How Spurious Features Are Memorized: Precise Analysis for Random and NTK Features

Simone Bombari [1] **Marco Mondelli** [1]

Abstract

Deep learning models are known to overfit and memorize spurious features in the training dataset. While numerous empirical studies have aimed at understanding this phenomenon, a rigorous theoretical framework to quantify it is still missing. In this paper, we consider spurious features that are uncorrelated with the learning task, and we provide a precise characterization of how they are memorized via two separate terms: *(i)* the *stability* of the model with respect to individual training samples, and *(ii)* the *feature alignment* between the spurious feature and the full sample. 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 gives a precise characterization of the feature alignment for the two prototypical settings of random features (RF) and neural tangent kernel (NTK) regression. We prove that the memorization of spurious features weakens as the generalization capability increases and, through the analysis of the feature alignment, we unveil the role of the model and of its activation function. Numerical experiments show the predictive power of our theory on standard datasets (MNIST, CIFAR-10).

1. Introduction

Neural networks often use features that are not inherently relevant for the intended task. This phenomenon can be caused by positive spurious correlations between certain patterns and the learning task (Geirhos et al., 2020; Xiao et al., 2021), but it occurs even when the patterns are rare (Yang et al., 2022) or simply irrelevant (Hermann & Lampinen, 2020), leading the model to *memorize* spurious relations present in the training data, which are not predictive for the sampling distribution. An extensive empirical effort has aimed at mitigating this phenomenon (Plumb et al., 2022; Chang et al., 2021). In fact, the benefits of solving this problem range from robustness to distribution-shift (Geirhos et al., 2019; Zhou et al., 2021), fairness (Zliobaite, 2015), and data-privacy (Leino & Fredrikson, 2020). However, avoiding to overfit spurious features is not always feasible, since memorization can be optimal for accuracy and over-parameterized models often exhibit their best performance when trained long enough to achieve 0 training error (Nakkiran et al., 2020; Feldman, 2020).

In this regard, a related (but separate) body of work has characterized the role of benign overfitting (Belkin, 2021; Bartlett et al., 2020), and it has precisely described the in-distribution generalization of interpolating models, such as random features and neural tangent kernels (Mei & Montanari, 2022; Ghorbani et al., 2021; Montanari & Zhong, 2022). However, this powerful theoretical machinery does not cover the memorization of spurious features, as noise is generally modelled to be in the labels, rather than in the input data. More generally, while practical work has tried to understand the impact of spurious features and disentangle them from core features in deep learning models (Hermann & Lampinen, 2020; Singla & Feizi, 2022), theoretical approaches remain predominantly directed to understand how learning is impacted by the complexity of the features (Qiu et al., 2023), or the degree of overparameterization (Sagawa et al., 2020), without capturing the role of the architecture.

Our paper bridges this gap, offering an analytically tractable framework to understand and quantify the memorization of spurious features. We consider a setting similar to (Yang et al., 2022), and we in particular look at the case where the spurious features are not correlated with the true label of the sample (thus the term *memorization*). Formally, we model the sample z as composed by two distinct parts, *i.e.*, $z \equiv [x, y]$, where x is the core feature and y the spurious one, see Figure 1 for an illustration. The memorization of spurious features is captured by the correlation between the true label g of the training sample and the output of the model evaluated on the spurious sample $z^s \equiv [-, y]$, where "$-$" corresponds to removing the core feature x (*e.g.*, replacing it with all zeros). In fact, z^s is independent of the label

[1]Institute of Science and Technology, Austria. Correspondence to: Simone Bombari <simone.bombari@ista.ac.at>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

g, as the spurious feature y is un-informative. However, due to memorization, the output of the model evaluated on z^s can still be correlated with g. Our analysis describes quantitatively this phenomenon in the setting of generalized linear regression. Surprisingly, it turns out that the emergence of memorization can be reduced to the separate effect of two distinct components:

1. The *feature alignment* $\mathcal{F}(z^s, z)$, see (8). This represents the similarity in feature space between the training sample z and the spurious one z^s; it depends on the feature map of the model and on the rest of the training dataset. 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 3.1. Similar notions of stability are provided in a rich line of work (Bousquet & Elisseeff, 2002), which relates them to generalization.

Our technical contributions can be summarized as follows:

- We connect the stability in generalized linear regression to the feature alignment between samples, see Lemma 4.1. Then, we show that this connection makes the memorization of spurious features a natural consequence of the generalization error of the model. This is the case when $\mathcal{F}(z^s, z)$ 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) (Rahimi & Recht, 2007), and *(ii)* the neural tangent kernel (NTK) (Jacot et al., 2018). Using tools from high dimensional probability, we prove the concentration of $\mathcal{F}(z^s, z)$ to a positive constant γ, see Theorems 5.4 and 6.3. For the NTK, we obtain a closed-form expression for γ, which unveils the role of the activation function in the memorization of spurious features.

In a nutshell, our results give a precise characterization of the feature alignment of RF and NTK models. This in turn establishes how the memorization of spurious features grows with the generalization error, and how it depends on the chosen model (RF/NTK) and, in particular, on the activation function. Finally, going beyond RF/NTK models trained on synthetic data, we empirically show that our theoretical predictions transfer to standard datasets (see Figure 3 that analyzes the impact of the activation function on MNIST and CIFAR-10) and different neural networks (see Figures 4 and 5 that consider fully connected, convolutional, and ResNet architectures).

2. Related work

Spurious features. Spurious correlations refer to signals that are correlated but not causally related to the learning task (Geirhos et al., 2020; Xiao et al., 2021), and they have been shown to lead to poor out-of-distribution robustness (Geirhos et al., 2019; Zhou et al., 2021) or biased predictors (Zliobaite, 2015; Seo et al., 2022; Ghosh et al., 2023). The phenomenon has been studied through the lens of overparameterization (Sagawa et al., 2020) and simplicity bias (Hermann & Lampinen, 2020; Shah et al., 2020; Qiu et al., 2023), where the latter refers to models that are inherently prone to learn "easy" patterns first (Kalimeris et al., 2019).

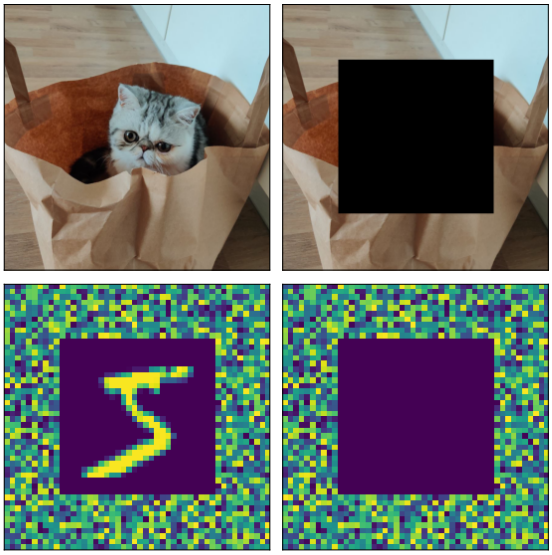

Figure 1. Example of a training sample z (top-left) and its spurious counterpart z^s (top-right). In experiments, we add a noise background (y) around the original images (x) before training (bottom-left). We then query the trained model only with the noise component (bottom-right).

Our paper considers spurious features that are independent of the learning task and, hence, focuses on their memorization. Spurious features in the training set can in fact be memorized also if they are irrelevant or rare (Yang et al., 2022; Bansal et al., 2022). This overfitting can then be used to retrieve information on the training set (Leino & Fredrikson, 2020; Bombari et al., 2022a).

Memorization and stability. Memorization measures the influence of a single sample on the final trained model. Feldman (2020); Feldman & Zhang (2020) point out its advantages on learning from heavy-tailed data, and Arpit et al. (2017); Stephenson et al. (2021) investigate its emergence in neural networks. A related concept is that of leave-one-out stability, which has been studied in a classical line of work: Hoaglin & Welsch (1978) focus on under-parameterized linear models; Bassily et al. (2021); Elisseeff & Pontil (2002); Mukherjee et al. (2006) link it to generalization; and Bousquet & Elisseeff (2002) discuss a wide range of variations on this object.

Random features and neural tangent kernel. The random features (RF) model (Rahimi & Recht, 2007; Pennington & Worah, 2017; Louart et al., 2018) 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, for example, the double descent phenomenon (Mei & Montanari, 2022). The neural tangent kernel (NTK) can be regarded as the kernel obtained by linearizing a neural network around the initialization (Jacot et al., 2018; Bartlett et al., 2021). A popular line of work has analyzed its spectrum (Fan & Wang, 2020; Adlam & Pennington, 2020; Wang & Zhu, 2021) and bounded its smallest eigenvalue (Soltanolkotabi et al., 2018; Nguyen et al., 2021; Montanari & Zhong, 2022; Bombari et al., 2022b). The behaviour of the NTK has been used in practical work to study adversarial training (Loo et al., 2022) and examples (Tsilivis & Kempe, 2022), and to understand reconstruction attacks for dataset distillation (Loo et al., 2024).

3. 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 (eventually noisy) 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 regression* model

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

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

$$\min_\theta \|\Phi\theta - G\|_2^2, \tag{2}$$

where $\Phi := [\varphi(z_1), \dots, \varphi(z_N)]^\top \in \mathbb{R}^{N\times p}$ is the feature matrix. We use the shorthand $K := \Phi\Phi^\top \in \mathbb{R}^{N\times N}$ for 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 (see equation (33) in (Bartlett et al., 2021)):

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

where θ^* is the gradient descent solution, θ_0 the initialization, $f(Z, \theta_0) := \Phi\theta_0$ 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

$$\mathcal{R} = \mathbb{E}_{z\sim\mathcal{P}_Z}\left[\left(f(z, \theta^*) - g\right)^2\right], \tag{4}$$

where g denotes the ground-truth label of the test sample z.

Stability. Let us introduce quantities related to "incomplete" datasets. 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 3.1. 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

$$\mathcal{S}_{z_1} := f(\cdot, \theta^*) - f(\cdot, \theta^*_{-1}). \tag{5}$$

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

$$\begin{aligned}\mathbb{E}_{z_1\sim\mathcal{P}_Z}\left[\mathcal{S}^2_{z_1}(z_1)\right] &= \mathbb{E}_{z_1\sim\mathcal{P}_Z}\left[\left(f(z_1, \theta^*_{-1}) - g_1\right)^2\right] \\ &= \mathbb{E}_{z\sim\mathcal{P}_Z}\left[\left(f(z, \theta^*_{-1}) - g\right)^2\right] =: \mathcal{R}_{Z_{-1}},\end{aligned} \tag{6}$$

where the purpose of the second step is just to match the notation used in (4), and $\mathcal{R}_{Z_{-1}}$ denotes the generalization error of the algorithm that uses Z_{-1} as training set.

Memorization of spurious features. The input samples are 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 *core feature* that is useful to accomplish the task (*e.g.*, the cat in top-left image of Figure 1), while y is the *spurious feature* containing noise (*e.g.*, the background). Formally, we assume that, for $i \in \{1, \dots, N\}$, $g_i = g(x_i)$, where g is a labelling function, *i.e.*, the label depends only

on the core feature x_i and it is independent of the spurious feature y_i (a similar setting was previously considered in (Loureiro et al., 2021)). Even if y_i is not useful for learning, the training algorithm may overfit it, memorizing its co-occurrence with the label g_i. This phenomenon could then lead to unpredictable behaviours at test time, such as poor performance on samples $z_t = [x_t, y_i]$ with different information but the same spurious feature (Yang et al., 2022), or data privacy breaches, through extraction of information about x_i via the knowledge of y_i (Leino & Fredrikson, 2020; Bombari et al., 2022a). Specifically, in computer vision, an unusual background could expose information about the object in the foreground (Leino & Fredrikson, 2020). In natural language processing, sensitive information (x_i) about an individual (y_i) could be extracted with proper prompting strategies: a language model might in fact be able to successfully auto-complete *"The address of y_i is..."* with *"...x_i"*, as shown by Bombari et al. (2022a) on question-answering tasks. With slight abuse of notation, during the discussion, we will use the term *feature* to indicate both elements in feature space $\varphi(z) \in \mathbb{R}^p$ (as in (1)), and portions of the samples x and y. This choice is common in the related literature, and we will elaborate whenever it could generate confusion.

To quantify the memorization of the spurious feature y_i, we will consider

$$\mathrm{Cov}\left(f(z_i^s, \theta^*), g_i\right), \tag{7}$$

which represents the covariance between the true label $g_i = g(x_i)$ and the output of the trained model (3) evaluated on $z_i^s = [x, y_i]$, which replaces the core feature x_i with an independent feature x. As x and x_i are independent, this correlation is due to the memorization of y_i. In the experiments, similarly to Singla & Feizi (2022); Yang et al. (2022), we report the *spurious accuracy*, *i.e.*, the fraction of queries $f(z_i^s, \theta^*)$ that returns the correct label $g(x_i)$. For MNIST and CIFAR-10, instead of sampling an independent x, we simply set it to 0, see Figure 1. Our code is publicly available at the GitHub repository `simone-bombari/spurious-features-memorization`.

4. Memorization and feature alignment

We start by relating the output $f(z_i^s(x), \theta^*)$ evaluated on the spurious sample $z_i^s = [x, y_i]$ with the output $f(z_i, \theta^*)$ evaluated on the original sample z_i. For generalized linear regression, this can be elegantly done via the notion of *stability* of Definition 3.1. By symmetry, from now on, we set $i = 1$ without loss of generality.

Lemma 4.1. *Let $\varphi : \mathbb{R}^d \to \mathbb{R}^p$ be a 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$\{\mathrm{rows}(\Phi_{-1})\}$ and $\mathcal{S}_{z_1}$ the stability with respect to z_1, as in Definition 3.1. Define*

$$\mathcal{F}_\varphi(z, z_1) := \frac{\varphi(z)^\top P^\perp_{\Phi_{-1}} \varphi(z_1)}{\left\| P^\perp_{\Phi_{-1}} \varphi(z_1) \right\|_2^2} \tag{8}$$

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

$$\mathcal{S}_{z_1}(z) = \mathcal{F}_\varphi(z, z_1)\, \mathcal{S}_{z_1}(z_1). \tag{9}$$

Classical work (Hoaglin & Welsch, 1978) considers a similar problem in the under-parametrized regime, exploiting the projector $H = \Phi(\Phi^\top\Phi)^{-1}\Phi^\top \in \mathbb{R}^{N\times N}$ ($p \le N$ is needed for $\Phi^\top\Phi$ to be invertible), known as the *hat matrix*. In contrast, Lemma 4.1 focuses on the over-parameterized regime and highlights the role of the projector $P_{\Phi_{-1}}$. The different behaviour of under-parameterized and over-parameterized models requires the proof of Lemma 4.1 to follow a different strategy, which is discussed in Appendix B.

In words, Lemma 4.1 relates the stability with respect to z_1 evaluated on the two samples 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. As a sanity check, the feature alignment between any sample and itself is equal to one, which trivializes (9). Then, as z and z_1 become less aligned, the stability $\mathcal{S}_{z_1}(z) = f(z, \theta^*) - f(z, \theta^*_{-1})$ starts to differ from $\mathcal{S}_{z_1}(z_1) = f(z_1, \theta^*) - f(z_1, \theta^*_{-1})$. We note that the feature alignment also depends on the rest of the training set Z_{-1}, as Z_{-1} implicitly appears in the projector $P^\perp_{\Phi_{-1}}$. We also remark that the invertibility of K directly implies that the denominator in (8) is different from zero, see Lemma B.1 in Appendix B.

Armed with Lemma 4.1, we now characterize the correlation between $f(z_1^s, \theta^*)$ and g_1. Let us replace $\mathcal{F}_\varphi(z_1^s, z_1)$ in (9) with a constant $\gamma_\varphi > 0$, independent from z_1. This is justified by Sections 5 and 6, where we prove the concentration of $\mathcal{F}_\varphi(z_1^s, z_1)$ for the RF and NTK model, respectively. Then, by using the definition of stability in (5), we get

$$\begin{aligned} f(z_1^s, \theta^*) &= f(z_1^s, \theta^*_{-1}) + \gamma_\varphi\, \mathcal{S}_{z_1}(z_1) \\ &= f(z_1^s, \theta^*_{-1}) + \gamma_\varphi \left(g_1 - f(z_1, \theta^*_{-1})\right). \end{aligned} \tag{10}$$

Note that $f(z_1^s, \theta^*_{-1})$ is independent from g_1, as it doesn't depend on x_1. In fact, $z_i^s = [x, y_i]$ is independent of x_1, and θ^*_{-1} is not trained on x_1. Thus, if the algorithm is stable, in the sense that $\mathcal{S}_{z_1}(z_1)$ is close to 0, we have that $f(z_1^s, \theta^*)$ has little dependence on g_1. Conversely, if $\mathcal{S}_{z_1}(z_1)$ grows, then $f(z_1^s, \theta^*)$ will start picking up the correlation with g_1. Concretely, we can look at the covariance between $f(z_1^s, \theta^*)$ and g_1, in the probability space of z_1:

$$\begin{aligned} &\mathrm{Cov}\left(f(z_1^s, \theta^*), g_1\right) = \gamma_\varphi \mathrm{Cov}\left(\mathcal{S}_{z_1}(z_1), g_1\right) \\ &\le \gamma_\varphi \sqrt{\mathrm{Var}\left(\mathcal{S}_{z_1}(z_1)\right) \mathrm{Var}\left(g_1\right)} \le \gamma_\varphi \sqrt{\mathcal{R}_{Z_{-1}}} \sqrt{\mathrm{Var}\left(g_1\right)}. \end{aligned} \tag{11}$$

Here, the first step uses (10) and the independence between $f(z_1^s, \theta_{-1}^*)$ and g_1, the second step is an application of Cauchy-Schwarz, and the last step follows from (6).

The terms γ_φ and $\sqrt{\mathcal{R}_{Z_{-1}}}$ in the RHS of (11) show that the memorization of spurious features is affected by two factors: *(i)* the similarity between z_1^s and z_1 (formalized by $\mathcal{F}_\varphi(z_1^s, z_1)$), and *(ii)* the generalization error of the model. While the dependence on the generalization error is not entirely surprising (overfitting causes both the inability of the model to generalize and the memorization of spurious correlations between g_1 and y_1), the key role of the feature alignment (in the form defined in (8)) is a-priori far from obvious. Furthermore, via the analysis of $\mathcal{F}_\varphi(z_1^s, z_1)$ in the next two sections, we will show how the memorization is affected by the choice of the feature map and, specifically, of the activation function.

Similarly, we can lower bound the *spurious loss*, defined as

$$\mathcal{L} := \mathbb{E}_{z_1}\left[\left(f(z_1^s, \theta^*) - g_1\right)^2\right]. \tag{12}$$

By introducing the shorthands $\bar{f} := \mathbb{E}_{z_1}[f(z_1^s, \theta^*)]$ and $\bar{g} := \mathbb{E}_{z_1}[g_1]$, we have the lower bound:

$$\begin{aligned}\mathcal{L} &= \mathbb{E}_{z_1}\left[\left(\left(f(z_1^s, \theta^*) - \bar{f}\right) - (g_1 - \bar{g}) + \left(\bar{f} - \bar{g}\right)\right)^2\right] \\ &= \mathbb{E}_{z_1}\left[\left(f(z_1^s, \theta^*) - \bar{f}\right)^2\right] + \mathbb{E}_{z_1}\left[(g_1 - \bar{g})^2\right] \\ &\quad + \mathbb{E}_{z_1}\left[\left(\bar{f} - \bar{g}\right)^2\right] - 2\,\mathrm{Cov}\left(f(z_1^s, \theta^*), g_1\right) \\ &\geq \mathbb{E}_{z_1}\left[(g_1 - \bar{g})^2\right] - 2\,\gamma_\varphi\sqrt{\mathcal{R}_{Z_{-1}}}\sqrt{\mathrm{Var}\,(g_1)}.\end{aligned} \tag{13}$$

The first term on the RHS of (13) is the minimal loss when no information about x_1 is available (and, thus, the best estimator is the expectation of g_1). The second term indicates how the memorization of the spurious feature y_1 improves the trivial guess $\bar{g}$, and it depends again on $\mathcal{R}_{Z_{-1}}$ and γ_φ.

5. Main result for random features

The *random features (RF) model* takes the form

$$f_{\text{RF}}(z, \theta) = \varphi_{\text{RF}}(z)^\top \theta, \qquad \varphi_{\text{RF}}(z) = \phi(Vz), \tag{14}$$

where 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 denote by μ_l the l-th Hermite coefficient of ϕ (see Appendix A.1 for details).

Assumption 5.1 (Data distribution)**.** The input data $(z_1, \dots, z_N)$ are N i.i.d. samples from $\mathcal{P}_Z = \mathcal{P}_X \times \mathcal{P}_Y$ s.t. $z_i \in \mathbb{R}^d$ 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 $\|y\|_2 = \sqrt{d_y}$.
2. $\mathbb{E}[x] = 0$ and $\mathbb{E}[y] = 0$.
3. $\mathcal{P}_X$ and $\mathcal{P}_Y$ satisfy *Lipschitz concentration*.

The first two assumptions are achieved by data preprocessing and could be relaxed as in Assumption 1 of (Bombari et al., 2022b) at the cost of a more involved argument. The third assumption (see Appendix A for details) corresponds to data having well-behaved tails, and it covers a number of important cases, *e.g.*, standard Gaussian (Vershynin, 2018), uniform on the sphere/hypercube (Vershynin, 2018), or data obtained via GANs (Seddik et al., 2020). This requirement is common in the related literature (Bombari et al., 2022b; Bubeck & Sellke, 2021; Nguyen et al., 2021) and it is often replaced by a stronger requirement (*e.g.*, data uniform on the sphere), see (Montanari & Zhong, 2022).

Assumption 5.2 (Over-parameterization and high-dimensional data)**.**

$$N \log^3 N = o(k),\ \sqrt{d}\log d = o(k),\ k\log^4 k = o(d^2). \tag{15}$$

The first condition in (15) 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 (Mei & Montanari, 2022). Combining the second and third conditions in (15), 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 5.3 (Activation function)**.** The activation function ϕ is a non-linear L-Lipschitz function.

Theorem 5.4. *Let Assumptions 5.1, 5.2, and 5.3 hold, and let $x \sim \mathcal{P}_X$ be sampled independently from everything. Consider querying the trained RF model* (14) *with $z_1^s = [x, y_1]$. Let $\alpha = d_y/d$ and $\mathcal{F}_{\text{RF}}(z_1^s, z_1)$ be the feature alignment between z_1^s and z_1, as defined in* (8)*. Then,*

$$|\mathcal{F}_{\text{RF}}(z_1^s, z_1) - \gamma_{\text{RF}}| = o(1), \tag{16}$$

with probability at least $1 - \exp(-c\log^2 N)$ over V, Z and x, where $\gamma_{\text{RF}} \leq 1$ does not depend on z_1 and x. Furthermore, letting μ_l be the l-th Hermite coefficient of the activation ϕ (see Appendix A.1), we have

$$\gamma_{\text{RF}} > \frac{\sum_{l=2}^{+\infty}\mu_l^2\alpha^l}{\sum_{l=1}^{+\infty}\mu_l^2} - o(1), \tag{17}$$

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

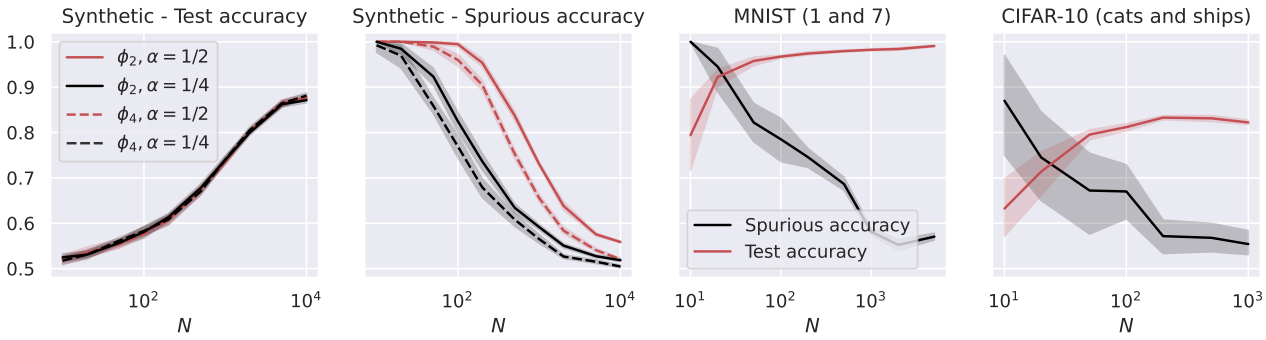

Figure 2. Test and spurious accuracies as a function of the number of training samples N, for various binary classification tasks. In the first two plots, we consider the RF model in (14) with $k = 10^5$ trained over Gaussian data with $d = 1000$. The labeling function is $g(x) = \text{sign}(u^\top x)$. 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 (see Appendix A.1). 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 1. The spurious accuracy is obtained by querying the model only with the noise background from the training set, replacing all the other pixels with 0, and taking 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.

Theorem 5.4 proves the concentration of the feature alignment $\mathcal{F}_{\text{RF}}(z_1^s, z_1)$ to a constant γ_{RF} between $\sum_{l=2}^{+\infty} \mu_l^2 \alpha^l / \sum_{l=1}^{+\infty} \mu_l^2 > 0$ and 1. The lower bound increases with α (as expected, since α is the fraction of the input given by the spurious feature), and it depends in a non-trivial way on the activation via its Hermite coefficients.

This result validates the argument in (10), where we replaced the feature alignment $\mathcal{F}_{\text{RF}}(z_1^s, z_1)$ with a constant $\gamma_{\text{RF}} > 0$. Thus, (11) now reads

$$\text{Cov}\left(f_{\text{RF}}(z_1^s, \theta^*), g_1\right) \leq \gamma_{\text{RF}} \sqrt{\mathcal{R}_{Z_{-1}}} \sqrt{\text{Var}\left(g_1\right)}, \quad (18)$$

which quantifies the memorization of spurious features in terms of the generalization error and the constant γ_{RF}.

These effects are clearly displayed in Figure 2 for binary classification on synthetic (first two plots) and image (last two plots) datasets. Specifically, as the number of samples N increases, the test accuracy increases and the spurious accuracy (obtained by querying the trained model with the spurious sample $f(z_i^s, \theta^*)$) decreases. Furthermore, for the synthetic dataset, while the test accuracy does not depend on α or the activation function, the spurious accuracy increases with α and by taking an activation function with dominant low-order Hermite coefficients, as predicted by (17). For an additional experiment highlighting the dependence on α, we refer the reader to Figure 7 in Appendix F.

Proof sketch. We set

$$\gamma_{\text{RF}} := \frac{\mathbb{E}_{z_1, z_1^s}[\varphi_{\text{RF}}(z_1^s)^\top P^\perp_{\Phi_{-1}} \varphi_{\text{RF}}(z_1)]}{\mathbb{E}_{z_1}[\|P^\perp_{\Phi_{-1}} \varphi_{\text{RF}}(z_1)\|_2^2]}. \quad (19)$$

Here, $P_{\Phi_{-1}}$ is the projector over $\text{Span}\{\text{rows}(\Phi_{\text{RF},-1})\}$ and $\Phi_{\text{RF},-1}$ the RF feature matrix after removing the first row. With this choice, the numerator and denominator of γ_{RF} equal the expectations of the corresponding quantities appearing in $\mathcal{F}_{\text{RF}}(z_1^s, z_1)$. Thus, the concentration result in (16) is obtained from the general form of the Hanson-Wright inequality in (Adamczak, 2015), see Lemma D.7. The upper bound $\gamma_{\text{RF}} \leq 1$ follows from an application of Cauchy-Schwarz inequality. In contrast, the lower bound is more involved and it is obtained via the three steps below.

Step 1: Centering the feature map φ_{RF}. We extract the term $\mathbb{E}_V[\phi(Vz)]$ from the expression of $\mathcal{F}_{\text{RF}}$ and show it can be neglected, due to the specific structure of $P^\perp_{\Phi_{-1}}$. Specifically, letting $\tilde{\varphi}_{\text{RF}}(z) := \varphi_{\text{RF}}(z) - \mathbb{E}_V[\varphi_{\text{RF}}(z)]$, we have

$$\begin{aligned} \mathcal{F}_{\text{RF}}(z_1^s, z_1) &\simeq \frac{\tilde{\varphi}_{\text{RF}}(z_1^s)^\top P^\perp_{\Phi_{-1}} \tilde{\varphi}_{\text{RF}}(z_1)}{\left\| P^\perp_{\Phi_{-1}} \tilde{\varphi}_{\text{RF}}(z_1) \right\|_2^2} \\ &= \frac{\tilde{\varphi}_{\text{RF}}(z_1^s)^\top \tilde{\varphi}_{\text{RF}}(z_1) - \tilde{\varphi}_{\text{RF}}(z_1^s)^\top P_{\Phi_{-1}} \tilde{\varphi}_{\text{RF}}(z_1)}{\left\| P^\perp_{\Phi_{-1}} \tilde{\varphi}_{\text{RF}}(z_1) \right\|_2^2}, \end{aligned} \quad (20)$$

where $\simeq$ denotes an equality up to a $o(1)$ term. This is formalized in Lemma D.3.

Step 2: Linearization of the centered feature map $\tilde{\varphi}_{\text{RF}}$. We consider the terms $\tilde{\varphi}_{\text{RF}}(z_1^s), \tilde{\varphi}_{\text{RF}}(z_1)$ that multiply $P_{\Phi_{-1}}$ in the RHS of (20), and we show that they are well approximated by their first-order Hermite expansions ($\mu_1 V z_1^s$ and $\mu_1 V z_1$, respectively). In fact, the rest of the Hermite series scales at most as N/d^2, which is negligible due to Assump-

Figure 3. We consider the NTK model in (24) with $k = 100$, trained on MNIST (digits 1 and 7, first and second plots), and CIFAR-10 (cats and ships, third and fourth plots). We repeat the experiments for activations whose derivatives are $\phi_2' = h_0 + h_1$ and $\phi_8' = h_0 + h_7$, where h_i denotes the i-th Hermite polynomial (see Appendix A.1). The rest of the setup is the same as that of Figure 2.

tion 5.2. Specifically, Lemma D.4 implies

$$\begin{aligned}&\frac{\tilde{\varphi}_{\text{RF}}(z_1^s)^\top \tilde{\varphi}_{\text{RF}}(z_1) - \tilde{\varphi}_{\text{RF}}(z_1^s)^\top P_{\Phi_{-1}} \tilde{\varphi}_{\text{RF}}(z_1)}{\left\| P^\perp_{\Phi_{-1}} \tilde{\varphi}_{\text{RF}}(z_1) \right\|_2^2} \\ &\simeq \frac{\tilde{\varphi}_{\text{RF}}(z_1^s)^\top \tilde{\varphi}_{\text{RF}}(z_1) - \mu_1^2 (V z_1^s)^\top P_{\Phi_{-1}} (V z_1)}{\left\| P^\perp_{\Phi_{-1}} \tilde{\varphi}_{\text{RF}}(z_1) \right\|_2^2}.\end{aligned} \tag{21}$$

Step 3: Lower bound in terms of α and $\{\mu_l\}_{l\geq 2}$. To conclude, we express the RHS of (21) as follows:

$$\begin{aligned}&\frac{\tilde{\varphi}_{\text{RF}}(z_1^s)^\top \tilde{\varphi}_{\text{RF}}(z_1) - \mu_1^2 (V z_1^s)^\top (V z_1) + \mu_1^2 (V z_1^s)^\top P^\perp_{\Phi_{-1}} (V z_1)}{\|\tilde{\varphi}_{\text{RF}}(z_1)\|_2^2 - \left\| P_{\Phi_{-1}} \tilde{\varphi}_{\text{RF}}(z_1) \right\|_2^2} \\ &\gtrsim \frac{\tilde{\varphi}_{\text{RF}}(z_1^s)^\top \tilde{\varphi}_{\text{RF}}(z_1) - \mu_1^2 (V z_1^s)^\top (V z_1)}{\|\tilde{\varphi}_{\text{RF}}(z_1)\|_2^2} \\ &\simeq \frac{\sum_{l=2}^{+\infty} \mu_l^2 \alpha^l}{\sum_{l=1}^{+\infty} \mu_l^2} > 0,\end{aligned} \tag{22}$$

where $\gtrsim$ denotes an inequality up to a $o(1)$ term. As $P_{\Phi_{-1}} = I - P^\perp_{\Phi_{-1}}$, the term in the first line equals the RHS of (21). Next, we show that $\mu_1^2 (V z_1^s)^\top P^\perp_{\Phi_{-1}} (V z_1)$ is equal to $\mu_1^2 (V[y_1, 0])^\top P^\perp_{\Phi_{-1}} (V[y_1, 0])$ (which corresponds to the common noise part in z_1, z_1^s) plus a vanishing term, see Lemma D.5. As $\mu_1^2 (V[y_1, 0])^\top P^\perp_{\Phi_{-1}} (V[y_1, 0]) \geq 0$, the inequality in the second line follows. The last step is obtained by showing concentration over V of numerator and denominator. The expression on the RHS of (22) is strictly positive as $\alpha > 0$ and ϕ is non-linear by Assumption 5.3.

6. Main result for NTK features

We consider the following two-layer neural network

$$f_{\text{NN}}(z, w) = \sum_{i=1}^{k} \phi\left(W_{i:} z\right). \tag{23}$$

The hidden layer contains k neurons; ϕ is an activation function applied component-wise; $W \in \mathbb{R}^{k \times d}$ denotes the weights of the hidden layer; $W_{i:}$ denotes the i-th row of W; and we set the k weights of the second layer to 1. We indicate with w the vector containing the parameters of this model, *i.e.*, $w = [\text{vec}(W)] \in \mathbb{R}^p$, with $p = kd$. We initialize the network with standard (*e.g.*, He's or LeCun's) initialization, *i.e.*, $[W_0]_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$.

The *NTK regression model* takes the form

$$\begin{aligned}f_{\text{NTK}}(z, \theta) &= \varphi_{\text{NTK}}(z)^\top \theta, \\ \varphi_{\text{NTK}}(z) &= \nabla_w f_{\text{NN}}(z, w)|_{w=w_0} = z \otimes \phi'(W_0 z),\end{aligned} \tag{24}$$

where $\nabla_w f_{\text{NN}}$ in the second line is computed via the chain rule. The vector of trainable parameters is $\theta \in \mathbb{R}^p$, with $p = kd$, which is initialized with $\theta_0 = w_0 = [\text{vec}(W_0)]$. We note that $f_{\text{NTK}}(z, \theta)$ is the linearization of $f_{\text{NN}}(z, w)$ around the initial point w_0 (Bartlett et al., 2021; Jacot et al., 2018), and the model in (24) corresponds to training the two-layer network (23) in the lazy regime (Chizat et al., 2019; Oymak & Soltanolkotabi, 2019). As such, it has received significant attention in the theoretical literature, see *e.g.* (Bombari et al., 2023; Dohmatob & Bietti, 2022; Montanari & Zhong, 2022).

Assumption 6.1 (Over-parameterization and topology)**.**

$$N \log^8 N = o(kd), \qquad N > d, \qquad k = \mathcal{O}(d). \tag{25}$$

The first condition is the smallest (up to log factors) over-parameterization that guarantees interpolation (Bombari et al., 2022b). 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 (24), and a stronger requirement, *i.e.*, the strict inequality $k < d$, has appeared in prior work (Nguyen & Hein, 2017; 2018; Nguyen & Mondelli, 2020).

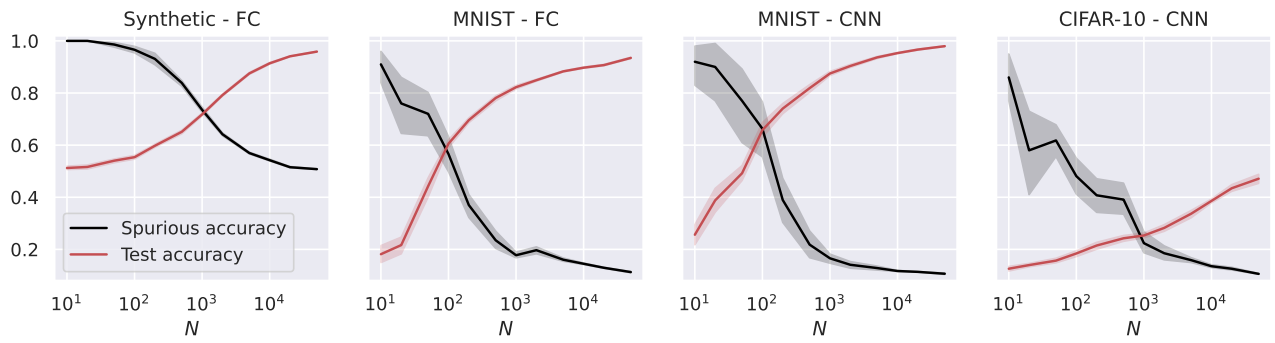

Figure 4. Test and spurious accuracies as a function of the number of training samples N, for a fully connected (FC, first two plots), and a small convolutional neural network (CNN, last two plots). In the first plot, we use synthetic (Gaussian) data with $d = 1000$, and the labeling function is $g(x) = \text{sign}(u^\top x)$. 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 1. 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.

Assumption 6.2 (Activation function)**.** The activation function ϕ is non-linear and its derivative ϕ' is L-Lipschitz.

We denote by μ_l' the l-th Hermite coefficient of ϕ'. We remark that the invertibility of the kernel K_{NTK} induced by the feature map (24) follows from Lemma E.1. At this point, we are ready to state our main result for the NTK model, whose full proof is contained in Appendix E.

Theorem 6.3. *Let Assumptions 5.1, 6.1, and 6.2 hold, and let $x \sim \mathcal{P}_X$ be sampled independently from everything. Consider querying the trained NTK model* (24) *with $z_1^s = [x, y_1]$. Let $\alpha = d_y/d \in (0,1)$ and $\mathcal{F}_{\text{NTK}}(z_1^s, z_1)$ be the feature alignment between z_1^s and z_1, as defined in* (8)*. Then, letting μ_l' be the l-th Hermite coefficient of the derivative of the activation ϕ' (see Appendix A.1), we have*

$$|\mathcal{F}_{\text{NTK}}(z_1^s, z_1) - \gamma_{\text{NTK}}| = o(1),$$
$$\text{where} \quad 0 < \gamma_{\text{NTK}} := \alpha \frac{\sum_{l=1}^{+\infty} {\mu_l'}^2 \alpha^l}{\sum_{l=1}^{+\infty} {\mu_l'}^2} < 1, \tag{26}$$

with probability at least $1 - N\exp(-c\log^2 k) - \exp(-c\log^2 N)$ over Z, x and W_0.

Theorem 5.4 proves the concentration of the feature alignment $\mathcal{F}_{\text{NTK}}(z_1^s, z_1)$ to a constant γ_{NTK}, which has an exact expression depending on α and the Hermite coefficients of the derivative of the activation.

This result validates the argument in (10), where we replaced the feature alignment $\mathcal{F}_{\text{NTK}}(z_1^s, z_1)$ with a constant $\gamma_{\text{NTK}} > 0$. Thus, (11) now reads

$$\text{Cov}\left(f_{\text{NTK}}(z_1^s, \theta^*), g_1\right) \le \gamma_{\text{NTK}} \sqrt{\mathcal{R}_{Z_{-1}}} \sqrt{\text{Var}\left(g_1\right)}, \tag{27}$$

which quantifies the memorization of spurious features in terms of the generalization error and the constant γ_{NTK}.

Figure 3 considers training on MNIST and CIFAR-10, and it shows that the predictions of Theorem 6.3 also hold for standard datasets: as N increases, the test accuracy improves and the spurious accuracy decreases; considering activations with dominant high-order Hermite coefficients reduces memorization. For additional experiments using the same setup as Figure 2 and highlighting the dependence on α, we refer the reader to Figures 6 and 7 in Appendix F.

Proof sketch. The argument is more direct than for the RF model since, in this case, we are able to express γ_{NTK} in closed form. We denote by $P_{\Phi_{-1}}$ the projector over the span of the rows of the NTK feature matrix $\Phi_{\text{NTK},-1}$ without the first row. Then, the *first step* is to *center the feature map* φ_{NTK}, which gives

$$\frac{\varphi_{\text{NTK}}(z_1^s)^\top P_{\Phi_{-1}}^\perp \varphi_{\text{NTK}}(z_1)}{\left\| P_{\Phi_{-1}}^\perp \varphi_{\text{NTK}}(z_1) \right\|_2^2} \simeq \frac{\tilde{\varphi}_{\text{NTK}}(z_1^s)^\top P_{\Phi_{-1}}^\perp \tilde{\varphi}_{\text{NTK}}(z_1)}{\left\| P_{\Phi_{-1}}^\perp \tilde{\varphi}_{\text{NTK}}(z_1) \right\|_2^2}, \tag{28}$$

where $\tilde{\varphi}_{\text{NTK}}(z) := z \otimes (\phi'(W_0 z) - \mathbb{E}_{W_0}[\phi'(W_0 z)])$. While a similar step appeared in the analysis of the RF model, its implementation for NTK requires a different strategy. In particular, we exploit that the samples z_1 and z_1^s are *approximately* contained in the span of the rows of Z_{-1} (see Lemma E.4). As the rows of Z_{-1} may not *exactly* span all $\mathbb{R}^d$, we resort to an *approximation* by adding a small amount of independent noise to every entry of Z_{-1}. The resulting perturbed dataset $\bar{Z}_{-1}$ satisfies $\text{Span}\{\text{rows}(\bar{Z}_{-1})\} = \mathbb{R}^d$ (see Lemma E.3), and we conclude via a continuity argument with respect to the magnitude of the perturbation (see Lemmas E.2 and E.5).

The *second step* is to upper bound the terms $\left|\tilde{\varphi}_{\text{NTK}}(z_1^s)^\top P_{\Phi_{-1}} \tilde{\varphi}_{\text{NTK}}(z_1)\right|$ and $\left\| P_{\Phi_{-1}} \tilde{\varphi}_{\text{NTK}}(z_1) \right\|_2^2$,

showing they have *negligible magnitude*, which gives

$$\frac{\tilde{\varphi}_{\text{NTK}}(z_1^s)^\top P^\perp_{\Phi_{-1}} \tilde{\varphi}_{\text{NTK}}(z_1)}{\left\| P^\perp_{\Phi_{-1}} \tilde{\varphi}_{\text{NTK}}(z_1) \right\|_2^2} \simeq \frac{\tilde{\varphi}_{\text{NTK}}(z_1^s)^\top \tilde{\varphi}_{\text{NTK}}(z_1)}{\|\tilde{\varphi}_{\text{NTK}}(z_1)\|_2^2}. \tag{29}$$

This is a consequence of the fact that, if $z \sim \mathcal{P}_Z$ is independent from Z_{-1}, then $\tilde{\varphi}(z)$ is roughly orthogonal to $\text{Span}\{\text{rows}(\Phi_{\text{NTK},-1})\}$, see Lemma E.8.

Finally, the *third step* is to show the *concentration* over W_0 of the numerator and denominator of the RHS of (29), see Lemma E.6. This allows us to conclude that

$$\frac{\tilde{\varphi}_{\text{NTK}}(z_1^s)^\top \tilde{\varphi}_{\text{NTK}}(z_1)}{\|\tilde{\varphi}_{\text{NTK}}(z_1)\|_2^2} \simeq \alpha \frac{\sum_{l=1}^{+\infty} \mu_l'^2 \alpha^i}{\sum_{l=1}^{+\infty} \mu_l'^2} > 0. \tag{30}$$

The RHS of (30) is strictly positive as $\alpha > 0$ and ϕ is non-linear by Assumption 6.2.

Discussion. As showed by (11) and (13), if $\mathcal{F}_\varphi(z_1, z_1^s)$ converges to 0, spurious correlations are *not* memorized by the model. However, Theorems 5.4 and 6.3 prove that, for the RF and NTK model respectively, a strictly positive geometric overlap ($\alpha > 0$) guarantees a strictly positive feature alignment. Thus, these results imply that, as long as the generalization error is not vanishing, spurious features are memorized and, in fact, we quantify the extent to which memorization occurs. Remarkably, our experiments on real datasets (see Figure 3) show that the effect of the activation function is in line with our predictions: taking ϕ' with higher order Hermite coefficients lead to models that are less prone to memorize spurious features. Finally, while we focus on generalized linear regression, the interplay between memorization of spurious features and generalization provided by our analysis holds in far more generality and, in particular, it is displayed by neural networks also capable of feature learning, see Figures 4 and 5.

7. Conclusions

In this work, we present a theoretical framework to quantify the memorization of spurious features. Our characterization hinges on *(i)* the classical notion of *stability* of the model w.r.t. a training sample, and *(ii)* a novel notion of *feature alignment* $\mathcal{F}(z^s, z)$ between two samples that share the same spurious feature y. By providing a precise analysis of the feature alignment in the prototypical settings of random and NTK features regression, we show that the memorization is proportional to the generalization error, and we characterize the proportionality constant, revealing how it depends on the model and its activation function. Our theoretical predictions are confirmed by numerical experiments on standard datasets (see Figure 3) and on different neural network architectures (see Figures 4 and 5).

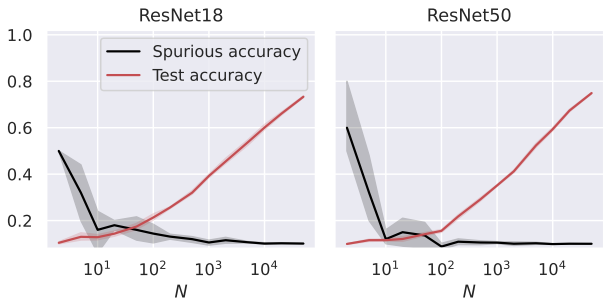

Figure 5. Test and spurious accuracies as a function of the number of training samples N, for two ResNet architectures. We use subsets of the CIFAR-10 dataset, with an external layer of noise added to images, see Figure 1. As we consider 10 classes, the accuracy of random guessing is 0.1. We plot the average over 10 independent trials and the confidence band at 1 standard deviation.

The approach we put forward is rather general, and our results could be extended to cases where the spurious feature y is correlated with the ground-truth label g. Another possible extension involves testing the trained model on a new spurious feature y', which is not present in the training set but is correlated with a feature y that has already been seen; or capturing the role of *simplicity bias* in this phenomenon. We also remark that the formalism introduced by Lemma 4.1 applies to any feature map φ (*e.g.*, with multiple fully-connected, convolutional or attention layers). Characterizing the feature alignment of such maps would allow to compare different models and establish which of those is less prone to memorizing spurious correlations.

Acknowledgements

The authors were partially supported by the 2019 Lopez-Loreta prize, and they would like to thank (in alphabetical order) Grigorios Chrysos, Simone Maria Giancola, Mahyar Jafari Nodeh, Christoph Lampert, Marco Miani, GuanWen Qiu, and Peter Súkeník for helpful discussions.

Impact statement

This paper presents work whose goal is to advance the field of Machine Learning. In particular, it provides an analytical framework to understand and quantify the memorization of spurious features. This would allow the design of machine learning models that are less prone to memorizing spurious correlations and, as such, more robust to distribution shifts, more fair and more private. While there are many potential societal consequences, our work is of theoretical nature and, therefore, we feel that none of these consequences has to be further discussed here.

References

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

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

Arpit, D., Jastrzundefinedbski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., and Lacoste-Julien, S. A closer look at memorization in deep networks. In *International Conference on Machine Learning (ICML)*, 2017.

Bansal, R., Pruthi, D., and Belinkov, Y. Measures of information reflect memorization patterns. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Bartlett, P. L., Montanari, A., and Rakhlin, A. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

Bassily, R., Nissim, K., Smith, A., Steinke, T., Stemmer, U., and Ullman, J. Algorithmic stability for adaptive data analysis. *SIAM Journal on Computing*, 50(3), 2021.

Belkin, M. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

Bombari, S., Achille, A., Wang, Z., Wang, Y.-X., Xie, Y., Singh, K. Y., Appalaraju, S., Mahadevan, V., and Soatto, S. Towards differential relational privacy and its use in question answering. *arXiv preprint arXiv:2203.16701*, 2022a.

Bombari, S., Amani, M. H., and Mondelli, M. Memorization and optimization in deep neural networks with minimum over-parameterization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022b.

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

Bousquet, O. and Elisseeff, A. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Bubeck, S. and Sellke, M. A universal law of robustness via isoperimetry. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Chang, C., Adam, G., and Goldenberg, A. Towards robust classification model by counterfactual and invariant data generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Dohmatob, E. and Bietti, A. On the (non-)robustness of two-layer neural networks in different learning regimes. *arXiv preprint arXiv:2203.11864*, 2022.

Elisseeff, A. and Pontil, M. 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.

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

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

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

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021.

Ghosh, B., Basu, D., and Meel, K. S. "How biased are your features?": Computing fairness influence functions with global sensitivity analysis. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023.

Hermann, K. and Lampinen, A. What shapes feature representations? Exploring datasets, architectures, and training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Hoaglin, D. C. and Welsch, R. E. The hat matrix in regression and anova. *The American Statistician*, 32(1):17–22, 1978.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Johnson, C. R. *Matrix Theory and Applications*. American Mathematical Society, 1990.

Kalimeris, D., Kaplun, G., Nakkiran, P., Edelman, B., Yang, T., Barak, B., and Zhang, H. Sgd on neural networks learns functions of increasing complexity. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Leino, K. and Fredrikson, M. Stolen memories: leveraging model memorization for calibrated white-box membership inference. In *Proceedings of the 29th USENIX Conference on Security Symposium*, 2020.

Loo, N., Hasani, R., Amini, A., and Rus, D. Evolution of neural tangent kernels under benign and adversarial training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Loo, N., Hasani, R., Lechner, M., Amini, A., and Rus, D. Understanding reconstruction attacks with the neural tangent kernel and dataset distillation. In *The Twelfth International Conference on Learning Representations*, 2024.

Louart, C., Liao, Z., and Couillet, R. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mezard, M., and Zdeborová, L. Learning curves of generic features maps for realistic datasets with a teacher-student model. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 18137–18151. Curran Associates, Inc., 2021.

Mei, S. and Montanari, A. 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.

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

Mukherjee, S., Niyogi, P., Poggio, T., and Rifkin, R. Learning theory: Stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Adv. Comput. Math.*, 25:161–193, 2006.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations (ICLR)*, 2020.

Nguyen, Q. and Hein, M. The loss surface of deep and wide neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

Nguyen, Q. and Hein, M. Optimization landscape and expressivity of deep CNNs. In *International Conference on Machine Learning (ICML)*, 2018.

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

Nguyen, Q., Mondelli, M., and Montufar, G. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep ReLU networks. In *International Conference on Machine Learning (ICML)*, 2021.

O'Donnell, R. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

Oymak, S. and Soltanolkotabi, M. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning (ICML)*, 2019.

Pennington, J. and Worah, P. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Plumb, G., Ribeiro, M. T., and Talwalkar, A. Finding and fixing spurious patterns with explanations. *Transactions on Machine Learning Research*, 2022.

Qiu, G., Kuang, D., and Goel, S. Complexity matters: Dynamics of feature learning in the presence of spurious correlations. In *NeurIPS Workshop on Mathematics of Modern Machine Learning*, 2023.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning (ICML)*, 2020.

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

Seddik, M. E. A., Louart, C., Tamaazousti, M., and Couillet, R. Random matrix theory proves that deep learning representations of GAN-data behave as gaussian mixtures. In *International Conference on Machine Learning (ICML)*, 2020.

Seo, S., Lee, J.-Y., and Han, B. Information-theoretic bias reduction via causal view of spurious correlation. In *AAAI Conference on Artificial Intelligence*, 2022.

Shah, H., Tamuly, K., Raghunathan, A., Jain, P., and Netrapalli, P. The pitfalls of simplicity bias in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Singla, S. and Feizi, S. Salient imagenet: How to discover spurious features in deep learning? In *International Conference on Learning Representations*, 2022.

Soltanolkotabi, M., Javanmard, A., and Lee, J. D. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

Stephenson, C., suchismita padhy, Ganesh, A., Hui, Y., Tang, H., and Chung, S. On the geometry of generalization and memorization in deep neural networks. In *International Conference on Learning Representations*, 2021.

Tropp, J. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, pp. 389–434, 2012.

Tsilivis, N. and Kempe, J. What can the neural tangent kernel tell us about adversarial robustness? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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

Wang, Z. and Zhu, Y. Deformed semicircle law and concentration of nonlinear random matrices for ultra-wide neural networks. *arXiv preprint arXiv:2109.09304*, 2021.

Xiao, K. Y., Engstrom, L., Ilyas, A., and Madry, A. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations (ICLR)*, 2021.

Yang, Y.-Y., Chou, C.-N., and Chaudhuri, K. Understanding rare spurious correlations in neural networks. In *ICML Workshop on Spurious Correlations, Invariance and Stability*, 2022.

Zhou, C., Ma, X., Michel, P., and Neubig, G. Examining and combating spurious features under distribution shift. In *International Conference on Machine Learning (ICML)*, 2021.

Zliobaite, I. On the relation between accuracy and fairness in binary classification. In *2nd ICML Workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML)*, 2015.

A. 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 (Vershynin, 2018)).

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 $t > 0$,

$$\mathbb{P}\left(|\tau(x) - \mathbb{E}_X[\tau(x)]| > t\right) \leq 2e^{-ct^2/\|\tau\|_{\text{Lip}}^2}. \tag{31}$$

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 $v^\top Av \geq 0$.

A.1. Hermite polynomials

In this subsection, we refresh standard notions on the Hermite polynomials. For a more comprehensive discussion, we refer to (O'Donnell, 2014). 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 A.1 (Proposition 11.31, (O'Donnell, 2014))**.** *Let ρ_1, ρ_2 be two standard Gaussian random variables, with correlation $\rho \in [-1, 1]$. Then,*

$$\mathbb{E}_{\rho_1,\rho_2}\left[h_i(\rho_1)h_j(\rho_2)\right] = \delta_{ij}\rho^i, \tag{32}$$

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

The first 5 Hermite polynomials are

$$h_0(\rho) = 1, \quad h_1(\rho) = \rho, \quad h_2(\rho) = \frac{\rho^2 - 1}{\sqrt{2}}, \quad h_3(\rho) = \frac{\rho^3 - 3\rho}{\sqrt{6}}, \quad h_4(\rho) = \frac{\rho^4 - 6\rho^2 + 3}{\sqrt{24}}. \tag{33}$$

Proposition A.2 (Definition 11.34, (O'Donnell, 2014))**.** *Every function $\phi \in L^2(\mathbb{R}, \gamma)$ is uniquely expressible as*

$$\phi(\rho) = \sum_{i\in\mathbb{N}} \mu_i^\phi h_i(\rho), \tag{34}$$

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,

$$\lim_{n\to+\infty} \left\| \left(\sum_{i=0}^{n} \mu_i^\phi h_i(\rho) \right) - \phi(\rho) \right\|_{L^2(\mathbb{R},\gamma)} = 0. \tag{35}$$

This readily implies the following result.

Proposition A.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,*

$$\mathbb{E}_{\rho_1,\rho_2}\left[\phi(\rho_1)\tau(\rho_2)\right] = \sum_{i\in\mathbb{N}} \mu_i^\phi \mu_i^\tau \rho^i. \tag{36}$$

B. Proof of Lemma 4.1

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 $O^\top$. This notation implies that if $O_1 u = 0$ 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

$$P_A = O_1^\top O_1. \tag{37}$$

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

$$\begin{aligned} P_A =& O_1^\top O_1 = O I_N O^\top \\ =& O D^\top D_s^{-2} D O^\top = O D^\top U^\top U D_s^{-2} U^\top U D O^\top \\ =& A^\top \left(U D_s^2 U^\top\right)^{-1} A = A^\top \left(U D O^\top O D^\top U^\top\right)^{-1} A \\ =& A^\top \left(A A^\top\right)^{-1} A \equiv A^+ A, \end{aligned} \tag{38}$$

where A^+ denotes the Moore-Penrose inverse.

Notice that this last form enables us to easily derive

$$P_{A_{-1}} A^+ v = A_{-1}^+ A_{-1} A^+ v = A_{-1}^+ I_{-1} A A^+ v = A_{-1}^+ I_{-1} v = A_{-1}^+ v_{-1}, \tag{39}$$

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 B.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,*

$$\left\| P_{\Phi_{-1}}^\perp \varphi(z_1) \right\|_2^2 \geq \lambda_{\min}\left(\Phi\Phi^\top\right). \tag{40}$$

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

$$\left\|\Phi^\top u\right\|_2^2 \geq \lambda_{\min}\left(\Phi\Phi^\top\right) \|u\|_2^2 \geq \lambda_{\min}\left(\Phi\Phi^\top\right). \tag{41}$$

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

$$\Phi^\top u = \varphi(z_1) + \Phi_{-1}^\top u_{-1} = \varphi(z_1) - P_{\Phi_{-1}}\varphi(z_1) = P_{\Phi_{-1}}^\perp \varphi(z_1). \tag{42}$$

Plugging this in (41), we get the thesis. □

At this point, we are ready to prove Lemma 4.1.

Proof of Lemma 4.1. 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

$$P_\Phi = P_{\Phi_{-1}} + \frac{P^\perp_{\Phi_{-1}}\varphi(z_1)\varphi(z_1)^\top P^\perp_{\Phi_{-1}}}{\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2}. \tag{43}$$

The above expression is a consequence of the Gram-Schmidt formula, and the quantity at the denominator is different from zero because of Lemma B.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 (38). Thus, again using (3), for any z we can write

$$\begin{aligned} f(z,\theta^*) - \varphi(z)^\top\theta_0 &= \varphi(z)^\top\Phi^+\left(G - f(Z,\theta_0)\right) \\ &= \varphi(z)^\top P_\Phi\Phi^+\left(G - f(Z,\theta_0)\right) \\ &= \varphi(z)^\top\left(P_{\Phi_{-1}} + \frac{P^\perp_{\Phi_{-1}}\varphi(z_1)\varphi(z_1)^\top P^\perp_{\Phi_{-1}}}{\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2}\right)\Phi^+\left(G - f(Z,\theta_0)\right). \end{aligned} \tag{44}$$

Notice that, thanks to (39), we can manipulate the first term in the bracket as follows

$$\begin{aligned} \varphi(z)^\top P_{\Phi_{-1}}\Phi^+\left(G - f(Z,\theta_0)\right) &= \varphi(z)^\top\Phi^+_{-1}\left(G_{-1} - f(Z_{-1},\theta_0)\right) \\ &= f(z,\theta^*_{-1}) - \varphi(z)^\top\theta_0. \end{aligned} \tag{45}$$

Thus, bringing the result of (45) on the LHS, (44) becomes

$$\begin{aligned} f(z,\theta^*) - f(z,\theta^*_{-1}) &= \frac{\varphi(z)^\top P^\perp_{\Phi_{-1}}\varphi(z_1)}{\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2}\varphi(z_1)^\top P^\perp_{\Phi_{-1}}\Phi^+\left(G - f(Z,\theta_0)\right) \\ &= \frac{\varphi(z)^\top P^\perp_{\Phi_{-1}}\varphi(z_1)}{\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2}\varphi(z_1)^\top\left(I - P_{\Phi_{-1}}\right)\Phi^+\left(G - f(Z,\theta_0)\right) \\ &= \frac{\varphi(z)^\top P^\perp_{\Phi_{-1}}\varphi(z_1)}{\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2}\left(f(z_1,\theta^*) - f(z_1,\theta^*_{-1})\right), \end{aligned} \tag{46}$$

where in the last step we again used (3) and (45). □

C. Useful lemmas

Lemma C.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,*

$$\begin{aligned} |\zeta(x+\delta,y) - \zeta(x,y)| &\le L\left\|\delta\right\|_2, \\ |\zeta(x,y+\delta) - \zeta(x,y)| &\le L\left\|\delta\right\|_2, \end{aligned} \tag{47}$$

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

$$\mathbb{P}_{xy}\left(|\tau\left(\zeta(x,y)\right) - \mathbb{E}_{xy}\left[\tau\left(\zeta(x,y)\right)\right]| > t\right) < e^{-ct^2}, \tag{48}$$

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

$$\begin{aligned}&|\tau\left(\zeta(x,y)\right) - \mathbb{E}_{xy}\left[\tau\left(\zeta(x,y)\right)\right]| \\ &\leq |\tau\left(\zeta(x,y)\right) - \mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right]| + |\mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right] - \mathbb{E}_{y}\mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right]| =: A + B.\end{aligned} \tag{49}$$

Thus, we can upper bound LHS of (48) as follows:

$$\mathbb{P}_{xy}\left(|\tau\left(\zeta(x,y)\right) - \mathbb{E}_{xy}\left[\tau\left(\zeta(x,y)\right)\right]| > t\right) \leq \mathbb{P}_{xy}\left(A + B > t\right). \tag{50}$$

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 (48) is also upper bounded by

$$\mathbb{P}_{xy}\left(|\tau\left(\zeta(x,y)\right) - \mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right]| > t/2\right) + \mathbb{P}_{xy}\left(|\mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right] - \mathbb{E}_{y}\mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right]| > t/2\right). \tag{51}$$

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

$$\mathbb{P}_{xy}\left(|\tau\left(\zeta(x,y)\right) - \mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right]| > t/2\right) < e^{-c_1 t^2}, \tag{52}$$

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

$$|\chi(y+\delta) - \chi(y)| = |\mathbb{E}_x\left[\tau\left(\zeta(x,y+\delta)\right) - \tau\left(\zeta(x,y)\right)\right]| \leq \mathbb{E}_x\left[|\tau\left(\zeta(x,y+\delta)\right) - \tau\left(\zeta(x,y)\right)|\right] \leq L\left\|\delta\right\|_2. \tag{53}$$

Then, we can write

$$\mathbb{P}_{xy}\left(|\mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right] - \mathbb{E}_{y}\mathbb{E}_{x}\left[\tau\left(\zeta(x,y)\right)\right]| > t/2\right) = \mathbb{P}_{y}\left(|\chi(y) - \mathbb{E}_{y}\left[\chi(y)\right]| > t/2\right) < e^{-c_2 t^2}, \tag{54}$$

for some absolute constant c_2. Thus,

$$\mathbb{P}_{xy}\left(|\tau\left(\zeta(x,y)\right) - \mathbb{E}_{xy}\left[\tau\left(\zeta(x,y)\right)\right]| > t\right) < e^{-c_1 t^2} + e^{-c_2 t^2} \leq e^{-ct^2}, \tag{55}$$

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

Lemma C.2. *Let $x \sim \mathcal{P}_X$, $y \sim \mathcal{P}_Y$ and $z = [x, y] \sim \mathcal{P}_Z$. Let Assumption 5.1 hold. Then, z is a Lipschitz concentrated random vector.*

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

$$\mathbb{P}_{z}\left(|\tau\left(z\right) - \mathbb{E}_{z}\left[\tau\left(z\right)\right]| > t\right) < e^{-ct^2}, \tag{56}$$

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

$$|\tau\left(z\right) - \tau\left(z'\right)| \leq \left\|z - z'\right\|_2 = \left\|x - x'\right\|_2, \tag{57}$$

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 C.1 gives the thesis. □

Lemma C.3. *Let τ and ζ be two Lipschitz functions. Let $z, z' \in \mathbb{R}^d$ be two fixed vectors such that $\left\|z\right\|_2 = \left\|z'\right\|_2 = \sqrt{d}$. 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 $t > 1$,*

$$\left|\tau(Vz)^\top \zeta(Vz') - \mathbb{E}_V\left[\tau(Vz)^\top \zeta(Vz')\right]\right| = \mathcal{O}\left(\sqrt{k}\log t\right), \tag{58}$$

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

$$\mathbb{E}_V\left[\left\|\tau(Vz)\right\|_2^2\right] = k\mathbb{E}_\rho\left[\tau^2(\rho)\right], \tag{59}$$

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

Proof. We have

$$\tau(Vz)^\top \zeta(Vz') = \sum_{j=1}^{k} \tau(v_j^\top z)\zeta(v_j^\top z'), \tag{60}$$

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 (Vershynin, 2018)), we have

$$\left|\tau(Vz)^\top \zeta(Vz') - \mathbb{E}_V\left[\tau(Vz)^\top \zeta(Vz')\right]\right| = \mathcal{O}\left(\sqrt{k}\log t\right). \tag{61}$$

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 C.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 5.1 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 $z_1^s := [x, y_1]$. Let $\alpha = d_y/d \in (0,1)$ and μ_l be the l-th Hermite coefficient of τ. Then, for any $t > 1$,*

$$\left|\tau(Vz_1^s)^\top \tau(Vz_1) - k\sum_{l=0}^{+\infty} \mu_l^2 \alpha^l\right| = \mathcal{O}\left(\sqrt{k}\left(\sqrt{\frac{k}{d}} + 1\right)\log t\right), \tag{62}$$

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

$$x' = \frac{\sqrt{d_x}\left(I - \frac{x_1 x_1^\top}{d_x}\right)x}{\left\|\left(I - \frac{x_1 x_1^\top}{d_x}\right)x\right\|_2}. \tag{63}$$

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

$$\frac{\left|x^\top x'\right|}{d_x} = \sqrt{1 - \left(\frac{x^\top x_1}{d_x}\right)^2} \geq 1 - \left(\frac{x^\top x_1}{d_x}\right)^2, \tag{64}$$

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

$$\|x - x'\|_2^2 = \|x\|_2^2 + \|x'\|_2^2 - 2x^\top x' \leq 2d_x\left(1 - \left(1 - \left(\frac{x^\top x_1}{d_x}\right)^2\right)\right) = 2\frac{\left(x^\top x_1\right)^2}{d_x}. \tag{65}$$

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

$$\mathbb{P}\left(\|x - x'\|_2 > t\right) \leq \mathbb{P}\left(|x^\top x_1| > \sqrt{d_x}t/\sqrt{2}\right) < \exp(-ct^2), \tag{66}$$

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 $\|z_1^s - z'\|_2 = \|x - x'\|_2$. We can write

$$\begin{aligned}
\left|\tau(Vz_1^s)^\top \tau(Vz_1) - \tau(Vz')^\top \tau(Vz_1)\right| &\leq \|\tau(Vz_1^s) - \tau(Vz')\|_2 \, \|\tau(Vz_1)\|_2 \\
&\leq C\,\|V\|_{\text{op}} \, \|z_1^s - z'\|_2 \, \|\tau(Vz_1)\|_2 \\
&\leq C_1\left(\sqrt{\frac{k}{d}} + 1\right)\|x - x'\|_2\sqrt{k} \\
&= \mathcal{O}\left(\sqrt{k}\left(\sqrt{\frac{k}{d}} + 1\right)\log t\right).
\end{aligned} \tag{67}$$

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 rTheorem 4.4.5 of (Vershynin, 2018) and Lemma C.3; the fourth step holds with probability at least $1-\exp(-c\log^2 t)$, and it uses (66). This probability is intended over V and x. We further have

$$\left|\tau(Vz')^\top\tau(Vz_1)-\mathbb{E}_V\left[\tau(Vz')^\top\tau(Vz_1)\right]\right|=\mathcal{O}\left(\sqrt{k}\log t\right), \tag{68}$$

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

We have

$$\mathbb{E}_V\left[\tau(Vz')^\top\tau(Vz_1)\right]=k\mathbb{E}_{\rho_1\rho_2}\left[\tau(\rho_1)\tau(\rho_2)\right], \tag{69}$$

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

$$\text{corr}(\rho_1,\rho_2)=\frac{z_1^\top z}{\|z_1\|_2\|z'\|_2}=\frac{d_y}{d}=\alpha. \tag{70}$$

Then, exploiting the Hermite expansion of τ, we have

$$\mathbb{E}_{\rho_1\rho_2}\left[\tau(\rho_1)\tau(\rho_2)\right]=\sum_{l=0}^{+\infty}\mu_l^2\alpha^l. \tag{71}$$

Putting together (67), (68), (69), and (71) gives the thesis. □

D. 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 5.1). 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 (14))

$$\varphi(z):=\phi(Vz)\in\mathbb{R}^k, \tag{72}$$

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 $z_1^s=[x,y_1]$, 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 Section 5. Again for the sake of compactness, we will not re-introduce such quantities in the statements or the proofs of the following lemmas.

The content of this section can be summarized as follows:

- In Lemma D.2 we prove a lower bound on the smallest eigenvalue of K, adapting to our settings Lemma C.5 of (Bombari et al., 2023). As our assumptions are less restrictive than those in (Bombari et al., 2023), we will crucially exploit Lemma D.1.
- In Lemma D.3, 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 5.4. This step is necessary only if $\mu_0\neq 0$.
- In Lemma D.4, we show that the non-linear component of the features $\tilde{\varphi}(z_1)-\mu_1 Vz_1$ and $\tilde{\varphi}(z_1^s)-\mu_1 Vz_1^s$ have a negligible component in the space spanned by the rows of Φ_{-1}.
- In Lemma D.7, we provide concentration results for $\varphi(z_1^s)^\top P_{\Phi_{-1}}^\perp\varphi(z_1)$, and we lower bound this same term in Lemma D.6, exploiting also the intermediate result provided in Lemma D.5.
- Finally, we prove Theorem 5.4.

Lemma D.1. *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 A. We have*

$$\lambda_{\min}\left(AA^\top\right) = \Omega(d^m), \tag{73}$$

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 (Schur, 1911):

$$\lambda_{\min}\left(AA^\top\right) = \lambda_{\min}\left(A_2 A_2^\top \circ A_m A_m^\top\right) \geq \lambda_{\min}\left(A_2 A_2^\top\right) \min_i \|(A_m)_{i:}\|_2^2 . \tag{74}$$

Note that the rows of Z are mean-0 and Lipschitz concentrated by Lemma C.2. Then, by following the argument of Lemma C.3 in (Bombari et al., 2023), we have

$$\lambda_{\min}\left(A_2 A_2^\top\right) = \Omega(d^2), \tag{75}$$

with probability at least $1 - \exp(-c \log^2 N)$ over Z. We remark that, for the argument of Lemma C.3 in (Bombari et al., 2023) 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 (Bombari et al., 2023)), which is implied by Assumption 5.2, despite it being milder than Assumption 4 in (Bombari et al., 2023).

For the second term of (74), we have

$$\|(A_m)_{i:}\|_2^2 = \|z_i\|_2^{2(m-2)} = d^{m-2}, \tag{76}$$

due to Assumption 5.1. Thus, the thesis readily follows. □

Lemma D.2. *We have that*

$$\lambda_{\min}(K) = \Omega(k), \tag{77}$$

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 (Bombari et al., 2023). In particular, we define a truncated version of Φ as follows

$$\bar{\Phi}_{:j} = \phi(Zv_j)\chi\left(\|\phi(Zv_j)\|_2^2 \leq R\right), \tag{78}$$

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 (Tropp, 2012)), 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

$$\lambda_{\min}(K) \geq \lambda_{\min}\left(\bar{\Phi}\bar{\Phi}^\top\right) \geq \lambda_{\min}\left(\bar{G}\right)/2 \geq \lambda_{\min}(G)/2 - o(k), \tag{79}$$

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 (Bombari et al., 2023)), and the third comes from Equation (C.45) in (Bombari et al., 2023). To perform these steps, our Assumptions 5.2 and 5.3 are enough, despite the second one being milder than Assumption 2 in (Bombari et al., 2023).

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

$$G = \mathbb{E}_V[K] = \mathbb{E}_V\left[\sum_{i=1}^{k} \phi(ZV_{i:}^\top)\phi(ZV_{i:}^\top)^\top\right] = k\mathbb{E}_v\left[\phi(Zv)\phi(Zv)^\top\right] := kM, \tag{80}$$

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

$$M_{ij} = \mathbb{E}_v\left[\phi(z_i^\top v)\phi(z_j^\top v)\right] = \sum_{l=0}^{+\infty} \mu_l^2 \frac{\left(z_i^\top z_j\right)^l}{d^l} = \sum_{l=0}^{+\infty} \mu_l^2 \frac{\left[\left(Z^{*l}\right)\left(Z^{*l}\right)^\top\right]_{ij}}{d^l}, \tag{81}$$

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

$$A := \left(Z^{*m}\right). \tag{82}$$

By Lemma D.1, the desired result readily follows. □

Lemma D.3. *Let $\mu_0 \neq 0$. Then,*

$$\left\|P_{\Phi_{-1}}^\perp \mathbf{1}_k\right\|_2 = o(\sqrt{k}), \tag{83}$$

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

$$\left\|\tilde{\Phi}_{:i}\right\|_{\psi_2} = \left\|\phi(ZV_{i:}) - \mathbb{E}_V[\phi(ZV_{i:})]\right\|_{\psi_2} \leq L\frac{\|Z\|_{\text{op}}}{\sqrt{d}} = \mathcal{O}\left(\sqrt{N/d} + 1\right), \tag{84}$$

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 (Bombari et al., 2022b).

Thus, another application of Lemma B.7 in (Bombari et al., 2022b) gives

$$\left\|\tilde{\Phi}_{-1}^\top\right\|_{\text{op}} = \mathcal{O}\left(\left(\sqrt{k} + \sqrt{N}\right)\left(\sqrt{N/d} + 1\right)\right) = \mathcal{O}\left(\sqrt{k}\left(\sqrt{N/d} + 1\right)\right), \tag{85}$$

where the first equality holds with probability at least $1 - e^{-cN}$ over V, and the second is a direct consequence of Assumption 5.2.

We can write

$$\Phi_{-1}^\top \frac{\mathbf{1}_{N-1}}{\mu_0(N-1)} = \left(\mu_0 \mathbf{1}_k \mathbf{1}_{N-1}^\top + \tilde{\Phi}_{-1}^\top\right)\frac{\mathbf{1}_{N-1}}{\mu_0(N-1)} = \mathbf{1}_k + \tilde{\Phi}_{-1}^\top \frac{\mathbf{1}_{N-1}}{\mu_0(N-1)} =: \mathbf{1}_k + v, \tag{86}$$

where

$$\|v\|_2 \leq \frac{1}{\mu_0(N-1)}\left\|\tilde{\Phi}_{-1}^\top\right\|_{\text{op}} \|\mathbf{1}_{N-1}\|_2 = \mathcal{O}\left(\sqrt{\frac{k}{N}}\left(\sqrt{N/d} + 1\right)\right) = o(\sqrt{k}). \tag{87}$$

Thus, we can conclude

$$\begin{aligned}\left\|P_{\Phi_{-1}}^\perp \mathbf{1}_k\right\|_2 &= \left\|P_{\Phi_{-1}}^\perp\left(\Phi_{-1}^\top \frac{\mathbf{1}_{N-1}}{\mu_0(N-1)} - v\right)\right\|_2 \\ &\leq \left\|P_{\Phi_{-1}}^\perp P_{\Phi_{-1}} \Phi_{-1}^\top \frac{\mathbf{1}_{N-1}}{\mu_0(N-1)}\right\|_2 + \|v\|_2 = o(\sqrt{k}),\end{aligned} \tag{88}$$

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

Lemma D.4. *Let $z \sim \mathcal{P}_Z$, sampled independently from Z_{-1}, and denote $\tilde{\phi}(x) := \phi(x) - \mu_0$. Then,*

$$\left\|P_{\Phi_{-1}}\left(\tilde{\phi}(Vz) - \mu_1 Vz\right)\right\|_2 = o(\sqrt{k}), \tag{89}$$

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

$$\begin{aligned}\left\|P_{\Phi_{-1}}\left(\tilde{\phi}(Vz)-\mu_1 Vz\right)\right\|_2 &\leq \left\|\Phi_{-1}^{+}\right\|_{\text{op}}\left\|\Phi_{-1}\left(\tilde{\phi}(Vz)-\mu_1 Vz\right)\right\|_2 \\ &= \mathcal{O}\left(\frac{\left\|\Phi_{-1}\left(\tilde{\phi}(Vz)-\mu_1 Vz\right)\right\|_2}{\sqrt{k}}\right),\end{aligned} \tag{90}$$

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 D.2.

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

$$|u_i - \mathbb{E}_V[u_i]| = \mathcal{O}\left(\sqrt{k}\log N\right), \tag{91}$$

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 $1 \leq i \leq N-1$, with probability at least $1-(N-1)\exp(-c\log^2 N) \geq 1-\exp(-c_1\log^2 N)$. Thus, we have

$$\|u - \mathbb{E}_V[u]\|_2 = \mathcal{O}\left(\sqrt{k}\sqrt{N}\log N\right) = o(k), \tag{92}$$

where the last equality holds because of Assumption 5.2.

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

$$\begin{aligned}|\mathbb{E}_V[u_i]| &\leq k\sum_{l=2}^{+\infty}\mu_l^2\left(\frac{\left|z^\top z_i\right|}{\|z\|_2\|z_i\|_2}\right)^l \\ &\leq k\max_l\mu_l^2\sum_{l=2}^{+\infty}\left(\frac{\left|z^\top z_i\right|}{\|z\|_2\|z_i\|_2}\right)^l \\ &= k\max_l\mu_l^2\left(\frac{z^\top z_i}{\|z\|_2\|z_i\|_2}\right)^2\frac{1}{1-\frac{|z^\top z_i|}{\|z\|_2\|z_i\|_2}} \\ &\leq 2k\max_l\mu_l^2\left(\frac{z^\top z_i}{\|z\|_2\|z_i\|_2}\right)^2 = \mathcal{O}\left(\frac{k\log^2 N}{d}\right),\end{aligned} \tag{93}$$

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 $1 \leq i \leq N-1$, with probability at least $1-(N-1)\exp(-c\log^2 N) \geq 1-\exp(-c_1\log^2 N)$. Then, we have

$$\|\mathbb{E}_V[u]\|_2 = \mathcal{O}\left(\sqrt{N}\,\frac{k\log^2 N}{d}\right) = o(k), \tag{94}$$

where the last equality is a consequence of Assumption 5.2.

Finally, (92) and (94) give

$$\left\|\Phi_{-1}\left(\tilde{\phi}(Vz)-\mu_1 Vz\right)\right\|_2 \leq \|\mathbb{E}_V[u]\|_2 + \|u-\mathbb{E}_V[u]\|_2 = o(k), \tag{95}$$

which plugged in (90) readily provides the thesis. □

Lemma D.5. *We have*

$$\left|(Vz_1^s)^\top P_{\Phi_{-1}}^\perp Vz_1 - \left\|P_{\Phi_{-1}}^\perp V_y y_1\right\|_2^2\right| = o(k), \tag{96}$$

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

$$Vz_1^s = V_x x + V_y y_1, \qquad Vz_1 = V_x x_1 + V_y y_1. \tag{97}$$

Thus, we can write

$$\begin{aligned}\left|(Vz_1^s)^\top P_{\Phi_{-1}}^\perp Vz_1 - \left\|P_{\Phi_{-1}}^\perp V_y y_1\right\|_2^2\right| &= \left|(V_x x)^\top P_{\Phi_{-1}}^\perp Vz_1 + (V_y y_1)^\top P_{\Phi_{-1}}^\perp V_x x_1\right| \\ &\le \left|x^\top V_x^\top P_{\Phi_{-1}}^\perp Vz_1\right| + \left|y_1^\top V_y^\top P_{\Phi_{-1}}^\perp V_x x_1\right|.\end{aligned} \tag{98}$$

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 (Vershynin, 2018). 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 $V_x^\top P_{\Phi_{-1}}^\perp Vz_1$, we have

$$\begin{aligned}\left|x^\top V_x^\top P_{\Phi_{-1}}^\perp Vz_1\right| &\le \log N \left\|V_x^\top P_{\Phi_{-1}}^\perp Vz_1\right\|_2 \\ &\le \log N \|V_x\|_{\text{op}} \left\|P_{\Phi_{-1}}^\perp\right\|_{\text{op}} \|V\|_{\text{op}} \|z_1\| \\ &= \mathcal{O}\left(\log N\left(\frac{k}{d}+1\right)\sqrt{d}\right) = o(k),\end{aligned} \tag{99}$$

where the first inequality holds with probability at least $1-\exp(-c\log^2 N)$ over x, and the last line holds because $\left\|P_{\Phi_{-1}}^\perp\right\|_{\text{op}} \le 1$, $\|z_1\| = \sqrt{d}$, and because of Assumption 5.2.

Similarly, exploiting the independence between x_1 and y_1, we can prove that $\left|y_1^\top V_y^\top P_{\Phi_{-1}}^\perp V_x x_1\right| = o(k)$, with probability at least $1-\exp(-c\log^2 N)$ over y_1. Plugging this and (99) in (98) readily gives the thesis. □

Lemma D.6. *We have*

$$\left|\varphi(z_1^s)^\top P_{\Phi_{-1}}^\perp \varphi(z_1) - \left(k\left(\sum_{l=2}^{+\infty}\mu_l^2\alpha^l\right) + \mu_1^2\left\|P_{\Phi_{-1}}^\perp V_y y_1\right\|_2^2\right)\right| = o(k), \tag{100}$$

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 C.3 and Assumption 5.2 gives

$$\begin{aligned}\|\varphi(z_1)\|_2 &= \mathcal{O}\left(\sqrt{k}\right), \qquad \|\varphi(z_1^s)\|_2 = \mathcal{O}\left(\sqrt{k}\right), \\ \|Vz_1\|_2 &= \mathcal{O}\left(\sqrt{k}\right), \qquad \|Vz_1^s\|_2 = \mathcal{O}\left(\sqrt{k}\right),\end{aligned} \tag{101}$$

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

$$\left|\varphi(z_1^s)^\top P_{\Phi_{-1}}^\perp \varphi(z_1) - \tilde{\phi}(Vz_1^s)^\top P_{\Phi_{-1}}^\perp \tilde{\phi}(Vz_1)\right| = o(k), \tag{102}$$

with probability at least $1-\exp(c_2\log^2 N)$ over V and Z, because of (101) and Lemma D.3. Note that (102) trivially holds even when $\mu_0=0$, as $\phi\equiv\tilde{\phi}$. Thus, (102) is true in any case with probability at least $1-\exp(c_2\log^2 N)$ over V and Z.

Furthermore, because of (101) and Lemma D.4, we have

$$\left|\tilde{\phi}(Vz_1^s)^\top P_{\Phi_{-1}}\tilde{\phi}(Vz_1)-\mu_1^2(Vz_1^s)^\top P_{\Phi_{-1}}(Vz_1)\right|=o(k), \tag{103}$$

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

Thus, putting (102) and (103) together, and using Lemma D.5, we get

$$\begin{aligned}
&\left|\varphi(z_1^s)^\top P^\perp_{\Phi_{-1}}\varphi(z_1)-\left(\tilde{\phi}(Vz_1^s)^\top\tilde{\phi}(Vz_1)-\mu_1^2(Vz_1^s)^\top(Vz_1)+\mu_1^2\left\|P^\perp_{\Phi_{-1}}V_y y_1\right\|_2^2\right)\right| \\
&\quad\le\left|\varphi(z_1^s)^\top P^\perp_{\Phi_{-1}}\varphi(z_1)-\tilde{\phi}(Vz_1^s)^\top P^\perp_{\Phi_{-1}}\tilde{\phi}(Vz_1)\right| \\
&\qquad+\left|-\tilde{\phi}(Vz_1^s)^\top P_{\Phi_{-1}}\tilde{\phi}(Vz_1)+\mu_1^2(Vz_1^s)^\top P_{\Phi_{-1}}(Vz_1)\right| \\
&\qquad+\left|\mu_1^2(Vz_1^s)^\top P^\perp_{\Phi_{-1}}(Vz_1)-\mu_1^2\left\|P^\perp_{\Phi_{-1}}V_y y_1\right\|_2^2\right|=o(k),
\end{aligned} \tag{104}$$

with probability at least $1-\exp(-c_4\log^2 N)$ over V and X and x. To conclude we apply Lemma C.4 setting $t=N$, together with Assumption 5.2, to get

$$\left|\tilde{\phi}(Vz_1^s)^\top\tilde{\phi}(Vz_1)-k\left(\sum_{l=1}^{+\infty}\mu_l^2\alpha^l\right)\right|=\mathcal{O}\left(\sqrt{k}\left(\sqrt{\frac{k}{d}}+1\right)\log N\right)=o(k), \tag{105}$$

and

$$\left|\mu_1^2(Vz_1^s)^\top(Vz_1)-k\mu_1^2\alpha\right|=\mathcal{O}\left(\sqrt{k}\left(\sqrt{\frac{k}{d}}+1\right)\log N\right)=o(k), \tag{106}$$

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

Applying the triangle inequality to (104), (105), and (106), we get the thesis. □

Lemma D.7. *We have that*

$$\left|\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2-\mathbb{E}_{z_1}\left[\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2\right]\right|=o(k), \tag{107}$$

$$\left|\varphi(z_1^s)^\top P^\perp_{\Phi_{-1}}\varphi(z_1)-\mathbb{E}_{z_1,z_1^s}\left[\varphi(z_1^s)^\top P^\perp_{\Phi_{-1}}\varphi(z_1)\right]\right|=o(k), \tag{108}$$

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 (Vershynin, 2018). This also implies that $\|V\|_{\text{op}}=\mathcal{O}\left(\sqrt{k/d}+1\right)$.

We indicate with $\nu:=\mathbb{E}_{z_1}\left[\varphi(z_1)\right]=\mathbb{E}_{z_1^s}\left[\varphi(z_1^s)\right]\in\mathbb{R}^k$, 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 Assumption5.2, we have

$$\left|\|\varphi(z_1)\|_2-\mathbb{E}_{z_1}\left[\|\varphi(z_1)\|_2\right]\right|=o\left(\sqrt{k}\right), \tag{109}$$

with probability at least $1-\exp(-c_2\log^2 N)$ over z_1 and V. In addition, by the last statement of Lemma C.3 and Assumption 5.2, 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

$$\mathbb{E}_{z_1}\left[\|\varphi(z_1)\|_2\right]=\mathcal{O}\left(\sqrt{k}\right), \tag{110}$$

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

$$\|\nu\|_2 = \|\mathbb{E}_{z_1}\left[\varphi(z_1)\right]\|_2 \leq \mathbb{E}_{z_1}\left[\|\varphi(z_1)\|_2\right] = \mathcal{O}\left(\sqrt{k}\right). \tag{111}$$

We can now rewrite the LHS of the first statement as

$$\begin{aligned}
&\left|\left\|P^{\perp}_{\Phi_{-1}}\varphi(z_1)\right\|_2^2 - \mathbb{E}_{z_1}\left[\left\|P^{\perp}_{\Phi_{-1}}\varphi(z_1)\right\|_2^2\right]\right| \\
&= \left|\left\|P^{\perp}_{\Phi_{-1}}\left(\hat{\varphi}(z_1)+\nu\right)\right\|_2^2 - \mathbb{E}_{z_1}\left[\left\|P^{\perp}_{\Phi_{-1}}\left(\hat{\varphi}(z_1)+\nu\right)\right\|_2^2\right]\right| \\
&= \left|\hat{\varphi}(z_1)^\top P^{\perp}_{\Phi_{-1}}\hat{\varphi}(z_1) + 2\nu^\top P^{\perp}_{\Phi_{-1}}\hat{\varphi}(z_1) - \mathbb{E}_{z_1}\left[\hat{\varphi}(z_1)^\top P^{\perp}_{\Phi_{-1}}\hat{\varphi}(z_1)\right]\right| \\
&\leq \left|\hat{\varphi}(z_1)^\top P^{\perp}_{\Phi_{-1}}\hat{\varphi}(z_1) - \mathbb{E}_{z_1}\left[\hat{\varphi}(z_1)^\top P^{\perp}_{\Phi_{-1}}\hat{\varphi}(z_1)\right]\right| + 2\left|\nu^\top P^{\perp}_{\Phi_{-1}}\hat{\varphi}(z_1)\right|.
\end{aligned} \tag{112}$$

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 (111). Thus, by Assumption5.2, we have that

$$\left|\nu^\top P^{\perp}_{\Phi_{-1}}\hat{\varphi}(z_1)\right| = o(k), \tag{113}$$

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 (Adamczak, 2015), we can write

$$\begin{aligned}
&\mathbb{P}\left(\left|\left\|P^{\perp}_{\Phi_{-1}}\hat{\varphi}(z_1)\right\|_2^2 - \mathbb{E}_{z_1}\left[\left\|P^{\perp}_{\Phi_{-1}}\hat{\varphi}(z_1)\right\|_2^2\right]\right| \geq k/\log N\right) \\
&\leq 2\exp\left(-c_6\min\left(\frac{k^2}{\log^2 N\left((k/d)^2+1\right)\left\|P^{\perp}_{\Phi_{-1}}\right\|_F^2}, \frac{k}{\log N\left(k/d+1\right)\left\|P^{\perp}_{\Phi_{-1}}\right\|_{\text{op}}}\right)\right) \\
&\leq 2\exp\left(-c_6\min\left(\frac{k}{\log^2 N\left((k/d)^2+1\right)}, \frac{k}{\log N\left(k/d+1\right)}\right)\right) \\
&\leq \exp\left(-c_7\log^2 N\right),
\end{aligned} \tag{114}$$

where the last inequality comes from Assumption 5.2.This, together with (112) and (113), proves the first part of the statement.

For the second part of the statement, we have

$$\begin{aligned}
&\left|\varphi(z_1^s)^\top P^{\perp}_{\Phi_{-1}}\varphi(z_1) - \mathbb{E}_{z_1,z_1^s}\left[\varphi(z_1^s)^\top P^{\perp}_{\Phi_{-1}}\varphi(z_1)\right]\right| \\
&\leq \left|\hat{\varphi}(z_1^s)^\top P^{\perp}_{\Phi_{-1}}\hat{\varphi}(z_1) - \mathbb{E}_{z_1,z_1^s}\left[\hat{\varphi}(z_1^s)^\top P^{\perp}_{\Phi_{-1}}\hat{\varphi}(z_1)\right]\right| + \left|\nu^\top P^{\perp}_{\Phi_{-1}}\hat{\varphi}(z_1)\right| + \left|\nu^\top P^{\perp}_{\Phi_{-1}}\hat{\varphi}(z_1^s)\right|.
\end{aligned} \tag{115}$$

Following the same argument that led to (113), we obtain

$$\left|\nu^\top P^{\perp}_{\Phi_{-1}}\hat{\varphi}(z_1^s)\right| = o(k), \tag{116}$$

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

$$P_2 := \frac{1}{2}\left(\begin{array}{c|c} 0 & P^{\perp}_{\Phi_{-1}} \\ \hline P^{\perp}_{\Phi_{-1}} & 0 \end{array}\right), \qquad V_2 := \left(\begin{array}{c|c|c} V_x & V_y & 0 \\ \hline 0 & V_y & V_x \end{array}\right), \tag{117}$$

and

$$\hat{\varphi}_2 := \phi\left(V_2[x_1, y_1, x]^\top\right) - \mathbb{E}_{x_1,y_1,x}\left[\phi\left(V_2[x_1, y_1, x]^\top\right)\right] \equiv [\hat{\varphi}(z_1), \hat{\varphi}(z_1^s)]^\top. \tag{118}$$

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^s, which follows from applying Lemma C.2 twice. Also, we have

$$\hat{\varphi}(z_1^s)^\top P_{\Phi_{-1}}^\perp \hat{\varphi}(z_1) = \hat{\varphi}_2^\top P_2 \hat{\varphi}_2. \tag{119}$$

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^s), again by the general version of the Hanson-Wright inequality given by Theorem 2.3 in (Adamczak, 2015), we can write

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{\varphi}_2^\top P_2 \hat{\varphi}_2 - \mathbb{E}_{z_1,z_1^s}\left[\hat{\varphi}_2^\top P_2 \hat{\varphi}_2\right]\right| \geq k/\log N\right) \\
&\quad\leq 2\exp\left(-c_9 \min\left(\frac{k^2}{\log^2 N\left((k/d)^2+1\right)\|P_2\|_F^2}, \frac{k}{\log N\left(k/d+1\right)\|P_2\|_{\text{op}}}\right)\right) \\
&\quad\leq 2\exp\left(-c_9 \min\left(\frac{k}{\log^2 N\left((k/d)^2+1\right)}, \frac{k}{\log N\left(k/d+1\right)}\right)\right) \\
&\quad\leq \exp\left(-c_{10}\log^2 N\right),
\end{aligned} \tag{120}$$

where the last inequality comes from Assumption 5.2. This, together with (115), (113), (116), and (119), proves the second part of the statement, and therefore the desired result.

□

Finally, we are ready to give the proof of Theorem 5.4.

Proof of Theorem 5.4. We will prove the statement for the following definition of γ_{RF}, independent from z_1 and z_1^s,

$$\gamma_{\text{RF}} := \frac{\mathbb{E}_{z_1,z_1^s}\left[\varphi(z_1^s)^\top P_{\Phi_{-1}}^\perp \varphi(z_1)\right]}{\mathbb{E}_{z_1}\left[\left\|P_{\Phi_{-1}}^\perp \varphi(z_1)\right\|_2^2\right]}. \tag{121}$$

By Lemma B.1 and D.2, we have

$$\left\|P_{\Phi_{-1}}^\perp \varphi(z)\right\|_2^2 = \Omega(k) \tag{122}$$

with probability at least $1 - \exp(-c_1 \log^2 N)$ over V, Z_{-1} and z. This, together with Lemma D.7, gives

$$\left|\frac{\varphi(z_1^s)^\top P_{\Phi_{-1}}^\perp \varphi(z_1)}{\left\|P_{\Phi_{-1}}^\perp \varphi(z_1)\right\|_2^2} - \frac{\mathbb{E}_{z_1,z_1^s}\left[\varphi(z_1^s)^\top P_{\Phi_{-1}}^\perp \varphi(z_1)\right]}{\mathbb{E}_{z_1}\left[\left\|P_{\Phi_{-1}}^\perp \varphi(z_1)\right\|_2^2\right]}\right| = o(1), \tag{123}$$

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

$$\begin{aligned}
\frac{\mathbb{E}_{z_1,z_1^s}\left[\varphi(z_1^s)^\top P_{\Phi_{-1}}^\perp \varphi(z_1)\right]}{\mathbb{E}_{z_1}\left[\left\|P_{\Phi_{-1}}^\perp \varphi(z_1)\right\|_2^2\right]} &\leq \frac{\mathbb{E}_{z_1,z_1^s}\left[\left\|P_{\Phi_{-1}}^\perp \varphi(z_1^s)\right\|_2 \left\|P_{\Phi_{-1}}^\perp \varphi(z_1)\right\|_2\right]}{\mathbb{E}_{z_1}\left[\left\|P_{\Phi_{-1}}^\perp \varphi(z_1)\right\|_2^2\right]} \\
&\leq \frac{\sqrt{\mathbb{E}_{z_1^s}\left[\left\|P_{\Phi_{-1}}^\perp \varphi(z_1^s)\right\|_2^2\right]}\sqrt{\mathbb{E}_{z_1}\left[\left\|P_{\Phi_{-1}}^\perp \varphi(z_1)\right\|_2^2\right]}}{\mathbb{E}_{z_1}\left[\left\|P_{\Phi_{-1}}^\perp \varphi(z_1)\right\|_2^2\right]} = 1.
\end{aligned} \tag{124}$$

Let's now focus on the lower bound. By Assumption5.2 and Lemma C.4 (in which we consider the degenerate case $\alpha = 1$ and set $t = N$), we have

$$\left| \left\| \tilde{\phi}\left(Vz_1\right) \right\|_2^2 - k \sum_{l=1}^{+\infty} \mu_l^2 \right| = o(k), \tag{125}$$

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

$$\begin{aligned}
\frac{\mathbb{E}_{z_1, z_1^s}\left[\varphi(z_1^s)^\top P_{\Phi_{-1}}^\perp \varphi(z_1)\right]}{\mathbb{E}_{z_1}\left[\left\| P_{\Phi_{-1}}^\perp \varphi(z_1) \right\|_2^2\right]} &\geq \frac{\varphi(z_1^s)^\top P_{\Phi_{-1}}^\perp \varphi(z_1)}{\left\| P_{\Phi_{-1}}^\perp \varphi(z_1) \right\|_2^2} - o(1) \\
&\geq \frac{\varphi(z_1^s)^\top P_{\Phi_{-1}}^\perp \varphi(z_1)}{\left\| P_{\Phi_{-1}}^\perp \tilde{\varphi}(z_1) \right\|_2^2} - o(1) \\
&\geq \frac{k\left(\sum_{l=2}^{+\infty} \mu_l^2 \alpha^l\right) + \mu_1^2 \left\| P_{\Phi_{-1}}^\perp V_y y_1 \right\|_2^2}{\|\tilde{\varphi}(z_1)\|_2^2} - o(1) \\
&\geq \frac{k\left(\sum_{l=2}^{+\infty} \mu_l^2 \alpha^l\right) + \mu_1^2 \left\| P_{\Phi_{-1}}^\perp V_y y_1 \right\|_2^2}{k \sum_{l=1}^{+\infty} \mu_l^2} - o(1) \\
&\geq \frac{\sum_{l=2}^{+\infty} \mu_l^2 \alpha^l}{\sum_{l=1}^{+\infty} \mu_l^2} - o(1),
\end{aligned} \tag{126}$$

where the first inequality is a consequence of (123), the second of Lemma D.3 and (122), the third of Lemma D.6 and again (122), and the fourth of (125), 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. □

E. Proofs for NTK features

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 5.1). 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 (24))

$$\varphi(z) := z \otimes \phi'(Wz) \in \mathbb{R}^{dk}, \tag{127}$$

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 A. 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 $z_1^s = [x, y_1]$, 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 Section 6. Again for the sake of compactness, we will not re-introduce such quantities in the statements or the proofs of the following lemmas.

The content of this section can be summarized as follows:

- In Lemma E.1, we prove the lower bound on the smallest eigenvalue of K, adapting to our settings the main result of (Bombari et al., 2022b).
- In Lemma E.5, we treat separately a term that derives from $\mathbb{E}_W\left[\phi'(Wz)\right] = \mu_0' \mathbf{1}_k$, showing that we can *center* the derivative of the activation function (Lemma E.9), without changing our final statement in Theorem 6.3. This step is necessary only if $\mu_0' \neq 0$. Our proof tackles the problem proving the thesis on a set of "perturbed" inputs $\bar{Z}_{-1}(\delta)$ (Lemma E.4), critically exploiting the non degenerate behaviour of their rows (Lemma E.3), and transfers the result on the original term, using continuity arguments with respect to the perturbation (Lemma E.2).

- In Lemma E.8, we show that the centered features $\tilde{\varphi}(z_1)$ and $\tilde{\varphi}(z_1^s)$ have a negligible component in the space spanned by the rows of Φ_{-1}. To achieve this, we exploit the bound proved in Lemma E.7.
- To conclude, we prove Theorem 6.3, exploiting also the concentration result provided in Lemma E.6.

Lemma E.1. *We have that*

$$\lambda_{\min}(K) = \Omega(kd), \tag{128}$$

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 (Bombari et al., 2022b). 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 (Bombari et al., 2022b)) does not change the result. Assumption 6.1, *i.e.*, $k = \mathcal{O}(d)$, satisfies the *loose pyramidal topology* condition (cf. Assumption 2.4 in (Bombari et al., 2022b)), and Assumption 6.1 is the same as Assumption 2.5 in (Bombari et al., 2022b). 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 E.2. *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*

$$\bar{Z}_{-1}(\delta) := Z_{-1} + \delta A, \tag{129}$$

$$\bar{\Phi}_{-1}(\delta) := \bar{Z}_{-1}(\delta) * \phi'\left(Z_{-1}W^\top\right). \tag{130}$$

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

$$\left\|\bar{\Phi}_{-1}(\delta) - \bar{\Phi}_{-1}(0)\right\|_{\text{op}} = \left\|\delta A * \phi'\left(Z_{-1}W^\top\right)\right\|_{\text{op}} \le \delta \|A\|_{\text{op}} \max_{2\le i\le N} \|\phi'(Wz_i)\|_2, \tag{131}$$

where the second step follows from Equation (3.7.13) in (Johnson, 1990).

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 E.1, $\det\left(\bar{\Phi}_{-1}(0)\bar{\Phi}_{-1}(0)^\top\right) \equiv \det\left(\Phi_{-1}\Phi_{-1}^\top\right) \neq 0$ 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 (38)), we also have the continuity of $\bar{P}_{\Phi_{-1}}(\delta)$ in $\delta = 0$, which gives the thesis. □

Lemma E.3. *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 6.1,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 E.4. *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*

$$\left\|\bar{P}^\perp_{\Phi_{-1}}(\delta)\left(z \otimes \mathbf{1}_k\right)\right\|_2 = o(\sqrt{dk}), \tag{132}$$

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

$$\bar{\Phi}^\top_{-1}(\delta)\zeta = \left(\bar{Z}_{-1}(\delta) * B_{-1}\right)^\top \zeta = \left(\bar{Z}^\top_{-1}(\delta)\zeta\right) \otimes \left(B^\top_{-1}\mathbf{1}_{N-1}\right). \tag{133}$$

Note that $B^\top_{-1} = \mu_0'\mathbf{1}_k\mathbf{1}^\top_{N-1} + \tilde{B}^\top_{-1}$, where $\tilde{B}^\top_{-1} = \phi'(WZ^\top_{-1}) - \mathbb{E}_W\left[\phi'(WZ^\top_{-1})\right]$ is a $k \times (N-1)$ matrix with i.i.d. and mean-0 rows. For an argument equivalent to the one used for (84) and (85), we have

$$\left\|\tilde{B}^\top_{-1}\right\|_{\text{op}} = \mathcal{O}\left(\left(\sqrt{k} + \sqrt{N}\right)\left(\sqrt{N/d} + 1\right)\right), \tag{134}$$

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

$$B^\top_{-1}\frac{\mathbf{1}_{N-1}}{\mu_0'(N-1)} = \left(\mu_0'\mathbf{1}_k\mathbf{1}^\top_{N-1} + \tilde{B}^\top_{-1}\right)\frac{\mathbf{1}_{N-1}}{\mu_0'(N-1)} = \mathbf{1}_k + \tilde{B}^\top_{-1}\frac{\mathbf{1}_{N-1}}{\mu_0'(N-1)} =: \mathbf{1}_k + v, \tag{135}$$

where we have

$$\|v\|_2 \leq \left\|\tilde{B}^\top_{-1}\right\|_{\text{op}}\left\|\frac{\mathbf{1}_{N-1}}{\mu_0'(N-1)}\right\|_2 = \mathcal{O}\left(\left(\sqrt{k/N} + 1\right)\left(\sqrt{N/d} + 1\right)\right) = o(\sqrt{k}), \tag{136}$$

where the last step is a consequence of Assumption 6.1. Plugging (135) in (133) we get

$$\frac{1}{\mu_0'(N-1)}\bar{\Phi}^\top_{-1}(\delta)\zeta = \frac{1}{\mu_0'(N-1)}\left(\bar{Z}_{-1}(\delta) * B_{-1}\right)^\top \zeta = \left(\bar{Z}^\top_{-1}(\delta)\zeta\right) \otimes \left(\mathbf{1}_k + v\right). \tag{137}$$

By Lemma E.3, 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 $z = \bar{Z}^\top_{-1}(\delta)\zeta$. We can therefore rewrite the previous equation as

$$\frac{1}{\mu_0'(N-1)}\bar{\Phi}^\top_{-1}(\delta)\zeta = z \otimes \mathbf{1}_k + z \otimes v. \tag{138}$$

Thus, we can conclude

$$\begin{aligned}\left\|\bar{P}^\perp_{\Phi_{-1}}(\delta)\left(z \otimes \mathbf{1}_k\right)\right\|_2 &= \left\|P^\perp_{\Phi_{-1}}\left(\frac{\Phi^\top_{-1}(\delta)\zeta}{\mu_0'(N-1)} - z \otimes v\right)\right\|_2 \\ &\leq \left\|\bar{P}^\perp_{\Phi_{-1}}(\delta)\Phi^\top_{-1}(\delta)\frac{\zeta}{\mu_0'(N-1)}\right\|_2 + \|z \otimes v\|_2 \\ &= \|z\|_2\,\|v\|_2 = o(\sqrt{dk}),\end{aligned} \tag{139}$$

where in the second step we use the triangle inequality, in the third step we use that $\Phi^\top_{-1}(\delta) = \bar{P}_{\Phi_{-1}}(\delta)\Phi^\top_{-1}(\delta)$, and in the last step we use (136). The desired result readily follows. □

Lemma E.5. *Let $\mu_0' \neq 0$. Then, for any $z \in \mathbb{R}^d$, we have*

$$\left\|P^\perp_{\Phi_{-1}}\left(z \otimes \mathbf{1}_k\right)\right\|_2 = o(\sqrt{dk}), \tag{140}$$

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 $\bar{P}_{\Phi_{-1}}(\delta) \in \mathbb{R}^{dk\times dk}$ be the projector over the Span of the rows of $\bar{\Phi}_{-1}(\delta)$.

By triangle inequality, we can write

$$\left\|P^\perp_{\Phi_{-1}}\left(z\otimes \mathbf{1}_k\right)\right\|_2 \le \left\|P^\perp_{\Phi_{-1}} - \bar{P}^\perp_{\Phi_{-1}}(\delta)\right\|_{\text{op}} \|z\otimes \mathbf{1}_k\|_2 + \left\|\bar{P}^\perp_{\Phi_{-1}}(\delta)\left(z\otimes \mathbf{1}_k\right)\right\|_2. \tag{141}$$

Because of Lemma E.2, 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 $\delta \in [0, \delta^*]$,

$$\left\|P^\perp_{\Phi_{-1}} - \bar{P}^\perp_{\Phi_{-1}}(\delta)\right\|_{\text{op}} \equiv \left\|\bar{P}^\perp_{\Phi_{-1}}(0) - \bar{P}^\perp_{\Phi_{-1}}(\delta)\right\|_{\text{op}} < \frac{1}{N}. \tag{142}$$

Hence, setting $\delta = \delta^*$ in (141), we get

$$\begin{aligned}\left\|P^\perp_{\Phi_{-1}}\left(z\otimes \mathbf{1}_k\right)\right\|_2 &\le \left\|P^\perp_{\Phi_{-1}} - \bar{P}^\perp_{\Phi_{-1}}(\delta^*)\right\|_{\text{op}} \|z\otimes \mathbf{1}_k\|_2 + \left\|\bar{P}^\perp_{\Phi_{-1}}(\delta^*)\left(z\otimes \mathbf{1}_k\right)\right\|_2 \\ &\le \|z\|_2 \|\mathbf{1}_k\|_2 / N + \left\|\bar{P}^\perp_{\Phi_{-1}}(\delta^*)\left(z\otimes \mathbf{1}_k\right)\right\|_2 \\ &= o(\sqrt{dk}),\end{aligned} \tag{143}$$

where the last step is a consequence of Lemma E.4, 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 E.6. *We have*

$$\left|\frac{\tilde{\varphi}(z_1^s)^\top \tilde{\varphi}(z_1)}{\|\tilde{\varphi}(z_1)\|_2^2} - \alpha\frac{\sum_{l=1}^{+\infty}\mu_l'^2\alpha^i}{\sum_{l=1}^{+\infty}\mu_l'^2}\right| = o(1), \tag{144}$$

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*

$$\tilde{\varphi}(z_1^s)^\top \tilde{\varphi}(z_1) = \Theta(dk), \qquad \|\tilde{\varphi}(z_1)\|_2^2 = \Theta(dk). \tag{145}$$

Proof. We have

$$\|\tilde{\varphi}(z_1)\|_2^2 = \left\|z_1\otimes\tilde{\phi}'\left(Wz_1\right)\right\|_2^2 = \|z_1\|_2^2\left\|\tilde{\phi}'\left(Wz_1\right)\right\|_2^2 = d\left\|\tilde{\phi}'\left(Wz_1\right)\right\|_2^2. \tag{146}$$

By Assumption 6.1 and Lemma C.4 (in which we consider the degenerate case $\alpha = 1$ and set $t = k$), we have

$$\left|\left\|\tilde{\phi}'\left(Wz_1\right)\right\|_2^2 - k\sum_{l=1}^{+\infty}\mu_l'^2\right| = o(k), \tag{147}$$

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

$$\left|\|\tilde{\varphi}(z_1)\|_2^2 - dk\sum_{l=1}^{+\infty}\mu_l'^2\right| = o(dk). \tag{148}$$

Notice that the second term in the modulus is $\Theta(dk)$, since the μ_l'-s cannot be all 0, because of Assumption 6.2; this shows that $\|\tilde{\varphi}(z_1)\|_2^2 = \Theta(dk)$.

Similarly, we can write

$$\tilde{\varphi}(z_1^s)^\top \tilde{\varphi}(z_1) = \left(z_1^\top z_1^s\right)\left(\tilde{\phi}'\left(Wz_1\right)^\top \tilde{\phi}'\left(Wz_1^s\right)\right). \tag{149}$$

We have

$$\left|z_1^\top z_1^s - \alpha d\right| = \left|x_1^\top x\right| \le \sqrt{d_x}\log d = o(d), \tag{150}$$

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 Assumption6.1 and Lemma C.4, we have

$$\left|\tilde{\phi}'\left(Wz_1\right)^\top\tilde{\phi}'\left(Wz_1^s\right)-k\sum_{l=1}^{+\infty}{\mu_l'}^2\alpha^l\right|=o(k), \tag{151}$$

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 6.2.

Thus, putting (149), (150) and (151) together, we get

$$\left|\tilde{\varphi}(z_1^s)^\top\tilde{\varphi}(z_1)-dk\alpha\sum_{l=1}^{+\infty}{\mu_l'}^2\alpha^l\right|=o(dk), \tag{152}$$

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 $\tilde{\varphi}(z_1^s)^\top\tilde{\varphi}(z_1)=\Theta(dk)$.

Finally, merging (152) with (148) and applying triangle inequality, (144) follows and the proof is complete. □

Lemma E.7. *Let $z\sim\mathcal{P}_Z$ be sampled independently from Z_{-1}. Then,*

$$\left\|\Phi_{-1}\tilde{\varphi}(z)\right\|_2=o\left(dk\right), \tag{153}$$

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.*,

$$\varphi(z_{i+1})^\top\tilde{\varphi}(z)=\left(z_{i+1}^\top z\right)\left(\phi'(Wz_{i+1})^\top\tilde{\phi}'(Wz)\right). \tag{154}$$

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 C.3, setting $t=N$, we have

$$\left|\phi'(Wz_{i+1})^\top\tilde{\phi}'(Wz)-\mathbb{E}_W\left[\phi'(Wz_{i+1})^\top\tilde{\phi}'(Wz)\right]\right|=\mathcal{O}\left(\sqrt{k}\log N\right), \tag{155}$$

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

$$\begin{aligned}\left|\mathbb{E}_W\left[\phi'(Wz_{i+1})^\top\tilde{\phi}'(Wz)\right]\right|&\leq k\sum_{l=1}^{+\infty}{\mu_l'}^2\left(\frac{\left|z_{i+1}^\top z\right|}{\left\|z_{i+1}\right\|_2\left\|z\right\|_2}\right)^l\\&\leq k\max_l{\mu_l'}^2\sum_{l=1}^{+\infty}\left(\frac{\left|z_{i+1}^\top z\right|}{\left\|z_{i+1}\right\|_2\left\|z\right\|_2}\right)^l\\&=k\max_l{\mu_l'}^2\frac{\left|z_{i+1}^\top z\right|}{\left\|z_{i+1}\right\|_2\left\|z\right\|_2}\frac{1}{1-\frac{\left|z_{i+1}^\top z\right|}{\left\|z_{i+1}\right\|_2\left\|z\right\|_2}}\\&\leq 2k\max_l{\mu_l'}^2\frac{\left|z_{i+1}^\top z\right|}{\left\|z_{i+1}\right\|_2\left\|z\right\|_2}=\mathcal{O}\left(\frac{k\log N}{\sqrt{d}}\right).\end{aligned} \tag{156}$$

Putting together (155) and (156), and applying triangle inequality, we get

$$\left|\phi'(Wz_{i+1})^\top\tilde{\phi}'(Wz)\right|=\mathcal{O}\left(\sqrt{k}\log N+\frac{k\log N}{\sqrt{d}}\right)=\mathcal{O}\left(\sqrt{k}\log N\right), \tag{157}$$

where the last step is a consequence of Assumption 6.1.Comparing this last result with (154), we obtain

$$\left|\varphi(z_{i+1})^\top \tilde{\varphi}(z)\right| = \mathcal{O}\left(\sqrt{dk}\log^2 N\right), \tag{158}$$

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

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

$$\begin{aligned}\left\|\Phi_{-1}\tilde{\varphi}(z)\right\|_2 \le& \sqrt{N-1}\max_i\left|\varphi(z_{i+1})^\top\tilde{\varphi}(z)\right| \\ =& \mathcal{O}\left(\sqrt{dk}\sqrt{N}\log^2 N\right) = o(dk),\end{aligned} \tag{159}$$

where the last step follows from Assumption 6.1. □

Lemma E.8. *We have*

$$\left|\frac{\tilde{\varphi}(z_1^s)^\top\tilde{\varphi}(z_1) - \tilde{\varphi}(z_1^s)^\top P_{\Phi_{-1}}\tilde{\varphi}(z_1)}{\left\|\tilde{\varphi}(z_1) - P_{\Phi_{-1}}\tilde{\varphi}(z_1)\right\|_2^2} - \frac{\tilde{\varphi}(z_1^s)^\top\tilde{\varphi}(z_1)}{\|\tilde{\varphi}(z_1)\|_2^2}\right| = o(1), \tag{160}$$

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*

$$\tilde{\varphi}(z_1^s)^\top\tilde{\varphi}(z_1) - \tilde{\varphi}(z_1^s)^\top P_{\Phi_{-1}}\tilde{\varphi}(z_1) = \Theta(dk), \qquad \left\|\tilde{\varphi}(z_1) - P_{\Phi_{-1}}\tilde{\varphi}(z_1)\right\|_2^2 = \Theta(dk). \tag{161}$$

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

$$\tilde{\varphi}(z_1^s)^\top\tilde{\varphi}(z_1) = \Theta(dk) \qquad \|\tilde{\varphi}(z_1)\|_2^2 = \Theta(dk). \tag{162}$$

by the second statement of Lemma E.6. Furthermore,

$$\begin{aligned}\left|\tilde{\varphi}(z_1^s)^\top P_{\Phi_{-1}}\tilde{\varphi}(z_1)\right| =& \left|\tilde{\varphi}(z_1^s)^\top\Phi_{-1}^\top K_{-1}^{-1}\Phi_{-1}\tilde{\varphi}(z_1)\right| \\ \le& \left\|\Phi_{-1}\tilde{\varphi}(z_1^s)\right\|_2 \lambda_{\min}\left(K_{-1}\right)^{-1}\left\|\Phi_{-1}\tilde{\varphi}(z_1)\right\|_2 \\ =& o(dk)\mathcal{O}\left(\frac{1}{dk}\right)o(dk) = o(dk),\end{aligned} \tag{163}$$

where the third step is justified by Lemmas E.1 and E.7, 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 (163) and (162), and a straightforward application of the triangle inequality, provides the thesis. □

Lemma E.9. *We have*

$$\left|\frac{\varphi(z_1^s)^\top P^\perp_{\Phi_{-1}}\varphi(z_1)}{\left\|P^\perp_{\Phi_{-1}}\varphi(z_1)\right\|_2^2} - \frac{\tilde{\varphi}(z_1^s)^\top P^\perp_{\Phi_{-1}}\tilde{\varphi}(z_1)}{\left\|P^\perp_{\Phi_{-1}}\tilde{\varphi}(z_1)\right\|_2^2}\right| = o(1), \tag{164}$$

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 E.5, 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 E.8. We will condition on the result of Lemma E.8 to hold until the end of the proof. Notice that this also implies

$$\tilde{\varphi}(z_1^s)^\top P^\perp_{\Phi_{-1}}\tilde{\varphi}(z_1) = \Theta(dk), \qquad \left\|P^\perp_{\Phi_{-1}}\tilde{\varphi}(z_1)\right\|_2^2 = \Theta(dk), \tag{165}$$

with probability at least $1-N\exp(-c\log^2 k)-\exp(-c\log^2 N)$ over Z, x, and W. Due to Lemma E.5, we jointly have

$$\left\|P^\perp_{\Phi_{-1}}\left(z_1\otimes\mathbf{1}_k\right)\right\|_2 = o(\sqrt{dk}), \qquad \left\|P^\perp_{\Phi_{-1}}\left(z_1^s\otimes\mathbf{1}_k\right)\right\|_2 = o(\sqrt{dk}), \tag{166}$$

with probability at least $1-\exp(c\log^2 N)$ over Z_{-1} and W. Also, by Lemma C.3 and Assumption 5.2, we jointly have

$$\left\|P^{\perp}_{\Phi_{-1}}\varphi(z_1^s)\right\|_2 \le \|\varphi(z_1^s)\|_2 = \|z_1^s\|_2\,\|\phi'(Wz_1^s)\|_2 = \mathcal{O}\left(\sqrt{dk}\right), \tag{167}$$

and

$$\left\|P^{\perp}_{\Phi_{-1}}\tilde{\varphi}(z_1)\right\|_2 \le \|\tilde{\varphi}(z_1)\|_2 = \|z_1\|_2\left\|\tilde{\phi}'(Wz_1)\right\|_2 = \mathcal{O}\left(\sqrt{dk}\right), \tag{168}$$

with probability at least $1-\exp(-c_1\log^2 N)$ over W. We will condition also on such high probability events ((166), (167), (168)) until the end of the proof. Thus, we can write

$$\begin{aligned}
&\left|\varphi(z_1^s)^\top P^{\perp}_{\Phi_{-1}}\varphi(z_1) - \tilde{\varphi}(z_1^s)^\top P^{\perp}_{\Phi_{-1}}\tilde{\varphi}(z_1)\right| \\
&\le \left|\varphi(z_1^s)^\top P^{\perp}_{\Phi_{-1}}\left(\varphi(z_1)-\tilde{\varphi}(z_1)\right)\right| + \left|\left(\varphi(z_1^s)-\tilde{\varphi}(z_1^s)\right)^\top P^{\perp}_{\Phi_{-1}}\tilde{\varphi}(z_1)\right| \\
&\le \left\|P^{\perp}_{\Phi_{-1}}\varphi(z_1^s)\right\|_2\left\|P^{\perp}_{\Phi_{-1}}\left(z_1\otimes\mu_0\mathbf{1}_k\right)\right\|_2 + \left\|P^{\perp}_{\Phi_{-1}}\tilde{\varphi}(z_1)\right\|_2\left\|P^{\perp}_{\Phi_{-1}}\left(z_1^s\otimes\mu_0\mathbf{1}_k\right)\right\|_2 = o(dk),
\end{aligned} \tag{169}$$

where in the last step we use (166), (167), and (168). Similarly, we can show that

$$\begin{aligned}
\left|\left\|P^{\perp}_{\Phi_{-1}}\varphi(z_1)\right\|_2 - \left\|P^{\perp}_{\Phi_{-1}}\tilde{\varphi}(z_1)\right\|_2\right| &\le \left\|P^{\perp}_{\Phi_{-1}}\varphi(z_1) - P^{\perp}_{\Phi_{-1}}\tilde{\varphi}(z_1)\right\|_2 \\
&\le \left\|P^{\perp}_{\Phi_{-1}}\left(z_1\otimes\mu_0\mathbf{1}_k\right)\right\|_2 = o(\sqrt{dk}).
\end{aligned} \tag{170}$$

By combining (165), (169), and (170), the desired result readily follows. □

Finally, we are ready to give the proof of Theorem 6.3.

Proof of Theorem 6.3. We have

$$\begin{aligned}
\left|\frac{\varphi(z_1^s)^\top P^{\perp}_{\Phi_{-1}}\varphi(z_1)}{\left\|P^{\perp}_{\Phi_{-1}}\varphi(z_1)\right\|_2^2} - \alpha\,\frac{\sum_{l=1}^{+\infty}\mu_l'^2\alpha^i}{\sum_{l=1}^{+\infty}\mu_l'^2}\right| &\le \left|\frac{\varphi(z_1^s)^\top P^{\perp}_{\Phi_{-1}}\varphi(z_1)}{\left\|P^{\perp}_{\Phi_{-1}}\varphi(z_1)\right\|_2^2} - \frac{\tilde{\varphi}(z_1^s)^\top P^{\perp}_{\Phi_{-1}}\tilde{\varphi}(z_1)}{\left\|P^{\perp}_{\Phi_{-1}}\tilde{\varphi}(z_1)\right\|_2^2}\right| \\
&\quad + \left|\frac{\tilde{\varphi}(z_1^s)^\top\tilde{\varphi}(z_1) - \tilde{\varphi}(z_1^s)^\top P_{\Phi_{-1}}\tilde{\varphi}(z_1)}{\left\|\tilde{\varphi}(z_1)-P_{\Phi_{-1}}\tilde{\varphi}(z_1)\right\|_2^2} - \frac{\tilde{\varphi}(z_1^s)^\top\tilde{\varphi}(z_1)}{\|\tilde{\varphi}(z_1)\|_2^2}\right| \\
&\quad + \left|\frac{\tilde{\varphi}(z_1^s)^\top\tilde{\varphi}(z_1)}{\|\tilde{\varphi}(z_1)\|_2^2} - \alpha\,\frac{\sum_{l=1}^{+\infty}\mu_l'^2\alpha^i}{\sum_{l=1}^{+\infty}\mu_l'^2}\right| \\
&= o(1),
\end{aligned} \tag{171}$$

where the first step is justified by the triangle inequality, and the second by Lemmas E.9, E.8, and E.6, and it holds with probability at least $1-N\exp(-c\log^2 k)-\exp(-c\log^2 N)$ over Z, x, and W. □

F. Additional experiments

Figure 6 reports the experiments on NTK features for the same setting considered in Figure 2 for random features. We consider binary classification tasks involving synthetic (first two plots) and standard (last two plots) datasets. As predicted by Theorem 6.3, when the number of samples N increases, the test accuracy increases and, correspondingly, the spurious accuracy decreases. Furthermore, for the synthetic dataset, while the test accuracy does not depend on α and on the activation function, the spurious accuracy increases with α and by taking an activation function with dominant low-order Hermite coefficients.

In Figure 7, we plot the test and the spurious accuracies as a function of $0<\alpha<1$. While the test accuracy does not depend on α, the spurious accuracy monotonically grows with α. This is in agreement with the results of Theorems 5.4 and 6.3.

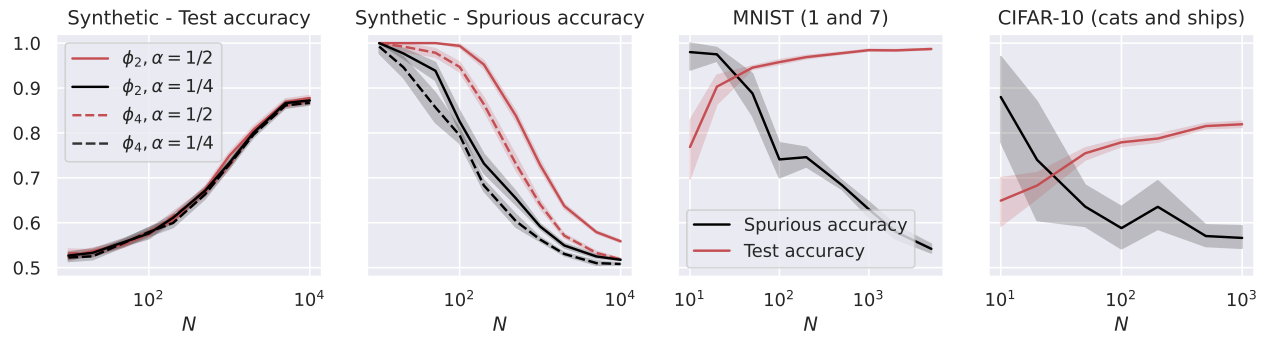

Figure 6. Test and spurious accuracies as a function of the number of training samples N, for various binary classification tasks. In the first two plots, we consider the NTK model in (24) with $k = 100$ trained over Gaussian data with $d = 1000$. The labeling function is $g(x) = \text{sign}(u^\top x)$. We repeat the experiments for $\alpha = \{0.25, 0.5\}$, and for the two activations 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 A.1). 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 1. The spurious accuracy is obtained by querying the model only with the noise background from the training set, replacing all the other pixels with 0, and taking 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.

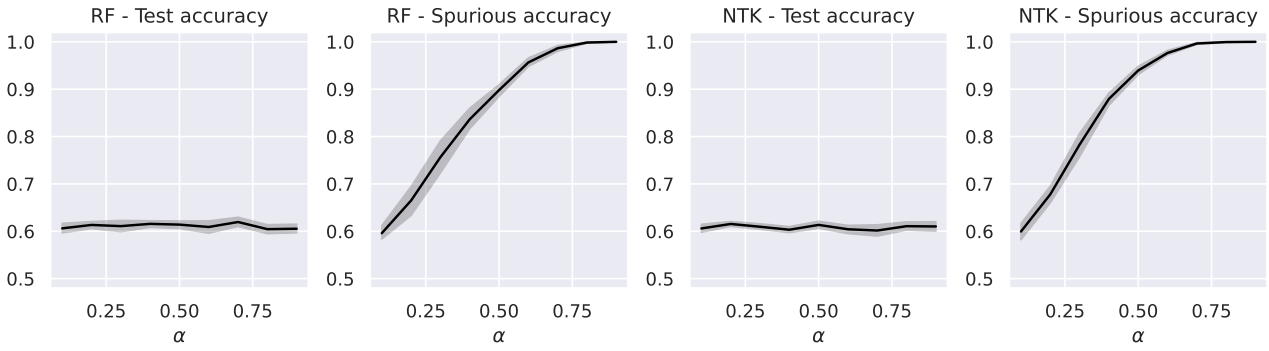

Figure 7. Test and spurious accuracies as a function of α. We consider RF (first and second plot) and NTK (third and fourth plot) models trained on a synthetic dataset. The settings are the same as in Figures 2 and 6, and we use a ReLU activation function. The number of training samples is fixed to $N = 200$.