How Two-Layer Neural Networks Learn, One (Giant) Step at a Time

Yatin Dandi
Information, Learning and Physics (IdePHICS) Lab
École Polytechnique Fédérale de Lausanne
Route Cantonale, 1015 Lausanne, Switzerland
yatin.dandi@epfl.ch

Florent Krzakala
Information, Learning and Physics (IdePHICS) Lab
École Polytechnique Fédérale de Lausanne
Route Cantonale, 1015 Lausanne, Switzerland
florent.krzakala@epfl.ch

Bruno Loureiro
Département d'Informatique
École Normale Supérieure - PSL & CNRS
45 rue d'Ulm, F-75230 Paris cedex 05, France
bruno.loureiro@di.ens.fr

Luca Pesce
Information, Learning and Physics (IdePHICS) Lab
École Polytechnique Fédérale de Lausanne
Route Cantonale, 1015 Lausanne, Switzerland
luca.pesce@epfl.ch

Ludovic Stephan
Information, Learning and Physics (IdePHICS) Lab
École Polytechnique Fédérale de Lausanne
Route Cantonale, 1015 Lausanne, Switzerland
ludovic.stephan@epfl.ch

Mathematics of Modern Machine Learning Workshop at NeurIPS 2023.

Abstract

We investigate theoretically how the features of a two-layer neural network adapt to the structure of the target function through a few large batch gradient descent steps, leading to improvement in the approximation capacity with respect to the initialization. We compare the influence of batch size and that of multiple (but finitely many) steps. For a single gradient step, a batch of size $n = \mathcal{O}(d)$ is both necessary and sufficient to align with the target function, although only a single direction can be learned. In contrast, $n = \mathcal{O}(d^2)$ is essential for neurons to specialize to multiple relevant directions of the target with a single gradient step. Even in this case, we show there might exist "hard" directions requiring $n = \mathcal{O}(d^\ell)$ samples to be learned, where ℓ is known as the leap index of the target. The picture drastically improves over multiple gradient steps: we show that a batch-size of $n = \mathcal{O}(d)$ is indeed enough to learn multiple target directions satisfying a staircase property, where more and more directions can be learned over time. Finally, we discuss how these directions allows to drastically improve the approximation capacity and generalization error over the initialization, illustrating a separation of scale between the random features/lazy regime, and the feature learning regime. Our technical analysis leverages a combination of techniques related to concentration, projection-based conditioning, and Gaussian equivalence which we believe are of independent interest. By pinning down the conditions necessary for specialization and learning, our results highlight the interaction between batch size and number of iterations, and lead to a hierarchical depiction where learning performance exhibits a stairway to accuracy over time and batch size, shedding new light on how neural networks adapt to features of the data.

A central property behind the success of neural networks is their capacity to adapt to the features in the training data. Indeed, many of the classical machine learning methods, e.g. linear or logistic regression, are specifically designed to a restricted class of functions (e.g. generalized linear functions). Others, such as kernel methods, can adapt to larger function classes (e.g. square-integrable functions), but sometimes at prohibitively many samples. Despite the limitations, these methods enjoy well-understood theoretical guarantees: they are convex (hence easy to train) and given a target function, it is well-understood how many samples are needed to achieve a target accuracy. The situation is dramatically different for neural networks: despite being universal approximators, little is known on how to optimally train them or how many hidden units and/or samples are required to learn a given class of functions. Nevertheless, they have proven to be flexible, efficient and easy to optimize in practice, properties which are often attributed to their capacity to adapt to features in the data. Curiously, most of our current theoretical understanding of neural networks stems from the investigation of their lazy regime where features are *not* learned during training. In this work, we take some (giant) steps forward from the lazy regime.

Our central goal is to paint a complete picture of how two-layer neural networks adapt to the features of training data $(\boldsymbol{z}^\nu, y^\nu)_{\nu=1}^n \in \mathbb{R}^{d+1}$ in the *early phase* of training after the first few steps of gradient descent. We recall the reader that for data is supported in a high-dimensional space, the curse of dimensionality prevents efficient learning even under standard regularity assumptions on the target function such as Lipschitzness Devroye et al. [2013]. Hence, understanding the efficient learning performance of neural networks observed in practice requires additional assumptions on the data distribution. In this work, we focus on a popular synthetic data model consisting of: a) independently drawn standard Gaussian covariates $\boldsymbol{z}^\nu \sim \mathcal{N}(0, I_d)$; b) a target function $y^\nu = f^\star(\boldsymbol{z}^\nu)$ depending only on a finite number of relevant directions, also known as a *multi-index model*. In other words, there exists a low-dimensional subspace $V^\star \subset \mathbb{R}^d$ with fixed dimension k and a function $g^\star : V^\star \to \mathbb{R}$ such that:

$$y = f^\star(\boldsymbol{z}) \coloneqq g^\star(\Pi^\star \boldsymbol{z}) = g^\star(\langle \boldsymbol{w_1^\star}, \boldsymbol{z}\rangle, \cdots, \langle \boldsymbol{w_r^\star}, \boldsymbol{z}\rangle). \tag{1}$$

where $\Pi^\star$ is the orthogonal projection on $V^\star$. Note that in this model the features are isotropic, with all the structure in the data being in the target. In contrast to other popular models for structured data , such as low-dimensional support of the inputs or smoothness of the target function, kernel methods do not adapt to target functions depending on low-dimensional projections Bach

[2017]. This makes it an ideal playground for quantifying the adaptativity of neural networks in the feature-learning regime. Given this class of structured targets, we consider supervised learning with the simplest universal approximator neural network: a fully-connected two-layer network with first and second layer weights $W \in \mathbb{R}^{p\times d}$ and $\boldsymbol{a} \in \mathbb{R}^p$ and activation function $\sigma : \mathbb{R} \to \mathbb{R}$:

$$\hat{f}(\boldsymbol{z}; W, \boldsymbol{a}) = \frac{1}{\sqrt{p}} \sum_{i=1}^{p} a_i \sigma(\langle \boldsymbol{w}_i, \boldsymbol{z} \rangle). \tag{2}$$

The central question in this work is to uncover how this two-layer neural network adapts to the low-dimensional target structure during training. In particular, we will be interested in quantifying how much data is required for the relevant directions of the target to be learned, and how this feature learning translates into the approximation capacity of the network with respect to kernel methods. We shall provide an informal presentation of our results and theorems, and refer the reader to Appendix A for the formal and precise presentation.

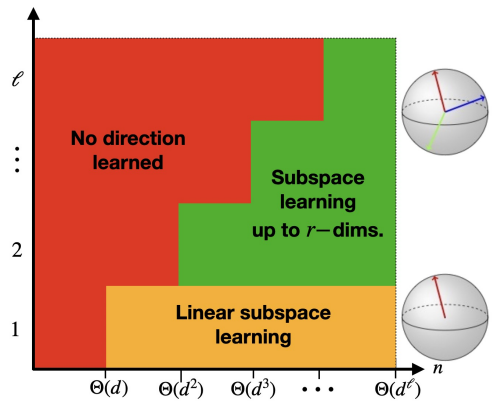

Figure 1: **Learning with a single gradient step.** Illustration of the relationship between batch size and target function complexity for learning multi-index functions with a single giant step in the $n = \Theta(d^k)$ regime (Thms. 1 & 2).

Informal summary of main results

To outperform the network at initialization (which can be regarded as a kernel method) the network must adapt to the data distribution. Mathematically, this translates to developing correlation in the first layer weights with the target subspace

$$V^\star := \mathrm{span}(\boldsymbol{w}_1^\star, \ldots, \boldsymbol{w}_r^\star)\,.$$

First we focus on **feature learning**, i.e. how the subspace $V^\star \subset \mathbb{R}^d$ is learned during training.

First GD step — First, we discuss **a single gradient step** of **full-batch gradient descent**, which turns out to be already non-trivial [Ba et al., 2022, Damian et al., 2022], with the update:

$$\boldsymbol{w}_i^1 = \boldsymbol{w}_i^0 - \frac{\eta}{2n} \sum_{\nu=1}^{n} \nabla_{\boldsymbol{w}_i} \left(y^\nu - \hat{f}(\boldsymbol{z}^\nu; W^0, \boldsymbol{a}^0) \right)^2 . \tag{3}$$

Theorems 1 and 2 identify a fundamental interplay between batch size and the complexity of the underlying target function. They are summarized in Fig. 1 (and a particular numerical example is shown in Fig. 4 in appendix). More precisely:

- We show that developing meaningful correlation with the target function requires a large batch size $n = \mathcal{O}(d)$ and learning rate $\eta = \Theta(p)$ when p, d, n are large. However, feature learning remains limited in this regime since we prove only a *single* direction can be learned. Thus, if the target depends on several relevant directions, only a "single neuron" approximation can be learned.
- Surpassing the single direction approximation with a single step *requires* a larger batch size with *at least* $n = \mathcal{O}(d^2)$ samples. This allows for the network weights to *specialize* to multiple target directions.
- Nonetheless, we show that there might be *hard* directions in the target which cannot be learned with $n = \mathcal{O}(d^2)$. Learning these directions necessitates a batch size of at least $n = \mathcal{O}(d^\ell)$, as well as suppressing the directions learned at $n = \mathcal{O}(d^{\ell-1})$, where ℓ is the *leap index* of the target (precisely defined in Def. 3) that thoroughly speaking corresponds to the lowest non-zero degree of the Hermite polynomials in the expansion of the target in this directions.
- Finally we note that the complexity of learning a low-dimensional target function is a topic that recently saw a surge of interest. For a single-index function with leap index ℓ, one-pass SGD has a sample complexity of $\mathcal{O}(d^{\ell-1})$ Ben Arous et al. [2021], and it has been recently shown Damian et al. [2023] that a smoothed version of SGD achieves a sample complexity of $\mathcal{O}(d^{\ell/2})$. This matches a lower bound from the correlation statistical query family, which encompasses all SGD-like methods.

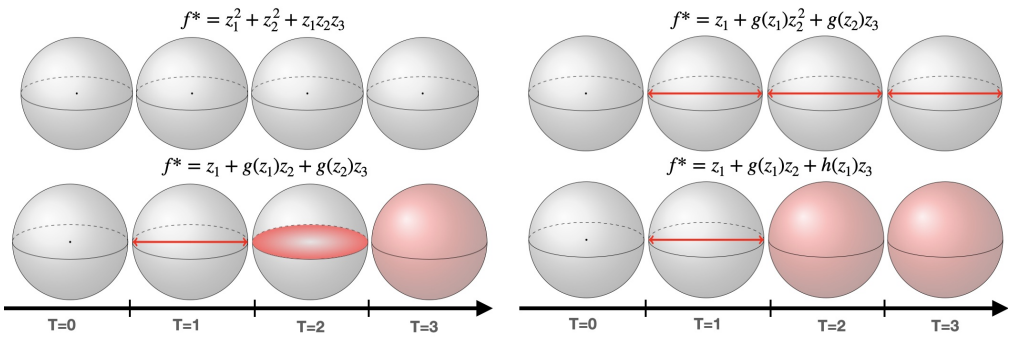

Figure 2: **Learning with multiple gradient steps.** An illustration of how neural networks trained using $n = \mathcal{O}(d)$ batches learn relevant directions after multiple GD training steps, allowing to learn more complex functions over time (see Thm. 3). Here we represent the space $V^\star$ of the relevant direction of the target function, and the (normalized) projection of learned direction W^t by the networks for different task $f^\star(\cdot)$ - the function $g(\cdot)$ and $h(\cdot)$ are assumed to have zero first two Hermite coefficients. During the early stage of training, neural networks first learn first a single direction associated to the linear part of the target, and then can learn over time other directions that are linear conditioned on the previous learned ones. Let $\{\boldsymbol{e}_i\}_{i\in[d]}$ be the standard basis of $\mathbb{R}^d$, the four examples show cases where: **Top left:** No directions is learned at all. **Top right:** The network can only learn a single direction $\boldsymbol{e}_1$ (single index regime). **Bottom left:** The network learns a new direction each time, $\boldsymbol{e}_1$, then $\boldsymbol{e}_2$ and finally $\boldsymbol{e}_3$. **Bottom right:** The network learns $\boldsymbol{e}_1$ at the first step and both $\boldsymbol{e}_2$ and $\boldsymbol{e}_3$ at the second steps.

In our single step setting, the sample complexity for large batch learning is $\mathcal{O}(d^\ell)$, which is worse than both the above methods. However, the time complexity of each algorithm paints a different picture. Both SGD algorithms are sequential, and require $\mathcal{O}(d)$ operations per step, which leads to a total time complexity of $\mathcal{O}(d^{\ell/2+1})$ at minimum. On the other hand, the computation of the update in eq. (1) is simply an average of independent terms, which is very easy to parallelize. Including the time to compute the average of each term e.g. using a *Gossip algorithm* Boyd et al. [2006], this sums up to a time complexity of $\mathcal{O}(d + \log(n))$. Such an algorithm is also amenable to decentralized learning schemes, where each agent only has access to a fraction of the overall data.

Learning over many GD iterations — The situation drastically improves when taking for **multiple gradient steps**. In this case, focusing on the linear batch size $n = \mathcal{O}(d)$ regime, and using a fresh batch of data at each GD iteration:

$$\boldsymbol{w}_i^{t+1} = \boldsymbol{w}_i^t - \frac{\eta}{2n} \sum_{\nu=1}^{n} \nabla_{\boldsymbol{w}_i} \left(y^\nu - \hat{f}(\boldsymbol{z}^\nu; W^t, \boldsymbol{a}^0) \right)^2 , \tag{4}$$

Note that splitting the training of the first and second layers and using a fresh batch of data at each iteration is a common approximation [Damian et al., 2022, Ba et al., 2022, Bietti et al., 2022]. In contrast to the recent works considering the population limit, we stress that here we take the batch size n to scale with the dimension d. In the realm of distributed and federated learning, scenarios with large batches, a single pass, and few iterations are often the norm Goyal et al. [2017], Li et al. [2020] (for instance this is the case when training large language models), further underlining the relevance of this scenario. In this case, Theorem 3 shows that more complex subspaces of the target directions *may* be progressively learned at each iteration, as we illustrate in Fig. 2. More precisely:

- Each additional gradient step allows for learning new perpendicular directions **upon the important condition that they are linearly connected to the previously learned directions** (see Sec .A.4 for precise definitions of this staircase property). Therefore, in contrast with a single step, taking multiple steps allows to learn a multiple-index target with only $n = \mathcal{O}(d)$ samples.
- Nonetheless, directions that are not coupled through the staircase property and with zero first Hermite coefficient cannot be learned in any finite number of steps. In fact, as discussed previously, they require a batch size of at least $n = \mathcal{O}(d^2)$. In other words, while multiple steps help specialization, it cannot help learning "hard" target directions.
- Similar to the *saddle-to-saddle* dynamics under gradient flow [Jacot et al., 2021, Abbe et al., 2023], the dynamics described through Thm.3 involves sequential learning of directions. However, in

the one-sample SGD regime [Ben Arous et al., 2021, Abbe et al., 2023] the gradient has vanishing correlation with new directions, thus requiring a polynomial number of updates to escape saddles. In contrast, the large-batch gradient updates contain a finite fraction of components along the new directions, allowing their learning through a single step. Moreover, we show that each update leads to a $\mathcal{O}(1)$ change in the components along direction in $V^\star$, obviating the need for coordinate-wise projections in Abbe et al. [2023].

The set of results described above provide a mathematical theory on how neural networks learn representations of the data over training. They corroborate, among others, the findings of Kalimeris et al. [2019], who observed that neural networks first fits the best linear classifier and subsequently learns functions of increasing complexity.

From features to generalization — Our last set of results connects feature learning with the approximation capacity of the network and illustrates that feature learning improves the learning of the target function $f^\star$ over random initialization.

- We show that a two-layer network with a finite second layer can *only* learn the part of the target function in the learned subspace (see Proposition 1). In fact, we conjecture that with p large enough (but still finite), it should be possible to approximate this part of $f^\star$ up to arbitrary precision (Conj. 1).
- As (possibly) universal kernels, large width two-layer networks at initialization enjoy of better approximation capacity. Indeed, Mei and Montanari [2022] proved that at initialization $W^{t=0}$, with $n = \Theta(d^k)$ only a degree k approximation of the target function can be learned. Our results characterize how feature learning allows us to improve over this sample complexity. In particular, we show that in the directions learned by the first layer, the target function $f^\star$ can be learned with less samples, while the component of $f^\star$ orthogonal to the learned features still requires $n = \mathcal{O}(d^k)$. While a complete mathematical control of the generalization error rates remains a difficult problem (see Conjecture 2), our results provide a clear separation on scales between two-layer networks and NTK-like methods, improving over the state-of-the art in the literature.
- In particular, in Corollary 1 we prove that with a single gradient step and $n, p = \mathcal{O}(d)$, one can *only* learn features in this one-dimensional subspace, doing no better than kernels in the orthogonal direction. This is illustrated in Fig.6 where we give an example where one step of gradient drastically improves generalization, and one where it does not. To prove these results, we provide a stronger *conditional* version of the Gaussian equivalence theorem [Mei and Montanari, 2022, Goldt et al., 2022, Hu and Lu, 2022], which is the backbone of Theorem 4, and we believe is of independent interest.

A more general overwiew of this broad field of research along with the formal statement of our main results is given in App. A. The code to reproduce our figures is available on GitHub, and we refer to App. B for details on the numerical implementations. Proofs are detailed in App. C and App. D.

Acknowledgements — We thank Denny Wu, Theodor Misiakiewicz, Loucas Pillaud-Vivien, Joan Bruna & Lenka Zdeborová for insightful discussions. We also acknowledge funding from the Swiss National Science Foundation grant SNFS OperaGOST, 200021_200390 and the *Choose France - CNRS AI Rising Talents* program.

References

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

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

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 5413–5452. PMLR, 02–05 Jul 2022.

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

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

S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, June 2006. ISSN 1557-9654. doi: 10.1109/TIT.2006.874516.

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 9768–9783. Curran Associates, Inc., 2022.

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

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

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

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

Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. Sgd on neural networks learns functions of increasing complexity. *Advances in neural information processing systems*, 32, 2019.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75 (4):667–766, 2022.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. The gaussian equivalence of generative models for learning with shallow neural networks. In Joan Bruna, Jan Hesthaven, and Lenka Zdeborova, editors, *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of *Proceedings of Machine Learning Research*, pages 426–471. PMLR, 16–19 Aug 2022.

Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 2022.

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

M. Opper and R. Urbanczik. Universal learning curves of support vector machines. *Phys. Rev. Lett.*, 86:4410–4413, May 2001. doi: 10.1103/PhysRevLett.86.4410.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural network. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14820–14830. Curran Associates, Inc., 2020.

Konstantin Donhauser, Mingqi Wu, and Fanny Yang. How rotational invariance of common kernels prevents generalization in high dimensions. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2804–2814. PMLR, 18–24 Jul 2021.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2021.12.003. Special Issue on Harmonic Analysis and Machine Learning.

Stefano Spigler, Mario Geiger, and Matthieu Wyart. Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124001, dec 2020. doi: 10.1088/1742-5468/abc61d.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1024–1034. PMLR, 13–18 Jul 2020.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12(1):2914, May 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-23103-1.

James B. Simon, Madeline Dickens, Dhruva Karkada, and Michael R. DeWeese. The eigenlearning framework: A conservation law perspective on kernel regression and wide neural networks, 2022.

Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10131–10143. Curran Associates, Inc., 2021.

Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Error rates for kernel classification under source and capacity conditions, 2022.

Lechao Xiao, Hong Hu, Theodor Misiakiewicz, Yue Lu, and Jeffrey Pennington. Precise learning curves and higher-order scalings for dot-product kernel regression. *Advances in Neural Information Processing Systems*, 35:4558–4570, 2022.

Rishabh Dudeja and Daniel Hsu. Learning single-index models in gaussian space. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1887–1930. PMLR, 06–09 Jul 2018. URL `https://proceedings.mlr.press/v75/dudeja18a.html`.

Gadi Naveh and Zohar Ringel. A self consistent theory of gaussian processes captures feature learning effects in finite cnns. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 21352–21364. Curran Associates, Inc., 2021.

Inbar Seroussi, Gadi Naveh, and Zohar Ringel. Separation of scales and a thermodynamic description of feature learning in some cnns. *Nature Communications*, 14(1):908, Feb 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-36361-y.

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. In *International Conference on Learning Representations*, 2022.

Blake Bordelon and Cengiz Pehlevan. Dynamics of finite width kernel and prediction fluctuations in mean field neural networks, 2023.

Leonardo Petrini, Francesco Cagnetta, Eric Vanden-Eijnden, and Matthieu Wyart. Learning sparse features can lead to overfitting in neural networks, 2022.

Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. Generalisation error in learning with random features and the hidden manifold model. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3452–3462. PMLR, 13–18 Jul 2020.

Oussama Dhifallah and Yue M. Lu. A precise performance analysis of learning with random features, 2020.

Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 18137–18151. Curran Associates, Inc., 2021.

Bruno Loureiro, Cedric Gerbelot, Maria Refinetti, Gabriele Sicuro, and Florent Krzakala. Fluctuations, bias, variance &; ensemble of learners: Exact asymptotics for convex losses in high-dimension. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 14283–14314. PMLR, 17–23 Jul 2022.

Dominik Schröder, Hugo Cui, Daniil Dmitriev, and Bruno Loureiro. Deterministic equivalent and error universality of deep random features learning, 2023.

David Bosch, Ashkan Panahi, and Babak Hassibi. Precise asymptotic analysis of deep random feature models, 2023.

Andrea Montanari and Basil N. Saeed. Universality of empirical risk minimization. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4310–4312. PMLR, 02–05 Jul 2022.

Yatin Dandi, Ludovic Stephan, Florent Krzakala, Bruno Loureiro, and Lenka Zdeborová. Universality laws for gaussian mixtures in generalized linear models, 2023.

Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pages 4782–4887. PMLR, 2022.

Emmanuel Abbe, Enric Boix-Adsera, Matthew S Brennan, Guy Bresler, and Dheeraj Nagaraj. The staircase property: How hierarchical structure can guide deep learning. *Advances in Neural Information Processing Systems*, 34:26989–27002, 2021.

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

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

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

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

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

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

Harold Grad. Note on N-dimensional hermite polynomials. *Communications on Pure and Applied Mathematics*, 2(4):325–330, 1949. ISSN 1097-0312. doi: 10.1002/cpa.3160020402.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On Lazy Training in Differentiable Programming. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks, 2023.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. *Advances in neural information processing systems*, 30, 2017.

Friedrich Gotze, Holger Sambale, and Arthur Sinulis. Concentration inequalities for polynomials in α-sub-exponential random variables. *Electronic Journal of Probability*, 2019.

W.H. Greub. *Multilinear Algebra*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2012. ISBN 9783662007952.

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

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

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Victor H. de la Pena and S. J. Montgomery-Smith. Decoupling Inequalities for the Tail Probabilities of Multivariate U-Statistics. *The Annals of Probability*, 23(2):806–816, April 1995. ISSN 0091-1798, 2168-894X. doi: 10.1214/aop/1176988291.

Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, and Bruno Loureiro. From high-dimensional & mean-field dynamics to dimensionless ODEs: A unifying approach to SGD in two-layers networks, February 2023. arXiv:2302.05882 [cond-mat, stat] type: article.

Svante Janson. *Gaussian hilbert spaces*. Number 129. Cambridge university press, 1997.

Michael Polyak. Feynman diagrams for pedestrians and mathematicians. *Graphs and patterns in mathematics and theoretical physics*, 73:15–42, 2005.

Boris Samuilovich Mityagin. The zero set of a real analytic function. *Mathematical Notes*, 107(3-4): 529–530, 2020.

Achim Klenke. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

— Appendices —

A Statement of main theoretical results

In this section we present the formal statements of our main results. Before doing that we present an additional overview on relevant related works, and needed mathematical preliminaries.

Other related works — The analysis of high-dimensional asymptotics of kernel regression has provided valuable insights into the advantages and limitations of kernel methods [Dietrich et al., 1999, Opper and Urbanczik, 2001, Ghorbani et al., 2019, 2020, Donhauser et al., 2021, Mei et al., 2022, Spigler et al., 2020, Bordelon et al., 2020, Canatar et al., 2021, Simon et al., 2022, Cui et al., 2021, 2022, Xiao et al., 2022]. In particular, a similar stairway picture as in Fig. 2 emerged from these works Xiao et al. [2022]. The key difference, however, is that at each regime $n = \mathcal{O}(d^k)$, kernels can only learn the k-th Hermite polynomial of the target. This should be contrasted with our feature learning regime where, once a direction is learned, all its Hermite coefficients are learned. Feature learning corrections to kernel methods have been investigated in Dudeja and Hsu [2018], Naveh and Ringel [2021], Seroussi et al. [2023], Atanasov et al. [2022], Bietti et al. [2022], Bordelon and Pehlevan [2023], Petrini et al. [2022]. On a complementary line of work, exact asymptotic results for the the random features models have been derived in the literature [Mei and Montanari, 2022, Gerace et al., 2020, Dhifallah and Lu, 2020, Loureiro et al., 2021, 2022, Schröder et al., 2023, Bosch et al., 2023]. A large part of these results are enabled by the Gaussian equivalence property [Goldt et al., 2022, Hu and Lu, 2022, Montanari and Saeed, 2022, Dandi et al., 2023].

Closer to us are [Ba et al., 2022, Damian et al., 2022]. Ba et al. [2022] showed that a sufficiently large single gradient step allows to beat kernel methods. Ba et al. [2022] showed that a single gradient step yields an approximately rank-one change on the weights which is enough to beat kernel methods, but did not characterize the impact on the generalization error. In our work, we prove that with a single gradient step and $n, p = \mathcal{O}(d)$, one can only learn features in this one-dimensional subspace, doing no better than kernels in the orthogonal direction. Additionally, their results are limited to single-index target and to a single gradient step. In contrast, Damian et al. [2022] showed that with $n = \omega(d^2)$ samples, two-layer neural networks can specialize to more than one direction of a multi-index target function with zero first Hermite coefficient (ℓ=2), and were again limited to a single step. Our work extends their sufficient conditions on the sample complexity to general $\ell \geq 1$. We also show they are also necessary, i.e. $\forall \epsilon > 0$, with less data $n = \Theta(d^{\ell-\epsilon})$ one cannot do better than random features. Thus, our results prove a clear separation between the class of functions learned within the $\Theta(d)$ batch-size setting of Ba et al. [2022] and the $\Theta(d^2)$ batch-size setting of Damian et al. [2023] and establish a general hierarchy of functions requiring increasing batch-size to be learned with a gradient step. Additionally, we characterize which class of multi-index targets can be instead learned with $n = \mathcal{O}(d)$ with *multiple* steps.

In a related but different vein, Abbe et al. [2022], building upon the earlier work of Abbe et al. [2021] show how a "staircase" property of target functions characterizes the sample complexity for infinite width two-layer networks trained by SGD, a.k.a. the mean-field regime Mei et al. [2018], Chizat and Bach [2018], Rotskoff and Vanden-Eijnden [2022], Sirignano and Spiliopoulos [2020]). Their focus, however, was on the peculiar case of *sparse boolean functions* depending solely on linear coefficients, without considering direction-specific specialization. More recently, Abbe et al. [2023] provided partial results for the case of isotropic Gaussian data, always in the regime of one-pass SGD, involving basis-dependent projections. Here we consider the more generic case of multi-index target functions with Gaussian data with large-batch SGD, without the use of a knowledge of the basis, and show that a "directional staircase" behavior arises when iterating a few giant gradient steps, while a related, but definitely different, picture arises with single steps depending on the batch size. We also provide a sharp characterization of when this phenomenon appears for multi-index targets and networks trained under large batch SGD, and provide a bound on the resulting generalization error.

Similar to the "summary statistics" approach in Saad and Solla [1995], Ben Arous et al. [2021], Ben Arous et al. [2022], our analysis is based upon the concentration of the overlaps of the neurons with the target subspace and their norms, instead of the concentration of the full gradient vector considered in recent works such as Abbe et al. [2022], Damian et al. [2022], removing any requirements on the constants in the sample complexity for alignment along the target subspace.

A.1 Preliminaries

Before stating our main results, 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

$$\langle f, g \rangle_\gamma = \int_{\mathbb{R}^m} fg \, \mathrm{d}\gamma_m = \mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, I_m)}[f(\boldsymbol{z}) g(\boldsymbol{z})]. \tag{5}$$

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

Definition 1 (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 $\boldsymbol{x} \in \mathbb{R}^m$,*

$$f(\boldsymbol{x}) = \sum_{j \in \mathbb{N}} \langle C_j(f), \mathcal{H}_j(\boldsymbol{x}) \rangle \tag{6}$$

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 2 (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*

$$C = \sum_{j_1, \dots, j_k = 1}^{r} S_{j_1, \dots, j_k} \boldsymbol{u}_{j_1} \otimes \cdots \otimes \boldsymbol{u}_{j_k} \tag{7}$$

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.

A.2 Setting and assumptions

Before stating our main results, we introduce the setting and main assumptions required. The first concerns the class of target functions we consider.

Assumption 1 (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 fixed number of orthonormal vectors $(\boldsymbol{w}_1, \dots, \boldsymbol{w}_r)$ and a fixed function $g^\star : \mathbb{R}^r \to \mathbb{R}$ such that*

$$y = f^\star(\boldsymbol{z}) \coloneqq g^\star(\langle \boldsymbol{w}_1^\star, \boldsymbol{z} \rangle, \dots, \langle \boldsymbol{w}_r^\star, \boldsymbol{z} \rangle). \tag{8}$$

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. Next, we define the leap index, a fundamental quantity which precisely parametrizes what are the first directions to be learned.

Definition 3 (Leap index)**.** *Since the input data is Gaussian $\boldsymbol{z} \sim \mathcal{N}(0, I_d)$, the target function admits a decomposition in terms of the Hermite decomposition (see Definition 1). We define the* leap index *of the target function $f^\star$ as the first integer $\ell > 0$ such that $C_\ell^\star \neq 0$:*

$$\ell = \min\{j \in \mathbb{N} : \langle f^\star, \mathcal{H}_j \rangle_\gamma \neq 0\} \tag{9}$$

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

Assumption 2 (Training procedure)**.** *Consider the following random initialization for the weights:*

$$\sqrt{p} \cdot a_i^0 \overset{i.i.d}{\sim} \mathrm{Unif}([-1, 1]) \quad \textit{and} \quad \boldsymbol{w}_i^0 \overset{i.i.d}{\sim} \mathrm{Unif}(\mathcal{S}^{d-1}). \tag{10}$$

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:

$$a_i^0 = -a_{p-i+1}^0 \quad \text{and} \quad \boldsymbol{w}_i^0 = \boldsymbol{w}_{p-i+1}^0 \quad \text{for all } i \in [p/2], \tag{11}$$

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 1, and the first layer weights are updated according to:*

$$\boldsymbol{w}_i^{t+1} = \boldsymbol{w}_i^t - \frac{\eta}{2n} \sum_{\nu=1}^n \nabla_{\boldsymbol{w}_i} \left(y^\nu - \hat{f}(\boldsymbol{z}^\nu; W^t, \boldsymbol{a}^0) \right)^2, \tag{12}$$

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:*

$$\hat{\boldsymbol{a}} = \arg\min_{\boldsymbol{a} \in \mathbb{R}^p} \frac{1}{2n} \sum_{\nu=1}^n \left(y^\nu - \hat{f}(\boldsymbol{z}^\nu; W^T, \boldsymbol{a}) \right)^2 + \lambda \|\boldsymbol{a}\|^2. \tag{13}$$

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.

We are now ready to state our main technical results.

A.3 Single gradient step

Our starting point is a *single* giant gradient step, and the phenomenology described in Fig. 1. The main goal is to determine under which conditions the relevant directions $\Pi^\star \boldsymbol{z}$ of the target function $f^\star$ can be learned with the training procedure introduced in Assumption 2. Hence, a crucial object in our analysis is given by the projection of the network weights in the space spanned by the target relevant directions:

$$\boldsymbol{\pi}_i^t = \Pi^\star \boldsymbol{w}_i^t; \tag{14}$$

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

Our first result is of a negative nature, showing the impossibility of learning in the data-scarce regime:

Theorem 1. *Let ℓ be the leap index of $f^\star$ (3), and assume that $n = \mathcal{O}(d^{\ell-\delta})$ for some $\delta > 0$. Then, with probability at least $1 - cpe^{-c(\delta)\log(d)^2}$, there exists a universal constant c such that for any $i \in [p]$,*

$$\frac{\|\boldsymbol{\pi}_i^1\|^2}{\|\boldsymbol{w}_i^1\|^2} \leq c \, \frac{\mathrm{polylog}(d)}{d^{(1 \wedge \delta)/2}}. \tag{15}$$

In other words, for *every* neuron i, only a vanishing fraction of the weight $\boldsymbol{w}_i^1$ lies in the target subspace $V^\star$. In particular, if $\delta > 1$, this large gradient step does not improve over the initial random feature weights.

On the other hand, when $n = \Omega(d^\ell)$, we are able to characterize exactly what is being learned in one gradient step.

Theorem 2. *Assume that the ℓ-th Hermite coefficient μ_ℓ of σ is nonzero, and set the learning rate*

$$\eta = pd^{\frac{\ell-1}{2}}. \tag{16}$$

Then, with probability at least $1 - ce^{-c\log(d)^2}$, there exists a random variable X independent of d with positive expectation such that

$$\frac{\|\boldsymbol{\pi}_i^1\|^2}{\|\boldsymbol{w}_i^1\|^2} \geq X_i, \tag{17}$$

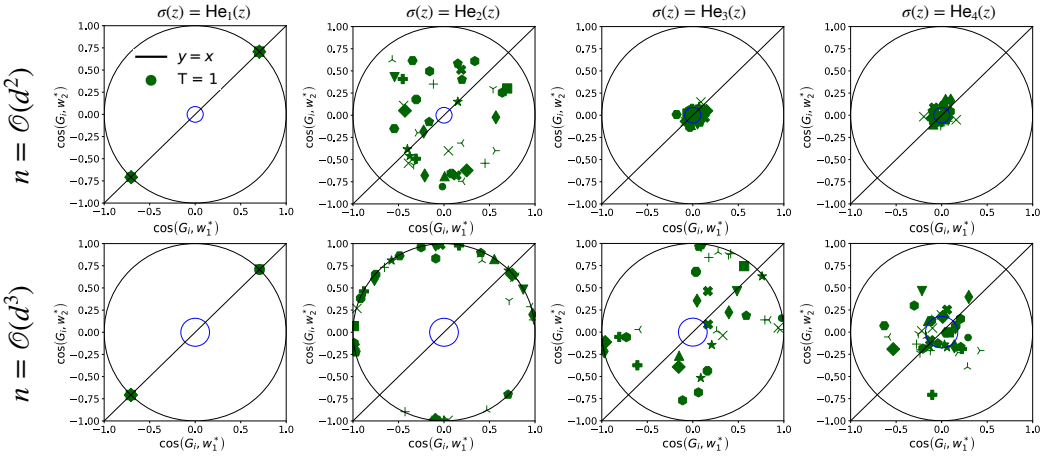

Figure 3: **Feature learning after a single step.** Specialization of hidden units in the $n = \mathcal{O}(d^k)$ regime ($k = 2, 3$). The plots show the cosine similarity of the gradient with respect to the target vectors $(\boldsymbol{w}_1^\star, \boldsymbol{w}_2^\star)$ for $p = 40$ different neurons, 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. See details in App. B.

where $X_1, \dots, X_p$ are i.i.d copies of X. Further, let $\boldsymbol{u}_1^\star, \dots, \boldsymbol{u}_{r_\ell}^\star$ be the higher-order singular vectors of $C_\ell^\star$, and define

$$V_\ell^\star = \mathrm{span}(\boldsymbol{u}_1^\star, \dots, \boldsymbol{u}_{r_\ell}^\star). \tag{18}$$

Then, the projections $\boldsymbol{\pi}_i^1$ asymptotically belong to $V_\ell^\star$, in the sense that there exists a constant c such that

$$\|(I - \Pi_{V_\ell^\star})\boldsymbol{\pi}_i^1\| \leq c\,\frac{\mathrm{polylog}(d)}{\sqrt{d}}, \tag{19}$$

and they span the space $V_\ell^\star$.

Note that in the case $\ell = 1$ the learned subspace $V_\ell^\star$ is one dimensional and is identified by the first Hermite coefficient of the target $C_1(f^\star)$: this corresponds to the "linear subspace learning" regime exemplified in Fig. 1. Some aspects of the results above were already present in previous works, with key differences: Ba et al. [2022] shows the existence of a rank-one property of the gradient at initialization for $n = \Theta(d)$, and Damian et al. [2022] implies the positive part of our result for $n = \Theta(d^2)$, provided that $V_2^\star = V^\star$ (which corresponds to their Assumption 2). Our theorem allows us to obtain the matching lower bounds, demonstrating their tightness, and provides the generic picture for *any higher* powers of d. In particular, our results prove a clear separation between the class of functions learned within the $\Theta(d)$ batch-size setting of Ba et al. [2022] and the $\Theta(d^2)$ batch-size setting of Damian et al. [2022] and establish a general hierarchy of functions requiring increasing batch-size to be learned with a single gradient step. We refer to Appendix C for the proofs of the theorems and a more detailed technical discussion.

A.4 Learning with many steps

We now move to the effect of multiple gradient steps, as described in Fig. 2. As we shall see, while the effect of multiple steps is limited to a subclass of functions, it allows to learn them meaningfully with few iterations. For simplicity, we restrict ourselves to the case $n = \mathcal{O}(d)$, although we expect the findings to hold in the other regimes. We unveil the intertwined dependence between representation learning efficiency and a "directional staircase" condition. Informally: once a direction is learned by the first-layer weights, the next gradient step uses it as a ladder to learn the next ones, upon the assumption that they are linearly connected (as we will make precise next). The resulting hierarchical learning picture extends to the large batch case the observations of Abbe et al. [2023] specialized in batch one single-pass SGD, using basis-dependent projections, and supressing the

interactions between neurons. Furthermore, unlike Abbe et al. [2023], both our definition 4 and the training algorithm are basis-independent. Lastly, our analysis incorporates the effect of the learned output $\hat{f}(\boldsymbol{z}; W^t, \boldsymbol{a})$ on the gradient updates, which leads to interaction between neurons.

We formalize the hierarchical learning by introducing the notion of *subspace conditioning*:

Definition 4 (Subspace conditioning)**.** *Let V be a vector space, and $U \subseteq V$ a subspace. For any function $f : V \to \mathbb{R}$, and $\boldsymbol{x} \in U$, we define the* conditional *function $f_{U,\boldsymbol{x}} : U^\perp \to \mathbb{R}$ as*

$$f_{U,\boldsymbol{x}}(\boldsymbol{x}^\perp) = f(\boldsymbol{x} + \boldsymbol{x}^\perp) \tag{20}$$

In short, the function $f_{U,\boldsymbol{x}}$ corresponds to f "conditioned" on the projection of its argument in U. Its first Hermite coefficient will be denoted as

$$\mu_{U,\boldsymbol{x}}(f) = \mathbb{E}_{\boldsymbol{x}^\perp \sim \mathcal{N}(0,I)} \left[\nabla_{\boldsymbol{x}^\perp} f_{U,\boldsymbol{x}}\right] \tag{21}$$

We are now equipped to state our main result describing sequential learning of directions under multiple gradient steps with batch size $n = \Theta(d)$. We denote by $W^\star \in \mathbb{R}^{r \times d}$ the matrix with rows $\boldsymbol{w_1^\star}, \cdots, \boldsymbol{w_r^\star}$. To avoid degeneracy issues, we restrict ourselves to polynomial activations and target functions.

Assumption 3. *Both the student activation $\sigma : \mathbb{R} \to \mathbb{R}$ and the teacher function $g^\star : \mathbb{R}^r \to \mathbb{R}$ are fixed polynomials with degrees independent of d, n. Furthermore,* $\deg(\sigma) \geq \deg(g^\star)$.

Theorem 3. *Assume that $n = \Theta(d)$, and $\eta > 0, p \in \mathbb{N}$ are fixed. Define a sequence of nested subspaces $U_0^\star \subseteq U_1^\star \subseteq \cdots \subseteq U_t^\star \subseteq \ldots$ as*

- $U_0^\star = \{0\}$,
- *for any $t \geq 0$, $U_{t+1}^\star = U_t^\star \oplus \operatorname{span}\left(\{\mu_{U_t^\star,\boldsymbol{x}}(f^\star) : \boldsymbol{x} \in U_t^\star\}\right)$.*

Then, under assumptions 2 and 3, after t gradient steps of the form (4)*, W^t satisfies the following with high-probability, for all $i \in [p]$:*

(i) Let $\boldsymbol{v} \in U_t^\star$, with $\|\boldsymbol{v}\| = 1$ be arbitrary. Let $\boldsymbol{v}^m = W^\star \boldsymbol{v}$ denote the components of $\boldsymbol{v}$ along the target basis. Then there exists an almost surely positive random variable $X_{t,\boldsymbol{v}^m,\boldsymbol{a}}$, independent of d, n such that $|\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle| \geq X_{t,\boldsymbol{v},\boldsymbol{a}} + O(\frac{\operatorname{polylog}(d)}{\sqrt{d}})$.

(ii) For any $\boldsymbol{v} \in {U^\perp}_t^\star \cap V^\star$, $|\langle \boldsymbol{w}_i, \boldsymbol{v}\rangle| = O(\frac{\operatorname{polylog}(d)}{\sqrt{d}})$.

Informally, the above statements imply that almost surely over $\boldsymbol{a}$, the first layer learns exactly $U_t^\star$.

It is easy to check that the above Theorem is consistent with Theorems 1 and 2 since $f_{\{0\},\boldsymbol{0}} = f$, and $v^\star$ is exactly the first hermite of $f^\star$.

Examples — As an illustration, we work out the subspaces $U_t^\star$ for simplified versions of the examples in Figure 5. For simplicity, we will take $\boldsymbol{w}_k^\star = \boldsymbol{e}_k$, the k-th vector of the standard basis, so that $\langle \boldsymbol{w}_k^\star, \boldsymbol{z}\rangle = z_k$.

• $f^\star(\boldsymbol{z}) = z_1 + z_2 + z_1^2 - z_2^2$: the normalized first Hermite coefficient of $f^\star$ is $\boldsymbol{v}^\star = (\boldsymbol{e}_2 + \boldsymbol{e}_2)/\sqrt{2}$, so $U_1^\star = \operatorname{span}(\boldsymbol{e}_1 + \boldsymbol{e}_2)$. Then, we can rewrite $f^\star$ in the new basis $(\boldsymbol{v}^\star, \boldsymbol{v}^\perp)$, to get

$$f^\star(\boldsymbol{z}) = \sqrt{2}\langle \boldsymbol{v}^\star, \boldsymbol{z}\rangle + 2\langle \boldsymbol{v}^\star, \boldsymbol{z}\rangle\langle \boldsymbol{v}^\perp, \boldsymbol{z}\rangle.$$

Hence, if $\boldsymbol{x} = \lambda \boldsymbol{v}^\star$, we have $\mu_{U_t^\star,\boldsymbol{x}}(f^\star) = 2\lambda \boldsymbol{v}^\perp$, and hence $U_2^\star = V^\star$.

• $f^\star(\boldsymbol{z}) = z_1 + z_2 + z_1^2 + z_2^2$: just as the above example, $U_1^\star = \operatorname{span}(\boldsymbol{e}_1 + \boldsymbol{e}_2)$, and hence we perform the same change of basis. However, this time,

$$f^\star(\boldsymbol{z}) = \sqrt{2}\langle \boldsymbol{v}^\star, \boldsymbol{z}\rangle + \langle \boldsymbol{v}^\star, \boldsymbol{z}\rangle^2 + \langle \boldsymbol{v}^\perp, \boldsymbol{z}\rangle^2.$$

This implies that for any $\boldsymbol{x} \in U_1^\star$, $\mu_{U_t^\star,\boldsymbol{x}}(f^\star) = \boldsymbol{0}$, and the direction of $\boldsymbol{v}^\perp$ is never learned.

We provide a proof of Theorem 3 in Appendix C. The notion of subspace conditioning (Definition 4) arises through an inductive decomposition of the projections of the gradient along the target subspace. Furthemore, our theory leads to a precise prediction of the orientation of the gradient in the target subspace after a finite number of steps. This is illustrated in Figure 5. Note that the non-linearity of the activation function allows different neurons to simultaneously specialize along different directions in contrast to the rank-one increase at each saddle in deep linear networks Jacot et al. [2021] under vanishing initialization. We refer to Appendix B for additional discussion.

To conclude this section, we provide an overview of our present understanding of the effect of batch-size (sample complexity), number of gradient-steps (iteration complexity), and the number of neurons (overparameterization) on learning different multi-index target functions in table 1. For simplicity, when discussing the leap index ℓ, we restrict to functions where all the directions in the target subspace have the same leap i.e when $V_\ell^* = V^*$. We define "staircase functions" as target functions such that the sequence of subspaces defined through subspace conditioning in Theorem 3 eventually span V^*.

Complexity of f^*	SGD with $n=1$	One-step GD with $n=\Theta(d^\ell)$	Multi-pass GD with $n=\Theta(d)$
Single-index, $\ell=1$	$\tau=n_{\mathrm{T}}=d$	$\tau=1, n_{\mathrm{T}}=n=\Theta(d)$	$\tau=1, n_{\mathrm{T}}=\Theta(d)$
Single-index, $\ell=2$	$\tau=n_{\mathrm{T}}=d\log d$	$\tau=1, n_{\mathrm{T}}=n=\Theta(d^2)$	$\tau=\Theta(\log d), n_{\mathrm{T}}=\Theta(d\log d)$
Single-index, $\ell>2$	$\tau=n_{\mathrm{T}}=d^{\ell-1}$	$\tau=1, n_{\mathrm{T}}=n=\Theta(d^\ell)$	$\tau=\Theta(d^{\ell-2}), n_{\mathrm{T}}=\Theta(d^{\ell-1})$
Staircase	$\tau=n_{\mathrm{T}}=d$	$\tau=1, n_{\mathrm{T}}=n=\Omega(d^2)$	$\tau=\Theta(1), n_{\mathrm{T}}=\Theta(d)$

Table 1: Number of iterations τ and sample complexity n_{T} needed to learn the features/directions of the function f^* with SGD one-sample at a time, using a single step-gradient descent, or using one-pass GD with large $n=O(d)$ batches. Results in (blue) are from Ben Arous et al. [2021], Abbe et al. [2023]. The upper bounds on the sample complexity are known to hold from (brown) Ba et al. [2022] and (orange) Damian et al. [2022], while the corresponding lower bound, as well as the results in black, are proven in the present paper. (red) are educated guesses, which are out of the scope of the paper since we only focus on a finite number of iterations τ. For staircase functions, one-step GD may require more than $\Theta(d^2)$ samples depending on the maximum leap across directions in the target subspace.

A.5 From feature learning to generalization bounds

We now investigate the consequence of Theorem 2 in the actual performance of the neural network. At a high-level, one expects that learning a subspace U by the first-layer effectively reduces the input-dimensionality to that of U, allowing the network to learn a function dependent on U using a significantly fewer number of examples and neurons. However, for learning functions dependent on directions not yet learned by W, we expect a requirement of a large number of neurons as well as samples, analogous to the setting of random features. In the following, we make this intuition precise by proving that the generalization error of training the second layer on the top of the learned features is lower bounded precisely by the directions which were not learned by the first-layer. Our results and conjectures incorporate the dependence on both the number of samples and the number of hidden neurons.

Finite p regime — This is the simplest case: in this setting, there is simply no way to fit anything beyond the learned directions, even with an infinite amount of data. This is formalized by the following proposition:

Proposition 1. *Assume that p is bounded as $n, d \to \infty$, and that the first layer W only learns a subspace $U \subseteq V^\star$. For any choice of second layer $\boldsymbol{a}$ such that $\|\boldsymbol{a}\|_\infty \leq \mathcal{O}(1)$, we have*

$$\mathbb{E}\left[\left(f^\star(\boldsymbol{z}) - \hat{f}(\boldsymbol{z}; W, \boldsymbol{a})\right)^2\right] \geq \mathbb{E}\left[\mathrm{Var}\left(f^\star(\boldsymbol{z}) \mid P_U \boldsymbol{z}\right)\right] - o(1) \tag{22}$$

where P_U is the orthogonal projection on U.

We refer to Appendix D for the proof of the proposition. Intuitevely, the right-hand side of the above inequality corresponds to the predictor $\hat{f}(\boldsymbol{z}) = \mathbb{E}\left[f^\star(\boldsymbol{z}) \mid P_U \boldsymbol{z}\right]$, which is the best predictor that depends only on $P_U \boldsymbol{z}$. Conversely, we expect that with enough neurons, it is possible to approximate this predictor:

Conjecture 1. *For any $\delta > 0$, there exists a $p_0 \in \mathbb{N}$ such that if $p \geq p_0$, there exists a choice of second layer $\boldsymbol{a}$ satisfying $\|\boldsymbol{a}\|_\infty \leq \mathcal{O}(1)$, such that as $n, p \to \infty$*

$$\mathbb{E}\left[\left(f^\star(\boldsymbol{z}) - \hat{f}(\boldsymbol{z}; W, \boldsymbol{a})\right)^2\right] \leq \mathbb{E}\left[\mathrm{Var}\left(f^\star(\boldsymbol{z}) \mid P_U \boldsymbol{z}\right)\right] + \delta + o(1) \tag{23}$$

Proving the above conjecture for generic target and activation functions requires proving a sufficient spread of W along the learned subspace U together with an approximation result for finite-dimensional neural networks having activation function σ. Such a conjecture is proven in Damian et al. [2022] for the specific case of $U = V^\star$ (in which case the first term in the RHS is zero) and $\sigma = \mathrm{relu}$, with an additional bias term. For generic σ, such approximation results are scarcer. In particular, since the projection of the weights W along U cannot be chosen arbitrarily, one cannot directly invoke classical approximation results such as the ones based on Barron spaces. However, the components of W along U are still randomly distributed across neurons. Therefore, one possible approach could be to use results on the generalization of random feature models on finite-dimensional inputs such as Rudi and Rosasco [2017]. We leave such an analysis for future work.

Beyond the fixed width regime — When the number of neurons is allowed to diverge, it is no longer impossible to learn funtions along the directions that are not present in U. Indeed, Mei et al. [2022] shows that even in the absence of feature learning (i.e. when the first layer is random), it is possible to learn a polynomial approximation of $f^\star$ of degree k as long as $n, p = \Omega(d^{k+\delta})$ for some $\delta > 0$. The exact performance of the network may heavily depend on the training process for the second layer. Intuitively, we expect the following behavior:

(i) On the directions present in W, with enough variety in the neurons, the model behaves similarly as a random feature one in this *finite*-dimensional space, and are thus able to learn "everything".
(ii) On the orthogonal directions, however, the models still behave as a random feature model in high dimension, and thus the polynomial limitations of Mei et al. [2022] still apply.

To formalize this conjecture, we introduce the following definition:

Definition 5. *Let V be a vector space, and $U \subseteq V$ a subspace. For any $k \geq 0$, we define the space $\mathcal{P}_{U,k}$ of functions $f : V \to \mathbb{R}$ such that for any $\boldsymbol{x} \in U$, the function $f_{U,\boldsymbol{x}}$ introduced in Definition 4 is a polynomial of degree at most k.*

Simply put, the space $\mathcal{P}_{U,k}$ consists of polynomials in $\boldsymbol{x}^\perp$ of degree at most k, whose coefficients can be functions of $\boldsymbol{x}$. We will denote by $P_{U,\leq k}$ and $P_{U,>k}$ the projections on $\mathcal{P}_{U,k}$ and $\mathcal{P}_{U,k}^\perp$ in $\ell^2(\mathbb{R}, \gamma)$, respectively, where γ is the Gaussian measure. This allows us to write the above intuition in the following precise conjecture:

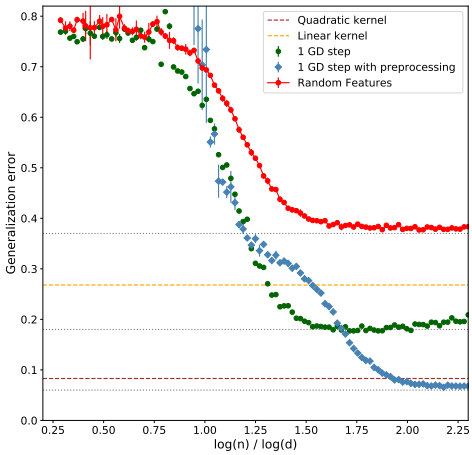

Figure 4: **Learning with training of the second layer.** Simulation illustrating the different regimes in Fig. 1, using $d = 512, p = 1024$, a symmetric two-index target function $f^\star(\boldsymbol{z}) = \sigma^\star(\langle \boldsymbol{w_1^\star}, \boldsymbol{z}\rangle) + \sigma^\star(\langle \boldsymbol{w_2^\star}, \boldsymbol{z}\rangle)$ 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). Details can be found in App. B.

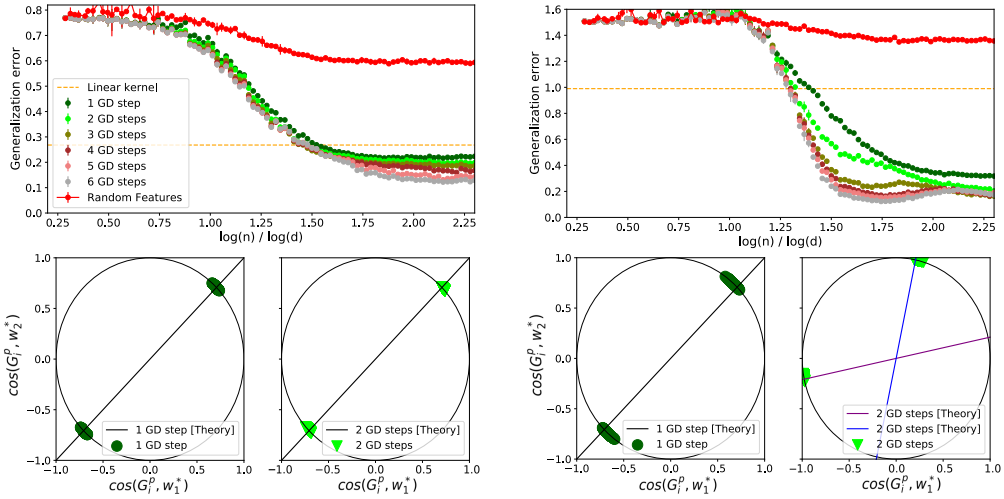

Figure 5: **Feature learning with multiple gradient steps. Top:** Generalization error as a function of n ($d = 512, p = 256$) after iterating the training procedure for six steps. **Bottom:** Cosine similarity of the projected gradient matrix G^p inside the target subspace for all the p neurons at a 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. See details in App. B.

Conjecture 2. *Assume that $p = \mathcal{O}(d^{\kappa_1})$ and $n = \mathcal{O}(d^{\kappa_2})$, and that the first layer W only learns a subspace $U^\star \subseteq V^\star$. Then, if $\hat{\boldsymbol{a}}$ is obtained as in Eq. (13) for any value of λ,*

$$\mathbb{E}\left[\left(f^\star(\boldsymbol{z}) - \hat{f}(\boldsymbol{z}; W, \hat{\boldsymbol{a}})\right)^2\right] \geq \|P_{U^\star,>\kappa} f^\star\|_\gamma^2 - o(1), \tag{24}$$

where $\kappa = \min(\kappa_1, \kappa_2)$ and $\|\cdot\|_\gamma$ is the norm in $\ell^2(\mathbb{R}, \gamma)$.

While a general proof for this conjecture would require significant additional work, we can prove some particular cases. First, it is easy to check that Proposition 1 corresponds to the $\kappa = \kappa_1 = 0$ case of the above conjecture. Second, in the next section, we prove the fairly delicate $\kappa_1 = \kappa_2$ case.

The linear case — We provide a proof of the $\kappa_1 = \kappa_2 = 1$ case of the above conjecture, also known as the *proportional regime*. In this setting, [Mei and Montanari, 2022, Gerace et al., 2020, Goldt et al., 2022, Hu and Lu, 2022] have shown that for the random features W^0 case, the generalization error of the minimizer for the second layer (13) is equivalent to the generalization error of an equivalent linear model; this result was extended in Ba et al. [2022] to a trained first layer W^1 for small, "lazy" stepsizes $\eta = \mathcal{O}(1)$. Ba et al. [2022] has also shown that when $\eta = \Theta(p)$, the equivalent linear description breaks down, due to the appearance of a spike term proportional to η in the trained weight matrix W^1; here, we show that this spike only allows the learning of non-linear functions parallel to the spike. Notice that the optimization problem (13) is equivalent to

$$\min_{\boldsymbol{a}\in\mathbb{R}^p} \frac{1}{n} \sum_{\nu=1}^{n} \left(\langle \boldsymbol{a}, \phi_{\mathrm{CK}}(\boldsymbol{z}^\nu)\rangle - f^\star(\boldsymbol{z}^\nu)\right)^2 + \lambda\|\boldsymbol{a}\|^2 \tag{25}$$

where $\phi_{\mathrm{CK}}(\boldsymbol{z}) = \sigma(W\boldsymbol{z})$ is the feature map corresponding to the *conjugate kernel* of the neural network. For a given direction $\boldsymbol{v}$, we define the *conditional linear* equivalent map $\phi_{\mathrm{CL}}(\boldsymbol{z}; \boldsymbol{v})$ as follows: given the decomposition $\boldsymbol{z} = z_{\boldsymbol{v}}\boldsymbol{v} + \boldsymbol{z}^\perp$,

$$\phi_{\mathrm{CL}}(\boldsymbol{z}; \boldsymbol{v}) = \mu(z_{\boldsymbol{v}}) + \Psi(z_{\boldsymbol{v}})z^\perp + \Phi(z_{\boldsymbol{v}})\boldsymbol{\xi} \tag{26}$$

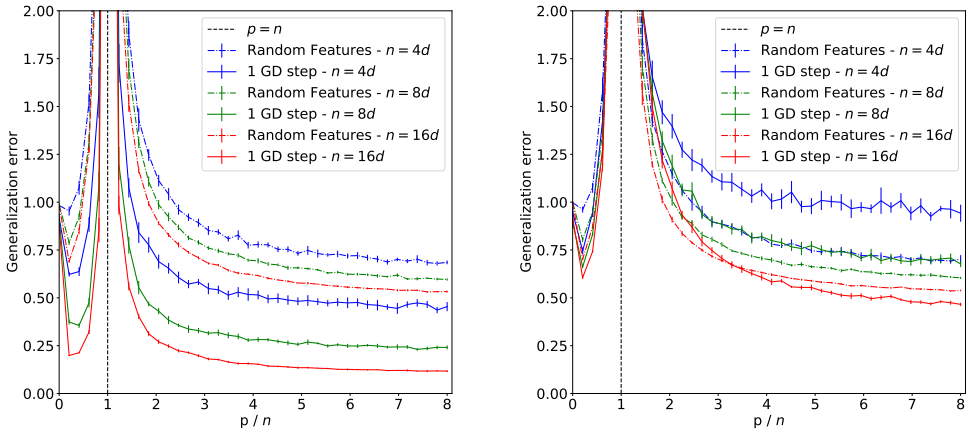

Figure 6: **Feature learning and generalization:** We illustrate how one step of gradient descent may (or may not) improve generalization over random features. The plot shows the generalization error as a function of the number of hidden neurons p normalized by the number of samples used for the first layer training $n = \{4d, 8d, 16d\}$ with fixed dimension $d = 2^8$. **Left**: $f^\star(\boldsymbol{z}) = z_1 + z_1 z_2$. While random features can only fit a linear model in the proportional regime considered, one step of gradient descent over W allows to fit the $z_1 z_2$ part, resulting in a much lower generalization error with respect to random features, despite only the direction z_1 being learned in W. **Right**: $f^\star(\boldsymbol{z}) = z_1 + z_2 z_3$. In this case, since the nonlinear part does not depend on z_1, one step of gradient descent on the W does not allow to improve generalization over random features (see Theorem 4 and Corollary 1). See App. B.

where $\boldsymbol{\xi} \sim \mathcal{N}(0, I_p - \boldsymbol{v}\boldsymbol{v}^\perp)$, and $\mu \in \mathbb{R}^p, \Psi \in \mathbb{R}^{p\times d}, \Phi \in \mathbb{R}^{p\times p}$ are chosen to match the first two conditional moments of ϕ_{CK}:

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

Then, the following theorem holds:

Theorem 4. *(informal) Assume that $n, p = \Theta(d)$, and that $V_1^\star$ defined in Theorem 2 is non-zero. Let $v^\star = C_1(f^\star)$, so that $V_1^\star = \text{span}(v^\star)$.Then, after one gradient step, there exists a vector $\boldsymbol{v} \in \mathbb{R}^d$ such that:*

- *the projection of $\boldsymbol{v}$ on $V^\star$ is proportional to $\boldsymbol{v}^\star$,*
- *the solutions $\hat{\boldsymbol{a}}, \tilde{\boldsymbol{a}}$ to the optimization problem in (25) with the feature maps $\phi_{CK}(\boldsymbol{z})$ and $\phi_{CL}(\boldsymbol{z}; \boldsymbol{v})$ yield the same generalization error.*

The full statement of Theorem 4 with the relevant hypotheses is discussed App. D. In particular, as we show in App. D, Theorem 4 provides a proof of the case $\kappa_1 = \kappa_2 = 1$ of Conjecture 2:

Corollary 1. *Conjecture 2 holds for $\kappa_1 = \kappa_2 = 1$.*

Informally, as for the random features model, features in the subspace orthogonal to $\boldsymbol{v}$ map to a linear model, and therefore only the linear part of $f^\star$ can be learned in this subspace. This precisely corresponds to the statement in Conjecture 2. This is illustrated in particular in Fig.6 where the generalization after one step of gradient descent is improved drastically over initialization for $f^\star(\boldsymbol{z}) = z_1 + z_1 z_2$, while it is not for $f^\star(\boldsymbol{z}) = z_1 + z_2 z_3$. We refer to App. B for additional discussion on the numerical validation of these claims.

General overview — We conclude this section by offering a summary on the generalization properties of two-layer networks. In table 2 we focus on the effect of overparameterization, i.e. number of neurons in the hidden layer.

Here, the function $f^\star(\boldsymbol{z}) = z_1 + z_1 z_2$ can either be learned in two steps through the staircase property or in a single step by increasing the number of hidden neurons. This is an example of

$f^*(\cdot)$	Random Features	GD $p=\mathcal{O}(1), n=\mathcal{O}(d)$	GD $p=\mathcal{O}(d), n=\mathcal{O}(d)$
$z_1+z_1z_2$	$\min(p,n)=\tilde{\Theta}(d^2)$	$\tau=2$	$\tau=1$
$z_1+z_2z_3$	$\min(p,n)=\tilde{\Theta}(d^2)$	No learn. in $\tau=\Theta(1)$	No learn. in $\tau=\Theta(1)$
$z_1+z_1z_2+\mathrm{He}_3(z_1)z_3$	$\min(p,n)=\tilde{\Theta}(d^4)$	$\tau=2$	$\tau=1$
$z_1+z_1z_2+\mathrm{He}_3(z_2)z_3$	$\min(p,n)=\tilde{\Theta}(d^4)$	$\tau=3$	$\tau=2$
$z_1+z_2z_3+\mathrm{He}_3(z_2)z_3$	$\min(p,n)=\tilde{\Theta}(d^4)$	No learn. in $\tau=\Theta(1)$	No learn. in $\tau=\Theta(1)$

Table 2: **Overparametrization helps learning multi-index targets.** Table summarizing the number of iterations needed to learn different multi-index targets $f^\star(\cdot)$ with underparametrized ($p=\mathcal{O}(1)$) and overparametrized ($p=\mathcal{O}(d)$) two-layer networks trained in the proportional batch-size regime $n=\mathcal{O}(d)$. We contrast this behaviour with the sample complexity of Random Features trained with n samples ($n_{\mathrm{T}}=n$), corresponding to $\tau=0$, i.e, the network at initialization. The time complexity can be significantly reduced by considering overparametrized networks, in accordance with Theorem 4.

a multi-index target function that can be learned with a single step of $\mathcal{O}(d)$ batch-size through overparameterization, exemplified in the left panel of Fig. 6. On the other hand, $f^\star(\boldsymbol{z})=z_1+z_2z_3$ is not linear conditioned on z_1 (learned at the first step), and this leads to a generalization pattern similar to Random Features in the overparametrized regime, see right panel of Fig. 6. We refer to App. B for the analysis of the last two example of Table 2 (see Fig. 9).
Conjecture. 2 extends the claims of Thm 4 to the polynomial scaling regime of $p=\mathcal{O}(d^{\kappa_1})$ and $d=\mathcal{O}(d^{\kappa_2})$. We exemplify the effect of overparametrization in this setting by considering the following target function:

$$f^\star(\boldsymbol{z})=z_1+\mathrm{He}_3(z_1)z_2+\mathrm{He}_2(z_1)\,\mathrm{He}_2(z_3). \tag{27}$$

Here, we can illustrate the intertwined dependence between batch-size, number of hidden neurons, and time iteration in determining learning efficiency. Indeed, changing any of these quantities can affect the components of the target function that are fitted by the network. Consider the following concrete scenarios, where $\hat{f}(\boldsymbol{z})$ denotes the predicted output function, after training the second layer:

(i) One-step GD with $n=\Theta(d), p=\mathcal{O}(1)$: $\hat{f}(\boldsymbol{z})=z_1$.

(ii) One-step GD with $n=\Theta(d), p=\mathcal{O}(d)$: $\hat{f}(\boldsymbol{z})=z_1+\mathrm{He}_3(z_1)z_2$.

(iii) One-step GD with $n=\Theta(d^2), p=\Theta(d^2)$: $\hat{f}(\boldsymbol{z})=z_1+\mathrm{He}_3(z_1)z_2+\mathrm{He}_2(z_1)\,\mathrm{He}_2(z_3)$.

Conclusion — In this work, we investigated the dynamics of two-layer neural networks as they learn a multi-index model using a single-pass, large-batch, gradient descent algorithm. We shed light on the intricate interactions between the task's structure, notably the complexity of the target function, the hyperparameters of SGD, such as the batch size and learning rate, and the network architecture, such as how the hidden layer width impact the approximation capacity of the network at fixed data budget. We highlight three key findings: a) The pronounced influence of a single gradient step on feature learning, underlining the nexus between batch size and the target's information exponent; b) The amplification of the network's approximation capacity over successive gradient steps and the learning of increasingly complex functions over time; c) The improvement in generalization when contrasted with the random feature/kernel regime. In conclusion, we presented a thorough mathematical framework detailing many nuances of data representation learning in two-layer neural networks during their early training phase. Finally, we note that while certain aspects of those results are left as conjectures, we believe that they capture both the dynamics of multiple gradient steps, as well as the structure of the remaining directions. Proving those conjectures requires fairly delicate concentration and random matrix arguments, which may be of independent mathematical interest.

B Numerical investigation

In this section we explain the procedures to get the different figures in the main text, along with the details behind the numerical experiments. We provide as well additional plots corroborating the theoretical results presented in the main manuscript. The code is available on GitHub.

Description of training algorithm and hyperparameters: 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. $\eta = \mathcal{O}(p\sqrt{\frac{n}{d}})$, and we take noiseless labels. If not stated otherwise, we consider fixed regularization strenghth $\lambda = 1$. We average over 10 different seeds to get the mean performance, and we use standard deviation for giving confidence intervals.

B.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 A.5 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.

Learning single-index targets: We start by analyzing the generalization performance of different two-layer networks after one giant GD step when learning single-index teacher functions (See Fig. 4). 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 $k > 1$), if the directions associated to teacher Hermite coefficients lower than k are not suppressed, or equivalently, if the leap index of the target is lower than k - see Fig. 1. 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 $n = \mathcal{O}(d)$:

$$\hat{c}_{j_1,\dots,j_d} \leftarrow \frac{1}{n}\sum_{\nu=1}^{n} y_\nu \, \mathrm{He}_{j_1}(\langle \boldsymbol{e}_1, \boldsymbol{z}_\nu\rangle)\cdots \mathrm{He}_{j_d}(\langle \boldsymbol{e}_d, \boldsymbol{z}_\nu\rangle) \tag{28}$$

$$y_\nu \leftarrow y_\nu - \sum_{j_1,\dots,j_d : j_1+\dots+j_d<k} \frac{\hat{c}_{j_1,\dots,j_d}}{j_1!\cdots j_d!}\, \mathrm{He}_{j_1}(\langle \boldsymbol{e}_1, \boldsymbol{z}_\nu\rangle)\cdots \mathrm{He}_{j_d}(\langle \boldsymbol{e}_d, \boldsymbol{z}_\nu\rangle) \tag{29}$$

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. 4 the inefficient estimation of eq. (28) 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. (29) and the suppressed part is injected back in the predictor only at test time:

$$\hat{f}(\boldsymbol{z}_\nu) = \frac{1}{\sqrt{p}}\hat{\boldsymbol{a}}^\top \sigma(W\boldsymbol{z}_\nu) + \sum_{j_1,\cdots,j_d : j_1+\dots+j_d<k} \frac{\hat{c}_{j_1,\dots,j_d}}{j_1!\cdots j_d!}\, \mathrm{He}_{j_1}(\langle \boldsymbol{e}_1, \boldsymbol{z}_\nu\rangle)\cdots \mathrm{He}_{j_d}(\langle \boldsymbol{e}_d, \boldsymbol{z}_\nu\rangle) \tag{30}$$

Comparison of different methods: The results presented in Fig. 4 verify the theoretical predictions of Thm. 2: in the $n = \mathcal{O}(d)$ regime vanilla Alg. 1 attains the "linear subspace learning"

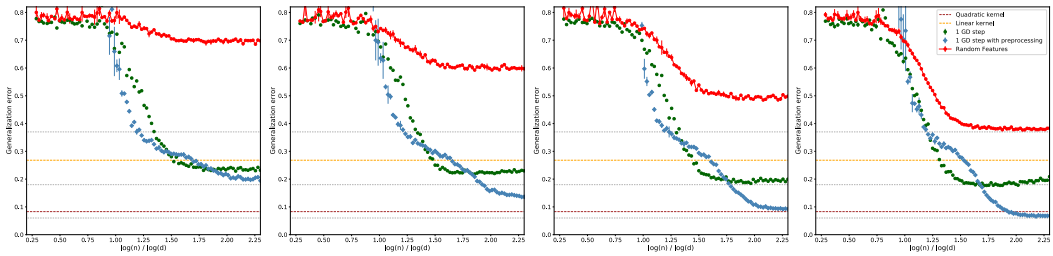

Figure 7: **Learning as a function of the number of neurons.** Simulations illustrating the different regimes in Fig. 1, using $d = 512$, a symmetric two-index target function $f^\star(\boldsymbol{z}) = \sigma^\star(\langle \boldsymbol{w_1^\star}, \boldsymbol{z}\rangle) + \sigma^\star(\langle \boldsymbol{w_2^\star}, \boldsymbol{z}\rangle)$ with activation $\sigma^\star(z) = \mathrm{He}_1(x) + \mathrm{He}_2(x)/2! + \mathrm{He}_4(x)/4!$, a relu student, and changing the value of $p \in (128, 256, 512, 1024)$ from left to right. (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 exemplified 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. The dotted black lines are a guide for the eyes referencing to the different plateaus of the rightmost plot.

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. 4, the manipulation in eq. (29) 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 neurons p on the generalization performance (see Fig. 7): 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. We again consider a single step of Alg. 1, focusing now on the analysis of the gradient - see Fig. 3. 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 $n = \mathcal{O}(d^k)$ regime (with $k > 1$) is necessary to analyze targets with leap index k in order to obtain specialized hidden units. Moreover, the leftmost section of Fig. 3 completes the picture offered by Figs. 4&7 about the lack of specialization in presence of teacher functions with non-zero first Hermite coefficient: the gradient is stuck in the single-index regime theoretically predicted by Thm. 2, regardless of the sample regime considered, preventing feature learning. To produce the plot, we consider the initialization of the second layer to be Gaussian, i.e. $\boldsymbol{a}^0 \sim \sqrt{p}\mathcal{N}(\boldsymbol{0}, I_p)$. This choice helps the spreading of the neurons in the teacher subspace, improving the figure's visualization clarity.

B.2 Learning with multiple steps

We move now the numerical investigation of the learning behavior of two layer networks 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. 5. 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. 5 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

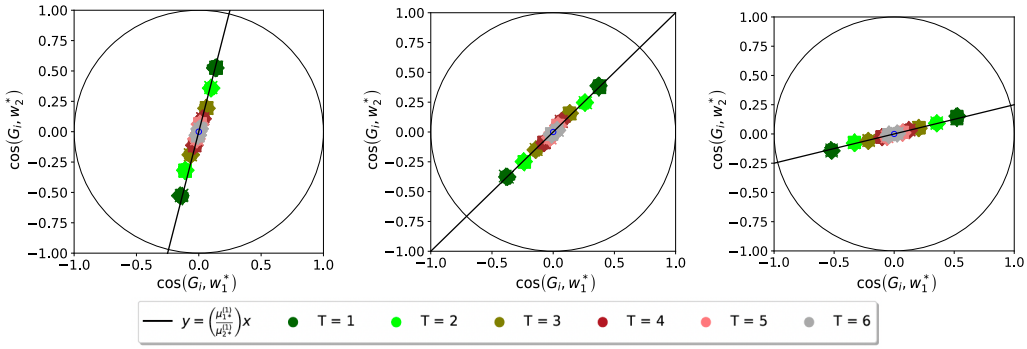

Figure 8: **Lack of feature learning after few GD steps.** The plots show the cosine similarity with respect to the teacher vectors $(\boldsymbol{w}_1^\star, \boldsymbol{w}_2^\star)$ for the gradient at different stages of the training. The predicted orientation (Thm. 2) of the gradient is shown as a 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.

does not improve relevantly over time, 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. 5 are closely related to the ones of Fig. 3. 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. 3, 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. 5 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. 3 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. 5 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. 8: the gradient is stuck in the single index regime even as the training proceeds, again in agreement with Thm. 3. Moreover, we illustrate by changing the teacher functions, that the theoretical prediction of Thm. 2 on the gradient orientation, are valid beyond the symmetric teachers.

Multiple stairs Exploiting the same visualization framework of Fig. 5, we complement the picture by studying a straightforward generalization in order to test Thm. 3 on functions having multiple linearly connected directions to the previously learned one, or informally, "multiple-stairs function". The results are presented in Fig. 10 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. 11 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

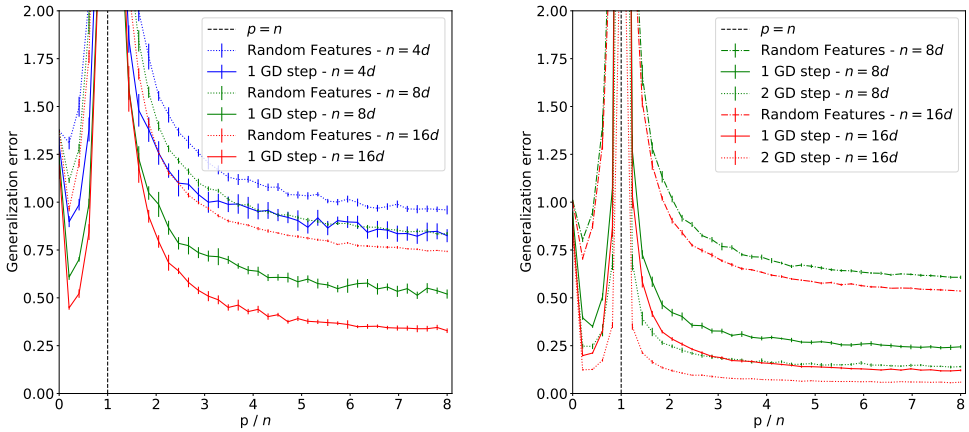

Figure 9: **Learning 3-index targets with overparametried networks**. We illustrate how overparametrization helps improving generalization over random features in accordance with Table 2. The plot shows the generalization error as a function of the number of hidden neurons p normalized by the number of samples used for the first layer training $n = \{4d, 8d, 16d\}$ with fixed dimension $d = 2^8$. **Left**: $f^\star(\boldsymbol{z}) = z_1 + z_1 z_2 + g_3(z_1) z_3$, where the auxiliary function has leap index $\ell = 3$, here $g(z) = \tanh z - z\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}[\xi \tanh \xi]$. While random features can only fit a linear model in this case, one step of gradient descent over W allows to fit the $z_1 z_2$ and $g(z_1) z_3$ parts, resulting in a much lower generalization error with respect to random features, despite only the direction z_1 being learned in W. **Right**: $f^\star(\boldsymbol{z}) = z_1 + z_1 z_2 + g_3(z_2) z_3$, where $g(\cdot)$ is defined above. In this case, two steps of gradient descent allow to improve generalization over both Random Features and one GD step by allowing the network to fit the $g(z_2) z_3$ term, linealry connected to z_2 that is learned only at the second step through the staircase property (Thm. 3).

the first step with respect to Fig. 10 (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.

Benefit of overparametrization In this paragraph we investigate benefit of overparametrization for two-layer networks trained with giant steps of GD when learning multi-index teacher functions (See Fig. 6). Theorem 4 precisely characterizes that in the diverging p limit it is not possible to fit functions that are non-linear conditioned on the knowledge of the spike $C_1(f_\star)$ learned at the first step. However, this conditional form of Gaussian Equivalence does not prevent the learning of functions in orthogonal directions that are linearly connected to $C_1(f_\star)$ (Thm. 3). The results in Fig. 6 clearly show these claims: when such linearly connected directions exist (left panel), one (giant) step of gradient descent training surpass the generalization performance of random features, while it is not possible otherwise (right panel). To produce the figure we fix the regularization strength to be infinitesimal $\lambda = 10^{-6}$ and we adapt the learning rate $\eta = 5p\sqrt{\frac{n}{d}}$ for different values of p. Moreover, we normalize the squared generalization risk by the variance of the target functions in order to have comparable y-axis for the two panel of Fig. 6. We complement these illustrations with an additional plot using the same numerical setup: Fig. 9 extend the claims above to the $3-$index target functions and conclude the analysis of the examples included in Table 2. In both cases displayed in Fig. 9 overparametrization helps improving the generalization with respect to random features and reduces the time iterations needed to learn the target compared to the underparametrized case. Although we do not consider the same example as Table 2 to improve numerical stability, the overall claims of the table are left unchanged by substituting the the $\ell-$th Hermite polynomial He_ℓ with the auxiliary function $g_\ell(\cdot)$ having leap index ℓ.

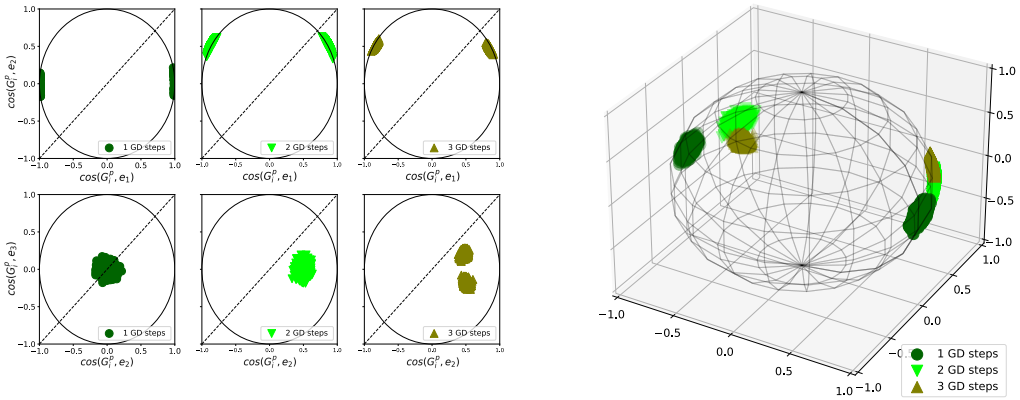

Figure 10: **Climbing multiple stairs.** Fix the teacher function $f_\star(\boldsymbol{z}) = {}^{z_1}/_3 + {}^{2z_1z_2}/_3 + z_2z_3$ and a relu student. The plots show the cosine similarity of the projected gradient matrix G^p inside the teacher subspace for all the p neurons at a fixed ratio ${}^n/_d = 4$, 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 $\boldsymbol{e}_3$.

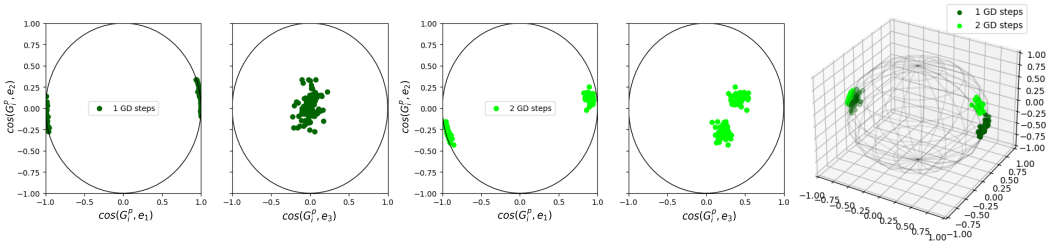

Figure 11: **Learning multiple directions at a time.** Fix the teacher function $f^\star(\boldsymbol{z}) = {}^{z_1}/_3 + 2\,\mathrm{He}_2(z_1)z_2 + z_1z_3$ and a relu student. The plots show the cosine similarity of the projected gradient matrix G^p inside the teacher subspace for all the p neurons at 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. (10). Fix a regularization parameter λ, 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:

$$\boldsymbol{G}_i^{(t)} \leftarrow \frac{a_{0,i}}{\sqrt{p}} \cdot \frac{1}{n} \sum_{\nu=1}^{n} \boldsymbol{x}^\nu \sigma'(\langle \boldsymbol{w}_i^{(t)}, \boldsymbol{z}^\nu \rangle) \left(\hat{f}(\boldsymbol{z}^\nu, W_t, \boldsymbol{a}_0) - f^\star(\boldsymbol{z}^\nu) \right) \tag{31}$$

$$W_{t+1} \leftarrow W_t - \eta G_t \tag{32}$$

    if $t == T_{max}$ **then**

      Train second layer Get the feature matrix $X_t \leftarrow \sigma(W_t Z)$, and compute estimator:

$$\hat{\boldsymbol{a}} \leftarrow \begin{cases} X_t^\top \left(X_t X_t^\top + \lambda I_n \right)^{-1} Y & \text{if } n < p \\ \left(X_t^\top X_t + \lambda I_p \right)^{-1} X_t^\top Y & \text{if } n > p \end{cases}$$

    end if

  end for

end for

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 1. 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 $\boldsymbol{i}, \boldsymbol{j} \in \mathbb{R}^k$,

$$\langle \mathcal{H}_{k,\boldsymbol{i}}(\boldsymbol{x}), \mathcal{H}_{k,\boldsymbol{j}}(\boldsymbol{x}) \rangle_\gamma = \frac{1}{|\mathfrak{o}(\boldsymbol{i})|} \mathbf{1}_{\boldsymbol{i} \text{ is a permutation of } \boldsymbol{j}} \tag{33}$$

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 (33) yields an immediate expression for the scalar product in $\ell^2(\mathbb{R}^m, \gamma_m)$:

$$\langle f, g \rangle_\gamma = \sum_{k \in \mathbb{N}} \langle C_k(f), C_k(g) \rangle. \tag{34}$$

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 $A \in \mathbb{R}^{r \times d}$ such that $AA^\top = I_r$. Then the Hermite coefficients of $f(\boldsymbol{x}) = g(A\boldsymbol{x})$ are*

$$C_k(f) = C_k(g) \cdot (A, \dots, A), \tag{35}$$

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*

$$\|X\|_{\psi_\alpha} = \inf \left\{ t > 0 \,:\, \mathbb{E}\left[\psi_\alpha\left(\frac{|X|}{t} \right) \right] \leq 1 \right\} \tag{36}$$

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,

$$\mathbb{P}(|X - \mathbb{E}X| > t\|X\|_{\psi_\alpha}) \leq 2e^{-t^\alpha}. \tag{37}$$

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*

$$\|XY\|_{\psi_{\alpha/2}} \leq K_\alpha \|X\|_{\psi_\alpha} \|Y\|_{\psi_\alpha} \tag{38}$$

Finally, we shall use the following theorem:

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

$$\left\| \sum_{i=1}^n X_i \right\|_{\psi_\alpha} \leq K_\alpha \log(n)^{1/\alpha} \left(\sqrt{\sum_{i=1}^n \sigma_i^2} + \max_i \|X_i\|_{\psi_\alpha} \right) \tag{39}$$

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 $k \in [r]$:*

$$\|\boldsymbol{\pi}_i^0\| \leq \frac{\sqrt{r}\log(d)}{\sqrt{d}} \tag{40}$$

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

$$\boldsymbol{g}_i = -\nabla_{\boldsymbol{w}_j} \mathcal{L}\left(\hat{f}(\boldsymbol{z}^\nu; W^0, \boldsymbol{a}), f^\star(\boldsymbol{z}^\nu) \right). \tag{41}$$

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

$$\boldsymbol{g}_i = \frac{a_i}{\sqrt{p}} \cdot \frac{1}{n} \sum_{\nu=1}^n \boldsymbol{z}^\nu \sigma'(\langle \boldsymbol{w}_i^0, \boldsymbol{z}^\nu \rangle) f^\star(\boldsymbol{z}^\nu) \tag{42}$$

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

$$\|\boldsymbol{w}_i^1\|^2 = 1 + 2\eta \langle \boldsymbol{w}_i^0, \boldsymbol{g_i} \rangle + \eta^2 \|\boldsymbol{g}_i\|^2 \tag{43}$$

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*

$$\mathbb{E}[\boldsymbol{g}_i] = \frac{a_i}{\sqrt{p}} \left(\sum_{k=0}^\infty c_{k+2} \langle (\boldsymbol{w}_i^0)^{\otimes k}, C_k^\star \rangle \boldsymbol{w}_i + \sum_{k=0}^\infty c_{k+1} C_{k+1}^\star \times_{1\dots k} (\boldsymbol{w}_i^0)^{\otimes k} \right) \tag{44}$$

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 (34), we find

$$\mathbb{E}\left[\boldsymbol{z}\sigma'(\langle \boldsymbol{w}, \boldsymbol{z} \rangle) f^\star(\boldsymbol{z})\right] = \sum_{k=0}^\infty c_{k+2} \langle \boldsymbol{w}^{\otimes k}, C_k^\star \rangle \boldsymbol{w} + \sum_{k=0}^\infty c_{k+1} C_{k+1}^\star \times_{1\dots k} \boldsymbol{w}^{\otimes k}. \tag{45}$$

□

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*

$$\left| \langle C_k^\star, (\boldsymbol{w}_i^0)^{\otimes k} \rangle \right| \leq c \left(\frac{\sqrt{r}\log(d)}{\sqrt{d}} \right)^k \quad \textit{and} \quad \left\| C_{k+1}^\star \times_{1\dots k} (\boldsymbol{w}_i^0)^{\otimes k} \right\| \leq c \left(\frac{\sqrt{r}\log(d)}{\sqrt{d}} \right)^k \tag{46}$$

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

$$\left\| \mathbb{E}[\boldsymbol{g}_i] - C_\ell^\star \times_{1\dots(\ell-1)} (\boldsymbol{w}_i^0)^{\otimes(\ell-1)} \right\| = \mathcal{O}\left(\frac{r^{\ell/2}\,\mathrm{polylog}(d)}{d^{\ell/2}} \right) \tag{47}$$

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

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

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 (43), in expectation. We begin with the cross-term:

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

$$\mathbb{E}\left[\langle \boldsymbol{w}_i^0, \boldsymbol{g}_i\rangle\right] = \mathcal{O}\left(\frac{r^{\ell/2}\,\mathrm{polylog}(d)}{pd^{\ell/2}}\right) \tag{48}$$

Proof. From eq. (47), we have

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

The first part of Lemma 5 gives

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

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

$$\begin{aligned}
\|\boldsymbol{g}_i\|^2 &= \frac{a_i^2}{n^2p^2}\sum_{\nu,\nu'=1}^{n} \langle \boldsymbol{z}^\nu, \boldsymbol{z}^{\nu'}\rangle \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}^\nu\rangle) f^\star(\boldsymbol{z}^\nu)\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}^{\nu'}\rangle) f^\star(\boldsymbol{z}^{\nu'}) \\
&= \frac{a_i^2}{n^2p^2}\Bigg(\sum_{\nu\neq\nu'} \langle \boldsymbol{z}^\nu, \boldsymbol{z}^{\nu'}\rangle \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}^\nu\rangle) f^\star(\boldsymbol{z}^\nu)\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}^{\nu'}\rangle) f^\star(\boldsymbol{z}^{\nu'}) \\
&\qquad + \sum_{\nu=1}^{n} \|\boldsymbol{z}^\nu\|^2 \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}^\nu\rangle)^2 f^\star(\boldsymbol{z}^\nu)^2\Bigg)
\end{aligned} \tag{49}$$

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

$$\mathbb{E}\left[\|\boldsymbol{g}_i\|^2\right] = \frac{n(n-1)}{n^2}\|\mathbb{E}[\boldsymbol{g}_i]\|^2 + \frac{a_i^2}{np^2}\mathbb{E}\left[\|\boldsymbol{z}^\nu\|^2\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}^\nu\rangle)^2 f^\star(\boldsymbol{z}^\nu)^2\right] \tag{50}$$

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* $1 - cpe^{-\log(d)^2}$*,*

$$\|\mathbb{E}[\boldsymbol{g}_i]\|^2 = a_i^2 X_i \cdot \frac{\left\|\boldsymbol{\pi}_i^0\right\|^{2(\ell-1)}}{p^2} + \mathcal{O}\left(\frac{r^{\ell/2}\,\mathrm{polylog}(d)}{d^{\ell/2}}\right) \tag{51}$$

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

$$c \cdot d \leq \mathbb{E}\left[\|\boldsymbol{z}\|^2\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}\rangle)^2 f^\star(\boldsymbol{z})^2\right] \leq C \cdot d \tag{52}$$

Proof. We begin with (51). Define the unit norm vectors

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

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

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

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 (51) then results from the expansion in (47).

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

$$\mathbb{E}\left[\|\boldsymbol{z}\|^2 \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}\rangle)^2 f^\star(\boldsymbol{z})^2\right] \le \sqrt{\mathbb{E}\left[\|\boldsymbol{z}\|^4\right]} \sqrt[4]{\mathbb{E}\left[\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}\rangle)^8\right] \mathbb{E}\left[f^\star(\boldsymbol{z})^8\right]} \le C \cdot d, \tag{53}$$

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 $\sigma'(x)^2 \ge \varepsilon$ if $x \in \mathcal{A}$ and $f^\star(\boldsymbol{z}^\star) > \varepsilon$ if $\boldsymbol{z}^\star \in \mathcal{B}$. From the fact that $\pi_i \le 1/2$ 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

$$\mathbb{E}\left[\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}\rangle)^2 f^\star(\boldsymbol{z})^2\right] \ge \gamma(\mathcal{C})\varepsilon^2, \tag{54}$$

which concludes the proof of Eq. (52). □

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* $i \in [p], k \in [r]$,

$$\left\|\boldsymbol{\pi}_i^1 - \mathbb{E}\left[\boldsymbol{\pi}_i^1\right]\right\| = \mathcal{O}\left(\frac{\eta\sqrt{r}\log(n)}{p\sqrt{n}}\right) \tag{55}$$

$$\left|\left\|\boldsymbol{w}_i^1\right\|^2 - \mathbb{E}\left[\left\|\boldsymbol{w}_i^1\right\|^2\right]\right| = \mathcal{O}\left(\frac{\eta\log(n)}{p\sqrt{n}} + \frac{\eta^2 d\log(n)^6}{p^2 n\sqrt{n}} + \frac{\eta^2\log(d)}{p^2 n\sqrt{d}} + \frac{\eta^2 r\log(n)^2}{p^2 n} + \frac{\eta^2\log(n)^\ell r^{(\ell-1)/2}}{p^2 d^{(\ell-1)/2}\sqrt{n}}\right) \tag{56}$$

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

$$\boldsymbol{X}^\nu = \boldsymbol{z}^\nu \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}^\nu\rangle) f^\star(\boldsymbol{z}^\nu). \tag{57}$$

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. (55) and the first term of Eq. (56).

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* $k \in [r]$,

$$|\langle \boldsymbol{w}, \boldsymbol{g}_i\rangle - \mathbb{E}[\langle \boldsymbol{w}, \boldsymbol{g}_i\rangle]| \le \frac{\log(n)}{p\sqrt{n}} \tag{58}$$

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

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{\nu=1}^{n}\langle \boldsymbol{w}, \boldsymbol{X}^\nu\rangle - \mathbb{E}\langle \boldsymbol{w}, \boldsymbol{X}\rangle\right| \geq \frac{\log(n)}{\sqrt{n}}\right) \leq 2e^{-c\log(n)^2}. \tag{59}$$

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

$$\|\boldsymbol{g}_i\|^2 - \mathbb{E}[\|\boldsymbol{g}_i\|^2] = \frac{a_i^2}{n^2p^2}\left(\underbrace{\sum_{\nu=1}^{n}\|\boldsymbol{X}^\nu\|^2 - n\mathbb{E}[\|\boldsymbol{X}\|^2]}_{S_1} + \underbrace{\sum_{\nu\neq\nu'}\langle \boldsymbol{X}^\nu, \boldsymbol{X}^{\nu'}\rangle - n(n-1)\|\mathbb{E}\boldsymbol{X}\|^2}_{S_2}\right) \tag{60}$$

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 $1 - e^{-c\log(n)^2}$,*

$$\mathbb{P}\big(|S_1| \geq \log(n)^6 d\sqrt{n}\big) \leq e^{-c\log(n)^2}. \tag{61}$$

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

$$\left\|\sum_{\nu=1}^{n}\|\boldsymbol{X}^\nu\|^2 - n\mathbb{E}\left[\|\boldsymbol{X}\|^2\right]\right\|_{\psi_{1/2}} \leq c\log(n)^2 d\sqrt{n}. \tag{62}$$

Equation (61) is then a consequence of the Orlicz concentration bound (37). □

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 $(X_1, \ldots, X_n), (Y_1, \ldots, Y_n)$ 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*

$$\mathbb{P}\left(\left\|\sum_{i\neq j} f_{ij}(X_i, X_j)\right\| \geq t\right) \leq C\mathbb{P}\left(\left\|\sum_{i\neq j} f_{ij}(X_i, Y_j)\right\| \geq \frac{t}{C}\right) \tag{63}$$

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

$$\mathbb{P}\left(\left|\sum_{\nu=1}^{n}\langle \boldsymbol{X}^\nu, \boldsymbol{Y}^\nu\rangle - n\bar{x}^2\right| \geq \sqrt{n}d\log(n)^6\right) \leq e^{-c\log(n)^2} \tag{64}$$

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

$$\tilde{S}_2 := \sum_{\nu,\nu'=1}^{n} \langle \boldsymbol{X}^{\nu}, \boldsymbol{Y}^{\nu'} \rangle - n^2 \bar{x}^2 \tag{65}$$

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

$$\boldsymbol{X}^{\nu} = \boldsymbol{X}_i^{\nu} + \boldsymbol{X}_{\perp}^{\nu} \quad \text{and} \quad \boldsymbol{Y}^{\nu} = \boldsymbol{Y}_i^{\nu} + \boldsymbol{Y}_{\perp}^{\nu} \tag{66}$$

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

$$\sum_{\nu,\nu'=1}^{n} \langle \boldsymbol{X}^{\nu}, \boldsymbol{Y}^{\nu'} \rangle - n^2 \bar{x}^2 = \underbrace{\langle \sum_{\nu=1}^{n} \boldsymbol{X}_i^{\nu}, \sum_{\nu=1}^{n} \boldsymbol{Y}_i^{\nu} \rangle - n^2 \bar{x}^2}_{S_2'} + \underbrace{\langle \sum_{\nu=1}^{n} \boldsymbol{X}_{\perp}^{\nu}, \sum_{\nu=1}^{n} \boldsymbol{Y}_{\perp}^{\nu} \rangle}_{S_2''} \tag{67}$$

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

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

$$\left\| \sum_{\nu=1}^{n} \boldsymbol{X}_i^{\nu} - n\mathbb{E}\boldsymbol{X} \right\| \leq C\sqrt{r}\log(n)\sqrt{n} \tag{68}$$

Proof. Let $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{r+1})$ 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

$$\mathbb{P}\left(\left| \langle \sum_{\nu=1}^{n} \boldsymbol{X}_i^{\nu} - n\mathbb{E}\boldsymbol{X}, \boldsymbol{u}_k \rangle \right| \geq \log(n)\sqrt{n} \right) \leq e^{-c\log(n)^2} \tag{69}$$

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* $1 - Ce^{-c\log(n)^2}$,

$$S_2' = \mathcal{O}\left(rn\log(n)^2 + \frac{\log(n)^{\ell} r^{(\ell-1)/2} n\sqrt{n}}{d^{(\ell-1)/2}} \right) \tag{70}$$

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*

$$|S_2''| \leq 2\log(d)n\sqrt{d} \tag{71}$$

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 $i \in [p]$:*

$$\left\| \boldsymbol{\pi}_i^1 - \left(\boldsymbol{\pi}_i^0 + \frac{\eta a_i}{\sqrt{p}} C_\ell^\star \times_{1\ldots(\ell-1)} (\boldsymbol{w}_i^0)^{\otimes(\ell-1)} \right) \right\| = \mathcal{O}\left(\frac{r^{\ell/2}\,\mathrm{polylog}(d)}{d^{\ell/2}} + \frac{\sqrt{r}\eta\log(d)}{p\sqrt{n}} \right) \tag{72}$$

$$\left\| \boldsymbol{w}_i^1 \right\|^2 = \Theta\left(1 + \frac{\eta^2 X_i \left\| \boldsymbol{\pi}_{\boldsymbol{i}}^{\boldsymbol{0}} \right\|^2}{p^2} + \frac{\eta^2 d}{np^2} \right) \tag{73}$$

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

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

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 (73) 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

$$\frac{\|\boldsymbol{\pi}_i^1\|}{\|\boldsymbol{w}_i^1\|} = \mathcal{O}\left(\max\left(\left\|\boldsymbol{\pi}_i^0\right\|, \frac{\left\|\boldsymbol{\pi}_i^0\right\|^{\ell-1}}{d^{\delta/2}} \right) \right) = \mathcal{O}\left(\frac{\mathrm{polylog}(d)}{d^{(1\wedge\delta)/2}} \right) \tag{74}$$

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,

$$\frac{\|\boldsymbol{\pi}_i^1\|}{\|\boldsymbol{w}_i^1\|} = \Omega\left(a_i^2 X_i \cdot (\sqrt{d}\|\pi_i^0\|)^{\ell-1} \right), \tag{75}$$

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 (42):

$$\boldsymbol{g}_i = \frac{a_i}{\sqrt{p}} \cdot \frac{1}{n} \sum_{\nu=1}^{n} \boldsymbol{z}^\nu \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}^\nu \rangle) f^\star(\boldsymbol{z}^\nu) \tag{76}$$

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

$$\sigma'_{>1}(u) = \sigma'(u) - \mu_1, \tag{77}$$

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

$$\boldsymbol{g}_i = \frac{a_j}{\sqrt{p}} \frac{1}{n} \mu_1 \sum_{i=1}^{n} y_i \boldsymbol{x}_i + \underbrace{\frac{1}{n} \frac{a_j}{\sqrt{p}} \mu_1 \sum_{i=1}^{n} \sigma'_{>1}(\boldsymbol{x}_i^\top \boldsymbol{w}^0) \boldsymbol{x}_i y_i}_{\Delta_j}, \tag{78}$$

or in matrix form:

$$\boldsymbol{G} = \boldsymbol{u}\boldsymbol{v}^\top + \Delta, \tag{79}$$

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 $\|\boldsymbol{v}\| = 1$, $\langle \Delta_j, \boldsymbol{v} \rangle = \mathcal{O}\left(\frac{\mathrm{polylog}(d)}{p\sqrt{d}} \right)$.

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

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

where we only consider the first half neurons due to the choice of the symmetric initialization in Equation (10)*.*

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 $\mathbb{E}\left[\Delta_j^\top \boldsymbol{v}\right] = \mathcal{O}\left(\frac{\text{polylog}(d)}{p\sqrt{d}}\right)$. Furthermore, from Lemma 8, we obtain that, with high probability:

$$|\Delta_j^\top \boldsymbol{v} - \mathbb{E}\left[\Delta_j^\top \boldsymbol{v}\right]| = \mathcal{O}\left(\frac{\log(n)}{p\sqrt{d}}\right). \tag{80}$$

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. (56)) implies that, with high probability,

$$\langle \boldsymbol{g_i}, \boldsymbol{g_j} \rangle = \mathbb{E}[\langle \boldsymbol{g_i}, \boldsymbol{g_j} \rangle] + \mathcal{O}\left(\frac{\text{polylog}(d)}{p^2\sqrt{d}}\right), \tag{81}$$

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 (50) still holds, and we get

$$\mathbb{E}[\langle \boldsymbol{g_i}, \boldsymbol{g_j} \rangle] \leq \|\mathbb{E}[\boldsymbol{g_i}]\|\|\mathbb{E}[\boldsymbol{g_j}]\| + \frac{a_i^2}{np^2}\mathbb{E}\left[\|\boldsymbol{z}\|^2 \sigma'(\langle \boldsymbol{w}_i^0, \boldsymbol{z} \rangle)\sigma'(\langle \boldsymbol{w}_j^0, \boldsymbol{z} \rangle) f^\star(\boldsymbol{z})^2\right] \tag{82}$$

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*

$$F(\Sigma) = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(0,\Sigma)}[f_1(z_1) \ldots f_N(z_N)] \tag{83}$$

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

$$|F(\Sigma) - F(\Sigma')| \leq C\|\Sigma - \Sigma'\|_\infty. \tag{84}$$

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:*

$$\|\boldsymbol{w}_i^1\|^2 = 1 + \eta C a_i^2 + \mathcal{O}(\frac{\text{polylog}\, d}{\sqrt{d}}) \tag{85}$$

Proof. Recall Equation 50:

$$\mathbb{E}\left[\|\boldsymbol{g}_i\|^2\right] = \frac{n(n-1)}{n^2}\|\mathbb{E}[\boldsymbol{g}_i]\|^2 + \frac{a_i^2}{np^2}\mathbb{E}\left[\|\boldsymbol{z}^\nu\|^2\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}^\nu\rangle)^2 f^\star(\boldsymbol{z}^\nu)^2\right] \tag{86}$$

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:

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

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 $j \in [r]$:

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

The result then follows by noting that Lemmas 6 and 8 imply that $\eta\langle \boldsymbol{w}, \boldsymbol{g}_i\rangle = \mathcal{O}(\frac{\operatorname{polylog} d}{\sqrt{d}})$ with high probability. □

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. We suppose that $\frac{n}{d}$ is fixed to a constant α. 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:

$$\boldsymbol{g}_j^1 = -\nabla_{\boldsymbol{w}_j}\mathcal{L}\left(\hat{f}(\boldsymbol{z}^\nu; W^1, \boldsymbol{a}), f^\star(\boldsymbol{z}^\nu)\right) \tag{87}$$

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:

$$\boldsymbol{w}_j^1 = \eta\frac{a_j\mu_1}{\sqrt{p}}\boldsymbol{v} + \boldsymbol{w}_j^0 + \Delta_j, \tag{88}$$

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 $z \in \mathbb{R}$:

$$\sigma_{\kappa,j}(z) = \sigma(c_j z + \kappa). \tag{89}$$

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

$$\sigma'_{>1,\kappa}(u) = \sigma'_{\kappa,j}(u) - \mu_{1,\kappa,j}. \tag{90}$$

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 (45), we have:

$$\begin{aligned}\mathbb{E}\left[\langle \boldsymbol{u}, \boldsymbol{g}_j^1\rangle\right] &= \mathbb{E}\left[(f^\star(\boldsymbol{z}) - \hat{f}(\boldsymbol{z}, \boldsymbol{W}^1, \boldsymbol{a}))\sigma'(\langle \boldsymbol{z}, \boldsymbol{w}_j^1\rangle)_j(\langle \boldsymbol{z}, \boldsymbol{u}\rangle)\right] \\ &= \mathbb{E}\left[(f^\star(\boldsymbol{z})\mu_{1,\langle \boldsymbol{z}, \boldsymbol{v}_j\rangle}(\langle \boldsymbol{z}_i, \boldsymbol{u}\rangle)\right] - \mathbb{E}\left[\hat{f}(\boldsymbol{z}, \boldsymbol{W}^1, \boldsymbol{a})\sigma'(\langle \boldsymbol{z}_i, \boldsymbol{w}^1\rangle)_j(\langle \boldsymbol{z}_i, \boldsymbol{u}\rangle)\right]\end{aligned} \tag{91}$$

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

The second term can be expressed as:

$$\langle (\mathbb{E}\left[\hat{f}(\boldsymbol{z}, \boldsymbol{W}^1, \boldsymbol{a})\sigma'(\langle \boldsymbol{z}, \boldsymbol{w}^1)_j \boldsymbol{z}_i\right]), \boldsymbol{u}\rangle \tag{92}$$

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:

$$\mathbb{E}\left[\hat{f}(\boldsymbol{z}, \boldsymbol{W}^1, \boldsymbol{a})\sigma'(\langle \boldsymbol{z}_i, \boldsymbol{w}^1\rangle)_j \boldsymbol{z}_i\right])^\top \boldsymbol{u} \xrightarrow{d\to\infty} 0 \tag{93}$$

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

$$y = f^\star(\boldsymbol{z}) = \sum_{j_1,\cdots,j_r=1}^{\infty} \frac{c^\star_{j_1,\cdots,j_r}}{j_1!j_2!\cdots j_r!} \operatorname{He}_{j_1}(\langle \boldsymbol{v}_u^\star, \boldsymbol{z}\rangle) \operatorname{He}_{j_2}(\langle \boldsymbol{u}, \boldsymbol{z}\rangle) \cdots \operatorname{He}_{j_r}(\langle \boldsymbol{u}_{r-2}, \boldsymbol{z}\rangle). \tag{94}$$

Lemma 4 and 12 imply that $\boldsymbol{v}_j \xrightarrow{\mathbb{P}} c'_j \boldsymbol{v}^\star$ where c'_j denotes the constant $\eta a_j \sqrt{p}\alpha\mu_1$ Since $\boldsymbol{u} \perp \boldsymbol{u}'_1, \cdots, \boldsymbol{u}'_{r-2}$, only the terms in y corresponding to products of the form $\operatorname{He}_{j_1}(\langle \boldsymbol{v}_u^\star, \boldsymbol{z}\rangle) \operatorname{He}_{j_2}(\langle \boldsymbol{u}, \boldsymbol{z}\rangle)$ 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:

$$\mathbb{E}\left[\operatorname{He}_{j_1}(\langle \boldsymbol{v}_u^\star, \boldsymbol{z}\rangle) \operatorname{He}_{j_2}(\langle \boldsymbol{u}, \boldsymbol{z}\rangle)\mu_{1,\langle \boldsymbol{z}, \boldsymbol{v}_j\rangle, j}\langle \boldsymbol{z}, \boldsymbol{u}\rangle\right] \to \mathbb{E}\left[\operatorname{He}_{j_1}(\langle \boldsymbol{v}_u^\star, \boldsymbol{z}\rangle) \operatorname{He}_{j_2}(\langle \boldsymbol{u}, \boldsymbol{z}\rangle)\mu_{1,\langle \boldsymbol{z}, c'_j \boldsymbol{v}^\star\rangle, j}\langle \boldsymbol{z}, \boldsymbol{u}\rangle\right] \tag{95}$$

Suppose $j_2 \neq 1$, then $\mathbb{E}\left[\operatorname{He}_{j_2}(\langle \boldsymbol{u}, \boldsymbol{z}\rangle)\langle \boldsymbol{z}_i, \boldsymbol{u}\rangle\right] = 0$. Therefore, the non-zero contributions arise from terms of the form $\operatorname{He}_{j_1}(\langle \boldsymbol{v}_u^\star, \boldsymbol{z}\rangle)\langle \boldsymbol{u}, \boldsymbol{z}\rangle$. 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 95 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 4.

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. $Q_{i,j}^t = \langle \boldsymbol{w}_i^t, \boldsymbol{w}_j^t\rangle \ \forall i, j \in [p]$. Let $M_t \in \mathbb{R}^{r\times p}$ denote the target-network overlap matrix i.e $.M_{i,j}^t = \langle \boldsymbol{w}_i^\star, \boldsymbol{w}_j^t\rangle \ \forall i \in [p], j \in [r]$. Let $\boldsymbol{W}^* \in \mathbb{R}^{r\times d}$ denote the matrix with rows $\boldsymbol{w_1^\star}, \cdots, \boldsymbol{w_r^\star}$.

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

Proof. Recall that:

$$\begin{aligned} Q_{i,j}^{t+1} &= \langle \boldsymbol{w}_i^{t+1}, \boldsymbol{w}_j^{t+1} \rangle \\ &= Q_{i,j}^t + \eta \langle \boldsymbol{g}_i^t, \boldsymbol{w}_j^t \rangle + \eta \langle \boldsymbol{w}_i^t, \boldsymbol{g}_j^t \rangle + \eta^2 \langle \boldsymbol{g}_i^t, \boldsymbol{g}_j^t \rangle. \\ M_{i,j}^{t+1} &= \langle \boldsymbol{w}_i^\star, \boldsymbol{w}_j^{t+1} \rangle \\ &= M_{i,j}^t + \eta \langle \boldsymbol{w}_i^\star, \boldsymbol{g}_j^t \rangle \end{aligned} \tag{96}$$

By the induction hypothesis, the entries of $\mathbb{E}[Q^t], \mathbb{E}[M^t]$ converge with high-probability to polynomial limits with error $\mathcal{O}\left(\frac{\text{polylog}\, d}{d}\right)$. Therefore, it suffices to show that $\mathbb{E}\left[\langle \boldsymbol{g}_i^t, \boldsymbol{w}_j^t \rangle\right]$, $\mathbb{E}\left[\langle \boldsymbol{w}_i^t, \boldsymbol{w}_j^t \rangle\right], \mathbb{E}\left[\langle \boldsymbol{g}_i^t, \boldsymbol{g}_j^t \rangle\right]$ converge to dimension-inpedendent polynomial limits. First, consider the case $i = j$. We have, analogous to Equation 50

$$\mathbb{E}\left[\|\boldsymbol{g}_i^t\|^2\right] = \underbrace{\frac{a_i^2}{np^2} \mathbb{E}\left[\|\boldsymbol{z}^\nu\|^2 \sigma'(\langle \boldsymbol{w}_i^t, \boldsymbol{z}^\nu \rangle)^2 (f^\star(\boldsymbol{z}^\nu) - \hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a}))^2\right]}_{T_1} + \underbrace{\frac{n(n-1)}{n^2} \|\mathbb{E}[\boldsymbol{g}_i]\|^2}_{T_2} \tag{97}$$

The first term T_1 can be decomposed as follows:

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

Similar to equation 53, 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 $\mathcal{F}_t$.

$$\frac{da_i^2}{np^2} \mathbb{E}\left[\sigma'(\langle \boldsymbol{w}_i^t, \boldsymbol{z}^\nu \rangle)^2 (f^\star(\boldsymbol{z}^\nu) - \hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a}))^2 | \mathcal{F}_t\right] \tag{98}$$

By assumption, $\frac{da_i^2}{np^2} = \frac{a_i^2\alpha}{p^2}$ 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 $\langle \boldsymbol{w}_1^t, \boldsymbol{z}^\nu\rangle, \cdots, \langle \boldsymbol{w}_p^t, \boldsymbol{z}^\nu\rangle$ $\langle \boldsymbol{w}_1^\star, \boldsymbol{z}^\nu,\rangle, \cdots, \langle \boldsymbol{w}_r^\star, \boldsymbol{z}^\nu\rangle$. By a change of variables the above term can therefore be expressed as an expectation w.r.t the $r+j$ correlated variables corresponding to the above overlaps.

Concretely, we have:

$$\frac{d}{np^2}\mathbb{E}\left[\sigma'(\langle \boldsymbol{w}_i^t, \boldsymbol{z}^\nu\rangle)^2 (f^\star(\boldsymbol{z}^\nu) - \hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a}))^2 | \mathcal{F}_t\right] = F_g(Q_t, M_t), \tag{99}$$

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 $\langle \boldsymbol{w}_1^t, \boldsymbol{z}^\nu\rangle, \cdots$ $\langle \boldsymbol{w}_p^t, \boldsymbol{z}^\nu\rangle$ and $\langle \boldsymbol{w}_1^\star, \boldsymbol{z}^\nu\rangle, \cdots$, $\langle \boldsymbol{w}_r^\star, \boldsymbol{z}^\nu\rangle$ respectively. Therefore, $\sigma'(\langle \boldsymbol{w}_i^t, \boldsymbol{z}^\nu\rangle)^2 (f^\star(\boldsymbol{z}^\nu) - \hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a}))^2$ is a polynomial in the zero mean correlated Gaussian variables $\langle \boldsymbol{w}_1^t, \boldsymbol{z}^\nu\rangle, \cdots, \langle \boldsymbol{w}_p^t, \boldsymbol{z}^\nu\rangle, \langle \boldsymbol{w}_1^\star, \boldsymbol{z}^\nu\rangle, \cdots, \langle \boldsymbol{w}_r^\star, \boldsymbol{z}^\nu\rangle$. Therefore, by Wick's/Isserlis' theorem [Janson, 1997, Polyak, 2005], 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 97. 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 $\boldsymbol{w}_1^*, \cdots, \boldsymbol{w}_r^*$. Concretely, we have:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{g}_i^t] &= \mathbb{E}\left[\boldsymbol{z}\sigma'(\langle \boldsymbol{w}_i^t, \boldsymbol{z}\rangle)(f^\star(\boldsymbol{z}) - \hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a}))\right] \\
&= \mathbb{E}\left[\boldsymbol{z}\sigma'(\langle \boldsymbol{w}_i^t, \boldsymbol{z}\rangle) f^\star(\boldsymbol{z})\right] - \mathbb{E}\left[\boldsymbol{z}\sigma'(\langle \boldsymbol{w}_i^t, \boldsymbol{z}\rangle)\hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a}))\right]
\end{aligned}$$

Consider the first-term, by Stein's Lemma, we obtain:

$$\begin{aligned}
\mathbb{E}\left[f^\star(\boldsymbol{z})\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}\rangle)\boldsymbol{z}\right] &= \mathbb{E}\left[\boldsymbol{z} g^\star(\langle \boldsymbol{w}_1^\star, \boldsymbol{z}\rangle, \cdots, \langle \boldsymbol{w}_r^\star, \boldsymbol{z}\rangle)\sigma'(\langle \boldsymbol{w}, \boldsymbol{z}\rangle)\right] \\
&= \mathbb{E}\left[\sigma'(\langle \boldsymbol{w}, \boldsymbol{z}\rangle)\nabla g^\star(\langle \boldsymbol{w}_1^\star, \boldsymbol{z}\rangle, \cdots, \langle \boldsymbol{w}_r^\star, \boldsymbol{z}\rangle)\right] \\
&+ \mathbb{E}\left[\boldsymbol{w}_i\sigma''(\langle \boldsymbol{w}_i, \boldsymbol{z}\rangle) g^\star(\langle \boldsymbol{w}_1^\star, \boldsymbol{z}\rangle, \cdots, \langle \boldsymbol{w}_r^\star, \boldsymbol{z}\rangle)\right]
\end{aligned}$$

By chain rule, $\nabla g^\star(\langle \boldsymbol{w}_1^\star, \boldsymbol{z}\rangle, \cdots, \langle \boldsymbol{w}_r^\star, \boldsymbol{z}\rangle)$ can be expressed as a linear combination of $\boldsymbol{w}_1^\star, \cdots$, $\boldsymbol{w}_r^\star$ and $\boldsymbol{w}_i^t$ with coefficients being polynomials in $\langle \boldsymbol{w}_1^\star, \boldsymbol{z}\rangle, \cdots, \langle \boldsymbol{w}_r^\star, \boldsymbol{z}\rangle$ independent of d.

Therefore, by Wick's theorem [Janson, 1997, Polyak, 2005], $\mathbb{E}\left[f^\star(\boldsymbol{z})\sigma'(\langle \boldsymbol{w}_i, \boldsymbol{z}\rangle)\boldsymbol{z}\right]$ can be expressed as $\sum_{k=1}^r p_k(Q_t, M_t)\boldsymbol{w}_k^* + h_i(Q_t, M_t)\boldsymbol{w}_i^t$, where $\{p_k\}_{k=1,\cdots,r}$ and h_i are polynomials independent of d. Similarly, we obtain $\mathbb{E}\left[\boldsymbol{z}\sigma'(\langle \boldsymbol{w}_i^t, \boldsymbol{z}\rangle)\hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a}))\right]$ as a linear combination of $\boldsymbol{w}_1^t, \cdots, \boldsymbol{w}_p^t$. Therefore, $\|\mathbb{E}[\boldsymbol{g}_i^t]\|^2$ conditioned on $\mathcal{F}_t$ is a polynomial in Q_t, M_t. Propagating errors from time t, we conclude that 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:

$$\boldsymbol{g}_j^t = -\nabla_{\boldsymbol{w}_j}\mathcal{L}\left(\hat{f}(\boldsymbol{z}^\nu;\boldsymbol{W}^t,\boldsymbol{a}), f^\star(\boldsymbol{z}^\nu)\right) = \frac{1}{n}a_j\sum_{\nu=1}^{n}\boldsymbol{z}^\nu(f^\star(\boldsymbol{z}^\nu) - \hat{f}(\boldsymbol{z}^\nu;\boldsymbol{W}^t,\boldsymbol{a}))\sigma'(\langle\boldsymbol{z}^\nu,\boldsymbol{w}_j^t\rangle) \tag{100}$$

□

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 (100), the expectation $\mathbb{E}\left[\langle\boldsymbol{v},\boldsymbol{g}_j^t\rangle\right]$ can be expressed as:

$$\mathbb{E}\left[\langle\boldsymbol{v},\boldsymbol{g}_j^t\rangle\right] = \mathbb{E}\left[a_j(f^\star(\boldsymbol{z}) - \hat{f}(\boldsymbol{z};\boldsymbol{W}^t,\boldsymbol{a}))\sigma'(\langle\boldsymbol{z},\boldsymbol{w}_j^t\rangle)\langle\boldsymbol{z},\boldsymbol{v}\rangle\right]. \tag{101}$$

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 [Janson, 1997, Polyak, 2005], 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 $\mathbb{E}\left[a_j f^\star(\boldsymbol{z})\sigma'(\langle\boldsymbol{z},\boldsymbol{w}_j^t\rangle)\langle\boldsymbol{z},\boldsymbol{v}\rangle\right]$. 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:

$$\mathbb{E}_{\boldsymbol{z}^\perp}\left[a_j f^\star(\boldsymbol{z})\sigma'(\langle\boldsymbol{z},\boldsymbol{w}_j^t\rangle)\langle\boldsymbol{z},\boldsymbol{v}\rangle\right] = \mathbb{E}_\kappa\left[a_j f^\star(\boldsymbol{z})f_1(\boldsymbol{a},\kappa)\langle\boldsymbol{z},\boldsymbol{v}\rangle\right], \tag{102}$$

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:

$$f_2(\kappa) = \mathbb{E}_{\langle\boldsymbol{z},\boldsymbol{v}\rangle}\left[y\langle\boldsymbol{z},\boldsymbol{v}\rangle|\kappa\right]. \tag{103}$$

Using the tower law of expectation, we obtain:

$$\mathbb{E}\left[a_j f^\star(\boldsymbol{z})\sigma'(\boldsymbol{z}^\top\boldsymbol{w}_j^t)\langle\boldsymbol{z},\boldsymbol{v}\rangle\right] = \mathbb{E}_\kappa\left[f_1(a_j,\kappa)f_2(\kappa)\right], \tag{104}$$

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 4), $\exists\kappa > 0$ 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:

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

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 $(c_{j,\boldsymbol{a}}u + \langle \kappa, \boldsymbol{v}_{j,\boldsymbol{a}}\rangle)^k$ is a non-constant polynomial in $\boldsymbol{a}$ for any k such that $k-q$ is even. Furthermore, the degree of the coefficient of κ^s in $(c_{j,\boldsymbol{a}}u + \langle \kappa, \boldsymbol{v}_{j,\boldsymbol{a}}\rangle)^k$ is strictly increasing in k. 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 104:

$$h(t, \boldsymbol{a}) = \mathbb{E}_{\kappa}\left[f_1(\boldsymbol{a}, \kappa) f_2(\kappa)\right]. \tag{105}$$

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:

$$h(t, \boldsymbol{a}) = \sum_{s\in\mathcal{S}_{y,t}} c_s p_s(\boldsymbol{a}), \tag{106}$$

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:

$$\frac{a_j}{\sqrt{p}}\mathbb{E}\left[\hat{f}(\boldsymbol{z}; \boldsymbol{W}^t, \boldsymbol{a}))\sigma'(\langle \boldsymbol{z}, \boldsymbol{w}_j^t\rangle)\langle \boldsymbol{z}, \boldsymbol{v}\rangle\right] = \frac{a_j}{p}\sum_{i=1}^{p} a_i\mathbb{E}\left[\sigma(\langle \boldsymbol{w}_i^t, \boldsymbol{z}\rangle)\sigma'(\langle \boldsymbol{z}, \boldsymbol{w}_j^t\rangle)\langle \boldsymbol{z}, \boldsymbol{v}\rangle\right] \tag{107}$$

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

$$\frac{a_j^2}{p}\mathbb{E}\left[\sigma(\langle \boldsymbol{w}_j^t, \boldsymbol{z}\rangle)\sigma'(\langle \boldsymbol{z}, \boldsymbol{w}_j^t\rangle)\langle \boldsymbol{z}, \boldsymbol{v}\rangle\right] \tag{108}$$

By Wick's theorem [Janson, 1997, Polyak, 2005] and the polynomial 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 $\mathcal{O}(\frac{\text{polylog}\, d}{\sqrt{d}})$. 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) $|\|\boldsymbol{g}_i^{t+1}\|^2 - \mathbb{E}\left[\|\boldsymbol{g}_i^{t+1}\|^2\right]| = \mathcal{O}\left(\frac{\text{polylog}\, d}{\sqrt{d}}\right)$

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

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

$$|\langle \boldsymbol{w}, \boldsymbol{g}_i\rangle - \mathbb{E}[\langle \boldsymbol{w}, \boldsymbol{g}_i\rangle]| = \mathcal{O}\left(\frac{\text{polylog}\, d}{\sqrt{d}}\right) \tag{109}$$

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:

$$\boldsymbol{g}_j^t = \frac{1}{n} a_j \sum_{\nu=1}^{n} \boldsymbol{z}_i (f^\star(\boldsymbol{z}^\nu) - \hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a}))\sigma'(\boldsymbol{z}_i^\top \boldsymbol{w}_j^t) \tag{110}$$

Define:

$$\boldsymbol{X}_i^\nu = \boldsymbol{z}^\nu \sigma'(\langle \boldsymbol{w}_i^t, \boldsymbol{z}^\nu\rangle)(f^\star(\boldsymbol{z}^\nu) - \hat{f}(\boldsymbol{z}^\nu; \boldsymbol{W}^t, \boldsymbol{a})). \tag{111}$$

Analogous to the proof of Proposition 2, we have:

$$\|\boldsymbol{g}_i\|^2 - \mathbb{E}[\|\boldsymbol{g}_i\|^2] = \frac{a_i^2}{n^2 p^2}\left(\underbrace{\sum_{\nu=1}^{n} \|\boldsymbol{X}_i^{\boldsymbol{\nu}}\|^2 - n\mathbb{E}[\|\boldsymbol{X}^{\boldsymbol{\nu}}\|^2]}_{S_1} + \underbrace{\sum_{\nu\neq\nu'} \langle \boldsymbol{X}_i^\nu, \boldsymbol{X}_i^{\nu'}\rangle - n(n-1)\|\mathbb{E}\langle \boldsymbol{X}_i^\nu, \boldsymbol{X}_i^{\nu'}\rangle\|^2}_{S_2}\right) \tag{112}$$

Similarly, we have:

$$\langle \boldsymbol{g}_i, \boldsymbol{g}_j\rangle - \mathbb{E}[\boldsymbol{g}_i, \boldsymbol{g}_j] = \frac{a_i^2}{n^2 p^2}\left(\underbrace{\sum_{\nu=1}^{n} \langle \boldsymbol{X}_i^{\boldsymbol{\nu}}, \boldsymbol{X}_j^{\boldsymbol{\nu}}\rangle - n\mathbb{E}[\langle \boldsymbol{X}_i^{\boldsymbol{\nu}}, \boldsymbol{X}_j^{\boldsymbol{\nu}}\rangle]}_{S_1'} + \underbrace{\sum_{\nu\neq\nu'} \langle \boldsymbol{X}_i^\nu, \boldsymbol{X}_j^{\nu'}\rangle - n(n-1)(\mathbb{E}\langle \boldsymbol{X}_i^\nu, \boldsymbol{X}_j^{\nu'}\rangle)^2}_{S_2'}\right) \tag{113}$$

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 repeated applications of 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 $\ell > 1$, $U_t^\star = 0$ for all $t \geq 1$. Applying the above arguments then implies that (i) and (iii) hold for all timesteps t.

Therefore, we may assume that $\ell = 1$. 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 Lemma 4 holds 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 target function defined in bottom right of Figure 5 i.e.:

$$f^\star(\boldsymbol{z}) = \sigma_1^\star(\langle \boldsymbol{w_1^\star}, \boldsymbol{z}\rangle) + \sigma_2^\star(\langle \boldsymbol{w_2^\star}, \boldsymbol{z}\rangle), \tag{114}$$

with $\sigma_1^\star(z) = z - z^2$ and $\sigma_2^\star(z) = z + z^2$. Equivalently, the above target function can be expressed in a rotated basis as:

$$f^\star(\boldsymbol{z}) = \sqrt{2}\boldsymbol{u_1^\star} + 2\boldsymbol{u_1^\star}\boldsymbol{u_2^\star}. \tag{115}$$

Where $\boldsymbol{u_1^\star} = \frac{1}{\sqrt{2}}(\boldsymbol{w_1^\star} + \boldsymbol{w_2^\star})$ and $\boldsymbol{u_2^\star} = \frac{1}{\sqrt{2}}(\boldsymbol{w_1^\star} - \boldsymbol{w_2^\star})$. Therefore $\boldsymbol{v}^* = \sqrt{2}\boldsymbol{u_1^\star}$ with $\|\boldsymbol{v}^*\| = \sqrt{2}$.

We follow the notation defined in the proof sketch in Section C.7 and assume that $a_i = \pm\frac{1}{\sqrt{p}}$, while $\alpha = \frac{n}{d} = 4$ as in Figure 5.

We have, using Equation (91):

$$\mathbb{E}\left[\langle \boldsymbol{u}, \boldsymbol{g}_j^1\rangle\right] \to \mathbb{E}\left[y\mu_{1,c_j'\langle \boldsymbol{z},\boldsymbol{v}^\star\rangle}(\langle \boldsymbol{z}_i, \boldsymbol{u}\rangle)\right] - \mathbb{E}\left[\hat{y}_i^1\sigma'(\langle \boldsymbol{z}_i, \boldsymbol{w}^1\rangle)_j(\langle \boldsymbol{z}_i, \boldsymbol{u}\rangle)\right]. \tag{116}$$

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

$$\langle \boldsymbol{g}^1, \boldsymbol{v}_u^\star\rangle \approx \mathbb{E}\left[y\mu_{1,c_j'\langle \boldsymbol{z},\boldsymbol{v}^\star\rangle,j}\langle \boldsymbol{z}_i, \boldsymbol{v}_u^\star\rangle\right], \tag{117}$$

Since the first Hermite coefficient μ_1 of the student activation for Relu equals 0.5, we obtain that $c'j = a_j\eta$.

We assume that $c_j \approx 1$. Therefore, $\mu_{1,\langle \boldsymbol{z},\boldsymbol{v}_u^\star\rangle}$ corresponds to the first Hermite coefficient of a translated Relu function and is given by:

$$\mu_{1,\kappa,j} = (1 - \Phi(-\kappa)) = \frac{1}{2}(1 + \text{erf}(-\kappa/\sqrt{2})). \tag{118}$$

Therefore, when $\eta = 2$, we obtain that $c'j = 2$:

$$\mu_{1,c_j'\langle \boldsymbol{z},\boldsymbol{v}^\star\rangle,j} = \frac{1}{2}(1 \pm \text{erf}(\langle \boldsymbol{z}, \boldsymbol{v}^\star\rangle/\sqrt{2})) = \frac{1}{2}(1 \pm \text{erf}(\langle \boldsymbol{z}, \boldsymbol{v}_u^\star\rangle)) \tag{119}$$

Let $\boldsymbol{v}_1^{(t=2)}$ and $\boldsymbol{v}_2^{(t=2)}$ denote the projections of the neurons $\boldsymbol{w}_j$ with $a_j = 1$ and $a_j = -1$ respectively. Therefore, using the Hermite decomposition of erf, we obtain the following predicted orientations in the setting considered in the right panel of Fig. 5:

$$\boldsymbol{v}_1^{(t=2)} = (1 - \frac{2}{\sqrt{3\pi}})\boldsymbol{w}_1^\star + (1 + \frac{2}{\sqrt{3\pi}})\boldsymbol{w}_2^\star \qquad \boldsymbol{v}_2^{(t=2)} = (1 + \frac{2}{\sqrt{3\pi}})\boldsymbol{w}_1^\star + (1 - \frac{2}{\sqrt{3\pi}})\boldsymbol{w}_2^\star \tag{120}$$

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, $\boldsymbol{v}_u^\star$ is given by $\boldsymbol{v}_u^\star = \frac{1}{\sqrt{r}}(\sum_{k=1}^{r} \boldsymbol{w}_k^\star)$. 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:

$$\begin{aligned} \mathbb{E}\left[f^\star(\boldsymbol{z}) H_k(\langle \boldsymbol{v}^\star, \boldsymbol{z}\rangle)(\langle \boldsymbol{u}_i, \boldsymbol{z}\rangle)\right] &= (\sum_{i=1}^{p}(u_i))(\mathbb{E}\left[f^\star(\boldsymbol{z}) H_k(\langle \boldsymbol{v}_u^\star, \boldsymbol{z}\rangle) z_1\right]) \\ &= 0. \end{aligned} \tag{121}$$

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

$$\mathcal{R}(W, \boldsymbol{a}) = \mathbb{E}_{\boldsymbol{z}}\left[\left(f^\star(\boldsymbol{z}) - \tilde{f}(P_U\boldsymbol{z}, P_{V^{\star\perp}}\boldsymbol{z})\right)^2\right] + O(\varepsilon_d), \tag{122}$$

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

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

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,

$$\mathcal{R}(W, \boldsymbol{a}) \geq \mathbb{E}_{P_U\boldsymbol{z}}\left[\mathrm{Var}\left(f^\star(\boldsymbol{z}) | P_U\boldsymbol{z}\right)\right] + O(\varepsilon_d), \tag{123}$$

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 (25).

Consider the sequence of vectors $\boldsymbol{v}_n \in \mathbb{R}^d$ defined as in Equation (79) by $\boldsymbol{v}_n = \frac{1}{n}\sum_{i=1}^n y_i \boldsymbol{z}_i$. 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:

$$\phi_{\text{CK}}(\boldsymbol{z}) = \sigma(W^{(1)}\boldsymbol{z}), \tag{124}$$

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:

$$\begin{aligned}&\mu(z_{\boldsymbol{v}}) = \mathbb{E}\left[\phi_{\text{CK}}(\boldsymbol{z}) \mid z_{\boldsymbol{v}}\right], \quad \Psi(z_{\boldsymbol{v}}) = \mathbb{E}\left[\phi_{\text{CK}}(\boldsymbol{z})(z^\perp)^\top \mid z_{\boldsymbol{v}}\right], \\ &\Phi(z_{\boldsymbol{v}}) = \mathrm{Cov}\left[\phi_{\text{CK}}(\boldsymbol{z}) \mid z_{\boldsymbol{v}}\right] - \Psi(z_{\boldsymbol{v}})\Psi(z_{\boldsymbol{v}})^\top\end{aligned} \tag{125}$$

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

$$\phi_{\text{CL}}(\boldsymbol{z}; \boldsymbol{v}) = \mu\left(z_{\boldsymbol{v}}\right) + \Psi(z_{\boldsymbol{v}})\boldsymbol{z}^\perp + \Phi(z_{\boldsymbol{v}})\boldsymbol{\xi}. \tag{126}$$

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

$$\mathbb{E}\left[\phi_{\text{CL}}(\boldsymbol{z}) \mid z_{\boldsymbol{v}}\right] = \mu(z_{\boldsymbol{v}}), \mathbb{E}\left[\phi_{\text{CL}}(\boldsymbol{z})(\boldsymbol{z}^{\perp})^{\top} \mid z_{\boldsymbol{v}}\right] = \Psi(z_{\boldsymbol{v}}),\ \text{Cov}\left[\phi_{\text{CL}}(\boldsymbol{z}) \mid z_{\boldsymbol{v}}\right] = \text{Cov}\left[\phi_{\text{CK}}(\boldsymbol{z}) \mid z_{\boldsymbol{v}}\right]. \tag{127}$$

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:

$$\min_{\boldsymbol{a}\in\mathbb{R}^p} \frac{1}{n}\sum_{\nu=1}^{n} \left(\langle \boldsymbol{a}, \phi_{\text{CK}}(\boldsymbol{z}^{\nu})\rangle - f^{\star}(\boldsymbol{z}^{\nu})\right)^2 + \lambda\|\boldsymbol{a}\|^2 \tag{128}$$

Define the following constraint set:

$$\mathcal{S}_p = \left\{\boldsymbol{a} \in \mathbb{R}^d \;\middle|\; \|\boldsymbol{a}\|_2 \le R, \quad \|\boldsymbol{a}\|_\infty \le Cp^{-\eta}\right\}. \tag{129}$$

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* (128) *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 $\widehat{\mathcal{R}}_n^{\star}(\boldsymbol{\Phi}, \boldsymbol{y}(\boldsymbol{Z})), \mathcal{R}_g^{\star}(\boldsymbol{\Phi}, \boldsymbol{y}(\boldsymbol{Z}))$ 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 $\boldsymbol{v}^{\star} = C_1(f^*)$ defined in Theorem 2 is nonzero. Then, the sequence of vectors $\boldsymbol{v}_n = \frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{z}_i \in \mathbb{R}^d$ satisfy:*

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

(ii) Under Assumption 6, the training and generalization errors obtained through the minimization of the objective (25) *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 $\Psi : \mathbb{R} \to \mathbb{R}$:

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

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

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

$$\begin{aligned} \widehat{\mathcal{R}}_n^{\star}(\boldsymbol{\Phi}_{CK}, \boldsymbol{y}(\boldsymbol{Z})) &\xrightarrow{\mathbb{P}} \mathcal{E} \quad \textit{if and only if} \quad \widehat{\mathcal{R}}_n^{\star}(\boldsymbol{\Phi}_{CL}, \boldsymbol{y}(\boldsymbol{Z})) \xrightarrow{\mathbb{P}} \mathcal{E} \\ \mathcal{R}_g^{\star}(\boldsymbol{\Phi}_{CK}, \boldsymbol{y}(\boldsymbol{Z})) &\xrightarrow{\mathbb{P}} \mathcal{E} \quad \textit{if and only if} \quad \mathcal{R}_g(\boldsymbol{\Phi}_{CL}, \boldsymbol{y}(\boldsymbol{Z})) \xrightarrow{\mathbb{P}} \mathcal{E}, \end{aligned} \tag{130}$$

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 $\varphi : \mathbb{R}^2 \to \mathbb{R}$, and $\forall k \in \mathbb{R}$:*

$$\lim_{n,p\to\infty} \sup_{\boldsymbol{\theta}_1 \in \mathcal{S}_p, \boldsymbol{\theta}_2 \in \mathcal{S}^{d-1}} \left| \mathbb{E}\left[\varphi(\boldsymbol{\theta}_1^\top \phi_{CK}(\boldsymbol{z}), \boldsymbol{\theta}_2^\top \boldsymbol{z}) \,\middle|\, z_{\boldsymbol{v}} = k\right] - \mathbb{E}\left[\varphi(\boldsymbol{\theta}^\top \phi_{CL}(\boldsymbol{z}), \boldsymbol{\theta}_2^\top \boldsymbol{z}) \,\middle|\, z_{\boldsymbol{v}} = k\right] \right| = 0 \tag{131}$$

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 $z_{\boldsymbol{v}} = k$, $\phi_{\text{CL}}(\boldsymbol{z}), \boldsymbol{z}_\perp$ is a Gaussian random variable. Next, consider $\phi_{\text{CK}}(\boldsymbol{z})$. Our proof relies on the observation that while the features $\phi_{\text{CK}}(\boldsymbol{z})$ 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 (79)):

$$\boldsymbol{W}^{(1)} = \eta \boldsymbol{u}\boldsymbol{v}^\top + \boldsymbol{W}^{(0)} + \eta\Delta, \tag{132}$$

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:

$$\sigma_{i,z_{\boldsymbol{v}}}(u) = \sigma(c_j u + \eta v_{\boldsymbol{z}}) - \mathbb{E}_u\left[\sigma(c_j u + \eta u_i z_{\boldsymbol{v}})\right], \tag{133}$$

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 (10), it suffices to restrict ourselves to the first half $p/2$ neurons.

For a fixed value of $z_{\boldsymbol{v}}$, the feature map $\phi_{\text{CK}}(\boldsymbol{z}) = \sigma(\boldsymbol{W}^1\boldsymbol{z})$ 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:

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

We have, using Lemma 12 and a union bound, that for $p, d = \Theta(n)$, $\Pr[\mathcal{A}_1] \overset{n,d\to\infty}{\longrightarrow} 1$. 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 (133) 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 $\phi_{\text{CK}}(\boldsymbol{z})$ 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:*

$$\Pr[\|\bar{\Phi}_{CK}\| \geq K\sqrt{p}] \leq 2\exp(-Cn) \tag{134}$$

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

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{a}_{CK}^{\top}\mu(z_{i,\boldsymbol{v}})\right)^2 \leq C \tag{135}$$

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:

$$\Pr[\frac{1}{n}\sum_{i=1}^{n}(y_i)^2 - \mathbb{E}\left[(y_i)^2\right] > K] \leq 2\exp(-\min(c_1K, c_2K^2)n). \tag{136}$$

For constants c_1, c_2. □

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

$$\mathcal{A}_{\boldsymbol{y}} = \left\{\frac{1}{n}\sum_{i=1}^{n}(y_i)^2 < C_1\right\}. \tag{137}$$

By Equation (136), 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:

$$\hat{\boldsymbol{a}} = \arg\min_{\boldsymbol{a}} \hat{\mathcal{R}}(W, \boldsymbol{a}) = \arg\min_{\boldsymbol{a}} \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{a}^{\top}\phi_k(\boldsymbol{z_i}))^2. \tag{138}$$

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

$$\hat{\mathcal{R}}(W, \boldsymbol{0}) = \frac{1}{n}\sum_{i=1}^{n}(y_i)^2. \tag{139}$$

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

$$\hat{\mathcal{R}}(W, \hat{\boldsymbol{a}}) \leq \hat{\mathcal{R}}(W, \boldsymbol{0}). \tag{140}$$

We obtain:

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

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

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

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

Applying Lemma 21 and $\Pr\left[\mathcal{A}_{\boldsymbol{y}}\right] \overset{n,d\to}{\longrightarrow} 1$ 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:

$$\mathcal{A}_1 = \left\{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{a}_{\text{CK}}^\top\mu(z_{i,\boldsymbol{v}})\right)^2 \leq C_1\right\} \tag{141}$$

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

$$\lim_{n,p\to\infty}\sup_{v_{\boldsymbol{z}_1},\cdots,v_{\boldsymbol{z}_n}}\left|\mathbb{E}\left[\mathbf{1}_{\mathcal{A}_1}\Phi\left(\widehat{\mathcal{R}}_n^\star(\boldsymbol{\Phi}_{CK},\boldsymbol{y}(\boldsymbol{Z}))\right) - \mathbf{1}_{\mathcal{A}_1}\Phi\left(\widehat{\mathcal{R}}_n^\star(\boldsymbol{\Phi}_{\text{CL}},\boldsymbol{y}(\boldsymbol{Z}))\right) \mid v_{\boldsymbol{z}_1},\cdots,v_{\boldsymbol{z}_n}\right]\right| = 0. \tag{142}$$

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 142, we obtain:

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

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

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

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 (126). 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:

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

where the last line follows from Fubini's theorem.

We note that the predictor $\hat{f}(\boldsymbol{z}) = \frac{1}{\sqrt{p}}\hat{a}_{\text{CL}}^\top \phi_{\text{CL}}(\boldsymbol{z})$ 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.