.

Subsequently, similar to Lemma 9, through Holder's inequality, Lemma 2 and Theorem 5, we obtain that $\|\boldsymbol{X}_i^\nu\|^2$, $\langle \boldsymbol{X}_i^\nu, \boldsymbol{X}_i^{\nu'} \rangle$, $\langle \boldsymbol{X}_i^\nu, \boldsymbol{X}_j^\nu \rangle$, $\langle \boldsymbol{X}_i^\nu, \boldsymbol{X}_j^{\nu'} \rangle$, have Orlicz norms of order $\mathcal{O}(d)$ with $\alpha = \alpha_t$ for some α_t independent of d.

The remaining proof follows by repeating the arguments in Lemmas 9, 10 for Orlicz norms of general order.

Similarly (iii) is obtained by replacing the application of Bernstein's inequality in Lemma 8 by Theorem 5. □

Lemmas 15 and 18 together with C.8 and the induction hypothesis imply statement (iii) at time $t+1$.

It remains to prove the base case i.e $t = 1$. If the leap index $\ell > 1$, $U_t^\star = 0$ for all $t \geq 1$. Applying the above arguments then implies that (i) and (iii) hold for all timesteps t.

Therefore, we consider the case $\ell = 1$ At $t = 1$, $U_1^\star$ is simply the subspace along $(C_1(f^\star))$. Let $\boldsymbol{v} = \pm\frac{1}{\|C_1(f^\star)\|}C_1(f^\star)$ be a vector as per (ii) let $i \in [p]$ be an arbitrary neuron. We have:

$$\begin{aligned}\langle \boldsymbol{v}, \boldsymbol{w}_i^1 \rangle &= \langle \boldsymbol{v}, \boldsymbol{w}_i^0 \rangle + \eta \langle \boldsymbol{v}, \boldsymbol{g}^i \rangle \\ &= \eta \langle \boldsymbol{v}, \boldsymbol{g}^i \rangle + \mathcal{O}(\frac{1}{\sqrt{d}})\end{aligned}$$

It is straightforward to check that 4 hold when σ, g^* are polynomials, while Lemma 14 holds in expectation. Applying the concentration results for Orlicz norms of general order as in Lemma 18 imply that Lemma 14 also holds in probability for polynomial σ, g^*. To establish (i) at time $t = 1$, we note that Lemma 14 implies that $\mathbb{E}\left[\boldsymbol{w}_i\right]^2$ converges to $1 + ca_i^2$ where c is independent of d, n. By Lemma 4 the first term equals with high-probability, $\pm\frac{a_i\mu_1}{p}\|C_1(f^\star)\| + \mathcal{O}\frac{1}{\sqrt{d}}$. Since, $\frac{a_i\mu_1}{p}\|C_1(f^\star)\|$ is a non-constant (linear) polynomial in a_i, this proves (ii) for $t = 1$. Lemmas 4 and part (iii) in Lemma 18 directly imply (iii) of the induction statements.

Points $(ii), (iii)$ of the induction statements directly imply Theorem 3.

C.9 Prediction of the alignment at the second step

We now utilize the analysis in the previous section to obtain a theoretical prediction for the gradient orientation after two steps for the above target function. We follow the notation defined in the proof sketch in Section C.7

We have, using Equation (89):

$$\mathbb{E}\left[\langle \boldsymbol{u}, \boldsymbol{g}_j^1\rangle\right] \to \mathbb{E}\left[y\mu_{1,\boldsymbol{x}^\top\boldsymbol{v}^\star}(\boldsymbol{x}_i^\top\boldsymbol{u})\right] - \mathbb{E}\left[\hat{y}_i^1\sigma'(\boldsymbol{x}_i^\top\boldsymbol{w}^1)_j(\boldsymbol{x}_i^\top\boldsymbol{u})\right]. \tag{111}$$

As explained in the previous section, the second term does not contribute to an alignment towards $\boldsymbol{v}^\perp$. Therefore we consider the ratio of the first term when $\boldsymbol{u} = \boldsymbol{v}^\star$ or $\boldsymbol{u} = \boldsymbol{v}^\perp$. We obtain:

$$\langle \boldsymbol{g}^1, \boldsymbol{v}^\star\rangle \approx \mathbb{E}\left[y\mu_{1,(\mu_1\boldsymbol{x}^\top\boldsymbol{v}^\star)/\sqrt{2p}}(\boldsymbol{x}_i^\top\boldsymbol{v}^\star)\right], \tag{112}$$

where μ_1 denotes the first Hermite coefficient of the student activation σ, given by 0.5 for Relu.

$\mu_{1,\boldsymbol{x}^\top\boldsymbol{v}^\star}$ corresponds to the first Hermite coefficient of a translated Relu function and is given by:

$$\mu_{1,\kappa,j} = (1 - \Phi(-a_j\eta\kappa)) = \frac{1}{2}(1 \pm \text{erf}(\eta\kappa/\sqrt{2})). \tag{113}$$

Therefore, when $\eta = 4\sqrt{p}$, we obtain:

$$\mu_{1,(\mu_1\boldsymbol{x}^\top\boldsymbol{v}^\star)/\sqrt{2p}} = \frac{1}{2}(1 \pm \text{erf}(\boldsymbol{x}^\top\boldsymbol{v}^\star)). \tag{114}$$

Therefore, using the Hermite decomposition of erf, we obtain the following predicted orientations in the setting considered in the right panel of Fig. 6:

$$\boldsymbol{v}_1^{(t=2)} = (1 - \frac{2}{\sqrt{3\pi}})\boldsymbol{w}_1^\star + (1 + \frac{2}{\sqrt{3\pi}})\boldsymbol{w}_2^\star \qquad \boldsymbol{v}_2^{(t=2)} = (1 + \frac{2}{\sqrt{3\pi}})\boldsymbol{w}_1^\star + (1 - \frac{2}{\sqrt{3\pi}})\boldsymbol{w}_2^\star \tag{115}$$

C.10 Limitations of the Staircase Structure

We show that a natural class of teacher functions, containing neurons with identical activation functions and uniform second-layer weights does not contain a staircase structure:

Proposition 4. *Let $y = f^\star(\boldsymbol{z}) = \sum_{k=1}^r \sigma^\star(\langle \boldsymbol{w}_k^\star, \boldsymbol{z}\rangle)$ for some $\sigma^\star$ having leap index 1, then $U_i^\star = U_1^\star$ for all $i \geq 1$.*

Proof. For any such target function, the direction $\boldsymbol{v}^\star$ is given by $\boldsymbol{v}^\star = \frac{1}{\sqrt{r}}(\sum_{k=1}^r \boldsymbol{w}_k^\star)$. Without loss of generality, assume that $\boldsymbol{w}_k^\star = \boldsymbol{e}_k$, where $\boldsymbol{e}_k$ denotes the unit vector corresponding to the k_{th} coordinate.

Now, consider any direction $\boldsymbol{u} \perp \boldsymbol{v}^\star$ in the teacher subspace. Such a vector satisfies $\sum_{k=1}^{r} u_i = 0$. Therefore, for any $k \geq 0$, we have:

$$\begin{aligned} \mathbb{E}\left[f^\star(\boldsymbol{z}) H_k((\boldsymbol{v}^\star)^\top \boldsymbol{z})((\boldsymbol{u}_i)^\top \boldsymbol{z})\right] &= (\sum_{i=1}^{p}(u_i))(\mathbb{E}\left[f^\star(\boldsymbol{z}) H_k((\boldsymbol{v}^\star)^\top \boldsymbol{z}) z_1\right]) \\ &= 0. \end{aligned} \tag{116}$$

Where we used the symmetry of $f^\star(\boldsymbol{z})$ w.r.t permutations of the first r coordinates. Therefore, the Hermite decomposition of $f^\star(\boldsymbol{z})$ does not contain any term that linearly couples $\boldsymbol{u}$ to $\boldsymbol{v}^\star$. □

Therefore, the presence of a staircase structure requires asymmetry between the the dependence of the target function on different directions in the teacher subspace.

D Learning the second layer

D.1 Proof of Proposition 1

We first prove the finite p case of Proposition 1. Let $\boldsymbol{a}$ be a second layer vector with $a_i \leq c/\sqrt{p}$, and assume that W only learns a subspace $U \subseteq V^\star$. We write $\mathbb{R}^d = U \oplus U^\perp \oplus V^{\star\perp}$, where $U^\perp$ is the orthogonal subspace of U in $V^\star$. By assumption, we have $\|P_{U^\perp}\boldsymbol{w}_i\| \leq \varepsilon_d$ for every i; where ε_d is going to zero as d grows.

For any $\boldsymbol{z} \in \mathbb{R}^d$, we have

$$\begin{aligned}\hat f(\boldsymbol{z}; W, \boldsymbol{a}) &= \sum_{i=1}^p \frac{a_i}{\sqrt{p}}\sigma(\langle \boldsymbol{w}_i, P_U\boldsymbol{z}\rangle + \langle \boldsymbol{w}_i, P_{U^\perp}\boldsymbol{z}\rangle + \langle \boldsymbol{w}_i, P_{V^{\star\perp}}\boldsymbol{z}\rangle) \\ &= \sum_{i=1}^p \frac{a_i}{\sqrt{p}}\sigma(\langle \boldsymbol{w}_i, P_U\boldsymbol{z}\rangle + \langle \boldsymbol{w}_i, P_{V^{\star\perp}}\boldsymbol{z}\rangle) + \frac{a_i}{\sqrt{p}}\varepsilon_d\tilde\sigma(\langle \boldsymbol{w}_i, P_{U^\perp}\boldsymbol{z}\rangle)\end{aligned}$$

where $\tilde\sigma$ is a Lipschitz function. We call the first term of the above expression $\tilde f(P_U\boldsymbol{z}, P_{V^{\star\perp}\boldsymbol{z}})$, forgetting the structure of the function $\hat f$. Then, we can write the risk as

$$\mathcal{R}(W, \boldsymbol{a}) = \mathbb{E}_{\boldsymbol{z}}\left[\left(f^\star(\boldsymbol{z}) - \tilde f(P_U\boldsymbol{z}, P_{V^{\star\perp}\boldsymbol{z}})\right)^2\right] + O(\varepsilon_d), \tag{117}$$

having used the Cauchy-Schwarz inequality to bound the contribution of $\tilde\sigma$. Then, by successive expectations,

$$\begin{aligned}\mathcal{R}(W, \boldsymbol{a}) &= \mathbb{E}_{P_{V^{\star\perp}}\boldsymbol{z}, P_U\boldsymbol{z}}\left[\mathbb{E}_{P_{U^\perp}\boldsymbol{z}}\left[\left(f^\star(\boldsymbol{z}) - \tilde f(P_U\boldsymbol{z}, P_{V^{\star\perp}\boldsymbol{z}})\right)^2 \middle| P_{V^{\star\perp}}\boldsymbol{z}, P_U\boldsymbol{z}\right]\right] + O(\varepsilon_d), \\ &\geq \mathbb{E}_{P_{V^{\star\perp}}\boldsymbol{z}, P_U\boldsymbol{z}}\left[\inf_f \mathbb{E}_{P_{U^\perp}\boldsymbol{z}}\left[(f^\star(\boldsymbol{z}) - f(P_U\boldsymbol{z}, P_{V^{\star\perp}\boldsymbol{z}}))^2 \middle| P_{V^{\star\perp}}\boldsymbol{z}, P_U\boldsymbol{z}\right]\right] + O(\varepsilon_d)\end{aligned}$$

where the infimum is taken over all measurable functions $f : U \times V^{\star\perp} \to \mathbb{R}$. But this infimum exactly corresponds to the definition of conditional expectation/conditional variance, which is independent from $P_{V^{\star\perp}}\boldsymbol{z}$ (since $f^\star$ is). As a result,

$$\mathcal{R}(W, \boldsymbol{a}) \geq \mathbb{E}_{P_U\boldsymbol{z}}\left[\mathrm{Var}\left(f^\star(\boldsymbol{z}) | P_U\boldsymbol{z}\right)\right] + O(\varepsilon_d), \tag{118}$$

which implies the statement of Proposition 1.

D.2 Full statement of Theorem 4

We now provide the full statement of Theorem 4. It establishes the asymptotic equivalence of the training and generalization errors of the original features and the conditional Gaussian features defined by equation (15).

Consider the sequence of vectors $\boldsymbol{v}_n \in \mathbb{R}^d$ defined as in Equation (76) by $\boldsymbol{v}_n = \frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{x}_i$. For simplicity, we omit the dependence of $\boldsymbol{v}_n$ on n and denote each entry by $\boldsymbol{v}$. For any vector $\boldsymbol{z} \in \mathbb{R}^d$, define the decomposition $\boldsymbol{z} = z_{\boldsymbol{v}}\boldsymbol{v} + \boldsymbol{z}^\perp$ and feature maps:

$$\phi_{\text{CK}}(\boldsymbol{z}) = \sigma(W^{(1)}\boldsymbol{z}), \tag{119}$$

where $W^{(1)}$ denotes the weight matrix obtained through the application of a single gradient step.

Then the random variable $\phi_{\text{CK}}(\boldsymbol{z})$ admits a regular conditional distribution conditioned on the values of $z_{\boldsymbol{v}}$ (Theorem 8.37 in Klenke (2013)). Therefore, the following mean, correlation, and covariance matrix are well-defined:

$$\begin{aligned} \mu(z_{\boldsymbol{v}}) &= \mathbb{E}\left[\phi_{\text{CK}}(\boldsymbol{z}) \mid z_{\boldsymbol{v}}\right], \quad \Psi(z_{\boldsymbol{v}}) = \mathbb{E}\left[\phi_{\text{CK}}(\boldsymbol{z})(z^{\perp})^{\top} \mid z_{\boldsymbol{v}}\right], \\ \Phi(z_{\boldsymbol{v}}) &= \text{Cov}\left[\phi_{\text{CK}}(\boldsymbol{z}) \mid z_{\boldsymbol{v}}\right] - \Psi(z_{\boldsymbol{v}})\Psi(z_{\boldsymbol{v}})^{\top} \end{aligned} \tag{120}$$

Now, for each value of $z_{\boldsymbol{v}}$, define the following random variable:

$$\phi_{\text{CL}}(\boldsymbol{z};\boldsymbol{v}) = \mu\left(z_{\boldsymbol{v}}\right) + \Psi(z_{\boldsymbol{v}})\boldsymbol{z}^{\perp} + \Phi(z_{\boldsymbol{v}})\boldsymbol{\xi}. \tag{121}$$

Then $\phi_{\text{CL}}(\boldsymbol{z};\boldsymbol{v})$ satisfies:

$$\mathbb{E}\left[\phi_{\text{CL}}(\boldsymbol{z}) \mid z_{\boldsymbol{v}}\right] = \mu(z_{\boldsymbol{v}}), \mathbb{E}\left[\phi_{\text{CL}}(\boldsymbol{z})(\boldsymbol{z}^{\perp})^{\top} \mid z_{\boldsymbol{v}}\right] = \Psi(z_{\boldsymbol{v}}),\ \text{Cov}\left[\phi_{\text{CL}}(\boldsymbol{z}) \mid z_{\boldsymbol{v}}\right] = \text{Cov}\left[\phi_{\text{CK}}(\boldsymbol{z}) \mid z_{\boldsymbol{v}}\right]. \tag{122}$$

Therefore, $\phi_{\text{CL}}(\boldsymbol{z};\boldsymbol{v})$ is a Gaussian variable having the same conditional mean, covariance as $\phi_{\text{CK}}(\boldsymbol{z};\boldsymbol{v})$ and the same corrrelation with $\boldsymbol{z}_{\perp}$ as $\phi_{\text{CK}}(\boldsymbol{z};\boldsymbol{v})$. Since $\boldsymbol{z}_{\perp}$ is Gaussian and independent of $z_{\boldsymbol{v}}$, this uniquely characterizes the conditional measure of $\phi_{\text{CL}}(\boldsymbol{z};\boldsymbol{v})$.

Now, consider a set of n training inputs $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n$. For each $i \in n$, generate an equivalent feature map $\boldsymbol{\Phi}_{\text{CL}}$ through equation (1), with $\boldsymbol{\xi}$ being independently sampled for each example. Let $\boldsymbol{\Phi}_{\text{CK}}$ and $\boldsymbol{\Phi}_{\text{CL}}$ denote matrices in $\mathbb{R}^{n\times p}$ with rows $\phi_{\text{CK}}(\boldsymbol{z_i})$ and $\phi_{\text{CL}}(\boldsymbol{z_i})$ respectively,

Consider the following minimization problem:

$$\min_{\boldsymbol{a}\in\mathbb{R}^p} \frac{1}{n}\sum_{\nu=1}^{n}\left(\langle \boldsymbol{a}, \phi_{\text{CK}}(\boldsymbol{z}^{\nu})\rangle - f^{\star}(\boldsymbol{z}^{\nu})\right)^2 + \lambda\|\boldsymbol{a}\|^2 \tag{123}$$

Define the following constraint set:

$$\mathcal{S}_p = \left\{\boldsymbol{\theta} \in \mathbb{R}^d \;\middle|\; \|\boldsymbol{\theta}\|_2 \le R, \quad \|\boldsymbol{\theta}\|_\infty \le Cp^{-\eta}\right\}. \tag{124}$$

We make the following assumption:

Assumption 6. *There exist constants R, C, η such that the minimizer $\hat{\boldsymbol{a}}_{CK}$ of the optimization problem defined by equation (123) lies in $\mathcal{S}_p$ with high probability as $n, d \to \infty$.*

The above assumption can be enforced by utilizing constrained minimization for the second layer. Alternatively, for overparameterized models i.e $p/n > 1$, one could utilize the arguments in Theorem 5 of Montanari and Saeed (2022). Let $\widehat{\mathcal{R}}^{\star}_n(\boldsymbol{\Phi}, \boldsymbol{y}(\boldsymbol{Z})), \mathcal{R}^{\star}_g(\boldsymbol{\Phi}, \boldsymbol{y}(\boldsymbol{Z}))$ denote the training and generalization errors respectively with features $\boldsymbol{\Phi}$ and labels $\boldsymbol{y}(\boldsymbol{Z})$.

Theorem 4. *Assume that $n, p = \Theta(d)$, and that the vector $V^{\star}_1 = \boldsymbol{v}^{\star}$ defined in Theorem 2 is nonzero. Then, the sequence of vectors $\boldsymbol{v}_n = \frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{x}_i \in \mathbb{R}^d$ satisfy:*

(i) *As $n, d \to \infty$, $P_{V^{\star}}\boldsymbol{v} \xrightarrow{\mathbb{P}} \frac{\mu}{\sqrt{p}}\boldsymbol{v}^{\star}$.*

(ii) *Under Assumption 6, the training and generalization errors obtained through the minimization of the objective (15) for training distribution defined by feature maps $\phi_{CK}(\boldsymbol{z})$ converge in distribution to the corresponding training and generalization errors for features $\phi_{CL}(\boldsymbol{z};\boldsymbol{v})$.*

Concretely, we have that for any bounded Lipschitz function $\Psi : \mathbb{R} \to \mathbb{R}$:

$$\lim_{n,p\to\infty}\left|\mathbb{E}\left[\Psi\left(\widehat{\mathcal{R}}^{\star}_n(\boldsymbol{\Phi}_{CK}, \boldsymbol{y}(\boldsymbol{Z}))\right)\right] - \mathbb{E}\left[\Psi\left(\widehat{\mathcal{R}}^{\star}_n(\boldsymbol{\Phi}_{CL}, \boldsymbol{y}(\boldsymbol{Z}))\right)\right]\right| = 0$$

$$\lim_{n,p\to\infty} \left|\mathbb{E}\left[\Psi\left(\mathcal{R}_g(\boldsymbol{\Phi}_{CK}, \boldsymbol{y}(\boldsymbol{Z}))\right)\right] - \mathbb{E}\left[\Psi\left(\mathcal{R}_g(\boldsymbol{\Phi}_{CL}, \boldsymbol{y}(\boldsymbol{Z}))\right)\right]\right| = 0$$

In particular, for any $\mathcal{E} \in \mathbb{R}$, and denoting $\xrightarrow{\mathbb{P}}$ the convergence in probability:

$$\begin{aligned} \widehat{\mathcal{R}}^{\star}_n(\boldsymbol{\Phi}_{CK}, \boldsymbol{y}(\boldsymbol{Z})) \xrightarrow{\mathbb{P}} \mathcal{E} \quad &\textit{if and only if} \quad \widehat{\mathcal{R}}^{\star}_n(\boldsymbol{\Phi}_{CL}, \boldsymbol{y}(\boldsymbol{Z})) \xrightarrow{\mathbb{P}} \mathcal{E} \\ \mathcal{R}^{\star}_g(\boldsymbol{\Phi}_{CK}, \boldsymbol{y}(\boldsymbol{Z})) \xrightarrow{\mathbb{P}} \mathcal{E} \quad &\textit{if and only if} \quad \mathcal{R}_g(\boldsymbol{\Phi}_{CL}, \boldsymbol{y}(\boldsymbol{Z})) \xrightarrow{\mathbb{P}} \mathcal{E}, \end{aligned} \tag{125}$$

Part (i) follows directly from Lemma 5. To prove the equivalence of training and generalization errors for the given direction, we rely on the framework of one-dimensional CLT (Central Limit Theorem), discussed in Goldt et al. (2022). One-dimensional CLT was recently shown to imply the universality of training and generalization errors for Random feature models in Hu and Lu (2022). However, in our setting where we train the model, and as verified empirically in Ba et al. (2022), a naive one-dimensional CLT with equivalent Gaussian features no longer holds.

Instead, we introduce a generalization termed "conditional one-dimensional CLT", given by the following Lemma:

Lemma 19. *For any Lipschitz function $\varphi : \mathbb{R}^2 \to \mathbb{R}$,*

$$\lim_{n,p\to\infty} \sup_{\boldsymbol{\theta}_1\in\mathcal{S}_p, \boldsymbol{\theta}_2\in\mathcal{S}^{d-1}} \left|\mathbb{E}\left[\varphi(\boldsymbol{\theta}_1^\top \phi_{\mathrm{CK}}(\boldsymbol{z}), \boldsymbol{\theta}_2^\top \boldsymbol{z}) \,\middle|\, z_{\boldsymbol{v}} = k\right] - \mathbb{E}\left[\varphi(\boldsymbol{\theta}^\top \phi_{\mathrm{CL}}(\boldsymbol{z}), \boldsymbol{\theta}_2^\top \boldsymbol{z}) \,\middle|\, z_{\boldsymbol{v}} = k\right]\right| = 0, \quad \forall k \in \mathbb{R}, \tag{126}$$

where $\mathcal{S}^{d-1}$ denotes the unit sphere in $\mathbb{R}^d$

Proof. For an input $\boldsymbol{z} \sim \mathcal{N}(0, I_d)$, we consider the decomposition $\boldsymbol{z} = z_{\boldsymbol{v}}\boldsymbol{v} + \boldsymbol{z}^\perp$ We note that conditioned on on $z_{\boldsymbol{v}} = k$, $\phi_{\mathrm{CL}}(\boldsymbol{z}), \boldsymbol{z}_\perp$ is a Gaussian random variable. Next, consider $\phi_{\mathrm{CK}}(\boldsymbol{z})$. Our proof relies on the observation that while the features $\phi_{\mathrm{CK}}(\boldsymbol{z})$ have complex non-linear dependence on $z_{\boldsymbol{v}}$, for a fixed value of $z_{\boldsymbol{v}}$, they are equivalent to a random-features mapping applied to $\boldsymbol{z}_\perp$. Concretely, we have by Lemma 12 that the weight matrix $\boldsymbol{W}^{(1)}$ has the following spike+bulk decomposition (equation (76)):

$$\boldsymbol{W}^{(1)} = \eta \boldsymbol{u}\boldsymbol{v}^\top + \boldsymbol{W}^{(0)} + \eta\Delta, \tag{127}$$

where $\boldsymbol{u} = \frac{\mu_1}{p}\boldsymbol{a}$

Let $\boldsymbol{W}^\perp$ denote the combined matrix $\boldsymbol{W}^{(0)} + \eta\Delta$ with rows $\boldsymbol{w}_i^\perp$ for $i \in [p]$.

Lemma 14 implies that there exist constants c_i for $i \in [p]$ depending only on a_i such that $\|\boldsymbol{w}_i^\perp\|^2 = c_i + \mathcal{O}(\frac{\operatorname{polylog} d}{\sqrt{d}})$ with high-probability. Define the following neuron-wise activation functions:

$$\sigma_{i,z_{\boldsymbol{v}}}(u) = \sigma(c_j u + \eta v_{\boldsymbol{z}}) - \mathbb{E}_u\left[\sigma(c_j u + \eta u_i z_{\boldsymbol{v}})\right], \tag{128}$$

where $i \in [p]$ denotes the index of the neuron and the expectation is w.r.t $z \sim \mathcal{N}(0, 1)$. Under the choise of symmetric initialization in Equation (26), it suffices to restrict ourselves to the first half $p/2$ neurons.

For a fixed value of $z_{\boldsymbol{v}}$, the feature map $\phi_{\mathrm{CK}}(\boldsymbol{z}) = \sigma(\boldsymbol{W}^1\boldsymbol{z})$ is equivalent to a random features mapping with neurons $\sigma_{i,v_{\boldsymbol{z}}}$ applies to inputs $\boldsymbol{z}^\perp \in \mathbb{R}^d$ with approximately orthogonal weights $\boldsymbol{W}^\perp$. Consider the following events for some positive constants C_1, C_2, C_3:

$$\mathcal{A}_1 = \left\{\sup_{i,j\in[p/2]} \left|\langle \boldsymbol{w}_i^\perp, \boldsymbol{w}_j^\perp\rangle - c_i\delta_{ij}\right| \le C_1\left(\frac{\operatorname{polylog} d}{d}\right)^{1/2}\right\} \quad \mathcal{A}_2 = \left\{\|\boldsymbol{W}^\perp\|_{\mathrm{op}} \le C_3(\operatorname{polylog} d)\right\}$$

We have, using Lemma 12 and a union bound, that for $p, d = \Theta(n)$, $\Pr[\mathcal{A}_1] \xrightarrow{n,d\to\infty} 1$. Furthermore, part (ii) of Lemma 14 in Ba et al. (2022) implies that $\Pr[\mathcal{A}_2] \to 1$. Next, we utilize Corollary 2 and Lemma 3 in Hu and Lu (2022). Note that the neuron wise activation functions (128) for a fixed value of $z_{\boldsymbol{v}}$ satisfy $\mathbb{E}_u\left[\sigma_{i,v_{\boldsymbol{z}}}(u)\right]=0$. We relax the requirement of odd-activation in Hu and Lu (2022) by noting that ϕ_{CK}, ϕ_{CL} have exactly equivalent means and covariances as in Theorem 6 of Dandi et al. (2023). □

The above Lemma states that the one-dimensional projections of $\phi_{\mathrm{CK}}(\boldsymbol{z})$ are asymptotically distributed as jointly Gaussian variables with $\boldsymbol{z}_{\perp}$.

D.3 Conditional GET

We now prove part (ii) of Theorem 4 using Lemma 19. This relies on the universality of training and generalization errors between the given distribution and the "conditional equivalent" distribution. The central idea of the proof again relies on a the isolation of the effects of the "spikes" and the "noise" in the features.

The technique presented here is also of independent interest for proving the universality of training, generalization errors in related setups such as with spiked-covariance inputs.

We utilize the following properties of the features :

Lemma 20. *For any fixed $z_{\boldsymbol{v}}$, the random variable $\phi_{CK}-\boldsymbol{\mu}(z_{\boldsymbol{v}})$ is sub-Gaussian with sub-Gaussian norm independent of $z_{\boldsymbol{v}}$ and n.*

Proof. The result follows from the assumption of uniform boundedness of the derivative of $\sigma^{\star}$ and the Lipschitz concentration of Gaussian variables. □

Lemma 21. *There exists a constant C such that the matrix $\bar{\Phi}_{CK}$ with rows $\phi_{CK}-\boldsymbol{\mu}(z_{\boldsymbol{v}})$ satisfies:*

$$\Pr[\|\bar{\Phi}_{CK}\| \geq K\sqrt{p}] \leq 2\exp(-Cn) \tag{129}$$

Proof. By Lemma 20, each row of $\bar{\Phi}_{CK}$ is sub-Gaussian. Therefore, the result follows from the concentration of spectral norm of matrices with independent sub-Gaussian rows (Theorem 5.39 in Vershynin (2010)). □

We start by proving certain properties of the optimal parameters a_i upon the training of the second layer:

Lemma 22. *Let $\hat{a}_{\mathrm{CK}}(\lambda)$ be the parameters obtained through ridge regression on features $\phi_{\mathrm{CK}}(\boldsymbol{z_i})_{\{i=1,\cdots n\}}$ with regularization strength λ. Then, there exists a constants C such that with high probability as $n, d\to\infty$:*

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{a}_{\mathrm{CK}}^{\top}\mu(z_{i,\boldsymbol{v}})\right)^2 \leq C \tag{130}$$

Proof. By assumption, $y_i(\boldsymbol{z})=\frac{1}{\sqrt{p}}a^{\top}\sigma(W\boldsymbol{z})$ with σ' uniformly bounded. Therefore, from the concentration of Lipschitz functions of gaussian variables, $y_i(\boldsymbol{z})$ is sub-Gaussian. Thus $y_i^2(\boldsymbol{z})$ are sub-exponential variables. Using Bernstein's inequality Vershynin (2018), we obtain:

$$\Pr[\frac{1}{n}\sum_{i=1}^{n}(y_i)^2-\mathbb{E}\left[(y_i)^2\right] > K] \leq 2\exp(-\min(c_1K, c_2K^2)n). \tag{131}$$

For constants c_1, c_2. □

Let $\mathcal{A}_{\boldsymbol{y}}$ denote the following event:

$$\mathcal{A}_{\boldsymbol{y}} = \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i)^2 < C_1 \right\}. \tag{132}$$

By Equation (131), we have $\Pr\left[\mathcal{A}_{\boldsymbol{y}}\right] \to 1$ as $n, d \to \infty$.

Let $\hat{\mathcal{R}}(W, \boldsymbol{a})$ denote the empirical risk at given values of $\boldsymbol{a}$, W. We have:

$$\hat{\boldsymbol{a}} = \arg\min_{\boldsymbol{a}} \hat{\mathcal{R}}(W, \boldsymbol{a}) = \arg\min_{\boldsymbol{a}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{a}^\top \phi_k(\boldsymbol{z}_i))^2. \tag{133}$$

We note that when $\boldsymbol{a} = \boldsymbol{0}$, we have:

$$\hat{\mathcal{R}}(W, \boldsymbol{0}) = \frac{1}{n} \sum_{i=1}^{n} (y_i)^2. \tag{134}$$

Since $\hat{\boldsymbol{a}}$ minimizes $\hat{\mathcal{R}}(W, \boldsymbol{a})$, we must have:

$$\hat{\mathcal{R}}(W, \hat{\boldsymbol{a}}) \le \hat{\mathcal{R}}(W, \boldsymbol{0}). \tag{135}$$

We obtain:

$$\begin{aligned}
&\frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{a}^\top \phi_k(\boldsymbol{z}_i))^2 \le \frac{1}{n} \sum_{i=1}^{n} (y_i)^2 \\
\implies\ & \frac{1}{2n} \sum_{i=1}^{n} (\boldsymbol{a}^\top \phi_k(\boldsymbol{z}_i))^2 \le \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{a}^\top \phi_k(\boldsymbol{z}_i) y_i \\
\implies\ & \frac{1}{2n} \sum_{i=1}^{n} (\hat{\boldsymbol{a}}^\top \phi_k(\boldsymbol{z}_i))^2 \le \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{a}^\top \phi_k(\boldsymbol{z}_i))^2} \sqrt{\frac{1}{n} \sum_{i=1}^{n} y_i^2},
\end{aligned}$$

where the last inequality follows from Cauchy-Schwarz. Therefore:

$$\sqrt{\frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{a}^\top \phi_k(\boldsymbol{z}_i))^2} \le 2 \sqrt{\frac{1}{n} \sum_{i=1}^{n} y_i^2}$$

$$\frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{a}^\top \mu(\boldsymbol{z}_i))^2 + \frac{1}{n} \|\bar{\Phi}_{\text{CK}}^\top \boldsymbol{a}\|_2^2 \le 4(\frac{1}{n} \sum_{i=1}^{n} y_i^2).$$

Applying Lemma 21 and $\Pr\left[\mathcal{A}_{\boldsymbol{y}}\right] \xrightarrow{n,d\to} 1$ then completes the proof.

Next, we prove the universality of the training, generalization error, conditioned on the values of the projections $z_{\boldsymbol{v}}$. This can be achieved through a number of techniques such as the Lindeberg's method in Hu and Lu (2022). We utilize the result of Montanari and Saeed (2022), who apply the interpolation technique to continuously transform the inputs $\boldsymbol{x}_i$ to equivalent Gaussian vectors $\boldsymbol{g}_i$.

Instead, we interpolate between the features $\phi_{CK}(\boldsymbol{z})$ and $\phi_{CL}(\boldsymbol{z})$. Define:

$$\boldsymbol{u}_{t,i} = \boldsymbol{\mu}(z_{i,\boldsymbol{v}}) + \cos(t)(\boldsymbol{\phi}_{CK}(\boldsymbol{z}) - \boldsymbol{\mu}(z_{i,\boldsymbol{v}})) + \sin(t)(\boldsymbol{\phi}_{CL}(\boldsymbol{z}) - \boldsymbol{\mu}(z_{i,\boldsymbol{v}})),$$

.

Let $\mathcal{A}_1$ denote the event:

$$\mathcal{A}_1 = \left\{ \frac{1}{n}\sum_{i=1}^{n} \left(\hat{a}_{\text{CK}}^{\top}\mu(z_{i,\boldsymbol{v}})\right)^2 \leq C_1 \right\} \tag{136}$$

Under the above interpolation path, we generalize Theorem 1 in Montanari and Saeed (2022) to obtain that for any bounded Lipschitz function $\Phi : \mathbb{R} \to \mathbb{R}$:

$$\lim_{n,p\to\infty} \sup_{v_{\boldsymbol{z}_1},\cdots,v_{\boldsymbol{z}_n}} \left| \mathbb{E}\left[\mathbf{1}_{\mathcal{A}_1}\Phi\left(\widehat{\mathcal{R}}_n^{\star}(\boldsymbol{\Phi}_{CK},\boldsymbol{y}(\boldsymbol{Z}))\right) \mid v_{\boldsymbol{z}_1},\cdots,v_{\boldsymbol{z}_n}\right] - \mathbb{E}\left[\mathbf{1}_{\mathcal{A}_1}\Phi\left(\widehat{\mathcal{R}}_n^{\star}(\boldsymbol{\Phi}_{\text{CL}},\boldsymbol{y}(\boldsymbol{Z}))\right) \mid v_{\boldsymbol{z}_1},\cdots,v_{\boldsymbol{z}_n}\right]\right| = 0. \tag{137}$$

Below, we explain the modifications to Theorem 1 in Montanari and Saeed (2022) that allow its applicability to our setting:

(i) We replace equation (12) in Assumption 5 tof Montanari and Saeed (2022) by the conditional 1d-CLT (Lemma 19). This is similar to the conditioning utilized in Dandi et al. (2023) for proving the universality in mixture models.

(ii) Our target function $y = f^{\star}(\boldsymbol{z})$ depends on the projection along the spike $v_{\boldsymbol{z}}$ as well as the orthogonal component $\boldsymbol{z}^{\perp}$. Since we condition on the values of $v_{\boldsymbol{z}}$, their dependence can be absorbed into the loss function for each input $\boldsymbol{z}_i^{\perp}$

(iii) While Theorem 1 in Montanari and Saeed (2022) does not allow a dependence of the labels on the latent variables $\boldsymbol{z}$, such a target function can be incorporated by considering the inputs to be the joint variables in $(\boldsymbol{\Phi}_{\text{CK}}(\boldsymbol{z}), \boldsymbol{z}) \in \mathbb{R}^{p+d}$ and constraining the parameters to have 0 components along the last d directions.

(iv) The event $\mathcal{A}_1$ and Lemma 21 ensure that Lemmas 5 and 6 in Montanari and Saeed (2022) hold under the presence of variable and unbounded means across samples $\boldsymbol{\mu}(z_{i,\boldsymbol{v}})$.

Next, using the Law of total expectation and Equation 137, we obtain:

$$\lim_{n,p\to\infty} \left| \mathbb{E}\left[\mathbf{1}_{\mathcal{A}_1}\Phi\left(\widehat{\mathcal{R}}_n^{\star}(\boldsymbol{\Phi}_{CK},\boldsymbol{y}(\boldsymbol{Z}))\right)\right] - \mathbb{E}\left[\mathbf{1}_{\mathcal{A}_1}\boldsymbol{\Phi}\left(\widehat{\mathcal{R}}_n^{\star}(\phi_{\text{CL}},\boldsymbol{y}(\boldsymbol{Z}))\right)\right]\right| = 0.$$

Finally, we note Lemma 22 implies that $\Pr[\mathcal{A}_1^c] \to 0$. Since Φ is bounded, we have that

$$\lim_{n,p\to\infty} \left| \mathbb{E}\left[\mathbf{1}_{\mathcal{A}_1^c}\Phi\left(\widehat{\mathcal{R}}_n^{\star}(\boldsymbol{\Phi}_{CK},\boldsymbol{y}(\boldsymbol{Z}))\right)\right] - \mathbb{E}\left[\mathbf{1}_{\mathcal{A}_1^c}\boldsymbol{\Phi}\left(\widehat{\mathcal{R}}_n^{\star}(\phi_{\text{CL}},\boldsymbol{y}(\boldsymbol{Z}))\right)\right]\right| = 0.$$

This completes the proof of Theorem 4.

D.4 Generalization Error Lower Bounds: Proof of Corollary 1

From Theorem 4, it is sufficient to prove the lower bound for the generalization error corresponding to the equivalent features $\phi_{\text{CL}}(\boldsymbol{z})$. Let $\boldsymbol{Z}$ denote the input design matrix with rows $\boldsymbol{z}_i$. Similarly, let $\boldsymbol{\Xi}$ denote the matrix with rows containing n independent Gaussian vectors, denoting the uncorrelated noise in the equivalent conditional Gaussian features defined by equation (121). We have that $\hat{a}_{\text{CL}}(\lambda, \boldsymbol{Z}, \boldsymbol{\Xi}) = \left(\boldsymbol{\Phi}_{CL}^{\top}\boldsymbol{\Phi}_{CL} + \frac{\lambda n}{N}\boldsymbol{I}\right)^{-1}\boldsymbol{\Phi}_{CL}^{\top}\boldsymbol{y}$. The generalization error can then be expressed as:

$$\begin{aligned}\mathcal{R}(W, \hat{a}_{\text{CL}}) &= \mathbb{E}_{\boldsymbol{z},\xi}\left[\left(f^{\star}(\boldsymbol{z}) - \hat{a}_{\text{CL}}(\lambda, \boldsymbol{Z}, \boldsymbol{\Xi})^{\top}\phi_{\text{CL}}(\boldsymbol{z})\right)^2\right] \\ &= \mathbb{E}_{\xi}\left[\mathbb{E}_{\boldsymbol{z}}\left[\left(f^{\star}(\boldsymbol{z}) - \hat{a}_{\text{CL}}(\lambda, \boldsymbol{Z}, \boldsymbol{\Xi})^{\top}\phi_{\text{CL}}(\boldsymbol{z})\right)^2\right]\right],\end{aligned}$$

where the last line follows from Fubini's theorem.

We note that the predictor $\hat{f}(\boldsymbol{z}) = \frac{1}{\sqrt{p}} \hat{a}_{\mathrm{CL}}^{\top} \phi_{\mathrm{CL}}(\boldsymbol{z})$ is a linear function of $\boldsymbol{z}^{\perp}$ with coefficients dependent on $z_{\boldsymbol{v}}$. Therefore, $\hat{f}(\boldsymbol{z}) \in \mathcal{P}_{\boldsymbol{v},1}$.

For a fixed value of ξ, we obtain the following expression for the generalization error:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{z}} \left[(f^{\star}(\boldsymbol{z}) - \hat{f}(\boldsymbol{z}))^2 \right] &= \| f^{\star} - \hat{f} \|^2. \\ &= \| P_{v,1}(f^{\star} - \hat{f}) \|^2 + \| P_{v,>1}(f^{\star} - \hat{f}) \|^2 \\ &\geq \| P_{v,>1}(f^{\star})^2 \|^2, \end{aligned}$$

where we used that $P_{v,>1}(f^{\star} - \hat{f}) = P_{v,>1}(f^{\star})$. Since the projection of $f^{\star}$ on the orthogonal complement of the teacher subspace is 0, Corollary 1 then follows using $P_{V^{\star}} \boldsymbol{v} \xrightarrow{\mathbb{P}} \frac{\mu}{\sqrt{p}} \boldsymbol{v}^{\star}$ and the dominated convergence theorem for the RHS.