Privacy at Interpolation: Precise Analysis for Random and NTK Features

Simone Bombari, Marco Mondelli
Institute of Science and Technology Austria (ISTA)
[firstname].[surname]@ista.ac.at

Abstract

Deep learning models are able to memorize the training set. This makes them vulnerable to recovery attacks, raising privacy concerns to users, and many widespread algorithms such as empirical risk minimization (ERM) do not directly enforce safety guarantees. In this paper, we study the safety of ERM models when the training samples are interpolated (i.e., *at interpolation*) against a family of powerful black-box information retrieval attacks. Our analysis quantifies this safety via two separate terms: *(i)* the model *stability* with respect to individual training samples, and *(ii)* the *feature alignment* between attacker query and original data. While the first term is well established in learning theory and it is connected to the generalization error in classical work, the second one is, to the best of our knowledge, novel. Our key technical result characterizes precisely the feature alignment for the two prototypical settings of random features (RF) and neural tangent kernel (NTK) regression. This proves that privacy strengthens with an increase in generalization capability, unveiling the role of the model and of its activation function. Numerical experiments show an agreement with our theory not only for RF/NTK models, but also for deep neural networks trained on standard datasets (MNIST, CIFAR-10).

1 Introduction

Deep learning models can memorize the training dataset [42], which becomes concerning if sensitive information can be extracted by adversarial users. Thus, a thriving research effort has aimed at addressing this issue, with differential privacy [21, 1] emerging as a safety criterion. Despite numerous improvements and provable privacy guarantees [5], this approach still comes at a significant performance cost [45], creating a difficult trade-off for users and developers. For this reason, many popular applications still rely on *empirical risk minimization* (ERM), with training times long enough to achieve 0 training loss. These settings, however, do not offer any theoretical guarantee for privacy protection, which leads to the following critical questions:

When do ERM-trained models interpolating the data offer privacy guarantees?
How do these guarantees depend on the model design and on its generalization performance?

In this work, we focus on a family of powerful black-box attacks in which the attacker has partial knowledge about a training sample z_1 and aims to recover information about the rest, without access to the model weights. This setting is of particular interest when the training samples contain both public and private information, and it is considered in [12], under the name of *relational privacy*.

Formally, the samples are modeled by two distinct components, *i.e.*, $z \equiv [x, y]$. Given knowledge on y, the attacker aims to retrieve information about x by querying the trained model with the *masked* sample $z^m := [-, y]$, see Figure 2 for an illustration, and Appendix B for a broader practical motivation of these settings. We consider generalized linear models trained with ERM, when the training algorithm completely fits the dataset. It turns out that in this setting the power of the attack can be *exactly* analyzed through two distinct components:

Mathematics of Modern Machine Learning Workshop at NeurIPS 2023.

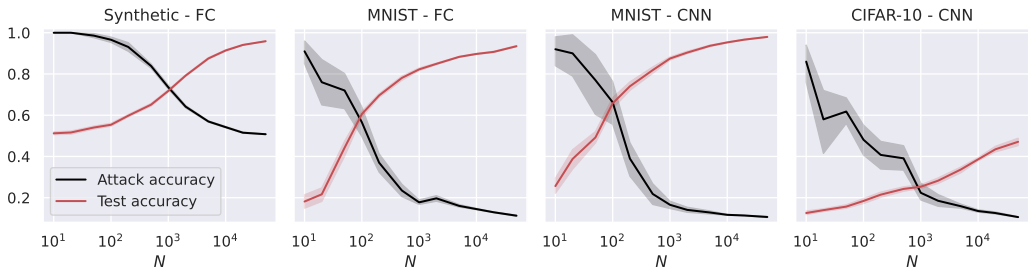

Figure 1: Test and attack accuracies as a function of the number of training samples N, for fully connected (FC, first two plots) and small convolutional neural networks (CNN, last two plots).

1. The *feature alignment* $\mathcal{F}(z^m, z)$, see (6). This captures the similarity in feature space between the training sample z and its masked counterpart z^m, and it depends on the feature map of the model. To the best of our knowledge, this is the first time that attention is raised over such an object.
2. The *stability* $\mathcal{S}_z$ of the model with respect to z, see Definition 2.1. Similar notions of stability are in the seminal work by [15], which draws a connection to generalization.

Our technical contributions can be summarized as follows:

- We connect the stability of generalized linear models to the feature alignment between samples, see Lemma 3.1. Then, we show that this connection makes the privacy of the model a natural consequence of its generalization capability, when $\mathcal{F}(z^m, z)$ can be well approximated by a constant $\gamma > 0$, independent of the original sample z.
- We focus on two settings widely analyzed in the theoretical literature, *i.e.*, *(i)* random features (RF) [39], and *(ii)* the neural tangent kernel (NTK) [29]. Here, under a natural scaling of the models, we prove the concentration of $\mathcal{F}(z^m, z)$ to a positive constant γ, see Theorems 4.2 and 4.3. For the NTK, we obtain a closed-form expression for γ, which connects the power of the attack to the activation function.

We experimentally show that both synthetic and standard datasets (MNIST, CIFAR-10) agree well with the theoretical predictions. Remarkably, we see the same proportionality between generalization and privacy for various neural network architectures (see Figure 1, and its further discussion in Appendix H). This provides empirical evidence of the wide generality of our findings, which appears to go beyond RF/NTK models.

In a nutshell, our results give a precise characterization of how the accuracy of the attack grows with the generalization error, unveiling the role of the model (and, specifically, of its activation). In contrast with the vast body of work relating differential privacy with generalization [20, 19, 8, 44], we focus on the widespread paradigm of empirical risk minimization, when all the training samples are interpolated. In this setting, there is no explicit assumption on algorithmic stability and no a-priori guarantee in terms of privacy, as training is not performed via a differentially private mechanism. Thus, no generalization or privacy bound similar to [15, 20, 27, 7] can be explicitely computed. For a more comprehensive discussion on the related work, we refer to Appendix A.

2 Preliminaries

Notation. Given a vector v, we denote by $\|v\|_2$ its Euclidean norm. Given $v \in \mathbb{R}^{d_v}$ and $u \in \mathbb{R}^{d_u}$, we denote by $v \otimes u \in \mathbb{R}^{d_v d_u}$ their Kronecker product. Given a matrix $A \in \mathbb{R}^{m \times n}$, we denote by $P_A \in \mathbb{R}^{n \times n}$ the projector over $\text{Span}\{\text{rows}(A)\}$. All the complexity notations $\Omega(\cdot)$, $\mathcal{O}(\cdot)$, $o(\cdot)$ and $\Theta(\cdot)$ are understood for sufficiently large data size N, input dimension d, number of neurons k, and number of parameters p. We indicate with $C, c > 0$ numerical constants, independent of N, d, k, p.

Setting. Let (Z, G) be a labelled training dataset, where $Z = [z_1, \dots, z_N]^\top \in \mathbb{R}^{N \times d}$ contains the training data (sampled i.i.d. from a distribution $\mathcal{P}_Z$) on its rows and $G = (g_1, \dots, g_N) \in \mathbb{R}^N$ contains the corresponding labels. We assume the label g_i to be a deterministic function of the sample z_i. Let $\varphi : \mathbb{R}^d \to \mathbb{R}^p$ be a generic feature map, from the input space to a feature space of dimension p. We consider the following *generalized linear model*

$$f(z, \theta) = \varphi(z)^\top \theta, \tag{1}$$

where $\varphi(z) \in \mathbb{R}^p$ is the feature vector associated with the input sample z, and $\theta \in \mathbb{R}^p$ are the trainable parameters of the model. We minimize the empirical risk with a quadratic loss:

$$\min_\theta \left\| \varphi(Z)^\top \theta - G \right\|_2^2 . \tag{2}$$

Here, $\varphi(Z) \in \mathbb{R}^{N\times p}$ is the feature matrix, containing $\varphi(z_i)$ in its i-th row. We use the shorthands $\Phi := \varphi(Z)$ and $K := \Phi\Phi^\top \in \mathbb{R}^{N\times N}$, where K denotes the kernel associated with the feature map. If K is invertible (i.e., the model can fit any set of labels G), gradient descent converges to the interpolator which is the closest in ℓ_2 norm to the initialization [26], i.e.,

$$\theta^* = \theta_0 + \Phi^+(G - f(Z, \theta_0)), \tag{3}$$

where θ^* is the gradient descent solution, θ_0 is the initialization, $f(Z, \theta_0) = \Phi^\top\theta_0$ the output of the model (1) at initialization, and $\Phi^+ := \Phi^\top K^{-1}$ the Moore-Penrose inverse. Let $z \sim \mathcal{P}_Z$ be an independent test sample. Then, we define the *generalization error* of the trained model as $\mathcal{R} = \mathbb{E}_{z\sim\mathcal{P}_Z}\left[(f(z,\theta^*) - g_z)^2\right]$, where g_z denotes the ground-truth label of the test sample z.

Stability. For our discussion, it is convenient to introduce quantities related to "incomplete" datasets. In particular, we indicate with $\Phi_{-1} \in \mathbb{R}^{(N-1)\times p}$ the feature matrix of the training set *without* the first sample z_1. For simplicity, we focus on the removal of the first sample, and similar considerations hold for the removal of any other sample. In other words, Φ_{-1} is equivalent to Φ, without the first row. Similarly, using (3), we indicate with $\theta^*_{-1} := \theta_0 + \Phi^+_{-1}(G_{-1} - f(Z_{-1}, \theta_0))$ the set of parameters the algorithm would have converged to if trained over (Z_{-1}, G_{-1}), the original dataset without the first pair sample-label (z_1, g_1). We can now proceed with the definition of our notion of "stability".

Definition 2.1. *Let θ^* (θ^*_{-1}) be the parameters of the model f given by* (1) *trained on the dataset Z (Z_{-1}), as in* (3)*. We define the* stability $\mathcal{S}_{z_1} : \mathbb{R}^d \to \mathbb{R}$ *with respect to the training sample z_1 as*

$$\mathcal{S}_{z_1} := f(\cdot, \theta^*) - f(\cdot, \theta^*_{-1}). \tag{4}$$

This quantity indicates how the trained model changes if we add z_1 to the dataset Z_{-1}. If the training algorithm completely fits the data (as in (3)), then $\mathcal{S}_{z_1}(z_1) = g_1 - f(z_1, \theta^*_{-1})$, which implies that

$$\mathbb{E}_{z_1\sim\mathcal{P}_Z}\left[\mathcal{S}^2_{z_1}(z_1)\right] = \mathbb{E}_{z_1\sim\mathcal{P}_Z}\left[\left(f(z_1,\theta^*_{-1}) - g_1\right)^2\right] = \mathbb{E}_{z\sim\mathcal{P}_Z}\left[\left(f(z,\theta^*_{-1}) - g_z\right)^2\right] =: \mathcal{R}_{Z_{-1}}, \tag{5}$$

where $\mathcal{R}_{Z_{-1}}$ denotes the generalization error of the algorithm that uses Z_{-1} as training set.

3 Stability, Generalization and Privacy

Stability and feature alignment. Our goal is to quantify how much information about $g(x_1)$ the attacker can recover through a generic query z. To do so, we relate $f(z, \theta^*)$ to the model evaluated on the original sample z_1. It turns out that, for generalized linear regression, under mild conditions on the feature map φ, this can be elegantly done via the notion of *stability* of Definition 2.1.

Lemma 3.1. *Let $\varphi : \mathbb{R}^d \to \mathbb{R}^p$ be a generic feature map, such that the induced kernel $K \in \mathbb{R}^{N\times N}$ on the training set is invertible. Let $z_1 \in \mathbb{R}^d$ be an element of the training dataset Z, and $z \in \mathbb{R}^d$ a generic test sample. Let $P_{\Phi_{-1}}$ be the projector over Span$\{\text{rows}(\Phi_{-1})\}$, and $\mathcal{S}_{z_1}$ be the stability with respect to z_1, as in Definition 2.1. Let us denote by*

$$\mathcal{F}_\varphi(z, z_1) := \frac{\varphi(z)^\top P^\perp_{\Phi_{-1}}\varphi(z_1)}{\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2} \tag{6}$$

the feature alignment *between z and z_1. Then, we have*

$$\mathcal{S}_{z_1}(z) = \mathcal{F}_\varphi(z, z_1)\,\mathcal{S}_{z_1}(z_1). \tag{7}$$

The idea of the argument is to express P_Φ as $P_{\Phi_{-1}}$ plus the projector over the span of $P^\perp_{\Phi_{-1}}\varphi(z_1)$, by leveraging the Gram-Schmidt decomposition of P_Φ. The proof is deferred to Appendix D. In words, Lemma 3.1 relates the stability with respect to z_1 evaluated on z and z_1 through the quantity $\mathcal{F}_\varphi(z, z_1)$, which captures the similarity between z and z_1 in the feature space induced by φ.

Generalization and privacy. Armed with Lemma 3.1, we now characterize the power of the attack query $f(z^m_1, \theta^*)$. Let us replace $\mathcal{F}_\varphi(z^m_1, z_1)$ in (7) with a constant $\gamma_\varphi > 0$ (the concentration for RF/NTK is proved in Section 4), independent from z_1. Then, by using (4), we get

$$f(z^m_1, \theta^*) = f(z^m_1, \theta^*_{-1}) + \gamma_\varphi\,\mathcal{S}_{z_1}(z_1) = f(z^m_1, \theta^*_{-1}) + \gamma_\varphi\left(g_1 - f(z_1, \theta^*_{-1})\right). \tag{8}$$

To quantify the power of the attack, we look at $\text{Cov}\left(f(z^m_1, \theta^*), g_1\right)$, in the probability space of z_1:

$$\text{Cov}\left(f(z^m_1, \theta^*), g_1\right) = \gamma_\varphi \text{Cov}\left(S_{z_1}(z_1), g_1\right) \le \gamma_\varphi\sqrt{\text{Var}\left(S_{z_1}(z_1)\right)\text{Var}\left(g_1\right)} \le \gamma_\varphi\sqrt{\mathcal{R}_{Z_{-1}}}\sqrt{\text{Var}\left(g_1\right)}. \tag{9}$$

Here, the first step uses (8) and the independence between $f(z_1^m, \theta^*_{-1})$ and g_1, the second step is an application of Cauchy-Schwarz, and the last step follows from (5). Let us focus on the RHS of (9). While $\sqrt{\text{Var}\,(g_1)}$ is a simple scaling factor, γ_φ and $\sqrt{\mathcal{R}_{Z_{-1}}}$ lead to an interesting interpretation: we expect the attack to become more powerful as the similarity between z_1^m and z_1 (formalized by $\mathcal{F}_\varphi(z_1^m, z_1)$) increases, and less effective as the generalization error of the model decreases. In fact, the potential threat hinges on the model overfitting the y-component at training time. This overfitting would both cause higher generalization error, and higher chances of recovering g_1 given only y_1.

4 Concentration Results for RF and NTK

Assumption 4.1 (Data distribution)**.** *The input data $(z_1, \dots, z_N)$ are N i.i.d. samples from $\mathcal{P}_Z = \mathcal{P}_X \times \mathcal{P}_Y$, such that $z_i \in \mathbb{R}^d$ can be written as $z_i = [x_i, y_i]$, with $x_i \in \mathbb{R}^{d_x}$, $y_i \in \mathbb{R}^{d_y}$ and $d = d_x + d_y$. We assume that $x_i \sim \mathcal{P}_X$ is independent of $y_i \sim \mathcal{P}_Y$, and the following holds:*

1. *$\|x\|_2 = \sqrt{d_x}$, and $\|y\|_2 = \sqrt{d_y}$, i.e., the data have normalized norm.*
2. *$\mathbb{E}[x] = 0$, and $\mathbb{E}[y] = 0$, i.e., the data are centered.*
3. *Both $\mathcal{P}_X$ and $\mathcal{P}_Y$ satisfy the* Lipschitz concentration property.

RF. The *random features (RF) model* takes the form $f_{\text{RF}}(z, \theta) = \varphi_{\text{RF}}(z)^\top \theta$, where $\varphi_{\text{RF}}(z) = \phi(Vz)$. V is a $k \times d$ matrix s.t. $V_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$, and ϕ is an activation applied component-wise. The number of parameters of this model is k, as V is fixed and $\theta \in \mathbb{R}^k$ contains trainable parameters. We consider the scalings $N \log^3 N = o(k)$, $\sqrt{d} \log d = o(k)$, $k \log^4 k = o\left(d^2\right)$, and ϕ to be L-Lipschitz, and we denote by μ_l its l-th Hermite coefficient.

Theorem 4.2. *Let $x \sim \mathcal{P}_X$ be sampled independently from everything, and $z_1^m = [x, y_1]$. Let $\alpha = d_y/d$ and $\mathcal{F}_{\text{RF}}(z_1^m, z_1)$ be the feature alignment between z_1^m and z_1, as defined in* (6)*. Then,*

$$|\mathcal{F}_{\text{RF}}(z_1^m, z_1) - \gamma_{\text{RF}}| = o(1), \tag{10}$$

with probability at least $1 - \exp(-c \log^2 N)$ over V, Z and x, where c is an absolute constant, and $\gamma_{\text{RF}} \leq 1$ does not depend on z_1 and x. Furthermore, we have

$$\gamma_{\text{RF}} > \frac{\sum_{l=2}^{+\infty} \mu_l^2 \alpha^l}{\sum_{l=1}^{+\infty} \mu_l^2} - o(1), \tag{11}$$

with probability at least $1 - \exp(-c \log^2 N)$ over V, and Z_{-1}, where c is an absolute constant, i.e.*, γ_{RF} is bounded away from 0 with high probability.*

NTK. We consider a linearized 2-layer neural network, with trainable parameters only in its hidden layer. The *NTK regression model* takes the form $f_{\text{NTK}}(z, \theta) = \varphi_{\text{NTK}}(z)^\top \theta$, where $\varphi_{\text{NTK}}(z) = z \otimes \phi'(Wz)$ [14, 32]. W is a $k \times d$ matrix s.t. $W_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$, and ϕ' is applied component-wise. The number of parameters of this model is $p = dk$, as W is fixed and $\theta \in \mathbb{R}^p$ contains trainable parameters. We consider the scalings $N \log^8 N = o(kd)$, $N > d$, $k = \mathcal{O}(d)$, and ϕ' to be L-Lipschitz and not constant, and we denote by μ'_l its l-th Hermite coefficient.

Theorem 4.3. *Let $x \sim \mathcal{P}_X$ be sampled independently from everything, and $z_1^m = [x, y_1]$. Let $\alpha = d_y/d \in (0, 1)$ and $\mathcal{F}_{\text{NTK}}(z_1^m, z_1)$ be the feature alignment between z_1^m and z_1, as defined in* (6)*. Then,*

$$|\mathcal{F}_{\text{NTK}}(z_1^m, z_1) - \gamma_{\text{NTK}}| = o(1), \qquad \textit{where} \quad 0 < \gamma_{\text{NTK}} := \alpha \frac{\sum_{l=1}^{+\infty} {\mu'_l}^2 \alpha^l}{\sum_{l=1}^{+\infty} {\mu'_l}^2} < 1, \tag{12}$$

with probability at least $1 - N \exp(-c \log^2 k) - \exp(-c \log^2 N)$ over Z, x, and W, where c is an absolute constant.

Remarks. The combination of Theorems 4.2/4.3 and (9) connects stability with privacy. In addition, for the NTK model, we are able to express the limit γ_{NTK} of the feature alignment in a closed form involving α and the Hermite coefficients of the derivative of the activation. The findings of both Theorems are clearly displayed in Figures 4 and 5 in Appendix H: as N increases, the test accuracy improves and the reconstruction attack becomes less effective. Furthermore, for the synthetic dataset, while the test accuracy does not depend on α and on the activation function, the success of the attack increases with α and by taking an activation function with dominant low-order Hermite coefficients, as predicted by (12) and suggested by (11).

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM SIGSAC Conference on Computer and Communications Security*, 2016.

[2] Radoslaw Adamczak. A note on the Hanson-Wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20:1–13, 2015.

[3] Ben Adlam and Jeffrey Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning (ICML)*, 2020.

[4] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning (ICML)*, 2019.

[5] Galen Andrew, Om Thakkar, Hugh Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[6] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

[7] Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[8] Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. *SIAM Journal on Computing*, 50(3), 2021.

[9] March Boedihardjo, Thomas Strohmer, and Roman Vershynin. Private measures, random walks, and synthetic data. *arXiv preprint arXiv:2204.09167*, 2022.

[10] March Boedihardjo, Thomas Strohmer, and Roman Vershynin. Private sampling: A noiseless approach for generating differentially private synthetic data. *SIAM Journal on Mathematics of Data Science*, 4(3):1082–1115, 2022.

[11] March Boedihardjo, Thomas Strohmer, and Roman Vershynin. Privacy of synthetic data: A statistical framework. *IEEE Transactions on Information Theory*, 69(1):520–527, 2023.

[12] Simone Bombari, Alessandro Achille, Zijian Wang, Yu-Xiang Wang, Yusheng Xie, Kunwar Yashraj Singh, Srikar Appalaraju, Vijay Mahadevan, and Stefano Soatto. Towards differential relational privacy and its use in question answering. *arXiv preprint arXiv:2203.16701*, 2022.

[13] Simone Bombari, Mohammad Hossein Amani, and Marco Mondelli. Memorization and optimization in deep neural networks with minimum over-parameterization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[14] Simone Bombari, Shayan Kiyani, and Marco Mondelli. Beyond the universal law of robustness: Sharper laws for random features and neural tangent kernels. In *International Conference on Machine Learning (ICML)*, 2023.

[15] Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

[16] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Conference on Security Symposium*, 2019.

[17] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Xiaodong Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *USENIX Conference on Security Symposium*, 2021.

[18] Simon S. Du, Jason D. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

[19] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[20] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *ACM Symposium on Theory of Computing (STOC)*, 2015.

[21] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[22] André Elisseeff and Massimiliano Pontil. Leave-one-out error and stability of learning algorithms with applications stability of randomized learning algorithms source. *International Journal of Systems Science (IJSySc)*, 6, 2002.

[23] Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[24] Vitaly Feldman. Does learning require memorization? A short tale about a long tail. In *ACM Symposium on Theory of Computing (STOC)*, pages 954–959, 2020.

[25] Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[26] Suriya Gunasekar, Blake E. Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[27] Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, 2016.

[28] Yiyun He, Roman Vershynin, and Yizhe Zhu. Algorithmically effective differentially private synthetic data. *arXiv preprint arXiv:2302.05552*, 2023.

[29] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[30] Charles R. Johnson. *Matrix Theory and Applications*. American Mathematical Society, 1990.

[31] 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.

[32] Andrea Montanari and Yiqiao Zhong. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *The Annals of Statistics*, 50(5):2816–2847, 2022.

[33] Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: Stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Adv. Comput. Math.*, 25:161–193, 2006.

[34] Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

[35] Quynh Nguyen and Matthias Hein. Optimization landscape and expressivity of deep CNNs. In *International Conference on Machine Learning (ICML)*, 2018.

[36] Quynh Nguyen and Marco Mondelli. Global convergence of deep networks with one wide layer followed by pyramidal topology. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[37] Quynh Nguyen, Marco Mondelli, and Guido Montufar. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep ReLU networks. In *International Conference on Machine Learning (ICML)*, 2021.

[38] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

[39] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[40] Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, 2021.

[41] Jssai Schur. Bemerkungen zur theorie der beschränkten bilinearformen mit unendlich vielen veränderlichen. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 1911(140):1–28, 1911.

[42] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (SP)*, 2017.

[43] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

[44] Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory (COLT)*, 2020.

[45] Florian Tramer and Dan Boneh. Differentially private learning needs better features (or much more data). In *International Conference on Learning Representations (ICLR)*, 2021.

[46] Joel Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, page 389–434, 2012.

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

[48] Zhichao Wang and Yizhe Zhu. Deformed semicircle law and concentration of nonlinear random matrices for ultra-wide neural networks. *arXiv preprint arXiv:2109.09304*, 2021.

[49] Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. Large scale private learning via low-rank reparametrization. In *International Conference on Machine Learning (ICML)*, 2021.

[50] Zheng Zhao, Shay B. Cohen, and Bonnie Webber. Reducing the frequency of hallucinated quantities in abstractive summaries. In *Findings of the Association for Computational Linguistics (EMNLP)*, 2020.

A Related Work

Private machine learning. Information retrieval via partial knowledge of the data is observed in question answering tasks by [12]. This setting is natural in language models, as they are prone to memorize the training set [17, 16], and to hallucinate it at test time [50, 40]. Differential privacy [21] enables training deep learning models maintaining privacy guarantees. This is achieved through the DPSGD algorithm [1] which, despite improvements [49, 5], still comes at a steep performance cost [1, 45]. To circumvent the problem, a recent line of work [10, 9, 11, 28] utilizes synthetic datasets, analyzing efficient algorithms via tools from high dimensional probability.

Stability. The leave-one-out (error) stability is linked to generalization in [8, 22, 33], and a wide range of variations on this object is discussed in the classical work by [15]. [24] takes a probabilistic viewpoint and shows that, when the data distribution is heavy-tailed, stability might be detrimental for learning. This is also supported empirically by [25]. In contrast, [33] proves that, if the generalization gap vanishes with the number of samples, the learning algorithm has to be leave-one-out stable.

Random Features and Neural Tangent Kernel. Random features (RF) are introduced by [39], and they can be regarded as a two-layer neural network with random first layer weights. This model is theoretically appealing, as it is analytically tractable and offers deep-learning-like behaviours, such as double descent [31]. The neural tangent kernel (NTK) can be regarded as the kernel obtained by linearizing a neural network around the initialization [29]. A popular line of work has analyzed its spectrum [23, 3, 48] and bounded its smallest eigenvalue [43, 37, 32, 13]. The behavior of the NTK is closely related to the memorization [32], optimization [4, 18], generalization [6] and adversarial robustness [14] of deep neural networks.

B Reconstruction Attack

Reconstruction attack. Let the input samples be decomposed in two *independent* components, *i.e.*, $z \equiv [x, y]$. With this notation, we mean that $z \in \mathbb{R}^d$ is the concatenation of $x \in \mathbb{R}^{d_x}$ and $y \in \mathbb{R}^{d_y}$ ($d_x + d_y = d$). Here, x is the part of the input that is useful to accomplish the task (*e.g.*, the cat in top-left image of Figure 2), while y is noise (*e.g.*, the background). Formally, we assume that, for $i \in \{1, \dots, N\}$, $g_i = g(x_i)$, where g is a deterministic labelling function, *i.e.*, the label depends only on x and it is independent of y. In practice, the algorithm may overfit to the noise component, learning the spurious correlations between y_i and the corresponding label g_i. An attacker might then exploit this phenomenon to reconstruct the label g_i, by simply querying the model with the noise component y_i. Without access to the model, this reconstruction would be impossible, as the noise y_i is independent from x_i, and therefore from $g(x_i)$. In our theoretical analysis, we assume the attacker to have access to a *masked* sample $z_i^m(x) = [x, y_i]$, *i.e.*, a version of z_i in which the component x_i is replaced with an independent sample x taken from the same distribution. We do the same in the synthetic setting of the experiments, while for MNIST and CIFAR-10 we just set x to 0, see Figure 2. Our goal is to understand whether the output of the model evaluated on such query, *i.e.*, $f(z_i^m, \theta^*)$, provides information on the ground-truth label $g(x_i)$. As the setting is symmetric with respect to the data ordering, without loss of generality, we assume the attack to be aimed towards the first sample z_1.

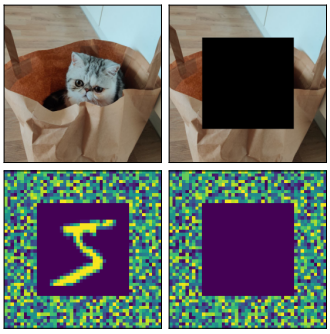

Figure 2: Example of a training sample z (top-left) and its masked counterpart z^m (top-right). In experiments, we add a noise background (y) around the original images (x) before training (bottom-left). The attack consists in querying the trained model only with the noise component (bottom-right).

The setting above in which an attacker tries to recover information on x_i from y_i is a known issue [12], and it is common when the sensitive information is not in the data itself, but rather in the relation among data points. A first motivating example comes from face recognition in computer vision. If the attacker wants to know if a certain individual (x) was at a certain compromising location (y), they could simply plug in the trained model a picture of the location without the individual (the empty shopping bag, in the first example of Figure 2). A second motivating example comes from NLP. Sensitive information (x) about an individual (y) is stored in a textual dataset. The attacker could guess that y was mentioned in the training dataset, and can try to recover x via the prompt *"The address of y is..."*. Similar experiments are performed for question-answering in [12], where the

tokens containing the information which is relevant to solve the task are masked, but the trained model can still hallucinate the correct answer.

C Additional Notations and Remarks

Given a sub-exponential random variable X, let $\|X\|_{\psi_1} = \inf\{t > 0 \ : \ \mathbb{E}[\exp(|X|/t)] \leq 2\}$. Similarly, for a sub-Gaussian random variable, let $\|X\|_{\psi_2} = \inf\{t > 0 \ : \ \mathbb{E}[\exp(X^2/t^2)] \leq 2\}$. We use the analogous definitions for vectors. In particular, let $X \in \mathbb{R}^n$ be a random vector, then $\|X\|_{\psi_2} := \sup_{\|u\|_2=1} \left\|u^\top X\right\|_{\psi_2}$ and $\|X\|_{\psi_1} := \sup_{\|u\|_2=1} \left\|u^\top X\right\|_{\psi_1}$. Notice that if a vector has independent, mean 0, sub-Gaussian (sub-exponential) entries, then it is sub-Gaussian (sub-exponential). This is a direct consequence of Hoeffding's inequality and Bernstein's inequality (see Theorems 2.6.3 and 2.8.2 in [47]).

We say that a random variable or vector respects the Lipschitz concentration property if there exists an absolute constant $c > 0$ such that, for every Lipschitz continuous function $\tau : \mathbb{R}^d \to \mathbb{R}$, we have $\mathbb{E}|\tau(X)| < +\infty$, and for all $t > 0$,

$$\mathbb{P}\left(|\tau(x) - \mathbb{E}_X[\tau(x)]| > t\right) \leq 2e^{-ct^2/\|\tau\|^2_{\text{Lip}}}. \tag{13}$$

When we state that a random variable or vector X is sub-Gaussian (or sub-exponential), we implicitly mean $\|X\|_{\psi_2} = \mathcal{O}(1)$, i.e. it doesn't increase with the scalings of the problem. Notice that, if X is Lipschitz concentrated, then $X - \mathbb{E}[X]$ is sub-Gaussian. If $X \in \mathbb{R}$ is sub-Gaussian and $\tau : \mathbb{R} \to \mathbb{R}$ is Lipschitz, we have that $\tau(X)$ is sub-Gaussian as well. Also, if a random variable is sub-Gaussian or sub-exponential, its p-th momentum is upper bounded by a constant (that might depend on p).

In general, we indicate with C and c absolute, strictly positive, numerical constants, that do not depend on the scalings of the problem, i.e. input dimension, number of neurons, or number of training samples. Their value may change from line to line.

Given a matrix A, we indicate with $A_{i:}$ its i-th row, and with $A_{:j}$ its j-th column. Given a square matrix A, we denote by $\lambda_{\min}(A)$ its smallest eigenvalue. Given a matrix A, we indicate with $\sigma_{\min}(A) = \sqrt{\lambda_{\min}(A^\top A)}$ its smallest singular value, with $\|A\|_{\text{op}}$ its operator norm (and largest singular value), and with $\|A\|_F$ its Frobenius norm ($\|A\|_F^2 = \sum_{ij} A_{ij}^2$).

Given two matrices $A, B \in \mathbb{R}^{m\times n}$, we denote by $A \circ B$ their Hadamard product, and by $A * B = [(A_{1:} \otimes B_{1:}), \dots, (A_{m:} \otimes B_{m:})]^\top \in \mathbb{R}^{m\times n^2}$ their row-wise Kronecker product (also known as Khatri-Rao product). We denote $A^{*2} = A * A$. We remark that $(A * B)(A * B)^\top = AA^\top \circ BB^\top$. We say that a matrix $A \in \mathbb{R}^{n\times n}$ is positive semi definite (p.s.d.) if it's symmetric and for every vector $v \in \mathbb{R}^n$ we have $v^\top Av \geq 0$.

C.1 Hermite Polynomials

In this subsection, we refresh standard notions on the Hermite polynomials. For a more comprehensive discussion, we refer to [38]. The (probabilist's) Hermite polynomials $\{h_j\}_{j\in\mathbb{N}}$ are an orthonormal basis for $L^2(\mathbb{R}, \gamma)$, where γ denotes the standard Gaussian measure. The following result holds.

Proposition C.1 (Proposition 11.31, [38])**.** *Let ρ_1, ρ_2 be two standard Gaussian random variables, with correlation $\rho \in [-1, 1]$. Then,*

$$\mathbb{E}_{\rho_1,\rho_2}[h_i(\rho_1)h_j(\rho_2)] = \delta_{ij}\rho^i, \tag{14}$$

where $\delta_{ij} = 1$ if $i = j$, and 0 otherwise.

The first 5 Hermite polynomials are

$$h_0(\rho) = 1, \quad h_1(\rho) = \rho, \quad h_2(\rho) = \frac{\rho^2 - 1}{\sqrt{2}}, \quad h_3(\rho) = \frac{\rho^3 - 3\rho}{\sqrt{6}}, \quad h_4(\rho) = \frac{\rho^4 - 6\rho^2 + 3}{\sqrt{24}}. \tag{15}$$

Proposition C.2 (Definition 11.34, [38])**.** *Every function $\phi \in L^2(\mathbb{R}, \gamma)$ is uniquely expressible as*

$$\phi(\rho) = \sum_{i\in\mathbb{N}} \mu_i^\phi h_i(\rho), \tag{16}$$

where the real numbers μ_i^ϕ's are called the Hermite coefficients of ϕ, and the convergence is in $L^2(\mathbb{R},\gamma)$. More specifically,

$$\lim_{n\to+\infty}\left\|\left(\sum_{i=0}^{n}\mu_i^\phi h_i(\rho)\right)-\phi(\rho)\right\|_{L^2(\mathbb{R},\gamma)}=0. \tag{17}$$

This readily implies the following result.

Proposition C.3. *Let ρ_1,ρ_2 be two standard Gaussian random variables with correlation $\rho\in[-1,1]$, and let $\phi,\tau\in L^2(\mathbb{R},\gamma)$. Then,*

$$\mathbb{E}_{\rho_1,\rho_2}\left[\phi(\rho_1)\tau(\rho_2)\right]=\sum_{i\in\mathbb{N}}\mu_i^\phi\mu_i^\tau\rho^i. \tag{18}$$

D Proof of Lemma 3.1

We start by refreshing some useful notions of linear algebra. Let $A\in\mathbb{R}^{N\times p}$ be a matrix, with $p\geq N$, and $A_{-1}\in\mathbb{R}^{(N-1)\times p}$ be obtained from A after removing the first row. We assume $AA^\top$ to be invertible, *i.e.,* the rows of A are linearly independent. Thus, also the rows of A_{-1} are linearly independent, implying that $A_{-1}A_{-1}^\top$ is invertible as well. We indicate with $P_A\in\mathbb{R}^{p\times p}$ the projector over Span$\{\text{rows}(A)\}$, and we correspondingly define $P_{A_{-1}}\in\mathbb{R}^{p\times p}$. As $AA^\top$ is invertible, we have that rank$(A)=N$.

By singular value decomposition, we have $A=UDO^\top$, where $U\in\mathbb{R}^{N\times N}$ and $O\in\mathbb{R}^{p\times p}$ are orthogonal matrices, and $D\in\mathbb{R}^{N\times p}$ contains the (all strictly positive) singular values of A in its "left" diagonal, and is 0 in every other entry. Let us define $O_1\in\mathbb{R}^{N\times p}$ as the matrix containing the first N rows of O. This notation implies that if $O_1u=0$ for $u\in\mathbb{R}^p$, then $Au=0$, *i.e.*, $u\in$ Span$\{\text{rows}(A)\}^\perp$. The opposite implication is also true, which implies that Span$\{\text{rows}(A)\}=$ Span$\{\text{rows}(O_1)\}$. As the rows of O_1 are orthogonal, we can then write

$$P_A=O_1^\top O_1. \tag{19}$$

We define $D_s\in\mathbb{R}^{N\times N}$, as the square, diagonal, and invertible matrix corresponding to the first N columns of D. Let's also define $I_N\in\mathbb{R}^{p\times p}$ as the matrix containing 1 in the first N entries of its diagonal, and 0 everywhere else. We have

$$\begin{aligned}P_A=&O_1^\top O_1=OI_NO^\top\\=&OD^\top D_s^{-2}DO^\top=OD^\top U^\top UD_s^{-2}U^\top UDO^\top\\=&A^\top\left(UD_s^2U^\top\right)^{-1}A=A^\top\left(UDO^\top OD^\top U^\top\right)^{-1}A\\=&A^\top\left(AA^\top\right)^{-1}A\equiv A^+A,\end{aligned} \tag{20}$$

where A^+ denotes the Moore-Penrose inverse.

Notice that this last form enables us to easily derive

$$P_{A_{-1}}A^+v=A_{-1}^+A_{-1}A^+v=A_{-1}^+I_{-1}AA^+v=A_{-1}^+I_{-1}v=A_{-1}^+v_{-1}, \tag{21}$$

where $v\in\mathbb{R}^N$, $I_{-1}\in\mathbb{R}^{(N-1)\times N}$ is the $N\times N$ identity matrix without the first row, and $v_{-1}\in\mathbb{R}^{N-1}$ corresponds to v without its first entry.

Lemma D.1. *Let $\Phi\in\mathbb{R}^{N\times k}$ be a matrix whose first row is denoted as $\varphi(z_1)$. Let $\Phi_{-1}\in\mathbb{R}^{(N-1)\times k}$ be the original matrix without the first row, and let $P_{\Phi_{-1}}$ be the projector over the* span *of its rows. Then,*

$$\left\|P_{\Phi_{-1}}^\perp\varphi(z_1)\right\|_2^2\geq\lambda_{\min}\left(\Phi\Phi^\top\right). \tag{22}$$

Proof. If $\lambda_{\min}\left(\Phi\Phi^\top\right)=0$, the thesis becomes trivial. Otherwise, we have that $\Phi\Phi^\top$, and therefore $\Phi_{-1}\Phi_{-1}^\top$, are invertible.

Let $u \in \mathbb{R}^N$ be a vector, such that its first entry $u_1 = 1$. We denote with $u_{-1} \in \mathbb{R}^{N-1}$ the vector u without its first component, i.e. $u = [1, u_{-1}]$. We have

$$\left\|\Phi^\top u\right\|_2^2 \geq \lambda_{\min}\left(\Phi\Phi^\top\right)\|u\|_2^2 \geq \lambda_{\min}\left(\Phi\Phi^\top\right). \tag{23}$$

Setting $u_{-1} = -\left(\Phi_{-1}\Phi_{-1}^\top\right)^{-1}\Phi_{-1}\varphi(z_1)$, we get

$$\Phi^\top u = \varphi(z_1) + \Phi_{-1}^\top u_{-1} = \varphi(z_1) - P_{\Phi_{-1}}\varphi(z_1) = P_{\Phi_{-1}}^\perp\varphi(z_1). \tag{24}$$

Plugging this in (23), we get the thesis. □

At this point, we are ready to prove Lemma 3.1.

Proof of Lemma 3.1. We indicate with $\Phi_{-1} \in \mathbb{R}^{(N-1)\times p}$ the feature matrix of the training set $\Phi \in \mathbb{R}^{N\times p}$ *without* the first sample z_1. In other words, Φ_{-1} is equivalent to Φ, without the first row. Notice that since $K = \Phi\Phi^\top$ is invertible, also $K_{-1} := \Phi_{-1}\Phi_{-1}^\top$ is.

We can express the projector over the span of the rows of Φ in terms of the projector over the span of the rows of Φ_{-1} as follows

$$P_\Phi = P_{\Phi_{-1}} + \frac{P_{\Phi_{-1}}^\perp\varphi(z_1)\varphi(z_1)^\top P_{\Phi_{-1}}^\perp}{\left\|P_{\Phi_{-1}}^\perp\varphi(z_1)\right\|_2^2}. \tag{25}$$

The above expression is a consequence of the Gram-Schmidt formula, and the quantity at the denominator is different from zero because of Lemma D.1, as K is invertible.

We indicate with $\Phi^+ = \Phi^\top K^{-1}$ the Moore–Penrose pseudo-inverse of Φ. Using (3), we can define $\theta_{-1}^* := \theta_0 + \Phi_{-1}^+\left(G_{-1} - f(Z_{-1}, \theta_0)\right)$, *i.e.*, the set of parameters the algorithm would have converged to if trained over (Z_{-1}, G_{-1}), the original data-set without the first pair sample-label (z_1, g_1).

Notice that $P_\Phi\Phi^\top = \Phi^\top$, as a consequence of (20). Thus, again using (3), for any z we can write

$$\begin{aligned} f(z,\theta^*) - \varphi(z)^\top\theta_0 &= \varphi(z)^\top\Phi^+\left(G - f(Z,\theta_0)\right) \\ &= \varphi(z)^\top P_\Phi\Phi^+\left(G - f(Z,\theta_0)\right) \\ &= \varphi(z)^\top\left(P_{\Phi_{-1}} + \frac{P_{\Phi_{-1}}^\perp\varphi(z_1)\varphi(z_1)^\top P_{\Phi_{-1}}^\perp}{\left\|P_{\Phi_{-1}}^\perp\varphi(z_1)\right\|_2^2}\right)\Phi^+\left(G - f(Z,\theta_0)\right). \end{aligned} \tag{26}$$

Notice that, thanks to (21), we can manipulate the first term in the bracket as follows

$$\begin{aligned} \varphi(z)^\top P_{\Phi_{-1}}\Phi^+\left(G - f(Z,\theta_0)\right) &= \varphi(z)^\top\Phi_{-1}^+\left(G_{-1} - f(Z_{-1},\theta_0)\right) \\ &= f(z,\theta_{-1}^*) - \varphi(z)^\top\theta_0. \end{aligned} \tag{27}$$

Thus, bringing the result of (27) on the LHS, (26) becomes

$$\begin{aligned} f(z,\theta^*) - f(z,\theta_{-1}^*) &= \frac{\varphi(z)^\top P_{\Phi_{-1}}^\perp\varphi(z_1)}{\left\|P_{\Phi_{-1}}^\perp\varphi(z_1)\right\|_2^2}\varphi(z_1)^\top P_{\Phi_{-1}}^\perp\Phi^+\left(G - f(Z,\theta_0)\right) \\ &= \frac{\varphi(z)^\top P_{\Phi_{-1}}^\perp\varphi(z_1)}{\left\|P_{\Phi_{-1}}^\perp\varphi(z_1)\right\|_2^2}\varphi(z_1)^\top\left(I - P_{\Phi_{-1}}\right)\Phi^+\left(G - f(Z,\theta_0)\right) \\ &= \frac{\varphi(z)^\top P_{\Phi_{-1}}^\perp\varphi(z_1)}{\left\|P_{\Phi_{-1}}^\perp\varphi(z_1)\right\|_2^2}\left(f(z_1,\theta^*) - f(z_1,\theta_{-1}^*)\right), \end{aligned} \tag{28}$$

where in the last step we again used (3) and (27). □

E Useful Lemmas

Lemma E.1. *Let x and y be two Lipschitz concentrated, independent random vectors. Let $\zeta(x, y)$ be a Lipschitz function in both arguments, i.e., for every δ,*

$$\begin{aligned} |\zeta(x+\delta, y) - \zeta(x, y)| &\leq L \|\delta\|_2 , \\ |\zeta(x, y+\delta) - \zeta(x, y)| &\leq L \|\delta\|_2 , \end{aligned} \tag{29}$$

for all x and y. Then, $\zeta(x, y)$ is a Lipschitz concentrated random variable, in the joint probability space of x and y.

Proof. To prove the thesis, we need to show that, for every 1-Lipschitz function τ, the following holds

$$\mathbb{P}_{xy}\left(|\tau\left(\zeta(x,y)\right) - \mathbb{E}_{xy}\left[\tau\left(\zeta(x,y)\right)\right]| > t\right) < e^{-ct^2}, \tag{30}$$

where c is a universal constant. An application of the triangle inequality gives

$$\begin{aligned} &|\tau\left(\zeta(x,y)\right) - \mathbb{E}_{xy}\left[\tau\left(\zeta(x,y)\right)\right]| \\ &\leq |\tau\left(\zeta(x,y)\right) - \mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right]| + |\mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right] - \mathbb{E}_{y}\mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right]| =: A + B. \end{aligned} \tag{31}$$

Thus, we can upper bound LHS of (30) as follows:

$$\mathbb{P}_{xy}\left(|\tau\left(\zeta(x,y)\right) - \mathbb{E}_{xy}\left[\tau\left(\zeta(x,y)\right)\right]| > t\right) \leq \mathbb{P}_{xy}\left(A + B > t\right). \tag{32}$$

If A and B are positive random variables, it holds that $\mathbb{P}(A + B > t) \leq \mathbb{P}(A > t/2) + \mathbb{P}(B > t/2)$. Then, the LHS of (30) is also upper bounded by

$$\begin{aligned} &\mathbb{P}_{xy}\left(|\tau\left(\zeta(x,y)\right) - \mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right]| > t/2\right) \\ &+ \mathbb{P}_{xy}\left(|\mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right] - \mathbb{E}_{y}\mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right]| > t/2\right). \end{aligned} \tag{33}$$

Since $\tau \circ \zeta$ is Lipschitz with respect to x for every y, we have

$$\mathbb{P}_{xy}\left(|\tau\left(\zeta(x,y)\right) - \mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right]| > t/2\right) < e^{-c_1 t^2}, \tag{34}$$

for some absolute constant c_1. Furthermore, $\chi(y) := \mathbb{E}_x\left[\tau\left(\zeta(x,y)\right)\right]$ is also Lipschitz, as

$$\begin{aligned} |\chi(y+\delta) - \chi(y)| &= |\mathbb{E}_x\left[\tau\left(\zeta(x, y+\delta)\right) - \tau\left(\zeta(x,y)\right)\right]| \\ &\leq \mathbb{E}_x\left[|\tau\left(\zeta(x, y+\delta)\right) - \tau\left(\zeta(x,y)\right)|\right] \leq L\|\delta\|_2 . \end{aligned} \tag{35}$$

Then, we can write

$$\begin{aligned} &\mathbb{P}_{xy}\left(|\mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right] - \mathbb{E}_{y}\mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right]| > t/2\right) \\ &= \mathbb{P}_{y}\left(|\chi(y) - \mathbb{E}_y\left[\chi(y)\right]| > t/2\right) < e^{-c_2 t^2}, \end{aligned} \tag{36}$$

for some absolute constant c_2. Thus,

$$\mathbb{P}_{xy}\left(|\tau\left(\zeta(x,y)\right) - \mathbb{E}_{xy}\left[\tau\left(\zeta(x,y)\right)\right]| > t\right) < e^{-c_1 t^2} + e^{-c_2 t^2} \leq e^{-ct^2}, \tag{37}$$

for some absolute constant c, which concludes the proof. □

Lemma E.2. *Let $x \sim \mathcal{P}_X$, $y \sim \mathcal{P}_Y$ and $z = [x, y] \sim \mathcal{P}_Z$. Let Assumption 4.1 hold. Then, z is a Lipschitz concentrated random vector.*

Proof. We want to prove that, for every 1-Lipschitz function τ, the following holds

$$\mathbb{P}_z\left(|\tau\left(z\right) - \mathbb{E}_z\left[\tau\left(z\right)\right]| > t\right) < e^{-ct^2}, \tag{38}$$

for some universal constant c. As we can write $z = [x, y]$, defining $z' = [x', y]$, we have

$$|\tau\left(z\right) - \tau\left(z'\right)| \leq \|z - z'\|_2 = \|x - x'\|_2 , \tag{39}$$

i.e., for every y, τ is 1-Lipschitz with respect to x. The same can be shown for y, with an equivalent argument. Since x and y are independent random vectors, both Lipschitz concentrated, Lemma E.1 gives the thesis. □

Lemma E.3. *Let τ and ζ be two Lipschitz functions. Let $z, z' \in \mathbb{R}^d$ be two fixed vectors such that $\|z\|_2 = \|z'\|_2 = \sqrt{d}$. Let V be a $k \times d$ matrix such that $V_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$. Then, for any $t > 1$,*

$$\left|\tau(Vz)^\top \zeta(Vz') - \mathbb{E}_V\left[\tau(Vz)^\top \zeta(Vz')\right]\right| = \mathcal{O}\left(\sqrt{k}\log t\right), \tag{40}$$

with probability at least $1 - \exp(-c\log^2 t)$ over V. Here, τ and ζ act component-wise on their arguments. Furthermore, by taking $\tau = \zeta$ and $z = z'$, we have that

$$\mathbb{E}_V\left[\|\tau(Vz)\|_2^2\right] = k\mathbb{E}_\rho\left[\tau^2(\rho)\right], \tag{41}$$

where $\rho \sim \mathcal{N}(0,1)$. This implies that $\|\tau(Vz)\|_2^2 = \mathcal{O}(k)$ with probability at least $1 - \exp(-ck)$ over V.

Proof. We have

$$\tau(Vz)^\top \zeta(Vz') = \sum_{j=1}^{k} \tau(v_j^\top z)\zeta(v_j^\top z'), \tag{42}$$

where we used the shorthand $v_j := V_{j:}$. As τ and ζ are Lipschitz, $v_j \sim \mathcal{N}(0, I/d)$, and $\|z\|_2 = \|z'\|_2 = \sqrt{d}$, we have that $\tau(Vz)^\top\zeta(Vz')$ is the sum of k independent sub-exponential random variables, in the probability space of V. Thus, by Bernstein inequality (cf. Theorem 2.8.1 in [47]), we have

$$\left|\tau(Vz)^\top \zeta(Vz') - \mathbb{E}_V\left[\tau(Vz)^\top \zeta(Vz')\right]\right| = \mathcal{O}\left(\sqrt{k}\log t\right). \tag{43}$$

with probability at least $1 - \exp(-c\log^2 t)$, over the probability space of V, which gives the thesis. The second statement is again implied by the fact that $v_j \sim \mathcal{N}(0, I/d)$ and $\|z\|_2 = \sqrt{d}$. □

Lemma E.4. *Let $x, x_1 \sim \mathcal{P}_X$ and $y_1 \sim \mathcal{P}_Y$ be independent random variables, with $x, x_1 \in \mathbb{R}^{d_x}$ and $y_1 \in \mathbb{R}^{d_y}$, and let Assumption 4.1 hold. Let $d = d_x + d_y$, V be a $k \times d$ matrix, such that $V_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$, and let τ be a Lipschitz function. Let $z_1 := [x_1, y_1]$ and $z_1^m := [x, y_1]$. Let $\alpha = d_y/d \in (0,1)$ and μ_l be the l-th Hermite coefficient of τ. Then, for any $t > 1$,*

$$\left|\tau(Vz_1^m)^\top\tau(Vz_1) - k\sum_{l=0}^{+\infty}\mu_l^2\alpha^l\right| = \mathcal{O}\left(\sqrt{k}\left(\sqrt{\frac{k}{d}}+1\right)\log t\right), \tag{44}$$

with probability at least $1 - \exp(-c\log^2 t) - \exp(-ck)$ over V and x, where c is a universal constant.

Proof. Define the vector x' as follows

$$x' = \frac{\sqrt{d_x}\left(I - \frac{x_1 x_1^\top}{d_x}\right)x}{\left\|\left(I - \frac{x_1 x_1^\top}{d_x}\right)x\right\|_2}. \tag{45}$$

Note that, by construction, $x_1^\top x' = 0$ and $\|x'\|_2 = \sqrt{d_x}$. Also, consider a vector y orthogonal to both x_1 and x. Then, a fast computation returns $y^\top x' = 0$. This means that x' is the vector on the $\sqrt{d_x}$-sphere, lying on the same plane of x_1 and x, orthogonal to x_1. Thus, we can easily compute

$$\frac{\left|x^\top x'\right|}{d_x} = \sqrt{1 - \left(\frac{x^\top x_1}{d_x}\right)^2} \geq 1 - \left(\frac{x^\top x_1}{d_x}\right)^2, \tag{46}$$

where the last inequality derives from $\sqrt{1-a} \geq 1 - a$ for $a \in [0,1]$. Then,

$$\|x - x'\|_2^2 = \|x\|_2^2 + \|x'\|_2^2 - 2x^\top x' \leq 2d_x\left(1 - \left(1 - \left(\frac{x^\top x_1}{d_x}\right)^2\right)\right) = 2\frac{\left(x^\top x_1\right)^2}{d_x}. \tag{47}$$

As x and x_1 are both sub-Gaussian, mean-0 vectors, with ℓ_2 norm equal to $\sqrt{d_x}$, we have that

$$\mathbb{P}\left(\|x - x'\|_2 > t\right) \leq \mathbb{P}\left(|x^\top x_1| > \sqrt{d_x}t/\sqrt{2}\right) < \exp(-ct^2), \tag{48}$$

where c is an absolute constant. Here the probability is referred to the space of x, for a fixed x_1. Thus, $\|x - x'\|_2$ is sub-Gaussian.

We now define $z' := [x', y_1]$. Notice that $z_1^\top z' = \|y_1\|_2^2 = d_y$ and $\|z_1^m - z'\|_2 = \|x - x'\|_2$. We can write

$$\begin{aligned}\left|\tau(Vz_1^m)^\top \tau(Vz_1) - \tau(Vz')^\top \tau(Vz_1)\right| &\le \|\tau(Vz_1^m) - \tau(Vz')\|_2 \, \|\tau(Vz_1)\|_2 \\ &\le C \, \|V\|_{\text{op}} \, \|z_1^m - z'\|_2 \, \|\tau(Vz_1)\|_2 \\ &\le C_1 \left(\sqrt{\frac{k}{d}} + 1\right) \|x - x'\|_2 \sqrt{k} \\ &= \mathcal{O}\left(\sqrt{k}\left(\sqrt{\frac{k}{d}} + 1\right) \log t\right).\end{aligned} \tag{49}$$

Here the second step holds as τ is Lipschitz; the third step holds with probability at least $1 - \exp(-c_1 \log^2 t) - \exp(-c_2 k)$, and it uses Theorem 4.4.5 of [47] and Lemma E.3; the fourth step holds with probability at least $1 - \exp(-c \log^2 t)$, and it uses (48). This probability is intended over V and x. We further have

$$\left|\tau(Vz')^\top \tau(Vz_1) - \mathbb{E}_V\left[\tau(Vz')^\top \tau(Vz_1)\right]\right| = \mathcal{O}\left(\sqrt{k} \log t\right), \tag{50}$$

with probability at least $1 - \exp(-c_3 \log^2 t) - \exp(-c_2 k)$ over V, because of Lemma E.3.

We have

$$\mathbb{E}_V\left[\tau(Vz')^\top \tau(Vz_1)\right] = k \mathbb{E}_{\rho_1 \rho_2}\left[\tau(\rho_1)\tau(\rho_2)\right], \tag{51}$$

where we indicate with ρ_1 and ρ_2 two standard Gaussian random variables, with correlation

$$\text{corr}(\rho_1, \rho_2) = \frac{z_1^\top z}{\|z_1\|_2 \, \|z'\|_2} = \frac{d_y}{d} = \alpha. \tag{52}$$

Then, exploiting the Hermite expansion of τ, we have

$$\mathbb{E}_{\rho_1 \rho_2}\left[\tau(\rho_1)\tau(\rho_2)\right] = \sum_{l=0}^{+\infty} \mu_l^2 \alpha^l. \tag{53}$$

Putting together (49), (50), (51), and (53) gives the thesis. □

F Proofs for Random Features

In this section, we indicate with $Z \in \mathbb{R}^{N \times d}$ the data matrix, such that its rows are sampled independently from $\mathcal{P}_Z$ (see Assumption 4.1). We denote by $V \in \mathbb{R}^{k \times d}$ the random features matrix, such that $V_{ij} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$. Thus, the feature map is given by (see Section 4)

$$\varphi(z) := \phi(Vz) \in \mathbb{R}^k, \tag{54}$$

where ϕ is the activation function, applied component-wise to the pre-activations Vz. We use the shorthands $\Phi := \phi(ZV^\top) \in \mathbb{R}^{N \times k}$ and $K := \Phi\Phi^\top \in \mathbb{R}^{N \times N}$, we indicate with $\Phi_{-1} \in \mathbb{R}^{(N-1) \times k}$ the matrix Φ without the first row, and we define $K_{-1} := \Phi_{-1}\Phi_{-1}^\top$. We call P_Φ the projector over the span of the rows of Φ, and $P_{\Phi_{-1}}$ the projector over the span of the rows of Φ_{-1}. We use the notations $\tilde{\varphi}(z) := \varphi(z) - \mathbb{E}_V[\varphi(z)]$ and $\tilde{\Phi}_{-1} := \Phi_{-1} - \mathbb{E}_V[\Phi_{-1}]$ to indicate the centered feature map and matrix respectively, where the centering is with respect to V. We indicate with μ_l the l-th Hermite coefficient of ϕ. We use the notation $z_1^m = [x, y_1]$, where $x \sim \mathcal{P}_X$ is sampled independently from V and Z. We denote by V_x (V_y) the first d_x (last d_y) columns of V, i.e., $V = [V_x, V_y]$. We define $\alpha = d_y/d$. Throughout this section, for compactness, we drop the subscripts "RF" from these quantities, as we will only treat the proofs related to the Random Features model. Again for the sake of compactness, we will not re-introduce such quantities in the statements or the proofs of the following lemmas.

Through the following Section, as mentioned in the main body of the paper, we will work under the following assumptions

Assumption F.1 (Over-parameterization and high-dimensional data)**.**

$$N \log^3 N = o(k), \qquad \sqrt{d} \log d = o(k), \qquad k \log^4 k = o\left(d^2\right). \tag{55}$$

The first condition in (55) requires the number of neurons k to scale faster than the number of data points N. This over-parameterization leads to a lower bound on the smallest eigenvalue of the kernel induced by the feature map, which in turn implies that the model interpolates the data, as required to write (3). This over-parameterized regime also achieves minimum test error [31]. Combining the second and third conditions in (55), we have that k can scale between $\sqrt{d}$ and d^2 (up to $\log$ factors). Finally, merging the first and third condition gives that d^2 scales faster than N. We notice that this holds for standard datasets (MNIST, CIFAR-10 and ImageNet).

Assumption F.2 (Activation function)**.** *The activation function ϕ is a non-linear L-Lipschitz function.*

This requirement is satisfied by common activations, e.g., ReLU, sigmoid, or $\tanh$.

Summary of this Section.

- In Lemma F.4 we prove a lower bound on the smallest eigenvalue of K, adapting to our settings Lemma C.5 of [14]. As our assumptions are less restrictive than those in [14], we will crucially exploit Lemma F.3.
- In Lemma F.5, we treat separately a term that derives from $\mathbb{E}_V\left[\phi(Vz)\right] = \mu_0 \mathbf{1}_k$, showing that we can *center* the activation function, without changing our final statement in Theorem 4.2. This step is necessary only if $\mu_0 \neq 0$.
- In Lemma F.6, we show that the non-linear component of the features $\tilde{\varphi}(z_1) - \mu_1 V z_1$ and $\tilde{\varphi}(z_1^m) - \mu_1 V z_1^m$ have a negligible component in the space spanned by the rows of Φ_{-1}.
- In Lemma F.9, we provide concentration results for $\varphi(z_1^m)^\top P^\perp_{\Phi_{-1}} \varphi(z_1)$, and we lower bound this same term in Lemma F.8, exploiting also the intermediate result provided in Lemma F.7.
- Finally, we prove Theorem 4.2.

Lemma F.3. *Let $A := (Z^{*m}) \in \mathbb{R}^{N \times d^m}$, for some natural $m \geq 2$, where $*$ refers to the Khatri-Rao product, defined in Appendix C. We have*

$$\lambda_{\min}\left(AA^\top\right) = \Omega(d^m), \tag{56}$$

with probability at least $1 - \exp(-c \log^2 N)$ over Z, where c is an absolute constant.

Proof. As $m \geq 2$, we can write $A = \left(Z^{*2}\right) * \left(Z^{*(m-2)}\right) =: A_2 * A_m$ (where $\left(Z^{*0}\right)$ is defined to be the vector full of ones $\mathbf{1}_N \in \mathbb{R}^N$). We can provide a lower bound on the smallest eigenvalue of such product through the following inequality [41]:

$$\lambda_{\min}\left(AA^\top\right) = \lambda_{\min}\left(A_2 A_2^\top \circ A_m A_m^\top\right) \geq \lambda_{\min}\left(A_2 A_2^\top\right) \min_i \|(A_m)_{i:}\|_2^2 . \tag{57}$$

Note that the rows of Z are mean-0 and Lipschitz concentrated by Lemma E.2. Then, by following the argument of Lemma C.3 in [14], we have

$$\lambda_{\min}\left(A_2 A_2^\top\right) = \Omega(d^2), \tag{58}$$

with probability at least $1 - \exp(-c \log^2 N)$ over Z. We remark that, for the argument of Lemma C.3 in [14] to go through, it suffices that $N = o(d^2 / \log^4 d)$ and $N \log^4 N = o(d^2)$ (see Equations (C.23) and (C.26) in [14]), which is implied by AssumptionF.1, despite it being milder than Assumption 4 in [14].

For the second term of (57), we have

$$\|(A_m)_{i:}\|_2^2 = \|z_i\|_2^{2(m-2)} = d^{m-2}, \tag{59}$$

due to Assumption 4.1. Thus, the thesis readily follows. □

Lemma F.4. *We have that*

$$\lambda_{\min}(K) = \Omega(k), \tag{60}$$

with probability at least $1 - \exp\left(-c\log^2 N\right)$ *over* V *and* Z*, where* c *is an absolute constant. This implies that* $\lambda_{\min}(K_{-1}) = \Omega(k)$.

Proof. The proof follows the same path as Lemma C.5 of [14]. In particular, we define a truncated version of Φ as follows

$$\bar{\Phi}_{:j} = \phi(Zv_j)\chi\left(\|\phi(Zv_j)\|_2^2 \leq R\right), \tag{61}$$

where χ is the indicator function and we introduce the shorthand $v_i := V_{i:}$. In this case, $\chi = 1$ if $\|\phi(Zv_j)\|_2^2 \leq R$, and $\chi = 0$ otherwise. As this is a column-wise truncation, it's easy to verify that $\Phi\Phi^\top \succeq \bar{\Phi}\bar{\Phi}^\top$. Over such truncated matrix, we can use Matrix Chernoff inequality (see Theorem 1.1 of [46]), which gives that $\lambda_{\min}\left(\bar{\Phi}\bar{\Phi}^\top\right) = \Omega(\lambda_{\min}\left(\bar{G}\right))$, where $\bar{G} := \mathbb{E}_V\left[\bar{\Phi}\bar{\Phi}^\top\right]$. Finally, we prove closeness between $\bar{G}$ and G, which is analogously defined as $G := \mathbb{E}_V\left[\Phi\Phi^\top\right]$.

To be more specific, setting $R = k/\log^2 N$, we have

$$\lambda_{\min}(K) \geq \lambda_{\min}\left(\bar{\Phi}\bar{\Phi}^\top\right) \geq \lambda_{\min}\left(\bar{G}\right)/2 \geq \lambda_{\min}(G)/2 - o(k), \tag{62}$$

where the second inequality holds with probability at least $1 - \exp(c_1 \log^2 N)$ over V, if $\lambda_{\min}(G) = \Omega(k)$ (see Equation (C.47) of [14]), and the third comes from Equation (C.45) in [14]. To perform these steps, our Assumptions F.1 and F.2 are enough, despite the second one being milder than Assumption 2 in [14].

To conclude the proof, we are left to prove that $\lambda_{\min}(G) = \Omega(k)$ with probability at least $1 - \exp(-c_2 \log^2 N)$ over V and Z.

We have that

$$G = \mathbb{E}_V[K] = \mathbb{E}_V\left[\sum_{i=1}^{k} \phi(ZV_{i:}^\top)\phi(ZV_{i:}^\top)^\top\right] = k\mathbb{E}_v\left[\phi(Zv)\phi(Zv)^\top\right] := kM, \tag{63}$$

where we use the shorthand v to indicate a random variable distributed as $V_{1:}$. We also indicate with z_i the i-th row of Z. Exploiting the Hermite expansion of ϕ, we can write

$$M_{ij} = \mathbb{E}_v\left[\phi(z_i^\top v)\phi(z_j^\top v)\right] = \sum_{l=0}^{+\infty} \mu_l^2 \frac{\left(z_i^\top z_j\right)^l}{d^l} = \sum_{l=0}^{+\infty} \mu_l^2 \frac{\left[\left(Z^{*l}\right)\left(Z^{*l}\right)^\top\right]_{ij}}{d^l}, \tag{64}$$

where μ_l is the l-th Hermite coefficient of ϕ. Note that the previous expansion was possible since $\|z_i\| = \sqrt{d}$ for all $i \in [N]$. As ϕ is non-linear, there exists $m \geq 2$ such that $\mu_m^2 > 0$. In particular, we have $M \succeq \frac{\mu_m^2}{d^m} AA^\top$ in a PSD sense, where we define

$$A := \left(Z^{*m}\right). \tag{65}$$

By Lemma F.3, the desired result readily follows. □

Lemma F.5. *Let* $\mu_0 \neq 0$*. Then,*

$$\left\|P_{\Phi_{-1}}^{\perp}\mathbf{1}_k\right\|_2 = o(\sqrt{k}), \tag{66}$$

with probability at least $1 - e^{-cd} - e^{-cN}$ *over* V *and* Z*, where* c *is an absolute constant.*

Proof. Note that $\Phi_{-1}^\top = \mu_0 \mathbf{1}_k \mathbf{1}_{N-1}^\top + \tilde{\Phi}_{-1}^\top$. Here, $\tilde{\Phi}_{-1}^\top$ is a $k \times (N-1)$ matrix with i.i.d. and mean-0 rows, whose sub-Gaussian norm (in the probability space of V) can be bounded as

$$\left\|\tilde{\Phi}_{:i}\right\|_{\psi_2} = \|\phi(ZV_{i:}) - \mathbb{E}_V[\phi(ZV_{i:})]\|_{\psi_2} \leq L\frac{\|Z\|_{\text{op}}}{\sqrt{d}} = \mathcal{O}\left(\sqrt{N/d} + 1\right), \tag{67}$$

where first inequality holds since ϕ is L-Lipschitz and $V_{i:}$ is a Gaussian (and hence, Lipschitz concentrated) vector with covariance I/d. The last step holds with probability at least $1 - e^{-cd}$ over Z, because of Lemma B.7 in [13].

Thus, another application of Lemma B.7 in [13] gives

$$\left\|\tilde{\Phi}_{-1}^\top\right\|_{\text{op}} = \mathcal{O}\left(\left(\sqrt{k}+\sqrt{N}\right)\left(\sqrt{N/d}+1\right)\right) = \mathcal{O}\left(\sqrt{k}\left(\sqrt{N/d}+1\right)\right), \tag{68}$$

where the first equality holds with probability at least $1-e^{-cN}$ over V, and the second is a direct consequence of Assumption F.1.

We can write

$$\Phi_{-1}^\top \frac{\mathbf{1}_{N-1}}{\mu_0(N-1)} = \left(\mu_0 \mathbf{1}_k \mathbf{1}_{N-1}^\top + \tilde{\Phi}_{-1}^\top\right)\frac{\mathbf{1}_{N-1}}{\mu_0(N-1)} = \mathbf{1}_k + \tilde{\Phi}_{-1}^\top \frac{\mathbf{1}_{N-1}}{\mu_0(N-1)} =: \mathbf{1}_k + v, \tag{69}$$

where

$$\|v\|_2 \le \frac{1}{\mu_0(N-1)}\left\|\tilde{\Phi}_{-1}^\top\right\|_{\text{op}} \|\mathbf{1}_{N-1}\|_2 = \mathcal{O}\left(\sqrt{\frac{k}{N}}\left(\sqrt{N/d}+1\right)\right) = o(\sqrt{k}). \tag{70}$$

Thus, we can conclude

$$\begin{aligned}\left\|P_{\Phi_{-1}}^\perp \mathbf{1}_k\right\|_2 &= \left\|P_{\Phi_{-1}}^\perp\left(\Phi_{-1}^\top \frac{\mathbf{1}_{N-1}}{\mu_0(N-1)} - v\right)\right\|_2 \\ &\le \left\|P_{\Phi_{-1}}^\perp P_{\Phi_{-1}} \Phi_{-1}^\top \frac{\mathbf{1}_{N-1}}{\mu_0(N-1)}\right\|_2 + \|v\|_2 = o(\sqrt{k}),\end{aligned} \tag{71}$$

where in the second step we use the triangle inequality, $\Phi_{-1}^\top = P_{\Phi_{-1}}\Phi_{-1}^\top$, and $\left\|P_{\Phi_{-1}}^\perp v\right\|_2 \le \|v\|_2$. □

Lemma F.6. *Let $z \sim \mathcal{P}_Z$, sampled independently from Z_{-1}, and denote $\tilde{\phi}(x) := \phi(x) - \mu_0$. Then,*

$$\left\|P_{\Phi_{-1}}\left(\tilde{\phi}(Vz) - \mu_1 Vz\right)\right\|_2 = o(\sqrt{k}), \tag{72}$$

with probability at least $1 - \exp\left(-c\log^2 N\right)$ over V, Z_{-1} and z, where c is an absolute constant.

Proof. As $P_{\Phi_{-1}} = \Phi_{-1}^+ \Phi_{-1}$, we have

$$\begin{aligned}\left\|P_{\Phi_{-1}}\left(\tilde{\phi}(Vz) - \mu_1 Vz\right)\right\|_2 &\le \left\|\Phi_{-1}^+\right\|_{\text{op}} \left\|\Phi_{-1}\left(\tilde{\phi}(Vz) - \mu_1 Vz\right)\right\|_2 \\ &= \mathcal{O}\left(\frac{\left\|\Phi_{-1}\left(\tilde{\phi}(Vz) - \mu_1 Vz\right)\right\|_2}{\sqrt{k}}\right),\end{aligned} \tag{73}$$

where the last equality holds with probability at least $1-\exp\left(-c\log^2 N\right)$ over V and Z_{-1}, because of Lemma F.4.

An application of Lemma E.3 with $t = N$ gives

$$|u_i - \mathbb{E}_V[u_i]| = \mathcal{O}\left(\sqrt{k}\log N\right), \tag{74}$$

where u_i is the i-th entry of the vector $u := \Phi_{-1}\left(\tilde{\phi}(Vz) - \mu_1 Vz\right)$. This can be done since both ϕ and $\tilde{\phi} \equiv \phi - \mu_0$ are Lipschitz, $v_j \sim \mathcal{N}(0, I/d)$, and $\|z\|_2 = \|z_{i+1}\|_2 = \sqrt{d}$. Performing a union bound over all entries of u, we can guarantee that the previous equation holds for every $1 \le i \le N-1$, with probability at least $1-(N-1)\exp(-c\log^2 N) \ge 1 - \exp(-c_1 \log^2 N)$. Thus, we have

$$\|u - \mathbb{E}_V[u]\|_2 = \mathcal{O}\left(\sqrt{k}\sqrt{N}\log N\right) = o(k), \tag{75}$$

where the last equality holds because of Assumption F.1.

Note that the function $f(x) := \tilde{\phi}(x) - \mu_1 x$ has the first 2 Hermite coefficients equal to 0. Hence, as $v_i^\top z$ and $v_i^\top z_i$ are standard Gaussian random variables with correlation $\frac{z^\top z_i}{\|z\|_2 \|z_i\|_2}$, we have

$$\begin{aligned}
|\mathbb{E}_V[u_i]| &\le k \sum_{l=2}^{+\infty} \mu_l^2 \left(\frac{\left|z^\top z_i\right|}{\|z\|_2 \|z_i\|_2} \right)^l \\
&\le k \max_l \mu_l^2 \sum_{l=2}^{+\infty} \left(\frac{\left|z^\top z_i\right|}{\|z\|_2 \|z_i\|_2} \right)^l \\
&= k \max_l \mu_l^2 \left(\frac{z^\top z_i}{\|z\|_2 \|z_i\|_2} \right)^2 \frac{1}{1 - \frac{|z^\top z_i|}{\|z\|_2 \|z_i\|_2}} \\
&\le 2k \max_l \mu_l^2 \left(\frac{z^\top z_i}{\|z\|_2 \|z_i\|_2} \right)^2 = \mathcal{O}\left(\frac{k \log^2 N}{d} \right),
\end{aligned} \tag{76}$$

where the last inequality holds with probability at least $1 - \exp\left(-c \log^2 N\right)$ over z and z_i, as they are two independent, mean-0, sub-Gaussian random vectors. Again, performing a union bound over all entries of $\mathbb{E}_V[u]$, we can guarantee that the previous equation holds for every $1 \le i \le N-1$, with probability at least $1 - (N-1)\exp(-c \log^2 N) \ge 1 - \exp(-c_1 \log^2 N)$. Then, we have

$$\|\mathbb{E}_V[u]\|_2 = \mathcal{O}\left(\sqrt{N} \frac{k \log^2 N}{d} \right) = o(k), \tag{77}$$

where the last equality is a consequence of Assumption F.1.

Finally, (75) and (77) give

$$\left\| \Phi_{-1}\left(\tilde{\phi}(Vz) - \mu_1 Vz \right) \right\|_2 \le \|\mathbb{E}_V[u]\|_2 + \|u - \mathbb{E}_V[u]\|_2 = o(k), \tag{78}$$

which plugged in (73) readily provides the thesis. □

Lemma F.7. *We have*

$$\left| (Vz_1^m)^\top P^\perp_{\Phi_{-1}} Vz_1 - \left\| P^\perp_{\Phi_{-1}} V_y y_1 \right\|_2^2 \right| = o(k), \tag{79}$$

with probability at least $1 - \exp(-c \log^2 N)$ *over* x, z_1 *and* V, *where* c *is an absolute constant.*

Proof. We have

$$Vz_1^m = V_x x + V_y y_1, \qquad Vz_1 = V_x x_1 + V_y y_1. \tag{80}$$

Thus, we can write

$$\begin{aligned}
\left| (Vz_1^m)^\top P^\perp_{\Phi_{-1}} Vz_1 - \left\| P^\perp_{\Phi_{-1}} V_y y_1 \right\|_2^2 \right| &= \left| (V_x x)^\top P^\perp_{\Phi_{-1}} Vz_1 + (V_y y_1)^\top P^\perp_{\Phi_{-1}} V_x x_1 \right| \\
&\le \left| x^\top V_x^\top P^\perp_{\Phi_{-1}} Vz_1 \right| + \left| y_1^\top V_y^\top P^\perp_{\Phi_{-1}} V_x x_1 \right|.
\end{aligned} \tag{81}$$

Let's look at the first term of the RHS of the previous equation. Notice that $\|V\|_{\text{op}} = \mathcal{O}\left(\sqrt{k/d} + 1 \right)$ with probability at least $1 - 2e^{-cd}$, because of Theorem 4.4.5 of [47]. We condition on such event until the end of the proof, which also implies having the same bound on $\|V_x\|_{\text{op}}$ and $\|V_y\|_{\text{op}}$. Since x is a mean-0 sub-Gaussian vector, independent from $V_x^\top P^\perp_{\Phi_{-1}} Vz_1$, we have

$$\begin{aligned}
\left| x^\top V_x^\top P^\perp_{\Phi_{-1}} Vz_1 \right| &\le \log N \left\| V_x^\top P^\perp_{\Phi_{-1}} Vz_1 \right\|_2 \\
&\le \log N \|V_x\|_{\text{op}} \left\| P^\perp_{\Phi_{-1}} \right\|_{\text{op}} \|V\|_{\text{op}} \|z_1\| \\
&= \mathcal{O}\left(\log N \left(\frac{k}{d} + 1 \right) \sqrt{d} \right) = o(k),
\end{aligned} \tag{82}$$

where the first inequality holds with probability at least $1 - \exp(-c \log^2 N)$ over x, and the last line holds because $\left\| P^{\perp}_{\Phi_{-1}} \right\|_{\text{op}} \leq 1$, $\|z_1\| = \sqrt{d}$, and because of Assumption F.1.

Similarly, exploiting the independence between x_1 and y_1, we can prove that $\left| y_1^\top V_y^\top P^{\perp}_{\Phi_{-1}} V_x x_1 \right| = o(k)$, with probability at least $1 - \exp(-c \log^2 N)$ over y_1. Plugging this and (82) in (81) readily gives the thesis. □

Lemma F.8. *We have*

$$\left| \varphi(z_1^m)^\top P^{\perp}_{\Phi_{-1}} \varphi(z_1) - \left(k \left(\sum_{l=2}^{+\infty} \mu_l^2 \alpha^l \right) + \mu_1^2 \left\| P^{\perp}_{\Phi_{-1}} V_y y_1 \right\|_2^2 \right) \right| = o(k), \tag{83}$$

with probability at least $1 - \exp(-c \log^2 N)$ *over* V *and* Z*, where* c *is an absolute constant.*

Proof. An application of Lemma E.3 and Assumption F.1 gives

$$\begin{aligned} \|\varphi(z_1)\|_2 &= \mathcal{O}\left(\sqrt{k}\right), \qquad \|\varphi(z_1^m)\|_2 = \mathcal{O}\left(\sqrt{k}\right), \\ \|V z_1\|_2 &= \mathcal{O}\left(\sqrt{k}\right), \qquad \|V z_1^m\|_2 = \mathcal{O}\left(\sqrt{k}\right), \end{aligned} \tag{84}$$

with probability at least $1 - \exp(-c_1 \log^2 N)$ over V, where c_1 is an absolute constant. We condition on such high probability event until the end of the proof.

Let's suppose $\mu_0 \neq 0$. Then, we have

$$\left| \varphi(z_1^m)^\top P^{\perp}_{\Phi_{-1}} \varphi(z_1) - \tilde{\phi}(V z_1^m)^\top P^{\perp}_{\Phi_{-1}} \tilde{\phi}(V z_1) \right| = o(k), \tag{85}$$

with probability at least $1 - \exp(c_2 \log^2 N)$ over V and Z, because of (84) and Lemma F.5. Note that (85) trivially holds even when $\mu_0 = 0$, as $\phi \equiv \tilde{\phi}$. Thus, (85) is true in any case with probability at least $1 - \exp(c_2 \log^2 N)$ over V and Z.

Furthermore, because of (84) and Lemma F.6, we have

$$\left| \tilde{\phi}(V z_1^m)^\top P_{\Phi_{-1}} \tilde{\phi}(V z_1) - \mu_1^2 (V z_1^m)^\top P_{\Phi_{-1}} (V z_1) \right| = o(k), \tag{86}$$

with probability at least $1 - \exp(-c_3 \log^2 N)$ over V and Z.

Thus, putting (85) and (86) together, and using Lemma F.7, we get

$$\begin{aligned} &\left| \varphi(z_1^m)^\top P^{\perp}_{\Phi_{-1}} \varphi(z_1) - \left(\tilde{\phi}(V z_1^m)^\top \tilde{\phi}(V z_1) - \mu_1^2 (V z_1^m)^\top (V z_1) + \mu_1^2 \left\| P^{\perp}_{\Phi_{-1}} V_y y_1 \right\|_2^2 \right) \right| \\ &\quad \leq \left| \varphi(z_1^m)^\top P^{\perp}_{\Phi_{-1}} \varphi(z_1) - \tilde{\phi}(V z_1^m)^\top P^{\perp}_{\Phi_{-1}} \tilde{\phi}(V z_1) \right| \\ &\qquad + \left| -\tilde{\phi}(V z_1^m)^\top P_{\Phi_{-1}} \tilde{\phi}(V z_1) + \mu_1^2 (V z_1^m)^\top P_{\Phi_{-1}} (V z_1) \right| \\ &\qquad + \left| \mu_1^2 (V z_1^m)^\top P^{\perp}_{\Phi_{-1}} (V z_1) - \mu_1^2 \left\| P^{\perp}_{\Phi_{-1}} V_y y_1 \right\|_2^2 \right| = o(k), \end{aligned} \tag{87}$$

with probability at least $1 - \exp(-c_4 \log^2 N)$ over V and X and x. To conclude we apply Lemma E.4 setting $t = N$, together with Assumption F.1, to get

$$\left| \tilde{\phi}(V z_1^m)^\top \tilde{\phi}(V z_1) - k \left(\sum_{l=1}^{+\infty} \mu_l^2 \alpha^l \right) \right| = \mathcal{O}\left(\sqrt{k} \left(\sqrt{\frac{k}{d}} + 1 \right) \log N \right) = o(k), \tag{88}$$

and

$$\left| \mu_1^2 (V z_1^m)^\top (V z_1) - k \mu_1^2 \alpha \right| = \mathcal{O}\left(\sqrt{k} \left(\sqrt{\frac{k}{d}} + 1 \right) \log N \right) = o(k), \tag{89}$$

which jointly hold with probability at least $1 - \exp(-c_5 \log^2 N)$ over V and x.

Applying the triangle inequality to (87), (88), and (89), we get the thesis. □

Lemma F.9. *We have that*

$$\left|\left\|P^{\perp}_{\Phi_{-1}}\varphi(z_1)\right\|_2^2 - \mathbb{E}_{z_1}\left[\left\|P^{\perp}_{\Phi_{-1}}\varphi(z_1)\right\|_2^2\right]\right| = o(k), \tag{90}$$

$$\left|\varphi(z_1^m)^\top P^{\perp}_{\Phi_{-1}}\varphi(z_1) - \mathbb{E}_{z_1,z_1^m}\left[\varphi(z_1^m)^\top P^{\perp}_{\Phi_{-1}}\varphi(z_1)\right]\right| = o(k), \tag{91}$$

jointly hold with probability at least $1-\exp(-c\log^2 N)$ *over* z_1*,* V *and* x*, where* c *is an absolute constant.*

Proof. Let's condition until the end of the proof on both $\|V_x\|_{\text{op}}$ and $\|V_y\|_{\text{op}}$ to be $\mathcal{O}\left(\sqrt{k/d}+1\right)$, which happens with probability at least $1-e^{-c_1 d}$ by Theorem 4.4.5 of [47]. This also implies that $\|V\|_{\text{op}} = \mathcal{O}\left(\sqrt{k/d}+1\right)$.

We indicate with $\nu := \mathbb{E}_{z_1}\left[\varphi(z_1)\right] = \mathbb{E}_{z_1^m}\left[\varphi(z_1^m)\right] \in \mathbb{R}^k$, and with $\hat{\varphi}(z) := \varphi(z) - \nu$. Note that, as φ is a $C\left(\sqrt{k/d}+1\right)$-Lipschitz function, for some constant C, and as z_1 is Lipschitz concentrated, by AssumptionF.1, we have

$$\left|\|\varphi(z_1)\|_2 - \mathbb{E}_{z_1}\left[\|\varphi(z_1)\|_2\right]\right| = o\left(\sqrt{k}\right), \tag{92}$$

with probability at least $1-\exp(-c_2\log^2 N)$ over z_1 and V. In addition, by the last statement of Lemma E.3 and Assumption F.1, we have that $\|\varphi(z_1)\|_2 = \mathcal{O}\left(\sqrt{k}\right)$ with probability $1-\exp(-c_3\log^2 N)$ over V. Thus, taking the intersection between these two events, we have

$$\mathbb{E}_{z_1}\left[\|\varphi(z_1)\|_2\right] = \mathcal{O}\left(\sqrt{k}\right), \tag{93}$$

with probability at least $1-\exp(-c_4\log^2 N)$ over z_1 and V. As this statement is independent of z_1, it holds with the same probability just over the probability space of V. Then, by Jensen inequality, we have

$$\|\nu\|_2 = \left\|\mathbb{E}_{z_1}\left[\varphi(z_1)\right]\right\|_2 \leq \mathbb{E}_{z_1}\left[\|\varphi(z_1)\|_2\right] = \mathcal{O}\left(\sqrt{k}\right). \tag{94}$$

We can now rewrite the LHS of the first statement as

$$\begin{aligned}
&\left|\left\|P^{\perp}_{\Phi_{-1}}\varphi(z_1)\right\|_2^2 - \mathbb{E}_{z_1}\left[\left\|P^{\perp}_{\Phi_{-1}}\varphi(z_1)\right\|_2^2\right]\right| \\
&= \left|\left\|P^{\perp}_{\Phi_{-1}}\left(\hat{\varphi}(z_1)+\nu\right)\right\|_2^2 - \mathbb{E}_{z_1}\left[\left\|P^{\perp}_{\Phi_{-1}}\left(\hat{\varphi}(z_1)+\nu\right)\right\|_2^2\right]\right| \\
&= \left|\hat{\varphi}(z_1)^\top P^{\perp}_{\Phi_{-1}}\hat{\varphi}(z_1) + 2\nu^\top P^{\perp}_{\Phi_{-1}}\hat{\varphi}(z_1) - \mathbb{E}_{z_1}\left[\hat{\varphi}(z_1)^\top P^{\perp}_{\Phi_{-1}}\hat{\varphi}(z_1)\right]\right| \\
&\leq \left|\hat{\varphi}(z_1)^\top P^{\perp}_{\Phi_{-1}}\hat{\varphi}(z_1) - \mathbb{E}_{z_1}\left[\hat{\varphi}(z_1)^\top P^{\perp}_{\Phi_{-1}}\hat{\varphi}(z_1)\right]\right| + 2\left|\nu^\top P^{\perp}_{\Phi_{-1}}\hat{\varphi}(z_1)\right|.
\end{aligned} \tag{95}$$

The second term is the inner product between $\hat{\varphi}(z_1)$, a mean-0 sub-Gaussian vector (in the probability space of z_1) such that $\|\hat{\varphi}(z_1)\|_{\psi_2} = \mathcal{O}\left(\sqrt{k/d}+1\right)$, and the independent vector $P^{\perp}_{\Phi_{-1}}\nu$, such that $\left\|P^{\perp}_{\Phi_{-1}}\nu\right\|_2 \leq \|\nu\|_2 = \mathcal{O}\left(\sqrt{k}\right)$, because of (94). Thus, by AssumptionF.1, we have that

$$\left|\nu^\top P^{\perp}_{\Phi_{-1}}\hat{\varphi}(z_1)\right| = o(k), \tag{96}$$

with probability at least $1-\exp(-c_5\log^2 N)$ over z_1 and V. Then, as $\left(\sqrt{k/d}+1\right)^{-1}\hat{\varphi}(z_1)$ is a mean-0, Lipschitz concentrated random vector (in the probability space of z_1), by the general version

of the Hanson-Wright inequality given by Theorem 2.3 in [2], we can write

$$\begin{aligned}
&\mathbb{P}\left(\left|\left\|P_{\Phi_{-1}}^{\perp}\hat{\varphi}(z_1)\right\|_2^2 - \mathbb{E}_{z_1}\left[\left\|P_{\Phi_{-1}}^{\perp}\hat{\varphi}(z_1)\right\|_2^2\right]\right| \geq k/\log N\right) \\
&\quad\leq 2\exp\left(-c_6\min\left(\frac{k^2}{\log^2 N\left((k/d)^2+1\right)\left\|P_{\Phi_{-1}}^{\perp}\right\|_F^2}, \frac{k}{\log N\,(k/d+1)\left\|P_{\Phi_{-1}}^{\perp}\right\|_{\text{op}}}\right)\right) \\
&\quad\leq 2\exp\left(-c_6\min\left(\frac{k}{\log^2 N\left((k/d)^2+1\right)}, \frac{k}{\log N\,(k/d+1)}\right)\right) \\
&\quad\leq \exp\left(-c_7\log^2 N\right),
\end{aligned} \tag{97}$$

where the last inequality comes from Assumption F.1.This, together with (95) and (96), proves the first part of the statement.

For the second part of the statement, we have

$$\begin{aligned}
&\left|\varphi(z_1^m)^\top P_{\Phi_{-1}}^{\perp}\varphi(z_1) - \mathbb{E}_{z_1,z_1^m}\left[\varphi(z_1^m)^\top P_{\Phi_{-1}}^{\perp}\varphi(z_1)\right]\right| \\
&\leq \left|\hat{\varphi}(z_1^m)^\top P_{\Phi_{-1}}^{\perp}\hat{\varphi}(z_1) - \mathbb{E}_{z_1,z_1^m}\left[\hat{\varphi}(z_1^m)^\top P_{\Phi_{-1}}^{\perp}\hat{\varphi}(z_1)\right]\right| + \left|\nu^\top P_{\Phi_{-1}}^{\perp}\hat{\varphi}(z_1)\right| + \left|\nu^\top P_{\Phi_{-1}}^{\perp}\hat{\varphi}(z_1^m)\right|.
\end{aligned} \tag{98}$$

Following the same argument that led to (96), we obtain

$$\left|\nu^\top P_{\Phi_{-1}}^{\perp}\hat{\varphi}(z_1^m)\right| = o(k), \tag{99}$$

with probability at least $1-\exp(-c_8\log^2 N)$ over z_1^m and V. Let us set

$$P_2 := \frac{1}{2}\left(\begin{array}{c|c} 0 & P_{\Phi_{-1}}^{\perp} \\ \hline P_{\Phi_{-1}}^{\perp} & 0 \end{array}\right), \qquad V_2 := \left(\begin{array}{c|c|c} V_x & V_y & 0 \\ \hline 0 & V_y & V_x \end{array}\right), \tag{100}$$

and

$$\hat{\varphi}_2 := \phi\left(V_2[x_1,y_1,x]^\top\right) - \mathbb{E}_{x_1,y_1,x}\left[\phi\left(V_2[x_1,y_1,x]^\top\right)\right] \equiv [\hat{\varphi}(z_1),\hat{\varphi}(z_1^m)]^\top. \tag{101}$$

We have that $\|P_2\|_{\text{op}} \leq 1$, $\|P_2\|_F^2 \leq k$, $\|V_2\|_{\text{op}} \leq 2\|V_x\|_{\text{op}} + 2\|V_y\|_{\text{op}} = \mathcal{O}\left(\sqrt{k/d}+1\right)$, and that $[x_1,y_1,x]^\top$ is a Lipschitz concentrated random vector in the joint probability space of z_1 and z_1^m, which follows from applying Lemma E.2 twice. Also, we have

$$\hat{\varphi}(z_1^m)^\top P_{\Phi_{-1}}^{\perp}\hat{\varphi}(z_1) = \hat{\varphi}_2^\top P_2\hat{\varphi}_2. \tag{102}$$

Thus, as $\left(\sqrt{k/d}+1\right)^{-1}\hat{\varphi}_2$ is a mean-0, Lipschitz concentrated random vector (in the probability space of z_1 and z_1^m), again by the general version of the Hanson-Wright inequality given by Theorem 2.3 in [2], we can write

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{\varphi}_2^\top P_2\hat{\varphi}_2 - \mathbb{E}_{z_1,z_1^m}\left[\hat{\varphi}_2^\top P_2\hat{\varphi}_2\right]\right| \geq k/\log N\right) \\
&\quad\leq 2\exp\left(-c_9\min\left(\frac{k^2}{\log^2 N\left((k/d)^2+1\right)\|P_2\|_F^2}, \frac{k}{\log N\,(k/d+1)\|P_2\|_{\text{op}}}\right)\right) \\
&\quad\leq 2\exp\left(-c_9\min\left(\frac{k}{\log^2 N\left((k/d)^2+1\right)}, \frac{k}{\log N\,(k/d+1)}\right)\right) \\
&\quad\leq \exp\left(-c_{10}\log^2 N\right),
\end{aligned} \tag{103}$$

where the last inequality comes from Assumption F.1. This, together with (98), (96), (99), and (102), proves the second part of the statement, and therefore the desired result.

□

Finally, we are ready to give the proof of Theorem 4.2.

Proof of Theorem 4.2. We will prove the statement for the following definition of γ_{RF}, independent from z_1 and z_1^m,

$$\gamma_{\text{RF}} := \frac{\mathbb{E}_{z_1, z_1^m}\left[\varphi(z_1^m)^\top P^\perp_{\Phi_{-1}} \varphi(z_1)\right]}{\mathbb{E}_{z_1}\left[\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2\right]}. \tag{104}$$

By Lemma D.1 and F.4, we have

$$\left\|P^\perp_{\Phi_{-1}}\varphi(z)\right\|_2^2 = \Omega(k) \tag{105}$$

with probability at least $1 - \exp(-c_1 \log^2 N)$ over V, Z_{-1} and z. This, together with Lemma F.9, gives

$$\left| \frac{\varphi(z_1^m)^\top P^\perp_{\Phi_{-1}} \varphi(z_1)}{\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2} - \frac{\mathbb{E}_{z_1, z_1^m}\left[\varphi(z_1^m)^\top P^\perp_{\Phi_{-1}} \varphi(z_1)\right]}{\mathbb{E}_{z_1}\left[\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2\right]} \right| = o(1), \tag{106}$$

with probability at least $1 - \exp(-c_2 \log^2 N)$ over V, Z and x, which proves the first part of the statement.

The upper-bound on γ_{RF} can be obtained applying Cauchy-Schwarz twice

$$\begin{aligned}
\frac{\mathbb{E}_{z_1, z_1^m}\left[\varphi(z_1^m)^\top P^\perp_{\Phi_{-1}} \varphi(z_1)\right]}{\mathbb{E}_{z_1}\left[\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2\right]} &\leq \frac{\mathbb{E}_{z_1, z_1^m}\left[\left\|P^\perp_{\Phi_{-1}}\varphi(z_1^m)\right\|_2 \left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2\right]}{\mathbb{E}_{z_1}\left[\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2\right]} \\
&\leq \frac{\sqrt{\mathbb{E}_{z_1^m}\left[\left\|P^\perp_{\Phi_{-1}}\varphi(z_1^m)\right\|_2^2\right]}\sqrt{\mathbb{E}_{z_1}\left[\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2\right]}}{\mathbb{E}_{z_1}\left[\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2\right]} = 1.
\end{aligned} \tag{107}$$

Let's now focus on the lower bound. By AssumptionF.1 and Lemma E.4 (in which we consider the degenerate case $\alpha = 1$ and set $t = N$), we have

$$\left| \left\|\tilde{\phi}\left(Vz_1\right)\right\|_2^2 - k\sum_{l=1}^{+\infty} \mu_l^2 \right| = o(k), \tag{108}$$

with probability at least $1 - \exp(-c_3 \log^2 N)$ over V and z_1. Then, a few applications of the triangle inequality give

$$\begin{aligned}
\frac{\mathbb{E}_{z_1, z_1^m}\left[\varphi(z_1^m)^\top P^\perp_{\Phi_{-1}} \varphi(z_1)\right]}{\mathbb{E}_{z_1}\left[\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2\right]} &\geq \frac{\varphi(z_1^m)^\top P^\perp_{\Phi_{-1}} \varphi(z_1)}{\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2} - o(1) \\
&\geq \frac{\varphi(z_1^m)^\top P^\perp_{\Phi_{-1}} \varphi(z_1)}{\left\|P^\perp_{\Phi_{-1}}\tilde{\varphi}(z_1)\right\|_2^2} - o(1) \\
&\geq \frac{k\left(\sum_{l=2}^{+\infty}\mu_l^2\alpha^l\right) + \mu_1^2\left\|P^\perp_{\Phi_{-1}}V_y y_1\right\|_2^2}{\|\tilde{\varphi}(z_1)\|_2^2} - o(1) \\
&\geq \frac{k\left(\sum_{l=2}^{+\infty}\mu_l^2\alpha^l\right) + \mu_1^2\left\|P^\perp_{\Phi_{-1}}V_y y_1\right\|_2^2}{k\sum_{l=1}^{+\infty}\mu_l^2} - o(1) \\
&\geq \frac{\sum_{l=2}^{+\infty}\mu_l^2\alpha^l}{\sum_{l=1}^{+\infty}\mu_l^2} - o(1),
\end{aligned} \tag{109}$$

where the first inequality is a consequence of (106), the second of Lemma F.5 and (105), the third of Lemma F.8 and again (105), and the fourth of (108), and they jointly hold with probability $1 - \exp(-c_4 \log^2 N)$ over V, Z_{-1} and z_1. Again, as the statement does not depend on z_1, we can conclude that it holds with the same probability only over the probability spaces of V and Z_{-1}, and the thesis readily follows. □

G Proofs for NTK Regression

In this section, we will indicate with $Z \in \mathbb{R}^{N \times d}$ the data matrix, such that its rows are sampled independently from $\mathcal{P}_Z$ (see Assumption 4.1). We denote by $W \in \mathbb{R}^{k \times d}$ the weight matrix at initialization, such that $W_{ij} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$. Thus, the feature map is given by (see Section 4)

$$\varphi(z) := z \otimes \phi'(Wz) \in \mathbb{R}^{dk}, \tag{110}$$

where ϕ' is the derivative of the activation function ϕ, applied component-wise to the vector Wz. We use the shorthands $\Phi := Z * \phi'(ZW^\top) \in \mathbb{R}^{N \times p}$ and $K := \Phi\Phi^\top \in \mathbb{R}^{N \times N}$, where $*$ denotes the Khatri-Rao product, defined in Appendix C. We indicate with $\Phi_{-1} \in \mathbb{R}^{(N-1) \times k}$ the matrix Φ without the first row, and we define $K_{-1} := \Phi_{-1}\Phi_{-1}^\top$. We call P_Φ the projector over the span of the rows of Φ, and $P_{\Phi_{-1}}$ the projector over the span of the rows of Φ_{-1}. We use the notations $\tilde{\varphi}(z) := \varphi(z) - \mathbb{E}_W[\varphi(z)]$ and $\tilde{\Phi}_{-1} := \Phi_{-1} - \mathbb{E}_W[\Phi_{-1}]$ to indicate the centered feature map and matrix respectively, where the centering is with respect to W. We indicate with μ_l' the l-th Hermite coefficient of ϕ'. We use the notation $z_1^m = [x, y_1]$, where $x \sim \mathcal{P}_X$ is sampled independently from V and Z. We define $\alpha = d_y/d$. Throughout this section, for compactness, we drop the subscripts "NTK" from these quantities, as we will only treat the proofs related to the NTK Regression model. Again for the sake of compactness, we will not re-introduce such quantities in the statements or the proofs of the following lemmas.

Through the following Section, as mentioned in the main body of the paper, we will work under the following assumptions

Assumption G.1 (Over-parameterization and topology)**.**

$$N \log^8 N = o(kd), \qquad N > d, \qquad k = \mathcal{O}(d). \tag{111}$$

The first condition is the smallest (up to $\log$ factors) over-parameterization that guarantees interpolation [13]. The second condition is rather mild (it is easily satisfied by standard datasets) and purely technical. The third condition is required to lower bound the smallest eigenvalue of the kernel induced by the feature map, and a stronger requirement, *i.e.*, the strict inequality $k < d$, has appeared in prior work [34, 35, 36].

Assumption G.2 (Activation function)**.** *The activation function ϕ is a non-linear function with L-Lipschitz first order derivative ϕ'.*

This requirements is satisfied by common activations, e.g. smoothed ReLU, sigmoid, or $\tanh$.

Summary of this Section.

- In Lemma G.3, we prove the lower bound on the smallest eigenvalue of K, adapting to our settings the main result of [13].
- In Lemma G.7, we treat separately a term that derives from $\mathbb{E}_W[\phi'(Wz)] = \mu_0' \mathbf{1}_k$, showing that we can *center* the derivative of the activation function (Lemma G.11), without changing our final statement in Theorem 4.3. This step is necessary only if $\mu_0' \neq 0$. Our proof tackles the problem proving the thesis on a set of "perturbed" inputs $\tilde{Z}_{-1}(\delta)$ (Lemma G.6), critically exploiting the non degenerate behaviour of their rows (Lemma G.5), and transfers the result on the original term, using continuity arguments with respect to the perturbation (Lemma G.4).
- In Lemma G.10, we show that the centered features $\tilde{\varphi}(z_1)$ and $\tilde{\varphi}(z_1^m)$ have a negligible component in the space spanned by the rows of Φ_{-1}. To achieve this, we exploit the bound proved in Lemma G.9.
- To conclude, we prove Theorem 4.3, exploiting also the concentration result provided in Lemma G.8.

Lemma G.3. *We have that*

$$\lambda_{\min}(K) = \Omega(kd), \tag{112}$$

with probability at least $1 - Ne^{-c\log^2 k} - e^{-c\log^2 N}$ *over* Z *and* W*, where* c *is an absolute constant.*

Proof. The result follows from Theorem 3.1 of [13]. Notice that our assumptions on the data distribution $\mathcal{P}_Z$ are stronger, and that our initialization of the very last layer (which differs from the Gaussian initialization in [13]) does not change the result. Assumption G.1, *i.e.*, $k = \mathcal{O}(d)$, satisfies the *loose pyramidal topology* condition (cf. Assumption 2.4 in [13]), and Assumption G.1 is the same as Assumption 2.5 in [13]. An important difference is that we do not assume the activation function ϕ to be Lipschitz anymore. This, however, stops being a necessary assumption since we are working with a 2-layer neural network, and ϕ doesn't appear in the expression of NTK. □

Lemma G.4. *Let* $A \in \mathbb{R}^{(N-1)\times d}$ *be a generic matrix, and let* $\bar{Z}_{-1}(\delta)$ *and* $\bar{\Phi}_{-1}(\delta)$ *be defined as*

$$\bar{Z}_{-1}(\delta) := Z_{-1} + \delta A, \tag{113}$$

$$\bar{\Phi}_{-1}(\delta) := \bar{Z}_{-1}(\delta) * \phi'\left(Z_{-1}W^\top\right). \tag{114}$$

Let $\bar{P}_{\Phi_{-1}}(\delta) \in \mathbb{R}^{dk\times dk}$ *be the projector over the Span of the rows of* $\bar{\Phi}_{-1}(\delta)$*. Then, we have that* $\bar{P}^{\perp}_{\Phi_{-1}}(\delta)$ *is continuous in* $\delta = 0$ *with probability at least* $1 - Ne^{-c\log^2 k} - e^{-c\log^2 N}$ *over* Z *and* W*, where* c *is an absolute constant and where the continuity is with respect to* $\|\cdot\|_{\text{op}}$*.*

Proof. In this proof, when we say that a matrix is continuous with respect to δ, we always intend with respect to the operator norm $\|\cdot\|_{\text{op}}$. Then, $\bar{\Phi}_{-1}(\delta)$ is continuous in 0, as

$$\left\|\bar{\Phi}_{-1}(\delta) - \bar{\Phi}_{-1}(0)\right\|_{\text{op}} = \left\|\delta A * \phi'\left(Z_{-1}W^\top\right)\right\|_{\text{op}} \le \delta \|A\|_{\text{op}} \max_{2\le i\le N} \|\phi'(Wz_i)\|_2, \tag{115}$$

where the second step follows from Equation (3.7.13) in [30].

By Weyl's inequality, this also implies that $\lambda_{\min}\left(\bar{\Phi}_{-1}(\delta)\bar{\Phi}_{-1}(\delta)^\top\right)$ is continuous in $\delta = 0$. Recall that, by Lemma G.3, $\det\left(\bar{\Phi}_{-1}(0)\bar{\Phi}_{-1}(0)^\top\right) \equiv \det\left(\Phi_{-1}\Phi_{-1}^\top\right) \neq 0$ with probability at least $1 - Ne^{-c\log^2 k} - e^{-c\log^2 N}$ over Z and W. This implies that $\left(\bar{\Phi}_{-1}(\delta)\bar{\Phi}_{-1}(\delta)^\top\right)^{-1}$ is also continuous, as for every invertible matrix M we have $M^{-1} = \text{Adj}(M)/\det(M)$ (where $\text{Adj}(M)$ denotes the Adjugate of the matrix M), and both $\text{Adj}(\cdot)$ and $\det(\cdot)$ are continuous mappings. Thus, as $\bar{P}_{\Phi_{-1}}(0) = \bar{\Phi}_{-1}(0)^\top\left(\bar{\Phi}_{-1}(0)\bar{\Phi}_{-1}(0)^\top\right)^{-1}\bar{\Phi}_{-1}(0)$ (see (20)), we also have the continuity of $\bar{P}_{\Phi_{-1}}(\delta)$ in $\delta = 0$, which gives the thesis. □

Lemma G.5. *Let* $A \in \mathbb{R}^{(N-1)\times d}$ *be a matrix with entries sampled independently (between each other and from everything else) from a standard Gaussian distribution. Then, for every* $\delta > 0$*, with probability 1 over* A*, the rows of* $\bar{Z}_{-1} := Z_{-1} + \delta A$ *span* $\mathbb{R}^d$*.*

Proof. As $N - 1 \ge d$, by Assumption G.1,negating the thesis would imply that the rows of $\bar{Z}_{-1}$ are linearly dependent, and that they belong to a subspace with dimension at most $d - 1$. This would imply that there exists a row of $\bar{Z}_{-1}$, call it $\bar{z}_j$, such that $\bar{z}_j$ belongs to the space spanned by all the other rows of $\bar{Z}_{-1}$, with dimension at most $d - 1$. This means that $A_{j:}$ has to belong to an affine space with the same dimension, which we can consider fixed, as it's not a function of the random vector $A_{j:}$, but only of Z_{-1} and $\{A_{i:}\}_{i\neq j}$. As the entries of $A_{j:}$ are sampled independently from a standard Gaussian distribution, this happens with probability 0. □

Lemma G.6. *Let* $A \in \mathbb{R}^{(N-1)\times d}$ *be a matrix with entries sampled independently (between each other and from everything else) from a standard Gaussian distribution. Let* $\bar{Z}_{-1}(\delta) := Z_{-1} + \delta A$ *and* $\bar{\Phi}_{-1}(\delta) := \bar{Z}_{-1}(\delta) * \phi'\left(Z_{-1}W^\top\right)$*. Let* $\bar{P}_{\Phi_{-1}}(\delta) \in \mathbb{R}^{dk\times dk}$ *be the projector over the Span of the rows of* $\bar{\Phi}_{-1}(\delta)$*. Let* $\mu_0' \neq 0$*. Then, for* $z \sim \mathcal{P}_Z$*, and for any* $\delta > 0$*, we have,*

$$\left\|\bar{P}^{\perp}_{\Phi_{-1}}(\delta)\left(z \otimes \mathbf{1}_k\right)\right\|_2 = o(\sqrt{dk}), \tag{116}$$

with probability at least $1 - \exp(-c\log^2 N)$ *over* Z*,* W*, and* A*, where* c *is an absolute constant.*

Proof. Let $B_{-1} := \phi'(Z_{-1}W^\top) \in \mathbb{R}^{(N-1)\times k}$. Notice that, for any $\zeta \in \mathbb{R}^{N-1}$, the following identity holds

$$\bar{\Phi}_{-1}^\top(\delta)\zeta = \left(\bar{Z}_{-1}(\delta) * B_{-1}\right)^\top \zeta = \left(\bar{Z}_{-1}^\top(\delta)\zeta\right) \otimes \left(B_{-1}^\top \mathbf{1}_{N-1}\right). \tag{117}$$

Note that $B_{-1}^\top = \mu_0' \mathbf{1}_k \mathbf{1}_{N-1}^\top + \tilde{B}_{-1}^\top$, where $\tilde{B}_{-1}^\top = \phi'(WZ_{-1}^\top) - \mathbb{E}_W\left[\phi'(WZ_{-1}^\top)\right]$ is a $k \times (N-1)$ matrix with i.i.d. and mean-0 rows. For an argument equivalent to the one used for (67) and (68), we have

$$\left\|\tilde{B}_{-1}^\top\right\|_{\text{op}} = \mathcal{O}\left(\left(\sqrt{k}+\sqrt{N}\right)\left(\sqrt{N/d}+1\right)\right), \tag{118}$$

with probability at least $1 - \exp(-c\log^2 N)$ over Z_{-1} and W. Thus, we can write

$$B_{-1}^\top \frac{\mathbf{1}_{N-1}}{\mu_0'(N-1)} = \left(\mu_0' \mathbf{1}_k \mathbf{1}_{N-1}^\top + \tilde{B}_{-1}^\top\right)\frac{\mathbf{1}_{N-1}}{\mu_0'(N-1)} = \mathbf{1}_k + \tilde{B}_{-1}^\top \frac{\mathbf{1}_{N-1}}{\mu_0'(N-1)} =: \mathbf{1}_k + v, \tag{119}$$

where we have

$$\|v\|_2 \le \left\|\tilde{B}_{-1}^\top\right\|_{\text{op}} \left\|\frac{\mathbf{1}_{N-1}}{\mu_0'(N-1)}\right\|_2 = \mathcal{O}\left(\left(\sqrt{k/N}+1\right)\left(\sqrt{N/d}+1\right)\right) = o(\sqrt{k}), \tag{120}$$

where the last step is a consequence of Assumption G.1. Plugging (119) in (117) we get

$$\frac{1}{\mu_0'(N-1)}\bar{\Phi}_{-1}^\top(\delta)\zeta = \frac{1}{\mu_0'(N-1)}\left(\bar{Z}_{-1}(\delta) * B_{-1}\right)^\top \zeta = \left(\bar{Z}_{-1}^\top(\delta)\zeta\right) \otimes (\mathbf{1}_k + v). \tag{121}$$

By Lemma G.5, we have that the rows of $\bar{Z}_{-1}(\delta)$ span $\mathbb{R}^d$, with probability 1 over A. Thus, conditioning on this event, we can set ζ to be a vector such that $z = \bar{Z}_{-1}^\top(\delta)\zeta$. We can therefore rewrite the previous equation as

$$\frac{1}{\mu_0'(N-1)}\bar{\Phi}_{-1}^\top(\delta)\zeta = z \otimes \mathbf{1}_k + z \otimes v. \tag{122}$$

Thus, we can conclude

$$\begin{aligned}
\left\|\bar{P}_{\bar{\Phi}_{-1}}^\perp(\delta)\left(z \otimes \mathbf{1}_k\right)\right\|_2 &= \left\|P_{\bar{\Phi}_{-1}}^\perp\left(\frac{\bar{\Phi}_{-1}^\top(\delta)\zeta}{\mu_0'(N-1)} - z \otimes v\right)\right\|_2 \\
&\le \left\|\bar{P}_{\bar{\Phi}_{-1}}^\perp(\delta)\bar{\Phi}_{-1}^\top(\delta)\frac{\zeta}{\mu_0'(N-1)}\right\|_2 + \|z \otimes v\|_2 \\
&= \|z\|_2 \|v\|_2 = o(\sqrt{dk}),
\end{aligned} \tag{123}$$

where in the second step we use the triangle inequality, in the third step we use that $\bar{\Phi}_{-1}^\top(\delta) = \bar{P}_{\bar{\Phi}_{-1}}(\delta)\bar{\Phi}_{-1}^\top(\delta)$, and in the last step we use (120). The desired result readily follows. □

Lemma G.7. *Let $\mu_0' \neq 0$. Then, for any $z \in \mathbb{R}^d$, we have,*

$$\left\|P_{\bar{\Phi}_{-1}}^\perp\left(z \otimes \mathbf{1}_k\right)\right\|_2 = o(\sqrt{dk}), \tag{124}$$

with probability at least $1 - Ne^{-c\log^2 k} - e^{-c\log^2 N}$ over Z and W, where c is an absolute constant.

Proof. Let $A \in \mathbb{R}^{(N-1)\times d}$ be a matrix with entries sampled independently (between each other and from everything else) from a standard Gaussian distribution. Let $\bar{Z}_{-1}(\delta) := Z_{-1} + \delta A$ and $\bar{\Phi}_{-1}(\delta) := \bar{Z}_{-1}(\delta) * \phi'\left(Z_{-1}W^\top\right)$. Let $\bar{P}_{\bar{\Phi}_{-1}}(\delta) \in \mathbb{R}^{dk\times dk}$ be the projector over the Span of the rows of $\bar{\Phi}_{-1}(\delta)$.

By triangle inequality, we can write

$$\left\|P_{\bar{\Phi}_{-1}}^\perp\left(z \otimes \mathbf{1}_k\right)\right\|_2 \le \left\|P_{\bar{\Phi}_{-1}}^\perp - \bar{P}_{\bar{\Phi}_{-1}}^\perp(\delta)\right\|_{\text{op}} \|z \otimes \mathbf{1}_k\|_2 + \left\|\bar{P}_{\bar{\Phi}_{-1}}^\perp(\delta)\left(z \otimes \mathbf{1}_k\right)\right\|_2. \tag{125}$$

Because of Lemma G.4, with probability at least $1 - Ne^{-c\log^2 k} - e^{-c\log^2 N}$ over Z and W, $\bar{P}^{\perp}_{\Phi_{-1}}(\delta)$ is continuous in $\delta = 0$, with respect to $\|\cdot\|_{\text{op}}$. Thus, there exists $\delta^* > 0$ such that, for every $\delta \in [0, \delta^*]$,

$$\left\| P^{\perp}_{\Phi_{-1}} - \bar{P}^{\perp}_{\Phi_{-1}}(\delta) \right\|_{\text{op}} \equiv \left\| \bar{P}^{\perp}_{\Phi_{-1}}(0) - \bar{P}^{\perp}_{\Phi_{-1}}(\delta) \right\|_{\text{op}} < \frac{1}{N}. \tag{126}$$

Hence, setting $\delta = \delta^*$ in (125), we get

$$\begin{aligned} \left\| P^{\perp}_{\Phi_{-1}}(z \otimes \mathbf{1}_k) \right\|_2 &\leq \left\| P^{\perp}_{\Phi_{-1}} - \bar{P}^{\perp}_{\Phi_{-1}}(\delta^*) \right\|_{\text{op}} \|z \otimes \mathbf{1}_k\|_2 + \left\| \bar{P}^{\perp}_{\Phi_{-1}}(\delta^*)(z \otimes \mathbf{1}_k) \right\|_2 \\ &\leq \|z\|_2 \|\mathbf{1}_k\|_2 / N + \left\| \bar{P}^{\perp}_{\Phi_{-1}}(\delta^*)(z \otimes \mathbf{1}_k) \right\|_2 \\ &= o(\sqrt{dk}), \end{aligned} \tag{127}$$

where the last step is a consequence of Lemma G.6, and it holds with probability at least $1 - \exp(-c\log^2 N)$ over Z, W, and A. As the LHS of the previous equation doesn't depend on A, the statements holds with the same probability, just over the probability spaces of Z and W, which gives the desired result. □

Lemma G.8. *We have*

$$\left| \frac{\tilde{\varphi}(z_1^m)^\top \tilde{\varphi}(z_1)}{\|\tilde{\varphi}(z_1)\|_2^2} - \alpha \frac{\sum_{l=1}^{+\infty} \mu_l'^2 \alpha^i}{\sum_{l=1}^{+\infty} \mu_l'^2} \right| = o(1), \tag{128}$$

with probability at least $1 - \exp(-c\log^2 N) - \exp(-c\log^2 k)$ *over* W *and* z_1*, where* c *is an absolute constant. With the same probability, we also have*

$$\tilde{\varphi}(z_1^m)^\top \tilde{\varphi}(z_1) = \Theta(dk), \qquad \|\tilde{\varphi}(z_1)\|_2^2 = \Theta(dk). \tag{129}$$

Proof. We have

$$\|\tilde{\varphi}(z_1)\|_2^2 = \left\| z_1 \otimes \tilde{\phi}'(Wz_1) \right\|_2^2 = \|z_1\|_2^2 \left\| \tilde{\phi}'(Wz_1) \right\|_2^2 = d \left\| \tilde{\phi}'(Wz_1) \right\|_2^2. \tag{130}$$

By Assumption G.1 and Lemma E.4 (in which we consider the degenerate case $\alpha = 1$ and set $t = k$), we have

$$\left| \left\| \tilde{\phi}'(Wz_1) \right\|_2^2 - k \sum_{l=1}^{+\infty} \mu_l'^2 \right| = o(k), \tag{131}$$

with probability at least $1 - \exp(-c\log^2 k)$ over W and z_1. Thus, we have

$$\left| \|\tilde{\varphi}(z_1)\|_2^2 - dk \sum_{l=1}^{+\infty} \mu_l'^2 \right| = o(dk). \tag{132}$$

Notice that the second term in the modulus is $\Theta(dk)$, since the μ_l'-s cannot be all 0, because of Assumption G.2; this shows that $\|\tilde{\varphi}(z_1)\|_2^2 = \Theta(dk)$.

Similarly, we can write

$$\tilde{\varphi}(z_1^m)^\top \tilde{\varphi}(z_1) = \left(z_1^\top z_1^m \right) \left(\tilde{\phi}'(Wz_1)^\top \tilde{\phi}'(Wz_1^m) \right). \tag{133}$$

We have

$$\left| z_1^\top z_1^m - \alpha d \right| = \left| x_1^\top x \right| \leq \sqrt{d_x} \log d = o(d), \tag{134}$$

where the inequality holds with probability at least $1 - \exp(-c_1 \log^2 d) \geq 1 - \exp(-c_2 \log^2 N)$ over x_1, as we are taking the inner product of two independent and sub-Gaussian vectors with norm $\sqrt{d_x}$. Furthermore, again by AssumptionG.1 and Lemma E.4, we have

$$\left| \tilde{\phi}'(Wz_1)^\top \tilde{\phi}'(Wz_1^m) - k \sum_{l=1}^{+\infty} \mu_l'^2 \alpha^l \right| = o(k), \tag{135}$$

with probability at least $1 - \exp(-c_3 \log^2 k)$ over W and z_1. Notice that the second term in the modulus is $\Theta(k)$, because of Assumption G.2.

Thus, putting (133), (134) and (135) together, we get

$$\left|\tilde{\varphi}(z_1^m)^\top \tilde{\varphi}(z_1) - dk\alpha \sum_{l=1}^{+\infty} {\mu_l'}^2 \alpha^l\right| = o(dk), \tag{136}$$

with probability at least $1 - \exp(-c_3 \log^2 k) - \exp(-c_2 \log^2 N)$ over W and z_1; this shows that $\tilde{\varphi}(z_1^m)^\top \tilde{\varphi}(z_1) = \Theta(dk)$.

Finally, merging (136) with (132) and applying triangle inequality, (128) follows and the proof is complete. □

Lemma G.9. *Let $z \sim \mathcal{P}_Z$ be sampled independently from Z_{-1}. Then,*

$$\|\Phi_{-1}\tilde{\varphi}(z)\|_2 = o\left(dk\right), \tag{137}$$

with probability at least $1 - \exp(-c \log^2 N)$ over W and z, where c is an absolute constant.

Proof. Let's look at the i-th entry of the vector $\Phi_{-1}\tilde{\varphi}(z)$, *i.e.*,

$$\varphi(z_{i+1})^\top \tilde{\varphi}(z) = \left(z_{i+1}^\top z\right)\left(\phi'(Wz_{i+1})^\top \tilde{\phi}'(Wz)\right). \tag{138}$$

As z and z_{i+1} are sub-Gaussian and independent with norm $\sqrt{d}$, we can write $\left|z^\top z_{i+1}\right| = \mathcal{O}\left(\sqrt{d}\log N\right)$ with probability at least $1 - \exp(-c\log^2 N)$ over z. We will condition on such high probability event until the end of the proof.

By Lemma E.3, setting $t = N$, we have

$$\left|\phi'(Wz_{i+1})^\top \tilde{\phi}'(Wz) - \mathbb{E}_W\left[\phi'(Wz_{i+1})^\top \tilde{\phi}'(Wz)\right]\right| = \mathcal{O}\left(\sqrt{k}\log N\right), \tag{139}$$

with probability at least $1 - \exp(-c_1 \log^2 N)$ over W. Exploiting the Hermite expansion of ϕ' and $\tilde{\phi}'$, we have

$$\begin{aligned}
\left|\mathbb{E}_W\left[\phi'(Wz_{i+1})^\top \tilde{\phi}'(Wz)\right]\right| &\leq k \sum_{l=1}^{+\infty} {\mu_l'}^2 \left(\frac{\left|z_{i+1}^\top z\right|}{\|z_{i+1}\|_2 \|z\|_2}\right)^l \\
&\leq k \max_l {\mu_l'}^2 \sum_{l=1}^{+\infty} \left(\frac{\left|z_{i+1}^\top z\right|}{\|z_{i+1}\|_2 \|z\|_2}\right)^l \\
&= k \max_l {\mu_l'}^2 \frac{\left|z_{i+1}^\top z\right|}{\|z_{i+1}\|_2 \|z\|_2} \frac{1}{1 - \frac{\left|z_{i+1}^\top z\right|}{\|z_{i+1}\|_2 \|z\|_2}} \\
&\leq 2k \max_l {\mu_l'}^2 \frac{\left|z_{i+1}^\top z\right|}{\|z_{i+1}\|_2 \|z\|_2} = \mathcal{O}\left(\frac{k\log N}{\sqrt{d}}\right).
\end{aligned} \tag{140}$$

Putting together (139) and (140), and applying triangle inequality, we get

$$\left|\phi'(Wz_{i+1})^\top \tilde{\phi}'(Wz)\right| = \mathcal{O}\left(\sqrt{k}\log N + \frac{k\log N}{\sqrt{d}}\right) = \mathcal{O}\left(\sqrt{k}\log N\right), \tag{141}$$

where the last step is a consequence of Assumption G.1.Comparing this last result with (138), we obtain

$$\left|\varphi(z_{i+1})^\top \tilde{\varphi}(z)\right| = \mathcal{O}\left(\sqrt{dk}\log^2 N\right), \tag{142}$$

with probability at least $1 - \exp(-c_2 \log^2 N)$ over W and z.

We want the previous equation to hold for all $1 \leq i \leq N-1$. Performing a union bound, we have that this is true with probability at least $1 - (N-1)\exp(-c_2\log^2 N) \geq 1 - \exp(-c_3 \log^2 N)$ over W and z. Thus, with such probability, we have

$$\begin{aligned}
\|\Phi_{-1}\tilde{\varphi}(z)\|_2 &\leq \sqrt{N-1}\max_i \left|\varphi(z_{i+1})^\top \tilde{\varphi}(z)\right| \\
&= \mathcal{O}\left(\sqrt{dk}\sqrt{N}\log^2 N\right) = o(dk),
\end{aligned} \tag{143}$$

where the last step follows from Assumption G.1. □

Lemma G.10. *We have*

$$\left|\frac{\tilde{\varphi}(z_1^m)^\top \tilde{\varphi}(z_1) - \tilde{\varphi}(z_1^m)^\top P_{\Phi_{-1}} \tilde{\varphi}(z_1)}{\left\|\tilde{\varphi}(z_1) - P_{\Phi_{-1}}\tilde{\varphi}(z_1)\right\|_2^2} - \frac{\tilde{\varphi}(z_1^m)^\top \tilde{\varphi}(z_1)}{\|\tilde{\varphi}(z_1)\|_2^2}\right| = o(1), \tag{144}$$

with probability at least $1 - N\exp(-c\log^2 k) - \exp(-c\log^2 N)$ *over* Z, x *and* W, *where* c *is an absolute constant. With the same probability, we also have*

$$\tilde{\varphi}(z_1^m)^\top \tilde{\varphi}(z_1) - \tilde{\varphi}(z_1^m)^\top P_{\Phi_{-1}} \tilde{\varphi}(z_1) = \Theta(dk), \qquad \left\|\tilde{\varphi}(z_1) - P_{\Phi_{-1}}\tilde{\varphi}(z_1)\right\|_2^2 = \Theta(dk). \tag{145}$$

Proof. Notice that, with probability at least $1 - \exp(-c\log^2 N) - \exp(-c\log^2 k)$ over W and z_1, we have both

$$\tilde{\varphi}(z_1^m)^\top \tilde{\varphi}(z_1) = \Theta(dk) \qquad \|\tilde{\varphi}(z_1)\|_2^2 = \Theta(dk). \tag{146}$$

by the second statement of Lemma G.8. Furthermore,

$$\begin{aligned}\left|\tilde{\varphi}(z_1^m)^\top P_{\Phi_{-1}} \tilde{\varphi}(z_1)\right| &= \left|\tilde{\varphi}(z_1^m)^\top \Phi_{-1}^\top K_{-1}^{-1} \Phi_{-1}\tilde{\varphi}(z_1)\right| \\ &\leq \|\Phi_{-1}\tilde{\varphi}(z_1^m)\|_2 \, \lambda_{\min}\left(K_{-1}\right)^{-1} \|\Phi_{-1}\tilde{\varphi}(z_1)\|_2 \\ &= o(dk)\mathcal{O}\left(\frac{1}{dk}\right) o(dk) = o(dk),\end{aligned} \tag{147}$$

where the third step is justified by Lemmas G.3 and G.9, and holds with probability at least $1 - Ne^{-c\log^2 k} - e^{-c\log^2 N}$ over Z, x, and W. A similar argument can be used to show that $\left\|P_{\Phi_{-1}}\tilde{\varphi}(z_1)\right\|_2^2 = o(dk)$, which, together with (147) and (146), and a straightforward application of the triangle inequality, provides the thesis. □

Lemma G.11. *We have*

$$\left|\frac{\varphi(z_1^m)^\top P_{\Phi_{-1}}^\perp \varphi(z_1)}{\left\|P_{\Phi_{-1}}^\perp \varphi(z_1)\right\|_2^2} - \frac{\tilde{\varphi}(z_1^m)^\top P_{\Phi_{-1}}^\perp \tilde{\varphi}(z_1)}{\left\|P_{\Phi_{-1}}^\perp \tilde{\varphi}(z_1)\right\|_2^2}\right| = o(1), \tag{148}$$

with probability at least $1 - N\exp(-c\log^2 k) - \exp(-c\log^2 N)$ *over* Z, x *and* W, *where* c *is an absolute constant.*

Proof. If $\mu_0' = 0$, the thesis is trivial, as $\varphi \equiv \tilde{\varphi}$. If $\mu_0' \neq 0$, we can apply Lemma G.7, and the proof proceeds as follows.

First, we notice that the second term in the modulus in the statement corresponds to the first term in the statement of Lemma G.10. We will condition on the result of Lemma G.10 to hold until the end of the proof. Notice that this also implies

$$\tilde{\varphi}(z_1^m)^\top P_{\Phi_{-1}}^\perp \tilde{\varphi}(z_1) = \Theta(dk), \qquad \left\|P_{\Phi_{-1}}^\perp \tilde{\varphi}(z_1)\right\|_2^2 = \Theta(dk), \tag{149}$$

with probability at least $1 - N\exp(-c\log^2 k) - \exp(-c\log^2 N)$ over Z, x, and W. Due to Lemma G.7, we jointly have

$$\left\|P_{\Phi_{-1}}^\perp \left(z_1 \otimes \mathbf{1}_k\right)\right\|_2 = o(\sqrt{dk}), \qquad \left\|P_{\Phi_{-1}}^\perp \left(z_1^m \otimes \mathbf{1}_k\right)\right\|_2 = o(\sqrt{dk}), \tag{150}$$

with probability at least $1 - \exp(c\log^2 N)$ over Z_{-1} and W. Also, by Lemma E.3 and Assumption F.1, we jointly have

$$\left\|P_{\Phi_{-1}}^\perp \varphi(z_1^m)\right\|_2 \leq \|\varphi(z_1^m)\|_2 = \|z_1^m\|_2 \, \|\phi'(W z_1^m)\|_2 = \mathcal{O}\left(\sqrt{dk}\right), \tag{151}$$

and

$$\left\|P_{\Phi_{-1}}^\perp \tilde{\varphi}(z_1)\right\|_2 \leq \|\tilde{\varphi}(z_1)\|_2 = \|z_1\|_2 \left\|\tilde{\phi}'(W z_1)\right\|_2 = \mathcal{O}\left(\sqrt{dk}\right), \tag{152}$$

with probability at least $1-\exp(-c_1 \log^2 N)$ over W. We will condition also on such high probability events ((150), (151), (152)) until the end of the proof. Thus, we can write

$$\begin{aligned}
&\left|\varphi(z_1^m)^\top P_{\Phi_{-1}}^\perp \varphi(z_1) - \tilde{\varphi}(z_1^m)^\top P_{\Phi_{-1}}^\perp \tilde{\varphi}(z_1)\right| \\
&\leq \left|\varphi(z_1^m)^\top P_{\Phi_{-1}}^\perp \left(\varphi(z_1) - \tilde{\varphi}(z_1)\right)\right| + \left|\left(\varphi(z_1^m) - \tilde{\varphi}(z_1^m)\right)^\top P_{\Phi_{-1}}^\perp \tilde{\varphi}(z_1)\right| \\
&\leq \left\|P_{\Phi_{-1}}^\perp \varphi(z_1^m)\right\|_2 \left\|P_{\Phi_{-1}}^\perp \left(z_1 \otimes \mu_0 \mathbf{1}_k\right)\right\|_2 + \left\|P_{\Phi_{-1}}^\perp \tilde{\varphi}(z_1)\right\|_2 \left\|P_{\Phi_{-1}}^\perp \left(z_1^m \otimes \mu_0 \mathbf{1}_k\right)\right\|_2 = o(dk),
\end{aligned} \tag{153}$$

where in the last step we use (150), (151), and (152). Similarly, we can show that

$$\begin{aligned}
\left|\left\|P_{\Phi_{-1}}^\perp \varphi(z_1)\right\|_2 - \left\|P_{\Phi_{-1}}^\perp \tilde{\varphi}(z_1)\right\|_2\right| &\leq \left\|P_{\Phi_{-1}}^\perp \varphi(z_1) - P_{\Phi_{-1}}^\perp \tilde{\varphi}(z_1)\right\|_2 \\
&\leq \left\|P_{\Phi_{-1}}^\perp \left(z_1 \otimes \mu_0 \mathbf{1}_k\right)\right\|_2 = o(\sqrt{dk}).
\end{aligned} \tag{154}$$

By combining (149), (153), and (154), the desired result readily follows. □

Finally, we are ready to give the proof of Theorem 4.3.

Proof of Theorem 4.3. We have

$$\begin{aligned}
\left|\frac{\varphi(z_1^m)^\top P_{\Phi_{-1}}^\perp \varphi(z_1)}{\left\|P_{\Phi_{-1}}^\perp \varphi(z_1)\right\|_2^2} - \alpha \frac{\sum_{l=1}^{+\infty} {\mu_l'}^2 \alpha^i}{\sum_{l=1}^{+\infty} {\mu_l'}^2}\right| &\leq \left|\frac{\varphi(z_1^m)^\top P_{\Phi_{-1}}^\perp \varphi(z_1)}{\left\|P_{\Phi_{-1}}^\perp \varphi(z_1)\right\|_2^2} - \frac{\tilde{\varphi}(z_1^m)^\top P_{\Phi_{-1}}^\perp \tilde{\varphi}(z_1)}{\left\|P_{\Phi_{-1}}^\perp \tilde{\varphi}(z_1)\right\|_2^2}\right| \\
&+ \left|\frac{\tilde{\varphi}(z_1^m)^\top \tilde{\varphi}(z_1) - \tilde{\varphi}(z_1^m)^\top P_{\Phi_{-1}} \tilde{\varphi}(z_1)}{\left\|\tilde{\varphi}(z_1) - P_{\Phi_{-1}} \tilde{\varphi}(z_1)\right\|_2^2} - \frac{\tilde{\varphi}(z_1^m)^\top \tilde{\varphi}(z_1)}{\left\|\tilde{\varphi}(z_1)\right\|_2^2}\right| \\
&+ \left|\frac{\tilde{\varphi}(z_1^m)^\top \tilde{\varphi}(z_1)}{\left\|\tilde{\varphi}(z_1)\right\|_2^2} - \alpha \frac{\sum_{l=1}^{+\infty} {\mu_l'}^2 \alpha^i}{\sum_{l=1}^{+\infty} {\mu_l'}^2}\right| \\
&= o(1),
\end{aligned} \tag{155}$$

where the first step is justified by the triangle inequality, and the second by Lemmas G.11, G.10, and G.8, and it holds with probability at least $1 - N\exp(-c\log^2 k) - \exp(-c\log^2 N)$ over Z, x, and W. □

H Experiments

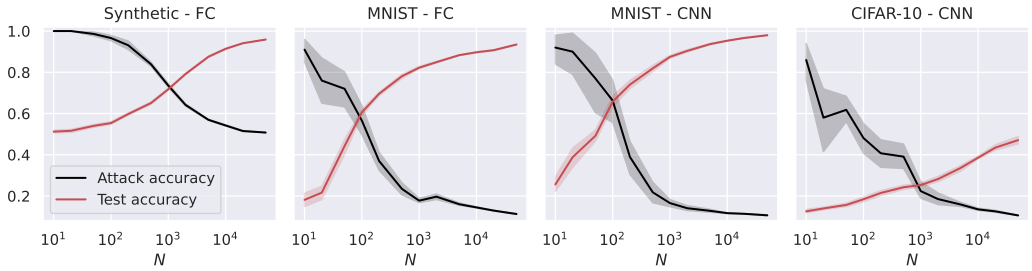

Figure 3: Test and attack accuracies as a function of the number of training samples N, for fully connected (FC, first two plots) and small convolutional neural networks (CNN, last two plots). In the first plot, we use synthetic (Gaussian) data with $d = 1000$, and the labeling function is $g(x) = \text{sign}(u^\top x)$. As we consider binary classification, the accuracy of random guessing is 0.5. The other plots use subsets of the MNIST and CIFAR-10 datasets, with an external layer of noise added to images, see Figure 2. As we consider 10 classes, the accuracy of random guessing is 0.1. We plot the average over 10 independent trials and the confidence band at 1 standard deviation.

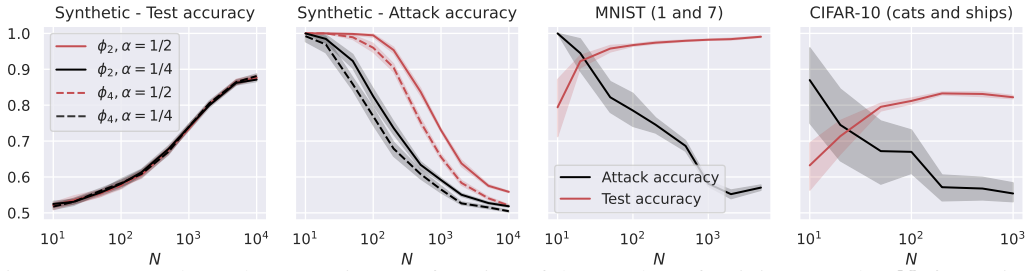

Figure 4: Test and attack accuracies as a function of the number of training samples N, for various binary classification tasks. In the first two plots, we consider the RF model with $k = 10^5$ trained over Gaussian data with $d = 1000$. The labeling function is $g(x) = \text{sign}(u^\top x)$. We repeat the experiments for $\alpha = \{0.25, 0.5\}$, and for the two activations $\phi_2 = h_1 + h_2$ and $\phi_4 = h_1 + h_4$, where h_i denotes the i-th Hermite polynomial. In the last two plots, we consider the same model with ReLU activation, trained over two MNIST and CIFAR-10 classes. The width of the noise background is 10 pixels for MNIST and 8 pixels for CIFAR-10, see Figure 2. The reconstruction attack queries the model only with the noise background, replacing all the other pixels with 0, and takes the sign of the output. As we consider binary classification, an accuracy of 0.5 is achieved by random guessing. We plot the average over 10 independent trials and the confidence band at 1 standard deviation.

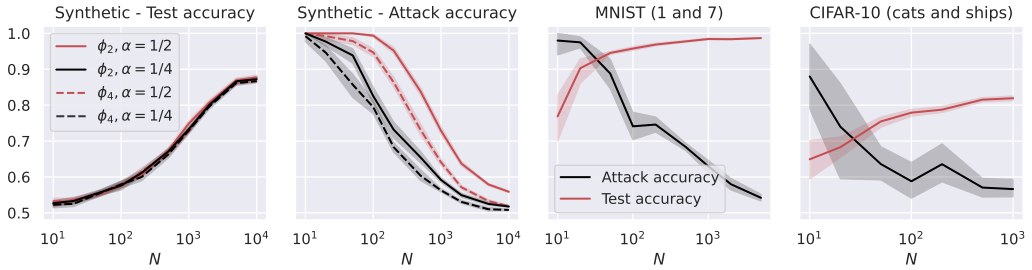

Figure 5: We consider the NTK model with $k = 100$ and, in the first two plots, we repeat the experiments for activations whose derivatives are $\phi_2' = h_0 + h_1$ and $\phi_4' = h_0 + h_3$, where h_i denotes the i-th Hermite polynomial (see Appendix C.1). The rest of the setup is similar to that of Figure 4.