Analysing feature learning of gradient descent using periodic functions

Jaehui Hwang HWCMI62@KAIST.AC.KR
School of Computing, KAIST, Daejeon, Korea

Taeyoung Kim TAEYOUNGKIM21@KAIST.AC.KR
School of Computing, KAIST, Daejeon, Korea

Hongseok Yang HONGSEOK.YANG@KAIST.AC.KR
School of Computing, KAIST, Daejeon, Korea

Abstract

We present the analysis of feature learning in neural networks when target functions are defined by periodic functions applied to one-dimensional projections of the input. Previously, Damian et al. [2] considered a similar question for target functions of the form $f^*(x) = p^*(\langle u_1, x\rangle, \ldots, \langle u_r, x\rangle)$ for some vectors $u_1, \ldots, u_r \in \mathbb{R}^d$ and polynomial p^*, and proved that feature learning occurs during the training of a shallow neural network, even when the first-layer weights of the network are updated only once during training. Here feature learning refers to a subset of the first-layer weights $w_1, \ldots, w_m \in \mathbb{R}^d$ of the trained network being in the same directions as $\{u_1, \ldots, u_r\}$. We show that for periodic target functions, the same single gradient-based update of the first-layer weights induces feature learning of a shallow neural network, despite the additional challenge that feature learning for periodic functions now involves both directions and magnitudes of $\{u_1, \ldots, u_r\}$: a useful feature of, say, $f^*(x) = \sin(\langle u, x\rangle)$ is a vector $w \in \mathbb{R}^d$ such that $\angle(w, u) \approx 0$ and $\|w\| \approx \|u\|$. Our theoretical result shows that the sample complexity for learning a periodic target function of limited form using a shallow neural network grows polynomially with the input dimension, due to feature learning of the gradient-based training. Experimental results further support our theoretical finding, and illustrate the benefits of feature learning for a broader class of periodic target functions.

1. Introduction

Recent advances in deep learning theory have provided answers to several fundamental questions regarding the training and generalisation of deep neural networks. For instance, for over-parameterised networks, researchers have shown that the gradient-based training achieves a global optimum although the training objective is not convex [13]. Also, researchers have found ways to parameterise neural networks and their training algorithms so as to promote feature learning [17], and have come up with theoretical explanations on several puzzling phenomena of network training, such as double descent [11] and grokking [10].

In this paper, we focus on the analysis of feature learning. We build on the work of Damian et al. [2] who analysed feature learning from the perspective of sample complexity and showed the importance of a large learning rate for feature learning. Concretely, they considered shallow neural networks with one hidden layer, and studied a training scheme where the first step of training updates the weights of the first-layer only based on gradients, and the subsequent steps change the weights of the second-layer only until convergence. They showed that although the weights of the first-layer get updated only once in their setup, if the learning rate is large enough, this single

Algorithm 1 Layerwise training algorithm with learning rates $\eta_1, \eta_2 > 0$, regularisation coefficients $\lambda_1, \lambda_2 > 0$, max update step $T > 0$, and training dataset $\{(x_i, y_i)\}_{i=1}^N$.

$W^{(1)} = W^{(0)} - \eta_1 \nabla_{W^{(0)}} \left(\frac{1}{2N} \sum_{j=1}^N (y_j - f(x_j; A^{(0)}, W^{(0)}))^2 + \frac{\lambda_1}{2} \|W^{(0)}\|^2 \right)$
for $t = 1, ..., T$ **do**
  $A^{(t)} = A^{(t-1)} - \eta_2 \nabla_{A^{(t-1)}} \left(\frac{1}{2N} \sum_{j=1}^N (y_j - f(x_j; A^{(t-1)}, W^{(1)}))^2 + \frac{\lambda_2}{2} \|A^{(t-1)}\|^2 \right)$
end

update is still good enough to induce the learning of useful features, which in turn lets their training algorithm achieve a better sample complexity than kernel-based alternatives. This result is for a particular class of target functions f^* that have the following form:

$$f^* : \mathbb{R}^d \to \mathbb{R}, \qquad f^*(x) = g^*(\langle u_1, x \rangle, \ldots, \langle u_r, x \rangle), \tag{1}$$

for some integers $r \ll d$, vectors $u_1, \ldots, u_r \in \mathbb{R}^d$, and a polynomial g^*. That is, f^* is a polynomial on a low-dimensional projection of the input.

Our goal is to analyse feature learning when a target function f^* to learn is a *periodic* function applied to a low-dimensional projection of the input. This means that in our case, the target function f^* has the form of Equation (1) except that g^* is not a polynomial but a periodic function. Note that since f^* is defined by means of a periodic function, the results of Damian et al. [2] do not apply here. More importantly, because of the use of a periodic function, learning a useful feature in our setup is more difficult than learning such a feature in the setup of Damian et al. [2]. In the latter setup, it is enough to approximate the directions of $u_1, \ldots, u_r$ in Equation (1) well, but in the former setup, we need to have good approximations of both the directions and magnitudes of $u_1, \ldots, u_r$.

In this paper, we report the preliminary results that assume the target function has the form $f^*(x) = h^*(\sin(\langle u, x \rangle))$ for some $u \in \mathbb{R}^d$ and a polynomial h^* with odd-degree terms only. For the case that h^* is the identity function, we formally show that one gradient-based update step of the first-layer weights of a neural network is still good enough to induce the learning of useful features, i.e., those that describe both the direction and the magnitude of u well. We prove that when the first-layer weights of the network are updated once based on gradients, the number of samples needed to learn the target function grows polynomially with respect to the input dimension d. We also report the results of our experiments that confirm our formal result for the case of the identity h^*, and also show experimentally the presence of feature learning and its benefit for more general h^*'s.

2. Setup

Let S^{d-1} be the unit sphere in $\mathbb{R}^d$ (i.e., $S^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$). We assume that the input x has the distribution $x \sim N(0, I_d)$, and the target function $f^* : \mathbb{R}^d \to \mathbb{R}$ is of the form $f^*(x) = h^*(\sin(\langle u, x \rangle))$ for some $u \in \mathbb{R}^d$ and a function $h^* : \mathbb{R} \to \mathbb{R}$. Note that the norm of u is not fixed, and changing u by $2u$ induces a dramatic change of f^* because $\sin(b)$ and $\sin(2b)$ have very different periods. Our theoretical results make a further assumption that h^* is the identity function, but our experimental results consider general h^*, polynomials consisting only odd-degree terms. This choice of odd polynomials allow us to rewrite $h^*(\sin(b))$ by linear combination of $\sin((2k-1)b)$ for $k = 1, 2, \ldots$ with application of the trigonometric law.

Our model is a shallow neural network defined by $f(x; \{w_i, a_i\}_{i=1}^m) = \sum_{i=1}^m a_i \sin(\langle w_i, x \rangle)$ for each input $x \in \mathbb{R}^d$. Here m is an even integer, denoting the width of the network, and the

$w_i \in \mathbb{R}^d$ and $a_i \in \mathbb{R}$ are network parameters. We assume the following initialisation of the network parameters: $w_i^{(0)} \overset{\text{i.i.d.}}{\sim} \mathcal{P}$, $w_{i+(m/2)}^{(0)} = w_i^{(0)}$, $a_i^{(0)} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\{\pm 1\})$, and $a_{i+m/2}^{(0)} = -a_i^{(0)}$ for all $i = 1, \ldots, m/2$. Here $\mathcal{P}$ is some spherically symmetric distribution; for the details about $\mathcal{P}$, see Section A of the supplementary materials. This initialisation is similar to the one used by Damian et al. [2], and it guarantees that f is the constant zero function at initialisation. For each training step $t \geq 0$, we write $W^{(t)} = (w_1^{(t)}, \ldots, w_m^{(t)})$ and $A^{(t)} = (a_1^{(t)}, \ldots, a_m^{(t)})$ to mean, respectively, a tuple of the w_j's and that of the a_j's at step t.

The training algorithm for network parameters is given in Algorithm 1, which is a simplified version of the training algorithm used by Damian et al. [2]. The algorithm minimises the L^2-regularised mean-squared error (MSE) using the full-batch gradient descent (GD). Concretely, it starts by updating the parameters W of the first-layer once based on the gradient of the loss. The learning rate of this update is η_1, and the hyperparameter λ_1, multiplied to the L^2 regulariser $\|W\|^2$ in the loss, is set to $1/\eta_1$. Then, the algorithm repeatedly updates the parameters of the second-layer A again based on the gradient of the regularised MSE loss. This time the regulariser is only for the second-layer parameters, and the algorithm uses a new learning rate η_2. The number of repetition T and the learning rate η_2 are chosen, respectively, large enough and small enough to ensure the convergence of the updated second-layer parameters (within a small fixed error). For the value of the other hyperparameter λ_2, in our experiments, we used the one found by the usual hyperparameter search with a separated validation set.

3. Theoretical Result

Our theoretical results are summarised by the following theorem. $\tilde{\Omega}_{d,\epsilon,\delta}$ is the standard big-Omega notation where subscript denotes the asymptotic variables, ignoring the (poly-)logarithmic factors.

Theorem 1 *Let $\{(x_i, y_i)\}_{i=1}^N$ be the training set. Assume that the target function has the form $f^*(x) = \sin(\langle u, x \rangle)$ for some $u \in \mathbb{R}^d$ with $\|u\| \leq 2$. Suppose that the x_i's are i.i.d. samples from the d-dimensional standard normal distribution $x_i \sim \mathcal{N}(0, I_d)$ and $y_i = f^*(x_i)$ for all i. Fix sufficiently small $\delta > 0$ and $\epsilon > 0$. If the learning rate η_1 of first step, the number N of samples, the width m of the neural network f satisfy $\eta_1 \geq \exp(\|u\|^2/2)$, $m \geq \tilde{\Omega}_{d,\epsilon,\delta}(\log(1/\delta)\epsilon^{-4})$, and $N \geq \tilde{\Omega}_{d,\epsilon,\delta}(d(\log(1/\delta))^2\epsilon^{-8} + \delta^{-2})$ for some large enough $C > 0$, then there exists some $T_0 > 0$ such that for all $T \geq T_0$, the parameters $(A^{(T)}, W^{(1)})$ of the network learnt by Algorithm 1 with T iterations of its for-loop satisfies*

$$\frac{1}{2} \cdot \mathbb{E}_{x \sim \mathcal{N}(0, I_d)} \left[\left(f^*(x) - f(x; A^{(T)}, W^{(1)}) \right)^2 \right] \leq \epsilon \tag{2}$$

with probability at least $1 - \delta$. Here the probability is over the randomness of the sampled training set $\{(x_i, y_i)\}_{i=1}^N$ and the initialisation of network parameters.

The left-hand side of the inequality in Equation 2 is often referred to as risk. The theorem asserts that a sufficiently wide neural network, trained with a sufficient number of training examples, exhibits low risk with high probability. Notably, the required network width is independent of the input dimension, indicating that a constant width is adequate for any input dimension. Additionally, the number of required training samples increases linearly with the input dimension. In consequence, a single gradient step suffices to alleviate the curse of dimensionality. The proof of the theorem can be found in Section C of the supplementary material.

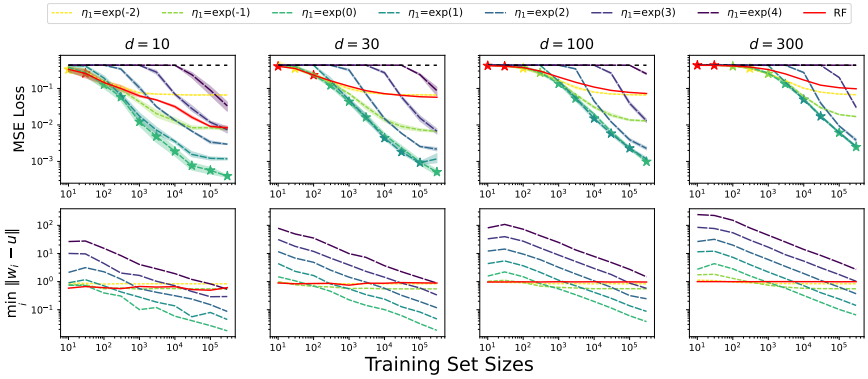

Figure 1: (Top) MSE losses of the networks trained by Algorithm 1 and the RF networks for multiple choices of input dimension d and learning rate η_1. The black dashed line represents the loss of the constant-zero predictor, and the stars the smallest losses achieved. (Bottom) Minimum distances between first-layer weights and target feature u, i.e., $\min_i \|w_i - u\|$.

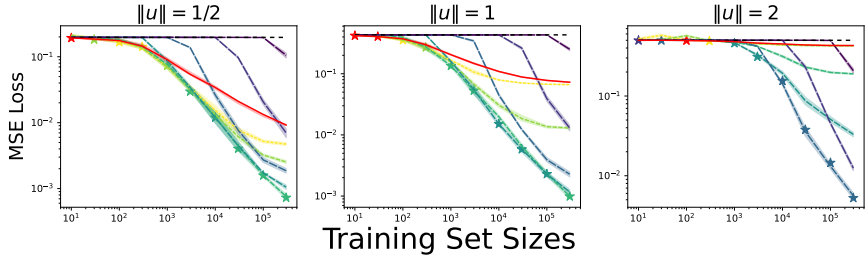

Figure 2: MSE losses of the networks trained by Algorithm 1 and the RF networks, for multiple target features u and learning rates η_1. We set $d = 100$. The black dashed line represents the loss of the constant-zero predictor, and the stars the smallest losses achieved.

We conjecture that if we omit the first one-step gradient update of the first-layer weights in Algorithm 1, which we call random-feature (RF) network, we cannot get the same width complexity as the one in Theorem 1. The formal statement of this conjecture is given as Conjecture 13 in Section D of the supplementary material, which also contains in-depth discussion about the conjecture.

4. Experimental Results

We describe the findings from our experiments. The first group of experiments are concerned with the target function $f^*(x) = \sin(\langle u, x\rangle)$ with $\|u\|$ being $1/2$, 1 or 2, and they aim at checking the claim of Theorem 1. The next group of experiments consider the target function $f^*(x) = (\sin(\langle u, x\rangle))^3$, or $\mathrm{He}_5(\sin(\langle u, x\rangle))/\sqrt{5!}$ with the same set of norms of u.[1] They check whether the claim of Theorem 1 holds for a wider class of target functions than those considered in the theorem.

1. $\mathrm{He}_5(t) = t^5 - 10t^3 + 15t$ is the sixth Hermite polynomial.

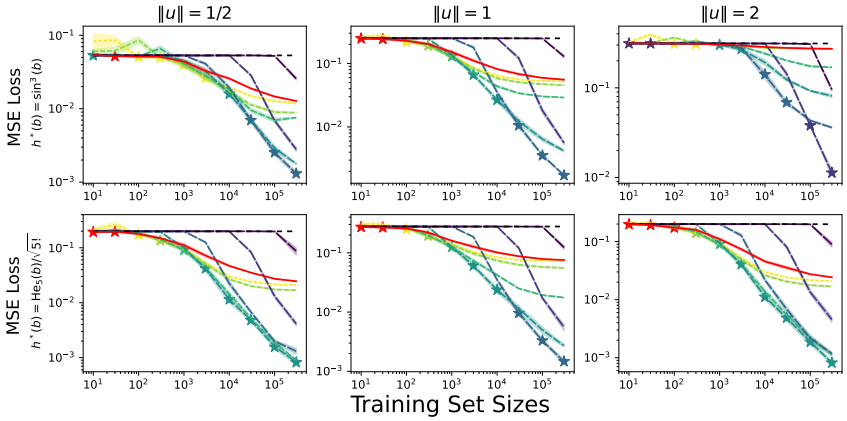

Figure 3: MSE losses of the networks trained by Algorithm 1 and the RF networks. The first row is about $f^*(x) = \sin(\langle u, x\rangle)^3$, and the second row $f^*(x) = \mathrm{He}_5(\sin(\langle u, x\rangle))/\sqrt{5!}$. The columns correspond to the $\|u\| = 1/2, 1, 2$ cases. The black dashed lines represent the loss of the constant-zero predictor, and the stars the smallest losses achieved.

In these two groups of experiments, we set the widths of neural networks to 1,000. By networks, we mean both those trained by Algorithm 1 and the RF networks (considered in Conjecture 13) whose first-layer weights are not trained. The experiments consider training sets of different sizes $N \in \{10^i, 3\times 10^i : i = 1, 2, 3, 4, 5\}$, and input dimensions $d \in \{10, 30, 100, 300\}$ for first group of experiments and $d = 100$ for second group. The iteration number T of the for-loop of Algorithm 1 is set to 10,000, but is shortened during the training when the algorithm detects the convergence of the training loss. The regularisation coefficient λ_2 is selected based on the hyperparameter search over the grid $\{2^{i/4}\}_{i=-48}^{72}$ with the validation set of size 10,000. We repeated the experiments 5 times and plotted their results with 95% bootstrap confidence intervals.

Figures 1 and 2 show the results of the experiments in the first group under seven different choices of the learning rate η_1 for the one-step first-layer training of Algorithm 1. The plots in the first row show the mean-squared-error (MSE) losses estimated with 100,000 samples, and those in the second show the minimum L_2 distance between the vector u of the target function (which can be viewed as a feature used by the target function) and the weight w_i of the neuron i in a neural network; the minimum is taken over all the neurons in the network. When η_1 is $\exp(0)$ or $\exp(1)$, the networks with trained first and second layers (by Algorithm 1) outperform the RF networks in terms of the MSE loss. Furthermore, the gaps between the losses of these two types of networks grow as the input dimension increases. As expected, the RF networks do not learn the u vector of the target function in their W parameters, but the networks trained by Algorithm 1 under appropriately large η_1 (namely, $\eta_1 \in \{\exp(0), \exp(1), \exp(2)\}$) learn the u vector in their W parameters. Since we measure $\|u - w_j\|$, the learning here is concerned with both the direction and the magnitude of u. Comparing the best performing learning rate in Figure 2, one can see that as the norm of u increases, the larger learning rate performs better, as indicated in the condition of η_1 of Theorem 1. Note the slow-down in the decreases of the MSE losses of the RF networks in the later part of training (of the second-layer parameters) when $d > 10$. This indicates the possibility that the RF networks have

exponential sample complexities as conjectured (since all polynomial relationships appear as lines in the log-log plots that we use here).

Figure 3 shows the results of the experiments in the second group, which are concerned with general target functions not covered by Theorem 1. It shows clear performance gaps in terms of the MSE loss between the networks trained by Algorithm 1 and the random feature networks.

5. Related Works and Future Works

After observing that theoretical predictions of over-parameterised neural networks often fail to describe the outcomes of practical network training [4, 7, 14], researchers studied alternative setups to over-parameterisation, which are still simple enough to facilitate theoretical analysis. One such alternative is a setup of Damian et al. [2] and Ba et al. [1], where the model is a shallow neural network $f(x) = \sum_{i=1}^{m} a_i \sigma(\langle w_i, x \rangle + b_i)$ with its parameters being trained as in our paper, and the target function has the form in Equation (1). Damian et al. [2] proved that while the models based on rotation-invariant kernels cannot exploit the low-dimensional structure of the target function (i.e., $r \ll d$) and their sample complexities increase as d^p where p is degree of g^*, their neural-network model can exploit the structure and have sample complexity $d^2 r + dr^p$. Ba et al. [1] considered a similar setup where $r = 1$ and g^* in Equation 1 is allowed to be a non-polynomial. They showed that while the so-called random feature model or the models based on over-parameterised networks do not perform better than the linear estimator asymptotically, their neural-network model (trained by an algorithm similar to ours) exhibits performance improvement over the linear estimator by learning higher-order features. Our work extends this line of research by studying periodic target functions.

While our work goes beyond the existing research on feature learning, it is limited in that we only analyse particular cases, and we have the same function for activation and target functions. As in the results of the Yehudai and Shamir [19], one of our future directions is giving general results for different activations and finding general conditions on the activation that allows feature learning. Also, we assume that the input is distributed as standard Gaussian, which allows us to use existing theory on the Gaussian expectations. It will be interesting to relax this assumption, like allowing the input distribution to be only spherically symmetric.

Our work was motivated by the recent uses of periodic activation functions in neural networks. Sitzmann et al. [15] used such functions in their work on the implicit representations of images and 3D objects, and showed that those activation functions lead to the improvement in the qualities of the reconstructed images and 3D objects over the baselines with ReLU and $\tanh$. The neural networks with periodic activation functions studied by Sitzmann et al. [15] are closely related to a kernel similar to the RBF kernel, which is known to model high-frequency functions well. Li and Pathak [8] proposed to regularise the learned function's frequencies by including a periodic embedding layer and controlling the initialisation of the parameters of the layer. Their approach has been applied to the problem of modelling the reward function in RL under high-frequency noises and mid- or low-frequency signals. Our results suggest the relevance of the theoretical and empirical analysis of feature learning to these works on periodic activation functions. Similar approaches suggest use of repeated application of Gabor or Fourier basis filter [3], and even infinite application of them [6]. We conjecture that such iterative nature of these architectures allow the network learn the high-frequency features thereby showing different feature learning behaviour compared to MLP. Extending our analysis to such models is interesting future direction.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (No. RS-2023-00279680).

References

[1] 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. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/f7e7fabd73b3df96c54a320862afcb78-Paper-Conference.pdf.

[2] 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. URL https://proceedings.mlr.press/v178/damian22a.html.

[3] Rizal Fathony, Anit Kumar Sahu, Devin Willmott, and J. Zico Kolter. Multiplicative filter networks. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=OmtmcPkkhT.

[4] Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5850–5861. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/405075699f065e43581f27d67bb68478-Paper.pdf.

[5] Hong Hu and Yue M. Lu. Universality laws for high-dimensional learning with random features. *IEEE Trans. Inf. Theor.*, 69(3):1932–1964, mar 2023. ISSN 0018-9448. doi: 10.1109/TIT.2022.3217698. URL https://doi.org/10.1109/TIT.2022.3217698.

[6] Zhichun Huang, Shaojie Bai, and J. Zico Kolter. (Implicit)2: Implicit layers for implicit representations. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 9639–9650. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4ffbd5c8221d7c147f8363ccdc9a2a37-Paper.pdf.

[7] Jaehoon Lee, Samuel Schoenholz, Jeffrey Pennington, Ben Adlam, Lechao Xiao, Roman Novak, and Jascha Sohl-Dickstein. Finite versus infinite neural networks: an empirical study. In

H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15156–15172. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/ad086f59924fffe0773f8d0ca22ea712-Paper.pdf.

[8] Alexander Li and Deepak Pathak. Functional regularization for reinforcement learning via learned fourier features. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 19046–19055. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/9f0609b9d45dd55bed75f892cf095fcf-Paper.pdf.

[9] Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018. ISSN 10505164, 21688737. URL https://www.jstor.org/stable/26542333.

[10] Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon Shaolei Du, Jason D. Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=XsHqr9dEGH.

[11] 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. doi: https://doi.org/10.1002/cpa.22008. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.22008.

[12] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012. ISBN 026201825X.

[13] Bartłomiej Polaczyk and Jacek Cyranka. Improved overparametrization bounds for global convergence of SGD for shallow neural networks. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=RjZq6W6FoE.

[14] Mariia Seleznova and Gitta Kutyniok. Analyzing finite neural networks: Can we trust neural tangent kernel theory? In Joan Bruna, Jan S. Hesthaven, and Lenka Zdeborová, editors, *Mathematical and Scientific Machine Learning, 16-19 August 2021, Virtual Conference / Lausanne, Switzerland*, volume 145 of *Proceedings of Machine Learning Research*, pages 868–895. PMLR, 2021. URL https://proceedings.mlr.press/v145/seleznova22a.html.

[15] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7462–7473. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/53c04118df112c13a8c34b38343b9c10-Paper.pdf.

[16] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

[17] Greg Yang and Edward J. Hu. Tensor programs iv: Feature learning in infinite-width neural networks. 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 11727–11737. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/yang21c.html.

[18] Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. 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. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/5481b2f34a74e427a2818014b8e103b0-Paper.pdf.

[19] Gilad Yehudai and Ohad Shamir. Learning a single neuron with gradient methods. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3756–3786. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/yehudai20a.html.

Appendix A. Experiment Details

The experiments were done in 4 NVIDIA GeForce RTX 2080 Ti GPU and Intel(R) Xeon(R) Gold 6234 CPU @ 3.30GHz with 2 CPU and 512 GB RAM.

The initialisation distribution $\mathcal{P}$ is designed so that in asymptotics, $w^{(1)}$ has its norm's distribution close to $\mathrm{Unif}([0, \eta \exp(-\|u\|^2/2)\|u\|])$. The implementation of sampling from this distribution is given as following:

$$\overline{w} \sim \mathrm{Unif}(S^{d-1}), \quad t \sim \mathrm{Exp}(\lambda = 0.5), \quad w = \overline{w} \times \sqrt{t}.$$

Appendix B. Additional Experiments

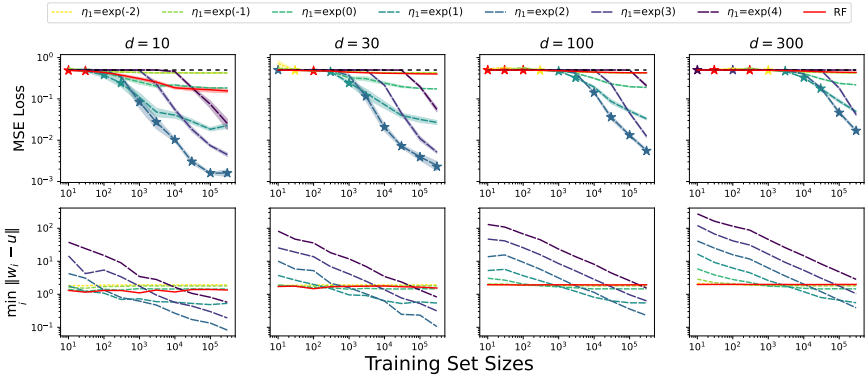

Figure 4: **Feature Learning under Larger Frequency**: (Top) The MSE loss of experiments same as Figure 1, but with $\|u\| = 2$. Black dashed line represents the risk of zero predictor. (Bottom) The minimum distance between learned weight and target feature $\min_i \|w_i^{(1)} - u\|$.

In this section, we report additional experiments. We first include the case where u is not unit norm. Figure 4 is the case when $\|u\| = 2$, and Figure 5 is the case when $\|u\| = 1/2$. As stated in Theorem 1, the required learning rate changes as $\|u\|$ changes, but as long as the learning rate is large enough, the feature learning occurs.

We finally give a preliminary version of our experiments to see whether our results generalise beyond the single-index setting, i.e., the multi-index case

$$f^*(x) = \sin(\langle u_1, x\rangle) + \sin(\langle u_2, x\rangle).$$

We randomly sample $u_1, u_2 \sim \mathrm{Unif}(S^{d-1})$ where the input dimension $d = 100$. This gives orthogonal u_1, u_2 with high probability, so this randomness does not affect the result largely. As shown in Figure 6, we can see the slight performance gap between networks trained with Algorithm 1 and random feature networks, but the difference is not so dramatic. Note that the y-axis is a linear scale, which is different to the figures in the main text.

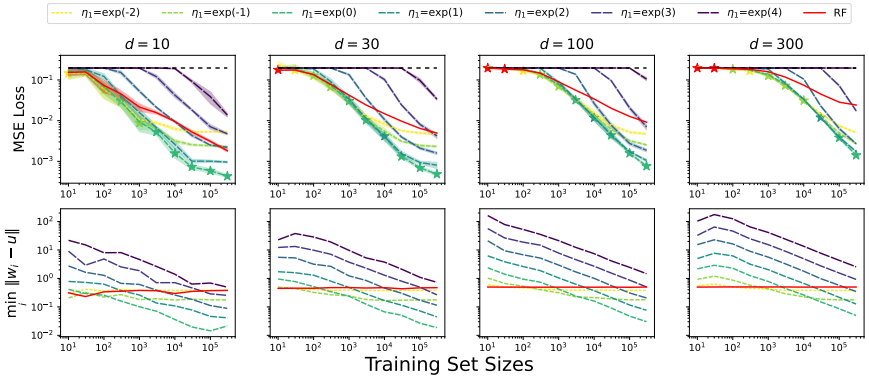

Figure 5: **Feature Learning under Smaller Frequency**: (Top) The MSE loss of experiments same as Figure 1, but with $\|u\| = 1/2$. Black dashed line represents the risk of zero predictor. The stars denote the smallest risk achieved. (Bottom) The minimum distance between learned weight and target feature $\min_i \|w_i^{(1)} - u\|$.

Appendix C. Proof of Theorem 1

In this section, we will give proof of Theorem 1. Throughout the proof, we will use universal constant C that is not consistently used, which means that multiple C can occur in the proof while being their values distinct. Before giving the proof, we give a more detailed proof idea.

First assume that there exist infinitely many training samples so that we can compute the true gradient of w:

$$\begin{aligned} g(w) :=& \frac{1}{2} \cdot \nabla_{w_i^{(0)}} \mathop{\mathbb{E}}_{x} \left[\left(f(x; A^{(0)}, W^{(0)}) - f^*(x) \right)^2 \right] \\ =& \mathop{\mathbb{E}}_{x} \left[\underbrace{\cos(\langle w, x \rangle)}_{\nabla_w f(x;A,W)} \underbrace{\sin(\langle u, x \rangle)}_{f^*(x)} x \right] \\ =& \mathop{\mathbb{E}}_{x} \left[-\sin(\langle w, x \rangle) \sin(\langle u, x \rangle) w + \cos(\langle w, x \rangle) \cos(\langle u, x \rangle) u \right], \end{aligned}$$

where we used Stein's lemma in the last equality.

Applying the analytic form of these expressions (see Section 5.3 of [9]), we can get a closed-form gradient

$$\exp\left(-\frac{\|u\|^2 + \|w\|^2}{2} \right) (\cosh(\langle w, u \rangle) u - \sinh(\langle w, u \rangle) w).$$

If one chooses w to be orthogonal to u or close to being orthogonal, one can see that

$$g(w) \approx \exp\left(-\frac{\|u\|^2 + \|w\|^2}{2} \right) u.$$

The initialisation scheme (A) was chosen so that $\exp(-\|w\|^2/2)$ has distribution $\mathrm{Unif}([0, 1])$, so with learning rate η_1, this ideal gradient gives $g(w)$ close to the Cu where

$$C \in [0, \eta_1 \exp(-\|u\|^2/2)].$$

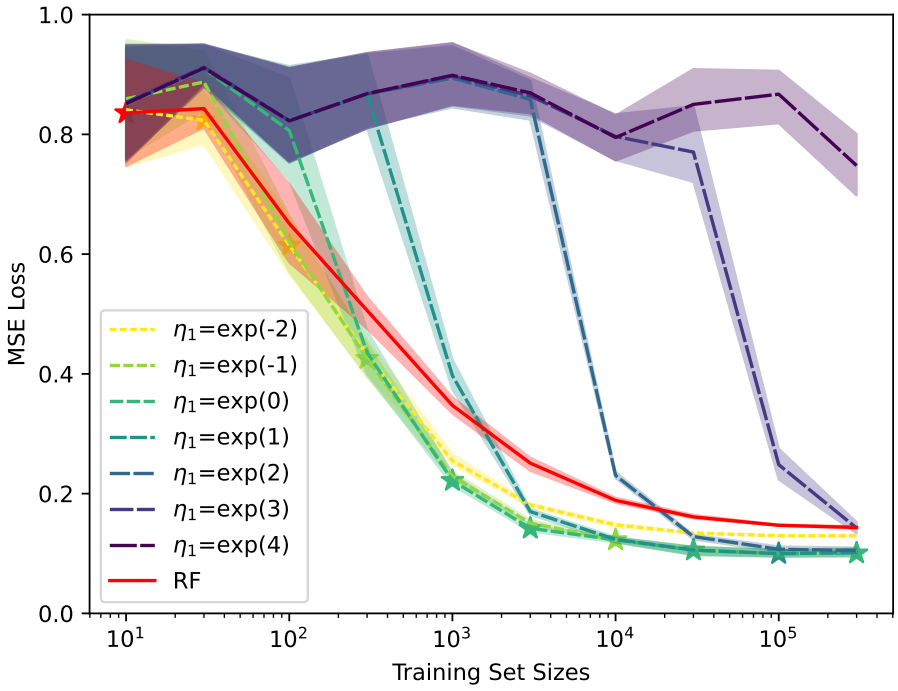

Figure 6: **Feature Learning for Multi-index**: The MSE Loss of experiments same as Figure 1, but with $f^*(x) = \sin(\langle u_1, x\rangle) + \sin(\langle u_2, x\rangle)$. The input dimension d is set to 100. The stars denote the smallest risk achieved.

To show this claim, we first show some properties of this g function.

Lemma 2 (Lipschitz-ness of Gradient) *Let*

$$g(w) = \mathbb{E}[\sigma'(\langle w, x\rangle) f^*(x) x],$$

then $g(w)$ is 1-Lipschitz.

Proof Consider the fixed x case, i.e.,

$$\hat{g}_1(w) = \cos(\langle w, x\rangle) \sin(\langle u, x\rangle) x.$$

We can take the derivative w.r.t. w, which gives

$$-\sin(\langle w, x\rangle) \sin(\langle u, x\rangle) x x^\intercal.$$

So, the Lipschitz constant of g can be upper bounded as

$$\|g\|_{\text{Lip}} \leq \left\| \mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,I_d)} [-\sin(\langle w, x\rangle) \sin(\langle u, x\rangle) x x^\intercal] \right\|$$

$$
\begin{aligned}
&\leq \left\| \mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,I_d)} \left[xx^\intercal\right] \right\| \\
&\leq \|I_d\| \\
&= 1.
\end{aligned}
$$

■

Lemma 3 (Boundedness of Gradient) *Let*

$$g(w) = \mathbb{E}[\sigma'(\langle w, x\rangle) f^*(x) x],$$

then $g(w)$ is bounded, i.e.,

$$\|g(w)\| \leq 2\|u\| + 1.$$

Proof We can upper bound each terms in the closed-form solution,

$$g(w) = \exp\left(-\frac{\|w\|^2 + \|u\|^2}{2}\right) \left(\cosh(\langle w, u\rangle) u - \sinh(\langle w, u\rangle) w\right).$$

From

$$\cosh(\langle w, u\rangle) = \cosh(|\langle w, u\rangle|) \leq \cosh(\|w\|\|u\|) \leq \exp\left(\|w\|\|u\|\right),$$

we have

$$
\begin{aligned}
&\left\| \exp\left(-\frac{\|w\|^2 + \|u\|^2}{2}\right) \cosh(\langle w, u\rangle) u \right\| \\
\leq &\exp\left(-\frac{\|w\|^2 + \|u\|^2 - 2\|w\|\|u\|}{2}\right) \|u\| \\
\leq &\|u\|.
\end{aligned}
$$

For the second term, we can first upper bound by similar computation:

$$
\begin{aligned}
\exp\left(-\frac{\|w\|^2 + \|u\|^2}{2}\right) \|\sinh(\langle w, u\rangle) w\| &\leq \exp\left(-\frac{\|w\|^2 + \|u\|^2}{2}\right) \sinh(\|w\|\|u\|)\|w\| \\
&\leq \frac{1}{2} \exp\left(-\frac{\|w\|^2 + \|u\|^2 - \|w\|\|u\|}{2}\right) \|w\|
\end{aligned}
$$

We can easily find the maximum of this term by taking the derivative w.r.t. $\|w\|$, by substituting $\|w\|$ with t,

$$
\begin{aligned}
&\frac{d}{dt} \exp\left(-\frac{t^2 + \|u\|^2 - 2\|u\|t}{2}\right) t \\
= &\exp\left(-\frac{t^2 - 2\|u\|t + \|u\|^2}{2}\right) \left(-t^2 + \|u\|t + 1\right),
\end{aligned}
$$

setting this value to zero gives

$$t^* = \frac{\|u\| + \sqrt{\|u\|^2 + 4}}{2}$$

which gives an upper bound

$$\exp\left(-\frac{1}{2}\left(\sqrt{\|u\|^2+4}-\|u\|\right)^2\right)\cdot\frac{\|u\|+\sqrt{\|u\|^2+4}}{2}\leq\frac{\|u\|+\sqrt{\|u\|^2+4}}{2}\leq\|u\|+1.$$

■

We now begin showing the first claim, that if w is close to being orthogonal to u, $g(w)$ is similar to u.

Lemma 4 (Gradient aligns with feature of target function) *Let $g(w)$ be the expected gradient of the first-layer weight w at initialisation,*

$$g(w)=\mathop{\mathbb{E}}_{x}\left[\sigma'(\langle w,x\rangle)f^*(x)x\right].$$

If $\delta\leq 1/\|u\|$, then the existence of some $v\in\mathbb{R}^d$ satisfying

$$\langle v,u\rangle=0,\quad \|v-w\|\leq\delta,$$

implies

$$\left\|g(w)-\exp\left(-\frac{\|w\|^2+\|u\|^2}{2}\right)u\right\|\leq 3\delta.$$

Proof We first have

$$\begin{aligned}&\left\|g(w)-\exp\left(-\frac{\|w\|^2+\|u\|^2}{2}\right)u\right\|\\ =&\exp\left(-\frac{\|w\|^2+\|u\|^2}{2}\right)\|\cosh(\langle u,w\rangle)u-u-\sinh(\langle u,w\rangle)w\|\\ \leq&\exp\left(-\frac{\|w\|^2+\|u\|^2}{2}\right)(\|\cosh(\langle u,w\rangle)u-u\|+\|\sinh(\langle u,w\rangle)w\|)\end{aligned}$$

To bound the first term, we can relate the value by existence of vector v:

$$\begin{aligned}&\|\cosh(\langle u,w\rangle)u-u\|\\ \leq&|\cosh(\langle u,w\rangle)-1|\,\|u\|\\ \leq&|\cosh(\langle u,w-v\rangle)-1|\,\|u\|\\ \leq&|\cosh(\|u\|\|w-v\|)-1|\,\|u\|\end{aligned}$$

where we used the fact that $\cosh$ is increasing function, and $\cosh(x)\geq 1$ for any x. From the assumption $\|w-v\|\leq\delta$, we can reduce further and

$$|\cosh(\|u\|\|w-v\|)-1|\,\|u\|\leq|\cosh(\delta\|u\|)-1|\,\|u\|\leq\delta\|u\|$$

assuming that $\delta\|u\|\leq 1$, using numerical inequality $x\geq\cosh(x)-1$ in $0\leq x\leq 1$.

For the second term, again we can introduce v and use the fact that $\sinh$ is increasing function to have

$$\|\sinh(\langle u,w\rangle)w\|$$

$$
\begin{aligned}
&\leq |\sinh(\langle u, w\rangle)|\, \|w\| \\
&= |\sinh(\langle u, w - v\rangle)|\, \|w\| \\
&\leq |\sinh(\|u\|\|w - v\|)|\, \|w\| \\
&\leq 2\|u\|\delta\|w\|,
\end{aligned}
$$

again assuming that $\delta\|u\| \leq 1$ and numerical inequality $\sinh(x) \leq 2x$ in $0 \leq x \leq 1$.

Combining these two results, we have

$$
\left\| g(w) - \exp\left(-\frac{\|w\|^2 + \|u\|^2}{2}\right) u \right\| \leq \exp\left(-\frac{\|w\|^2 + \|u\|^2}{2}\right) (\|u\|^2 + 2\|u\|\|w\|)\delta.
$$

One can see that both

$$
\begin{aligned}
\exp\left(-\frac{\|w\|^2 + \|u\|^2}{2}\right) \|u\|^2 \leq 1, \\
\exp\left(-\frac{\|w\|^2 + \|u\|^2}{2}\right) \|u\|\|w\| \leq 1,
\end{aligned}
$$

so we have

$$
\left\| g(w) - \exp\left(-\frac{\|w\|^2 + \|u\|^2}{2}\right) u \right\| \leq 3\delta.
$$

■

Lemma 5 *Suppose that w follows distribution described in Section 2. Let $\|u\|$ and η_1 fixed, and satisfy $\epsilon \leq 2\|u\|$. If*

$$
\eta_1 \geq \exp\left(\frac{\|u\|^2}{2}\right),
$$

then

$$
P(\|\eta_1 g(w) - u\| \leq \epsilon) \geq \frac{\epsilon^2}{12\pi^2\eta_1^2\sqrt{2\log\eta_1}}.
$$

for p only depending on η_1 and ϵ.

Proof We use the result of Lemma 4. We bound the probability to choose 'nearly orthogonal' direction of u, and the probability of choosing appropriate η_1 each. Formally, we introduce w' which is dependent on w and proceed as following:

$$
\begin{aligned}
P(\|\eta_1 g(w) - u\| \leq \epsilon) &= P(\|\eta_1 g(w) - \eta_1 g(w') + \eta_1 g(w') - u\| \leq \epsilon) \\
&\geq P(\|\eta_1 g(w) - \eta_1 g(w')\| + \|\eta_1 g(w') - u\| \leq \epsilon) \\
&\geq P(\|\eta_1 g(w) - \eta_1 g(w')\| \leq \epsilon/2 \wedge \|\eta_1 g(w') - u\| \leq \epsilon/2) \geq p.
\end{aligned}
$$

Here, we define w' as follows.

$$
w' = \frac{\|w\|}{\|w_\perp\|} w_\perp, \quad w_\perp = w - \frac{\langle w, u\rangle}{\|u\|^2} u.
$$

Then, w' satisfies $\langle w', u\rangle = 0$ and $\|w'\| = \|w\|$. We decompose the probability by introducing conditional probability,

$$\begin{aligned}&P(\|\eta_1 g(w) - \eta_1 g(w')\| \le \epsilon/2 \wedge \|\eta_1 g(w') - u\| \le \epsilon/2)\\ =&P(\|\eta_1 g(w') - u\| \le \epsilon/2) P\left(\|\eta_1 g(w) - \eta_1 g(w')\| \le \frac{\epsilon}{2} \middle| \|\eta_1 g(w') - u\| \le \frac{\epsilon}{2}\right).\end{aligned}$$

We will handle the first term, then consider the conditional probability. For the first term, we can compute as

$$\begin{aligned}\left\|\eta_1 g(w') - u\right\| &= \left\|\eta_1 \exp\left(-\frac{\|w\|^2 + \|u\|^2}{2}\right)(\cosh(0)u - \sinh(0)w) - u\right\|\\ &= \left|\eta_1 \exp\left(-\frac{\|w\|^2 + \|u\|^2}{2}\right)\right| \|u\|.\end{aligned}$$

Since $\|w\|^2 \sim \exp(1/2)$, we have

$$\begin{aligned}P\left(\exp\left(-\frac{\|w\|^2}{2}\right) \le t\right) &= P(\|w\|^2 \ge -2\log x)\\ &= \exp\left(-\frac{1}{2}\log x^{-2}\right)\\ &= x,\end{aligned}$$

which showing

$$\exp(\|w\|^2/2) \sim \mathrm{Unif}([0,1]).$$

Therefore,

$$\|\eta_1 g(w') - u\| \sim \mathrm{Unif}\left(\left[-\|u\|, \eta_1 \exp\left(-\frac{\|u\|^2}{2}\right)\|u\| - \|u\|\right]\right),$$

and from the assumption of $\eta_1 \ge \exp(\|u\|^2/2)$, we can see that this support includes 0.

Assuming that $\epsilon \le 2\|u\|$, we have

$$P\left(\|\eta_1 g(w') - u\| \le \frac{\epsilon}{2}\right) \ge \frac{\epsilon}{2\eta_1 \exp(-\|u\|^2/2)\|u\|} \ge \frac{\epsilon}{2\eta_1},$$

where we used that $\exp(-t^2/2)t \le 1$ for any $t \ge 0$. Now for the conditional probability, note that by construction of w', we have

$$g(w') = \exp\left(-\frac{\|w\|^2 + \|u\|^2}{2}\right)u.$$

We can see that applying $\delta = \epsilon/(6\eta_1)$ implies the error bound of $\epsilon/2$, by Lemma 4. Using these, we can see that

$$\|w - w'\| \le \frac{\epsilon}{6\eta_1} \quad \Rightarrow \quad \|\eta_1 g(w) - \eta_1 g(w')\| \le \frac{\epsilon}{2}.$$

From our previous conditioning, we have

$$\eta_1 \exp\left(-\frac{\|w\|^2}{2}\right) \in \left[1 - \frac{\epsilon}{2\|u\|}, 1\right]$$

which gives

$$\|w\| \leq \sqrt{2\log \eta_1}. \tag{3}$$

From this, it is enough to have

$$\|w - w'\| \leq \|w\| \angle(w, w') \leq \frac{\epsilon}{6\eta_1},$$

which is equivalent to

$$\angle(w, w') \leq \frac{\epsilon}{6\eta_1\sqrt{2\log \eta_1}}.$$

Since $\angle(w, w') = \pi/2 - \angle(w, u) \leq \frac{\pi}{2}\cos(\angle(w, u))$, we can weaken the inequality to

$$\frac{|\langle w, u\rangle|}{\|w\| \cdot \|u\|} \leq \frac{\epsilon}{3\pi\eta_1\sqrt{2\log \eta_1}}.$$

Due to the spherical symmetry of w's distribution, we can see that this probability increases as the input dimension d increases, so it is enough to consider the case when $d = 2$. In such cases, we can replace the inner product to some $z \sim \text{Unif}([-\pi, \pi])$ and have

$$P\left(|z| \leq \frac{\epsilon}{3\pi\eta_1\sqrt{2\log \eta_1}}\right) = \frac{\epsilon}{3\pi^2\eta_1\sqrt{2\log \eta_1}}.$$

■

What we have shown in Lemma 4 is that if we assume infinitely many samples, so that we can compute the exact gradient, the weight aligns with the feature of the target function. Moreover, we have shown that such an event happens with constant probability for any input dimension, hence with enough width not depending on the input dimension, we get close to the u.

Now we relate this result to $W^{(1)}$ in Algorithm 1 by showing that the empirical gradient concentrates around its expectation. We write $\hat{g}_N(w)$ as the empirical gradient computed in Algorithm 1, whose expression is

$$\hat{g}_N(w) = \frac{1}{N}\sum_{i=1}^{N} \sigma'(\langle w, x_i\rangle) f^*(x_i) x_i = \frac{1}{N}\sum_{i=1}^{w} \cos(\langle w, x_i\rangle)\sin(\langle u, x_i\rangle) x_i.$$

Lemma 6 (Concentration of gradient) *Over the randomness of sampling of training data,*

$$\sup_{\|w\| \leq L} \|\hat{g}_N(w) - g(w)\| \leq 64\sqrt{\frac{d\log(dNL)}{N}}$$

with probability $\geq 1 - 1/\sqrt{N}$.

Proof We first introduce the truncation on $\hat{g}_N(w)$ to work on the bounded space. We define $\tilde{g}_N(w)$ as

$$\tilde{g}_N(w) = \sum_{i=1}^{w} \cos(\langle w, x_i\rangle)\sin(\langle u, x_i\rangle) x_i \mathbb{I}_{\|x_i\| \leq M}$$

for some constant M to be defined later.

Using this truncation, we will upper bound the difference as

$$\|\hat{g}_N(w) - g(w)\| \leq \|\hat{g}_N(w) - \tilde{g}_N(w)\| + \left\|\tilde{g}_N(w) - \mathop{\mathbb{E}}_x[\tilde{g}_N(w)]\right\| + \left\|\mathop{\mathbb{E}}_x[\tilde{g}_N(w)] - g(w)\right\|.$$

The first term is zero if all the inputs are bounded $\|x_i\| \leq M$. For $x \sim \mathcal{N}(0, I_d)$, we have

$$P\left(\|x\| \geq \sqrt{d+t}\right) \leq \exp(-t)$$

for any $t \geq 0$. Using this for independent x, we can apply Union-bound to have

$$\begin{aligned} P\left(\sup_{i\in[N]} \|x_i\| \geq \sqrt{d+t}\right) &= P\left(\exists(i \in [N]).\, \|x_i\| \geq \sqrt{d+t}\right) \\ &\leq N\exp(-t). \end{aligned}$$

Upon rewriting, we have

$$P\left(\sup_{i\in[N]} \|x_i\| \geq M\right) \leq N\exp(-M^2 + d)$$

so we have

$$\hat{g}_N(w) = \tilde{g}_N(w) \tag{4}$$

with probability $\geq 1 - N\exp(-M^2 + d)$.

For the second term, we will consider the concentration of this truncated gradient. We can first rewrite the norm of the difference by inner product with unit vectors,

$$\left\|\tilde{g}_N(w) - \mathop{\mathbb{E}}_x[\tilde{g}_N(w)]\right\| = \sup_{u\in S^{d-1}} \left\langle \tilde{g}_N(w) - \mathop{\mathbb{E}}_x[\tilde{g}_N(w)], u\right\rangle.$$

Let $\mathcal{N}_{1/2}$ be the ϵ-net on the sphere S^{d-1}, then

$$\sup_{u\in S^{d-1}} \left\langle \tilde{g}_N(w) - \mathop{\mathbb{E}}_x[\tilde{g}_N(w)], u\right\rangle \leq 2 \sup_{u\in \mathcal{N}_{1/2}} \left\langle \tilde{g}_N(w) - \mathop{\mathbb{E}}_x[\tilde{g}_N(w)], u\right\rangle.$$

Now if we recall the definition of $\tilde{g}_N(w)$,

$$\tilde{g}_N = \frac{1}{N}\sum_{i=1}^{N} \cos(\langle w, x_i\rangle)\sin(\langle u, x_i\rangle)x_i\mathbb{I}_{\|x_i\|\leq M},$$

$\cos(\langle w, x_i\rangle), \sin(\langle u, x_i\rangle)$ are bounded by 1, so the sub-Gaussian norm of this random vector is determined by the sub-Gaussian norm of $x_i\mathbb{I}_{\|x_i\|\leq M}$.

For the spherically symmetric distribution with its norm bounded by M, we can see that the random variable with the largest sub-Gaussian norm is $\mathrm{Unif}(MS^{d-1})$, whose sub-Gaussian norm is $2M/\sqrt{d}$. So we have

$$\left\langle \cos(\langle w, x_i\rangle)\sin(\langle u, x_i\rangle)x_i\mathbb{I}_{\|x_i\|\leq M}, u\right\rangle$$

is $2M/\sqrt{d}$-sub-Gaussian random variable. Applying the centering lemma [16], we have

$$\left\langle \cos(\langle w, x_i\rangle)\sin(\langle u, x_i\rangle)x_i\mathbb{I}_{\|x_i\|\leq M} - \mathop{\mathbb{E}}_x\left[\tilde{g}_N(w)\right], u\right\rangle$$

as $4M/\sqrt{d}$-sub-Gaussian random variable. Applying Hoeffding's inequality shows that

$$\left\langle \tilde{g}_N(w) - \mathop{\mathbb{E}}_x \left[\tilde{g}_N(w)\right], u \right\rangle \leq 4M\sqrt{\frac{2z}{dN}}$$

with probability $\geq 1 - 2\exp(-z)$, for any $z \geq 0$. By applying the union-bound over $u \in \mathcal{N}_{1/2}$, we have

$$\left\| \tilde{g}_N(w) - \mathop{\mathbb{E}}_x \left[\tilde{g}_N(w)\right] \right\| \leq 8M\sqrt{\frac{2z}{dN}}$$

with probability $\geq 1 - 2 \cdot 5^d \cdot \exp(-z)$.

Now let $\mathcal{N}_\epsilon$ be the ϵ-net of $\{w : \|w\| \leq L\}$, whose cardinality is smaller than $(1 + 2L/\epsilon)^d$. Let $\pi : \{w : \|w\| \leq L\} \to \mathcal{N}_\epsilon$ as the projection function that satisfy $\|\pi(w) - w\| \leq \epsilon$. Then we have

$$\begin{aligned}
&\sup_{\|w\| \leq L} \left\| \tilde{g}_N(w) - \mathop{\mathbb{E}}_x \left[\tilde{g}_N(w)\right] \right\| \\
&\leq \sup_{\|w\| \leq L} \|\tilde{g}_N(w) - \tilde{g}_N(\pi(w))\| + \sup_{w \in \mathcal{N}_\epsilon} \left\| \tilde{g}_N(w) - \mathop{\mathbb{E}}_x \left[\tilde{g}_N(w)\right] \right\| + \sup_{\|w\| \leq L} \left\| \mathop{\mathbb{E}}_x [\tilde{g}_N(\pi(w))] - \mathop{\mathbb{E}}_x [\tilde{g}_N(w)] \right\|.
\end{aligned}$$

The second term can be upper bounded by taking the union-bound on the ϵ-net $\mathcal{N}_\epsilon$, i.e.,

$$\sup_{w \in \mathcal{N}_\epsilon} \left\| \tilde{g}_N(w) - \mathop{\mathbb{E}}_x \left[\tilde{g}_N(w)\right] \right\| \leq 8M\sqrt{\frac{2z}{dN}}$$

with probability $\geq 1 - 2 \cdot 5^d + (1 + 2L/\epsilon)^d \exp(-z)$.

By similar computation as Lemma 2, we can show that

$$w \mapsto \cos(\langle w, x \rangle) \sin(\langle u, x \rangle) x \mathbb{I}_{\|x\| \leq M}$$

is M^2-Lipschitz, hence both $\tilde{g}_N(w)$ and $\mathbb{E}_x[\tilde{g}_N(w)]$ are M^2-Lipschitz. Therefore,

$$\begin{aligned}
\sup_{\|w\| \leq L} \|\tilde{g}_N(w) - \tilde{g}_N(\pi(w))\| &\leq M^2\epsilon, \\
\sup_{\|w\| \leq L} \|\mathbb{E}\left[\tilde{g}_N(w)\right] - \mathbb{E}\left[\tilde{g}_N(\pi(w))\right]\| &\leq M^2\epsilon.
\end{aligned}$$

Summing up, we have

$$\sup_{\|w\| \leq L} \left\| \tilde{g}_N(w) - \mathop{\mathbb{E}}_x \left[\tilde{g}_N(w)\right] \right\| \leq 2M^2\epsilon + 8M\sqrt{\frac{2z}{dN}}$$

with probability $\geq 1 - 2 \cdot 5^d \cdot (1 + 2L/\epsilon)^d \cdot \exp(-z)$.

Set

$$z = 10d\log(L/\epsilon) + t,$$

then

$$\sup_{\|w\| \leq L} \left\| \tilde{g}_N(w) - \mathop{\mathbb{E}}_x \left[\tilde{g}_N(w)\right] \right\| \leq 2M^2\epsilon + 8\sqrt{10}M\sqrt{\frac{d\log(L/\epsilon) + t}{dN}}$$

holds with probability $\geq 1 - \exp(-t)$. Here, we used the constant number as C again.

The bound between the expectations can be derived as

$$\begin{aligned}\sup_{\|w\|\le L}\|\mathbb{E}\left[\tilde{g}_N(w)\right]-g(w)\| &= \sup_{\|w\|\le L}\left\|\mathbb{E}\left[\cos(\langle w,x\rangle)\sin(\langle u,x\rangle)x\mathbb{I}_{\|x\|>M}\right]\right\| \\ &\le \mathbb{E}\left[\|x\|\mathbb{I}_{\|x\|\ge M}\right] \\ &\le \mathbb{E}\left[\|x\|^2\right]^{1/2}\mathbb{E}\left[\mathbb{I}_{\|x\|\ge M}\right]^{1/2} \\ &\le \sqrt{d}\sqrt{P(\|x\|\ge M)} \\ &\le \sqrt{d}\exp(-M^2+d).\end{aligned}$$

Combining these results, we have

$$\sup_{\|w\|\le L}\|\hat{g}_N(w)-g(w)\| \le 2M^2\epsilon + 8\sqrt{10}M\sqrt{\frac{d\log(L/\epsilon)+t}{dN}} + \sqrt{d}\exp(-M^2+d)$$

with probability $\ge 1-\exp(-t)-n\exp(-M^2+d)$.

Now, set $M=\sqrt{d}+\sqrt{t}$, $\epsilon = 1/\sqrt{dN}$, and $t=(\log N)/2$, which gives $-M^2+d\le -t$, so

$$\begin{aligned}&\sup_{\|w\|\le L}\|\hat{g}_N(w)-g(w)\| \\ \le& 6\frac{d+t}{\sqrt{dt}} + 8\sqrt{10}(\sqrt{d}+\sqrt{t})\sqrt{\frac{d\log L + d\log N + d\log d + t}{dN}} + \sqrt{d}\exp(-t) \\ \le& (13+16\sqrt{10})\sqrt{\frac{d\log N + d\log d + d\log L}{N}} \\ \le& 64\sqrt{\frac{d\log N + d\log d + d\log L}{N}}\end{aligned}$$

with probability $\ge 1-1/\sqrt{N}$. ■

Combining Lemma 5 and Lemma 6 shows that with high probability, the one-step gradient has some feature $w^{(1)}$ close to the feature of target function u, which is summarised as following Corollary.

Corollary 7 *Let the error threshold ϵ, and failure probability δ be given. Suppose that the learning rate η_1, the width of network m and the number of training sample N satisfies*

$$\begin{aligned}&\eta_1 \ge \exp\left(\frac{\|u\|^2}{2}\right), \\ &m \ge 2\frac{\log(\delta/2)}{\log\left(1-\frac{\epsilon^2}{12\pi^2\eta_1^2\sqrt{2\log\eta_1}}\right)}, \\ &N \ge \frac{4}{\delta^2}, \\ &N \ge (\sqrt{2}\log\eta_1)^{1/4}, \\ &N \ge \frac{64d\eta_1^2}{\epsilon^2}\left(\log\frac{64d\eta_1^2}{\epsilon^2}\right)^2,\end{aligned}$$

then there exists feature

$$\left\| w_i^{(1)} - u \right\| \leq \epsilon$$

with probability $\geq 1 - \delta$.

Proof One can plug in (3) to L of Lemma 6, then solve each required conditions. ■

Now we proceed by showing that such an approximation of feature will conclude the proof of Theorem 1. To do so, we first show that if there is some feature w that is close enough to the feature of target function u, there is some choice of second-layer weight A^* that has a population risk small.

Lemma 8 (Approximation with learned feature) *Assume that* $\|u\| \leq 2$. *Suppose that there exists* $w^{(i)}$ *such that* $\|w_i - u\| \leq \epsilon$ *with* $\epsilon \leq \frac{1}{10000} \min\left(\|u\|^2, \frac{1}{\|u\|^2}\right)$. *Then, there exists* $A^* \in \mathbb{R}^m$ *with*

$$\frac{1}{2} \cdot \mathbb{E}\left[\left(\sin(\langle u, x\rangle) - f(x; A^*, W)\right)^2\right] \leq 4\sqrt{\epsilon},$$

and $\|A^*\| = \sqrt{\sum_{i=1}^m a_i^2} \leq 3$.

Proof Consider single-neuron case first:

$$\frac{1}{2} \min_a \mathop{\mathbb{E}}_x \left[\left(\sin(\langle u, x\rangle) - a \sin(\langle w, x\rangle)\right)^2\right].$$

Using the analytic expression, we can evaluate the expectation to obtain

$$\begin{aligned}
&\min_a \mathop{\mathbb{E}}_x \left[\left(\sin(\langle u, x\rangle) - a \sin(\langle w, x\rangle)\right)^2\right] \\
&= \exp(-\|u\|^2) \sinh(\|u\|^2) - 2a \exp\left(-\frac{\|u\|^2 + \|w\|^2}{2}\right) \sinh(\langle u, w\rangle) + a^2 \exp(-\|w\|^2) \sinh(\|w\|)^2.
\end{aligned}$$

Since this is a quadratic equation on a, we can directly solve and find minimum value with

$$a^* = \exp\left(\frac{\|w\|^2 - \|u\|^2}{2}\right) \frac{\sinh(\langle u, w\rangle)}{\sinh(\|w\|^2)}.$$

Applying this, we can get

$$\begin{aligned}
&\frac{1}{2} \mathop{\mathbb{E}}_x \left[\left(\sin(\langle u, x\rangle) - a^* \sin(\langle w, x\rangle)\right)^2\right] \\
=&\frac{1}{2} \left|\exp\left(-\|u\|^2\right) \sinh(\|u\|^2) - \exp(-\|w\|^2) \sinh(\|w\|^2)(a^*)^2\right| \\
\leq&\frac{1}{2} \left|\exp\left(-\|u\|^2\right) \sinh(\|u\|^2) - \exp(-\|w\|^2) \sinh(\|w\|^2)\right| \\
&+ \frac{1}{2} \left|\exp(-\|w\|^2) \sinh(\|w\|^2) - \exp(-\|w\|^2) \sinh(\|w\|^2)(a^*)^2\right|.
\end{aligned}$$

For the first term, since $\frac{d}{dt} e^{-t} \sinh(t) = e^{-2t} \leq 1$ for all $t \geq 0$, we can upper bound as

$$\left|\exp\left(-\|u\|^2\right) \sinh(\|u\|^2) - \exp(-\|w\|^2) \sinh(\|w\|^2)\right|$$

$$
\begin{aligned}
&= \left| \int_{\|u\|^2}^{\|w\|^2} \frac{d}{dt} e^{-t} \sinh(t) dt \right| \\
&\leq \left| \int_{\|u\|^2}^{\|w\|^2} \left| \frac{d}{dt} e^{-t} \sinh(t) \right| dt \right| \\
&\leq \left| \|u\|^2 - \|w\|^2 \right| \\
&= (\|u\| + \|w\|)\epsilon. \qquad (5)
\end{aligned}
$$

Similarly, we can use $e^{-t}\sinh(t) \leq 1/2$ for all $t \geq 0$ which holds since it is monotonically increasing and its limit at $t \to \infty$ is 1/2,

$$
\begin{aligned}
&\left| \exp(-\|w\|^2)\sinh(\|w\|^2) - \exp(-\|w\|^2)\sinh(\|w\|^2)(a^*)^2 \right| \\
\leq &\frac{1}{2}\left|1 - (a^*)^2\right| \\
\leq &\frac{1}{2}|1 - a^*| \cdot |1 + a^*|.
\end{aligned}
$$

We first focus on the term $|1 - a^*|$. To upper bound the difference, we can compute as

$$
\begin{aligned}
&|1 - a^*| \\
= &\left| 1 - \exp\left(\frac{\|w\|^2 - \|u\|^2}{2} \right) \frac{\sinh(\langle u, w \rangle)}{\sinh(\|w\|^2)} \right| \\
&\leq \left| \exp\left(\frac{\|w\|^2 - \|u\|^2}{2} \right) - 1 \right| + \exp\left(\frac{\|w\|^2 - \|u\|^2}{2} \right) \left| \frac{\sinh(\langle u, w \rangle)}{\sinh(\|w\|^2)} - 1 \right| \\
&\leq \left| \exp\left(\|u\|\epsilon + \frac{\epsilon^2}{2} \right) - 1 \right| + \exp\left(\|u\|\epsilon + \frac{\epsilon^2}{2} \right) \left| \frac{\sinh(\langle u, w \rangle)}{\sinh(\|w\|^2)} - 1 \right|
\end{aligned}
$$

From the assumption $\epsilon \leq 1/(2\|u\|)$ and $\epsilon \leq 1$, we have

$$
\begin{aligned}
\left| \exp\left(\|u\|\epsilon + \frac{\epsilon^2}{2} \right) - 1 \right| &\leq 2\|u\|\epsilon + \epsilon^2, \\
\exp\left(\|u\|\epsilon + \frac{\epsilon^2}{2} \right) &\leq 2,
\end{aligned}
$$

with numerical inequality $e^t - 1 \leq 2t$ for $0 \leq t \leq 1$. Applying these we get

$$
|\exp(\|u\|\epsilon + \epsilon^2/2) - 1| + \exp\left(\|u\|\epsilon + \frac{\epsilon^2}{2} \right) \left| 1 - \frac{\sinh(\langle u, w \rangle)}{\sinh(\|w\|^2)} \right| \leq 2\|u\|\epsilon + \epsilon^2 + 2\left| 1 - \frac{\sinh(\langle u, w \rangle)}{\sinh(\|w\|^2)} \right|.
$$

We will upper bound the difference between the ratio and 1 on two side, depending on its sign.

When $\sinh(\langle u, w \rangle) \geq \sinh(\|w\|^2)$, we can upper bound the difference as

$$
\begin{aligned}
\frac{\sinh(\langle u, w \rangle)}{\sinh(\|w\|^2)} - 1 &\leq \frac{\sinh(\|u\|^2 + \|u\|\epsilon)}{\sinh((\|u\| - \epsilon)^2)} - 1 \\
&\leq \frac{\sinh(\|u\|^2 + \|u\|\epsilon)}{\sinh(\|u\|^2 - 2\|u\|\epsilon)} - 1,
\end{aligned}
$$

Here, using that $f(x) = 50k - \big(\sinh(x^2(1+k))/\sinh(x^2(1-2k)) - 1\big)$ is decreasing function on $0 \le x \le 2$ and $k \le 0.001$, and also using $f(4) = f(2^2) > 0$, assuming $\epsilon \le \frac{\|u\|}{1000}$ gives

$$\frac{\sinh(\langle u, w\rangle)}{\sinh(\|w\|^2)} - 1 \le 50\frac{\epsilon}{\|u\|},$$

for $\|u\| \le 2$. On the other hand, when $\sinh(\langle u, w\rangle) < \sinh(\|w\|^2)$, we can upper bound the difference as

$$\begin{aligned} 1 - \frac{\sinh(\langle u, w\rangle)}{\sinh(\|w\|^2)} &\le 1 - \frac{\sinh(\|u\|^2 - \|u\|\epsilon)}{\sinh((\|u\| + \epsilon)^2)} \\ &\le 1 - \frac{\sinh(\|u\|^2 - \|u\|\epsilon)}{\sinh(\|u\|^2 + 2\|u\|\epsilon)}, \end{aligned}$$

and similarly using that $f(x) = 50k - \big(1 - \sinh(x^2(1-k))/\sinh(x^2(1+2k))\big)$ is decreasing function on $0 \le x \le 2$ and $k \le 0.001$, and also using $f(4) = f(2^2) > 0$, assuming $\epsilon \le \frac{\|u\|}{1000}$ gives

$$\frac{\sinh(\langle u, w\rangle)}{\sinh(\|w\|^2)} - 1 \le 50\frac{\epsilon}{\|u\|},$$

for $\|u\| \le 2$. Thus, we can conclude

$$\left|1 - \frac{\sinh(\langle u, w\rangle)}{\sinh(\|w\|^2)}\right| \le 50\frac{\epsilon}{\|u\|},$$

for $\|u\| \le 2$ and $\epsilon \le \frac{\|u\|}{1000}$. Applying these, we get

$$|1 - a^*| \le 2\|u\|\epsilon + \epsilon^2 + \frac{100\epsilon}{\|u\|}.$$

Assuming $\epsilon \le \frac{1}{10000}\min\left(\frac{1}{\|u\|^2}, \|u\|^2\right)$, by using $\sqrt{\epsilon} \le \frac{\|u\|}{100}$ and $\sqrt{\epsilon} \le \frac{1}{100\|u\|}$, we get

$$\begin{aligned} |1 - a^*| &\le 2|u\|\frac{1}{100\|u\|}\sqrt{\epsilon} + \epsilon^2 + \frac{100}{\|u\|}\frac{\|u\|}{100}\sqrt{\epsilon} \\ &= \frac{\sqrt{\epsilon}}{50} + \epsilon^2 + \sqrt{\epsilon} \le 3\sqrt{\epsilon}, \end{aligned}$$

for enough small ϵ. Assuming $\epsilon < 1/9$, we get $a^* \le 3$ and $|1 + a^*| \le 4$. Applying these, we get

$$|\exp(\|u\|\epsilon + \epsilon^2/2) - 1| + \exp\left(\|u\|\epsilon + \frac{\epsilon^2}{2}\right)\left|1 - \frac{\sinh(\langle u, w\rangle)}{\sinh(\|w\|^2)}\right| \le \frac{4}{2}3\sqrt{\epsilon} = 6\sqrt{\epsilon}. \tag{6}$$

by using (5), (6) we get

$$\begin{aligned} \frac{1}{2}\,\mathbb{E}\left[(\sin(\langle u, x\rangle) - f(x; A^*, W))^2\right] &\le \frac{(\|u\| + \|w\|)\epsilon}{2} + 3\sqrt{\epsilon} \\ &\le \frac{\sqrt{\epsilon}}{100} + \frac{\epsilon^2}{2} + 3\sqrt{\epsilon} \end{aligned}$$

$$\leq 4\sqrt{\epsilon},$$

with using $\sqrt{\epsilon} \leq \frac{1}{100\|u\|}$. ■

What this Lemma shows is that there exists some weight A^* that achieves small norm and small population risk. We will now relate this crafted weight with the optimised weight $A^{(T)}$. To do so, we will rely on Rademacher complexity to show that the population risk and risk on the training data are close enough.

Lemma 9 (Rademacher Complexity of Simple Neural Networks) *Let $\mathcal{F}$ be the neural network class of bounded first and second-layer weights, with 1-Lipschitz activation σ:*

$$\mathcal{F} = \left\{ x \mapsto \sum_{j=1}^{m} a_j \sigma(\langle x, w_j \rangle) : \|a\|_2 \leq B_a, \|w_j\|_2 \leq B_w, \|a\|_1 \leq B'_a \right\}.$$

Then, the Rademacher complexity w.r.t. standard Gaussian density satisfies

$$\mathfrak{R}_N(\mathcal{F}) \leq 2B_a B_w \sqrt{\frac{md}{N}}.$$

Proof

$$\begin{aligned}
\mathfrak{R}_N(\mathcal{F}) &= \mathop{\mathbb{E}}_{x,\sigma}\left[\sup_{\|a\|_2 \leq B_a, \|w_j\| \leq B_w}\left[\frac{1}{N}\sum_{i=1}^{N}\sigma_i \sum_{j=1}^{m} a_j \sigma(\langle x_i, w_j\rangle + b_j)\right]\right] \\
&= \frac{B_a}{N}\mathop{\mathbb{E}}_{x,\sigma}\left[\sup_{\|w_j\| \leq B_w}\left[\sqrt{\sum_{j=1}^{m}\left(\sum_{i=1}^{N}\sigma_i \sigma(\langle x_i, w_j\rangle + b_j)\right)^2}\right]\right] \\
&\leq \frac{B_a\sqrt{m}}{N}\mathop{\mathbb{E}}_{x,\sigma}\left[\sup_{\|w\| \leq B_w}\left[\left|\sum_{i=1}^{N}\sigma_i \sigma(\langle x_i, w_j\rangle + b_j)\right|\right]\right] \\
&\leq \frac{B_a\sqrt{m}}{N}\mathop{\mathbb{E}}_{x,\sigma}\left[\left(\sup_{\|w\| \leq B_w}\left[\sum_{i=1}^{N}\sigma_i \sigma(\langle x_i, w_j\rangle + b_j)\right] + \sup_{\|w\| \leq B_w}\left[\sum_{i=1}^{N} -\sigma_i \sigma(\langle x_i, w_j\rangle + b_j)\right]\right)\right] \\
&\leq \frac{2B_a\sqrt{m}}{N}\mathop{\mathbb{E}}_{x,\sigma}\left[\sup_{\|w\| \leq B_w}\left[\sum_{i=1}^{N}\sigma_i \sigma(\langle x_i, w\rangle + b_j)\right]\right] \\
&\leq \frac{2B_a\sqrt{m}}{N}\mathop{\mathbb{E}}_{x,\sigma}\left[\sup_{\|w\| \leq B_w}\left[\sum_{i=1}^{N}\sigma_i (\langle x_i, w\rangle + b_j)\right]\right] \\
&= \frac{2B_a\sqrt{m}}{N}\mathop{\mathbb{E}}_{x,\sigma}\left[\sup_{\|w\| \leq B_w}\left[\sum_{i=1}^{N}\sigma_i \langle x_i, w\rangle\right]\right] \\
&= \frac{2B_a\sqrt{m}}{N}\mathop{\mathbb{E}}_{x,\sigma}\left[\sup_{\|w\| \leq B_w}\left[\left\langle w_j, \sum_{i=1}^{N}\sigma_i x_i\right\rangle\right]\right]
\end{aligned}$$

$$= \frac{2B_a B_w \sqrt{m}}{N} \mathop{\mathbb{E}}_{x,\sigma} \left[\left\| \sum_{i=1}^{N} \sigma_i x_i \right\| \right]$$
$$\leq 2B_a B_w \sqrt{\frac{md}{N}}.$$

■

Corollary 10 *For any $\delta > 0$, with probability at least $1 - \delta$ over a sample S of size N, each of the following inequalities holds for all network $f \in \mathcal{F}$:*

$$\frac{1}{2} \cdot \mathop{\mathbb{E}}_{x} \left[(f(x) - f^*(x))^2 \right] \leq \frac{1}{2N} \sum_{i=1}^{N} (f(x_i) - f^*(x_i))^2 + 2(B'_a + 1)\mathfrak{R}_n(\mathcal{F}) + \frac{1}{2}(B'_a + 1)^2 \sqrt{\frac{\log 1/\delta}{2N}},$$
$$\frac{1}{2N} \sum_{i=1}^{N} (f(x_i) - f^*(x_i))^2 \leq \frac{1}{2} \cdot \mathop{\mathbb{E}}_{x} \left[(f(x) - f^*(x))^2 \right] + 2(B'_a + 1)\mathfrak{R}_n(\mathcal{F}) + \frac{1}{2}(B'_a + 1)^2 \sqrt{\frac{\log 1/\delta}{2N}}.$$

Proof This is direct application of Theorem 10.2 of Mohri et al. [12] with the fact that $|f| \leq B'_a$ and $|f^*| \leq 1$. ■

Using this result, we can relate the population risk and empirical risk.

Lemma 11 (Error Bound of Training the Second-Layer) *Suppose that the feature of target function u has its norm bounded, $\|u\| \leq 2$. If the weights $w_i^{(1)}$ satisfy*

$$\min_i \|w_i^{(1)} - u\| \leq \epsilon,$$
$$\max_i \|w_i^{(1)}\| \leq B_W,$$

then there exists some regularisation coefficient λ_2 such that for some large T^*, any time step $T \geq T^*$,

$$\frac{1}{2} \cdot \mathop{\mathbb{E}}_{x \sim \mathcal{N}(0, I_d)} \left[\left(f(x; A^{(1)}, W^{(T)}) - f^*(x) \right)^2 \right]$$
$$\leq 5\sqrt{\epsilon} + (144\sqrt{m} + 72) B_W \sqrt{\frac{md}{N}} + (18m + 4\sqrt{2}) \sqrt{\frac{\log 1/\delta}{N}}.$$

with probability at least $1 - \delta$.

Proof First, Lemma 8 shows that there exists A^* such that

$$\frac{1}{2} \cdot \mathop{\mathbb{E}}_{x \sim \mathcal{N}(0, I_d)} \left[\left(f(x; A^*, W^{(1)}) - f^*(x) \right)^2 \right] \leq 4\sqrt{\epsilon}$$

satisfying

$$w_i^{(1)} \leq B_W, \quad \|A^*\|_2 \leq 3, \quad \|A^*\|_1 \leq 3$$

Now applying Corollary 10, we have

$$\frac{1}{2N}\sum_{i=1}^{N}(f(x_i;W^{(1)},A^*)-f^*(x_i))^2$$
$$\leq \frac{1}{2}\cdot\mathop{\mathbb{E}}_{x}\left[(f(x)-f^*(x))^2\right]+48B_W\sqrt{\frac{md}{N}}+4\sqrt{2}\sqrt{\frac{\log 1/\delta}{N}}.$$

Now since we have A^* satisfying $\|A^*\|_2 \leq 3$, we can view this as instance in the feasible set of following constrained regression problem:

$$\min_{\|A\|_2\leq 3}\frac{1}{2N}\sum_{i=1}^{N}(f(x_i;W^{(1)},A)-f^*(x_i))^2.$$

One can rewrite this constrained regression problem with Lagrange multiplier,

$$\min_{A}\max_{\lambda\geq 0}\frac{1}{2N}\sum_{i=1}^{N}(f(x_i;W^{(1)},A)-f^*(x_i))^2+\lambda(\|A\|^2-3).$$

Since this problem is convex, we can apply minimax duality to have equivalent problem

$$\max_{\lambda\geq 0}\min_{A}\frac{1}{2N}\sum_{i=1}^{N}(f(x_i;W^{(1)},A)-f^*(x_i))^2+\lambda(\|A\|^2-3).$$

The inner problem is now Ridge regression problem with regularisation parameter λ, so by dual attainment, we have some pair $(\lambda^*, A_{\lambda^*})$ that achieves optimum on both problem:

$$\frac{1}{2N}\sum_{i=1}^{N}\left(f(x_i;W^{(1)},A_{\lambda^*})-f^*(x_i)\right)^2=\frac{1}{2N}\sum_{i=1}^{N}\left(f(x_i;W^{(1)},A^*)-f^*(x_i)\right)^2$$
$$\leq 4\sqrt{\epsilon}+48B_W\sqrt{\frac{md}{N}}+4\sqrt{2}\sqrt{\frac{\log 1/\delta}{N}}.$$

Since the Ridge regression problem is strongly convex, the gradient descent for some large enough steps $T \geq T_0$ with sufficiently small learning rate η_2 will give solution $A^{(T)}$ such that

$$\frac{1}{2N}\sum_{i=1}^{N}(f(x_i;W^{(1)},A^{(T)})-f^*(x_i))^2 \leq 5\sqrt{\epsilon}+48B_W\sqrt{\frac{md}{N}}+4\sqrt{2}\sqrt{\frac{\log 1/\delta}{N}},$$
$$\|A^{(T)}\|_2 \leq 6,$$
$$\|A^{(T)}\|_1 \leq 6\sqrt{m}.$$

Applying Corollary 10 back, we have

$$\frac{1}{2}\cdot\mathop{\mathbb{E}}_{x}\left[\left(f(x;W^{(1)},A^{(T)})-f^*(x)\right)^2\right]$$
$$\leq 5\sqrt{\epsilon}+(144\sqrt{m}+72)B_W\sqrt{\frac{md}{N}}+(18m+4\sqrt{2})\sqrt{\frac{\log 1/\delta}{N}}.$$

■

We are now ready to give detailed form of Theorem 1.

Theorem 12 *Suppose that the feature of target vector u has its norm bounded, $\|u\| \leq 2$. Let $\epsilon > 0$ be the error threshold, and $\delta > 0$ the failure probability. If the number of training sample n, the network width m, and the first-layer learning rate η_1 satisfies following conditions,*

$$\begin{aligned}
\eta_1 &\geq \exp\left(\frac{\|u\|^2}{2}\right), \\
m &\geq 2\frac{\log(\delta/4)}{\log\left(1 - \frac{\epsilon^4/15^4}{12\pi^2\eta_1^2\sqrt{2\log\eta_1}}\right)}, \\
N &\geq \frac{16}{\delta^2}, \\
N &\geq (\sqrt{2}\log\eta_1)^{1/4}, \\
N &\geq \frac{64\cdot 15^2 d\eta_1^2}{\epsilon^4}\left(\log\frac{64\cdot 15^2 d\eta_1^2}{\epsilon^4}\right)^2, \\
\epsilon &\leq 2\cdot 15^2\cdot\exp\left(-\frac{\|u\|^2}{2}\right)\|u\|, \\
N &\geq \left(\frac{2^3\cdot 3^4\cdot\eta_1^2 m}{\epsilon}\right)^2 d, \\
N &\geq \frac{(72m)^2\log(1/\delta)}{\epsilon^2}
\end{aligned}$$

then for any $\eta_2 \leq c$ for some sufficiently small $c > 0$, there exists λ_2 such that for some large T^*, any time step $T \geq T^*$ satisfies

$$\frac{1}{2}\cdot\mathop{\mathbb{E}}_{x}\left[\left(f(x; W^{(1)}, A^{(T)}) - f^*(x)\right)^2\right] \leq \epsilon$$

with probability at least $1-\delta$.

Proof We will set N, η_1, and m to satisfy assumptions of Corollary 7 and Lemma 11.

Let $\epsilon' = \epsilon^2/15^2$, and plug this to Corollary 7, which assures that the first term, $5\sqrt{\epsilon'}$ of Lemma 11 is less than $\epsilon/3$. The rest of probability conditions are satisfied by first five assumptions on the statement of this Theorem.

To apply the result of Lemma 11, we should upper bound B_W. We can upper bound $\|w_i^{(1)}\|$ by relating with the expected value of gradient $g(w_i^{(0)})$

$$\begin{aligned}
\|w_i^{(1)}\| &\leq \|\eta g(w_i^{(0)})\| + \eta\|g(w_i^{(0)}) - \hat{g}_N(w_i^{(0)})\| \\
&\leq \eta_1\exp\left(-\frac{\|u\|^2}{2}\right)\|u\| + \eta_1\frac{\epsilon^2}{2\cdot 15^2} \\
&\leq 2\eta_1\exp\left(-\frac{\|u\|^2}{2}\right)\|u\| \\
&\leq 2\eta_1,
\end{aligned}$$

where we used the intermediate result of Lemma 6 used in the proof of Corollary 7, and the assumption on ϵ and η_1. Using this, last two assumptions on N ensures that the rest of two terms in Lemma 11 is also less than $\epsilon/3$. ■

Appendix D. Conjecture on Random Feature Networks

In this section, we list several supporting claims the conjecture is probably true.

Conjecture 13 *Consider the setting of Theorem 1. Assume that we omit the first update step of the first-layer weights in Algorithm 1, and run the algorithm. If there exists some T_0 such that for all $T \geq T_0$, the parameters $(A^{(T)}, W^{(0)})$ satisfy*

$$\frac{1}{2} \cdot \mathbb{E}_{x\sim\mathcal{N}(0,I_d)} \left[\left(f^*(x) - f(x; A^{(T)}, W^{(0)}) \right)^2 \right] \leq \epsilon,$$

with probability at least $1-\delta$, then we must have $m = \Omega(\exp(d))$ or $N = \Omega(\exp(d))$.

Lemma 8 argues that if there exists a learned feature $w_i^{(1)}$ close to the feature of target function u, we can construct the second-layer weight A^* to have a small risk. The natural idea is that this is a sufficient and necessary condition. We give some intuitive derivation as to why this is reasonable.

Lemma 14 (Decomposition of risk) *The risk (2) admits following decomposition:*

$$\begin{aligned}
&\frac{1}{2} \cdot \mathbb{E}_{x\sim\mathcal{N}(0,I_d)} \left[\left(f^*(x) - f(x; A^{(T)}, W^{(0)}) \right)^2 \right] \leq \epsilon \\
=&\frac{1}{2} \cdot \mathop{\mathbb{E}}_{x_\perp} \left[\mathop{\mathbb{E}}_{g} \left[\left(\sin(g) - \sum_{i=1}^m a_i \sin(g\langle u, w_i\rangle + \langle x_\perp, w_i\rangle) \right)^2 \right] \right] \\
=&\underbrace{\frac{1}{2} \cdot \left(\mathop{\mathbb{E}}_{g} \left[\sin(g) - \sum_{i=1}^m a_i \mathop{\mathbb{E}}_{x_\perp} \left[\sin(g\langle u, w_i\rangle + \langle x_\perp, w_i\rangle) \right] \right] \right)^2}_{\textit{Bias}^2} \\
&+ \underbrace{\frac{1}{2} \cdot \mathop{\mathbb{E}}_{g} \left[\mathrm{Var}_{x_\perp} \left(\sum_{i=1}^m a_i \sin(g\langle u, w_i\rangle + \langle x_\perp, w_i\rangle) \right) \right]}_{\textit{Variance}}.
\end{aligned}$$

Proof We prove this by formula similar to the bias-variance decomposition.

We first rewrite the expectation as nested expectation,

$$\mathbb{E}\left[\left(\sin(\langle u, x\rangle) - f(x; \{(w_i, a_i)\}_{i=1}^m) \right)^2 \right] = \mathop{\mathbb{E}}_{x_\perp} \left[\mathop{\mathbb{E}}_{x_u} \left[\left(\sin(\langle u, x\rangle) - f(x; \{(w_i, a_i)\}_{i=1}^m) \right)^2 \right] \right]$$

where

$$x_u = \langle x, u\rangle u, \quad x_\perp = x - x_u.$$

Since $\langle x, u\rangle \overset{d}{=} \mathcal{N}(0,1)$, we can rewrite as

$$\begin{aligned}
&\mathbb{E}\left[\left(\sin(\langle u, x\rangle) - f(x; \{(w_i, a_i)\}_{i=1}^m) \right)^2 \right] \\
=& \mathop{\mathbb{E}}_{x_\perp} \left[\mathop{\mathbb{E}}_{g} \left[\left(\sin(g) - \sum_{i=1}^m a_i \sin(g\langle u, w_i\rangle + \langle x_\perp, w_i\rangle) \right)^2 \right] \right]
\end{aligned}$$

with $g \sim \mathcal{N}(0,1)$.

Now if we see this equation, the $x_\perp$ term correspond to the randomness of prediction function $\sum_{i=1}^m a_i \sin(g(\langle u, w_i\rangle + \langle x_\perp, w_i\rangle))$, and g can be viewed as the input x where we measure the MSE loss of prediction function. So, by bias-variance decomposition, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_\perp}\left[\mathop{\mathbb{E}}_{g}\left[\left(\sin(g) - \sum_{i=1}^m a_i \sin(g\langle u, w_i\rangle + \langle x_\perp, w_i\rangle)\right)^2\right]\right] \\
&= \underbrace{\left(\mathop{\mathbb{E}}_{g}\left[\sin(g) - \sum_{i=1}^m a_i \mathop{\mathbb{E}}_{x_\perp}\left[\sin(g\langle u, w_i\rangle + \langle x_\perp, w_i\rangle)\right]\right]\right)^2}_{\text{Bias}^2} \\
&+ \underbrace{\mathop{\mathbb{E}}_{g}\left[\mathrm{Var}_{x_\perp}\left(\sum_{i=1}^m a_i \sin(g\langle u, w_i\rangle + \langle x_\perp, w_i\rangle)\right)\right]}_{\text{Variance}}.
\end{aligned}$$

■

Intuitively, we view this prediction as 1D regression in the direction of u, and consider the rest of the direction as a random predictor. Now, for some A to have a small error, we need a small variance term, which requires $\langle x_\perp, w_i\rangle$ to be small. One can show that the variance of this inner product vanishes only when w_i and u are parallel. However if we assume that $\|w_i - u\| \geq \epsilon$ for any i, we need $w_i = u(1+C)$ for some $C \geq \epsilon$ or $C \leq -\epsilon$, which makes it impossible to decrease the bias term.

Under this intuition, it is enough to show that random feature requires an exponentially wide network. This is illustrated in the following Lemma.

Lemma 15 (Random feature is far from true feature) *Fix an unit vector $u \in S^{d-1}$, the failure probability δ, the error constant ϵ. For the weights $w_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d/d)$, if $N \leq \frac{\sqrt{\pi}\delta}{\epsilon^{d-1}}$,*

$$\min_{i\in[N]} \|w_i - u\| \geq \epsilon$$

with probability $\geq 1-\delta$.

Proof

Suppose that $\|w - u\| \leq \epsilon$. Note that for any $\|w\| \in \mathbb{R}$, This implies that

$$\angle(u, w) \leq \sin\angle(u, w) \leq \epsilon.$$

Using the fact that $w/\|w\|$ is uniformly distributed over S^{d-1}, we can simplify the probability as

$$P(\angle(u,w) \leq \epsilon) = \frac{\text{Area}(\{w \in S^{d-1} : \angle(u,w) \leq \epsilon\})}{\text{Area}(S^{d-1})}.$$

Now we can relate the areas as

$$\frac{1}{2}\text{Area}(S^{d-1}) = \int_0^{\pi/2} \text{Area}(\sin\theta S^{d-2}) d\theta,$$

and

$$\text{Area}(\{w \in S^{d-1} : \angle(u,w) \le \epsilon\}) = \int_0^\epsilon \text{Area}(\sin\theta S^{d-2}) d\theta,$$

where $\sin\theta S^{d-2}$ is the sphere in $(d-1)$-dimension with radius $\sin\theta$.

Since the area of hypersphere is proportional to $(d-2)$-power of radius, we have

$$P(\angle(u,w) \le \epsilon) = \frac{\int_0^\epsilon \sin^{d-2}\theta d\theta}{2\int_0^{\pi/2}\sin^{d-2}\theta d\theta} \le \frac{\int_0^\epsilon \theta^{d-2} d\theta}{2\int_0^{\pi/2}\sin^{d-2}\theta d\theta}$$

using the approximation $\sin t \le t$ for $t \ge 0$.

The denominator can be approximated as

$$2\int_0^{\pi/2}\sin^{d-2}\theta d\theta = \sqrt{\pi}\frac{\Gamma((d-1)/2)}{\Gamma(d/2)} \ge \sqrt{\pi}\frac{\Gamma(d/2-1)}{\Gamma(d/2)} = \frac{\sqrt{\pi}}{d/2-1} \ge \frac{2\sqrt{\pi}}{d}$$

and the numerator can be computed as

$$\int_0^\epsilon \theta^{d-2} d\theta = \frac{1}{d-1}\epsilon^{d-1},$$

so the probability is upper bounded by

$$P(\angle(u,w) \le \epsilon) \le \frac{d}{d-1}\frac{1}{2\sqrt{\pi}}\epsilon^{d-1} \le \frac{1}{\sqrt{\pi}}\epsilon^{d-1}.$$

So we have

$$P(\|w-u\| \le \epsilon) \le \frac{1}{\sqrt{\pi}}\epsilon^{d-1}.$$

Since each draw of w_i are independent, we have

$$P(\min_i \|w_i - u\|^2 \ge \epsilon) \ge \left(1 - \frac{1}{\sqrt{\pi}}\epsilon^{d-1}\right)^N.$$

Expanding this power, by writing $t = \frac{1}{\sqrt{\pi}}\epsilon^{d-1}$,

$$\begin{aligned}(1-t)^N &= \sum_{i=0}^N (-t)^i \binom{N}{i} \\ &= 1 - Nt + \sum_{i=2}^N (-t)^i \binom{N}{i} \\ &\ge 1 - Nt\end{aligned}$$

assuming

$$t \le \frac{\binom{N}{i}}{\binom{N}{i+1}} = \frac{i+1}{N-i}$$

for $i \ge 2$.

This holds if $Nt \le 1$, and if it is not the case, the lower bound $1 - Nt$ is trivial. ■

Now if we consider the second-layer optimisation, one can think of it as linear regression where the input dimension is $\exp(d)$, and standard results show that gradient descent requires $\exp(d)$ number of samples to learn the target function.

We also note that several related works prove similar impossibility results. Yehudai and Shamir [18] have shown that the random feature model can not model even the single ReLU neuron without the exponentially large number of neurons. While this result does not directly relate to our result, they show that the periodic triangle wave function ψ has output having nearly zero correlation with any other function, i.e., $\mathbb{E}_{w\sim S^{d-1}}\left[f(x)^2\psi(\langle w,x\rangle)\right] \to 0$ as $d \to \infty$, i.e., the signal does not give any information on w. Similar results apply to any periodic function, however, the result requires $\|u\| = \Theta_d(d)$, which does not fit to our case, since we let $\|u\| = \Theta_d(1)$. The Gaussian equivalent theorem [5] applies to our result, showing that the random feature model performs the same as the linear predictor in their proportional limit, which gives non-zero population risk as the target function is non-linear. Since the best linear predictor of $\sin(\langle u,x\rangle)$ is $f(x) = 0$, this shows that as long as the number of the sample remains linear on the input dimension, random feature model performs equivalently to the trivial predictor. While this result applies to our result, they do not give exact sample complexity, which we presume to be exponential, and also only work on the proportional limit setting.