Beyond the Universal Law of Robustness: Sharper Laws for Random Features and Neural Tangent Kernels

Simone Bombari [1] **Shayan Kiyani** [2] **Marco Mondelli** [1]

Abstract

Machine learning models are vulnerable to adversarial perturbations, and a thought-provoking paper by Bubeck and Sellke has analyzed this phenomenon through the lens of over-parameterization: interpolating smoothly the data requires significantly more parameters than simply memorizing it. However, this "universal" law provides only a necessary condition for robustness, and it is unable to discriminate between models. In this paper, we address these gaps by focusing on empirical risk minimization in two prototypical settings, namely, random features and the neural tangent kernel (NTK). We prove that, for random features, the model is not robust for any degree of over-parameterization, even when the necessary condition coming from the universal law of robustness is satisfied. In contrast, for even activations, the NTK model meets the universal lower bound, and it is robust as soon as the necessary condition on over-parameterization is fulfilled. This also addresses a conjecture in prior work by Bubeck, Li and Nagaraj. Our analysis decouples the effect of the kernel of the model from an "interaction matrix", which describes the interaction with the test data and captures the effect of the activation. Our theoretical results are corroborated by numerical evidence on both synthetic and standard datasets (MNIST, CIFAR-10).

1. Introduction

Despite the deployment of deep neural networks in a wide set of applications, these models are known to be vulnerable to adversarial perturbations (Biggio et al., 2013; Szegedy et al., 2014), which raises serious concerns about their robustness guarantees. To address these issues, a wide variety of adversarial training methods has been developed (Goodfellow et al., 2015; Kurakin et al., 2017; Madry et al., 2018; Raghunathan et al., 2018; Wong & Kolter, 2018). In a parallel effort, a recent line of theoretical work has focused on providing a principled understanding to the phenomenon of adversarial robustness, see e.g. (Dohmatob, 2022; Zhu et al., 2022; Javanmard et al., 2020; Donhauser et al., 2021) and the review in Section 2. Specifically, the recent paper by Bubeck & Sellke (2021) has highlighted the role of over-parameterization as a *necessary* condition to achieve robustness: while interpolation requires the number of parameters p of the model to be at least linear in the number of data samples N, *smooth* interpolation (namely, robustness) requires $p > dN$, where d is the input dimension. We note that the law of robustness put forward by Bubeck & Sellke (2021) is "universal" in the sense that it holds for *any* parametric model. Furthermore, an earlier conjecture by Bubeck et al. (2021) posits that there exists a model of two-layer neural network such that robustness is successfully achieved with this minimal number of parameters; i.e., the condition $p > dN$ is both *necessary* and *sufficient* for robustness. This state of affairs leads to the following natural question:

When does over-parameterization become a sufficient condition to achieve adversarial robustness?

In this paper, we consider a *sensitivity*[1] measure proportional to $\|\nabla_z f(z, \theta)\|_2$, where z is the test data point and f is the model with parameters θ whose robustness is investigated. This quantity is closely related to notions of sensitivity appearing in related works (Dohmatob, 2022; Dohmatob & Bietti, 2022; Bubeck & Sellke, 2021), and it leads to a more stringent requirement on the robustness than the perturbation stability considered by Zhu et al. (2022), see also the discussion in Section 3. For such a sensitivity measure, we show that the answer to the question above depends on the specific model at hand. Specifically, we will focus on the solution yielded by empirical risk minimization (ERM) in two prototypical settings widely analyzed in the theoretical literature: *(i)* Random Features (RF) (Rahimi & Recht, 2007), and the *(ii)* Neural Tangent Kernel (NTK) (Jacot et al., 2018).

[1]Institute of Science and Technology, Austria [2]University of Pennsylvania. Correspondence to: Simone Bombari <simone.bombari@ista.ac.at>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1]The sensitivity can be interpreted as the non-robustness.

Main contribution. Our key results are summarized below.

- For random features, we show that ERM leads to a model which is *not robust* for *any* degree of over-parameterization. Specifically, we tackle a regime in which the universal law of robustness by Bubeck & Sellke (2021) trivializes, and we provide a more refined bound.
- For NTK with an even activation function, we give an upper bound on the ERM solution that *matches* the lower bound by Bubeck & Sellke (2021). As the NTK model approaches the behavior of gradient descent for a suitable initialization (Chizat et al., 2019), this also shows that a class of two-layer neural networks has sensitivity of order $\sqrt{Nd/p}$. This addresses Conjecture 2 of (Bubeck et al., 2021), albeit for a slightly different notion of sensitivity (see the comparison in Section 3).

At the technical level, our analysis involves the study of the spectra of RF and NTK random matrices, and it provides the following insights.

- We introduce in (19) a new quantity, dubbed the *interaction matrix*. Studying this matrix in isolation is a key step in all our results, as its different norms are intimately related to the sensitivity of both the RF and the NTK model. To the best of our knowledge, this is the first time that attention is raised over such an object, which we deem relevant to the theoretical characterization of adversarial robustness.
- Our analytical characterization captures the role of the activation function: it turns out that a specific symmetry (e.g., being even) boosts the adversarial robustness of the models.

Finally, we experimentally show that the robustness behavior of both synthetic and standard datasets (MNIST, CIFAR-10) agrees well with our theoretical results.

2. Related work

From the vast literature studying adversarial examples, we succintly review related works focusing on linear models, random features and NTK models.

Linear regression. In light of the universal law of robustness, in the interpolation regime, linear regression cannot achieve adversarial robustness (as $p = d$). For linear models, Donhauser et al. (2021); Javanmard et al. (2020) have focused on adversarial training (as opposed to ERM, which is considered in this work), showing that over-parameterization hurts the robust generalization error. We also point out that, in some settings, even the standard generalization error is maximized when the model is under-parametrized (Hastie et al., 2022). Precise asymptotics on the robust error in the classification setting are provided in (Javanmard & Soltanolkotabi, 2022; Taheri et al., 2020). A different approach is pursued by Tsipras et al. (2019), who study a linear model that exhibits a trade-off between generalization and robustness.

Random features (RF). The RF model introduced by Rahimi & Recht (2007) can be regarded as a two-layer neural network with random first layer weights, and it solves the lack of freedom in the number of trainable parameters, which is constrained to be equal to the input dimension for linear regression. The popularity of random features derives from their analytical tractability, combined with the fact that the model still reproduces behaviors typical of deep learning, such as the double-descent curve (Mei & Montanari, 2022). Our analysis crucially relies on the spectral properties of the kernel induced by this model. Such properties have been studied by Liao & Couillet (2018), and tight bounds on the smallest eigenvalue of the feature kernel have been provided by Nguyen et al. (2021) for ReLU activations. Theoretical results on the adversarial robustness of random features have been shown for both adversarial training (Hassani & Javanmard, 2022) and the ERM solution (Dohmatob, 2022; Dohmatob & Bietti, 2022). We will discuss the comparison with (Dohmatob, 2022; Dohmatob & Bietti, 2022) in the next paragraph.

Neural tangent kernel (NTK). The NTK can be regarded as the kernel obtained by linearizing the neural network around the initialization (Jacot et al., 2018). 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., 2022). The behavior of the NTK is closely related to memorization (Montanari & Zhong, 2022), optimization (Allen-Zhu et al., 2019; Du et al., 2019) and generalization (Arora et al., 2019) properties of deep neural networks. The NTK has also been recently exploited to identify non-robust features learned with standard training (Tsilivis & Kempe, 2022) and gain insight on adversarial training (Loo et al., 2022). More closely related to our work, the robustness of the NTK model is studied in (Zhu et al., 2022; Dohmatob & Bietti, 2022; Dohmatob, 2022). Specifically, Zhu et al. (2022) analyze the interplay between width, depth and initialization of the network, thus improving upon results by Huang et al. (2021); Wu et al. (2021). However, the perturbation stability considered by Zhu et al. (2022) captures an average robustness, rather than an adversarial one (see the detailed comparison in Section 3), which makes these results not immediately comparable to ours. In contrast, our notion of sensitivity is close to that analyzed in (Dohmatob & Bietti, 2022; Dohmatob, 2022), which establish trade-offs between robustness and memorization for both the RF and

the NTK model. However, Dohmatob & Bietti (2022) study a teacher-student model with quadratic teacher and infinite data, which also does not allow for a direct comparison. Finally, Dohmatob (2022) studies the ERM solution and focuses on *(i)* infinite-width networks, and *(ii)* finite-width networks where the number of neurons scales linearly with the input dimension. We remark that this last setup does not lead to lower bounds on the sensitivity tighter than those by Bubeck & Sellke (2021), and our results on the NTK model provide the first upper bounds on the sensitivity (which also match the lower bounds by Bubeck & Sellke (2021)).

3. Preliminaries

Notation. Given a vector v, we indicate with $\|v\|_2$ 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, let $\|A\|_{\text{op}}$ be its operator norm, $\|A\|_F$ its Frobenius norm and $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ its smallest and largest eigenvalues respectively. We denote by $\|\varphi\|_{\text{Lip}}$ the Lipschitz constant of the function φ. 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.

Generalized linear regression. Let (X, Y) be a labelled training dataset, where $X = [x_1, \dots, x_N]^\top \in \mathbb{R}^{N \times d}$ contains the training samples on its rows and $Y = (y_1, \dots, y_N) \in \mathbb{R}^N$ contains the corresponding labels. Let $\Phi : \mathbb{R}^d \to \mathbb{R}^p$ be a generic *feature map*, from the input space to a feature space of dimension p. We consider the following *generalized linear model*

$$f(x, \theta) = \Phi(x)^\top \theta, \tag{1}$$

where $\Phi(x) \in \mathbb{R}^p$ is the feature vector associated with the input sample x, and $\theta \in \mathbb{R}^p$ is the set of the trainable parameters of the model. Our supervised learning setting involves solving the optimization problem

$$\min_\theta \left\|\Phi(X)^\top \theta - Y\right\|_2^2, \tag{2}$$

where $\Phi(X) \in \mathbb{R}^{N \times p}$ is the feature matrix, containing $\Phi(x_i)$ in its i-th row. From now on, we use the shorthands $\Phi := \Phi(X)$ and $K := \Phi\Phi^\top \in \mathbb{R}^{N \times N}$, where K denotes the kernel associated with the feature map. If we assume K to be invertible (i.e., the model is able to fit any set of labels Y), it is well known that gradient descent converges to the interpolator which is the closest in ℓ_2 norm to the initialization, see e.g. (Gunasekar et al., 2017). In formulas,

$$\theta^* = \theta_0 + \Phi^\top K^{-1}(Y - f(X, \theta_0)), \tag{3}$$

where θ^* is the gradient descent solution, θ_0 is the initialization and $f(X, \theta_0) = \Phi(X)^\top \theta_0$ is the output of the model (1) at initialization.

Sensitivity. We measure the robustness of a model $f(z, \theta)$ as a function of the test sample z. In particular, we are interested in a quantity that expresses how *sensitive* is the output when small perturbations are applied to the input z. More specifically, we could imagine an adversarial example "built around" z as

$$z_{\text{adv}} = z + \Delta_{\text{adv}}, \quad \text{with } \|\Delta_{\text{adv}}\|_2 \le \delta \|z\|_2 . \tag{4}$$

Here, Δ_{adv} is a *small* adversarial perturbation, in the sense that its ℓ_2 norm is only a fraction δ of the ℓ_2 norm of z. Crucially, we expect δ not to depend on the scalings of the problem. For example, if z represents an image with d pixels, this can correspond to perturbing every pixel by a given amount. In this case, the robustness depends on the amount of perturbation per pixel, and not on the number of pixels d. This in turn implies that δ is a numerical constant independent of d.

We are interested in the response of the model to z_{adv}, and how this relates to the natural scaling of the output, which we assume to be $\Theta(1)$.[2] Up to first order, we can write

$$\begin{aligned} |f(z_{\text{adv}}, \theta) - f(z, \theta)| &\simeq \left|\nabla_z f(z, \theta)^\top \Delta_{\text{adv}}\right| \\ &\le \|\Delta_{\text{adv}}\| \, \|\nabla_z f(z, \theta)\|_2 \\ &\le \delta \|z\|_2 \, \|\nabla_z f(z, \theta)\|_2 . \end{aligned} \tag{5}$$

The first inequality is saturated if the attacker builds an adversarial perturbation that perfectly aligns with $\nabla_z f(z, \theta)$, and the second inequality follows from (4). Hence, we define the *sensitivity* of the model evaluated in z as

$$\mathcal{S}_{f_\theta}(z) := \|z\|_2 \, \|\nabla_z f(z, \theta)\|_2 . \tag{6}$$

Recall that both δ and the output of the model are constants, independent of the scaling of the problem. Hence, a robust model needs to respect $\mathcal{S}_{f_\theta}(z) = \mathcal{O}(1)$ for its test samples z. In contrast, if $\mathcal{S}_{f_\theta}(z) \gg 1$ (e.g., the sensitivity grows with the number of pixels of the image), we expect the model to be adversarially vulnerable when evaluated in z. In a nutshell, the goal of this paper is to provide bounds on $\mathcal{S}_{f_\theta}(z)$ for random features and NTK models, thus establishing their robustness.

Related notions of sensitivity. Measures of robustness similar to (6) have been used in the related literature. In particular, Dohmatob & Bietti (2022) study $\mathbb{E}_{z \sim P_X}[\mathcal{S}^2_{f_\theta}(z)]$ in the context of a teacher-student setting with a quadratic target. Dohmatob (2022) considers the model obtained from solving the generalized regression problem (2) and characterizes a sensitivity measure given by the Sobolev seminorm. This quantity is again similar to $\mathbb{E}_{z \sim P_X}[\mathcal{S}^2_{f_\theta}(z)]$, the

[2] This is a natural choice, as e.g. the output label is not expected to grow with the input dimension. However, any scaling is in principle allowed and would not change our final results.

only difference being that the gradient is projected onto the sphere (namely, onto the data manifold) before taking the norm. Zhu et al. (2022) study a notion of perturbation stability given by $\mathbb{E}_{z,\hat{z},\theta} \left|\nabla_z f(z,\theta)^\top (z-\hat{z})\right|$, where z is sampled from the data distribution and $\hat{z}$ is uniform in a ball centered at z with radius δ. This would be similar to averaging over Δ_{adv} in (5), instead of choosing it adversarially. Hence, Zhu et al. (2022) capture a weaker notion of *average robustness*, as opposed to this work which studies the *adversarial robustness*. Finally, Bubeck et al. (2021); Bubeck & Sellke (2021) consider $\|f(z,\theta)\|_{\text{Lip}}$, where the Lipschitz constant is with respect to z, which is equivalent to $\max_z \mathcal{S}_{f_\theta}(z)/\|z\|_2$. While this is a more stringent condition than (6), we remark that the adversarial robustness corresponds to choosing adversarially Δ_{adv} (and not z), as done in (5). We also note that our main results will hold with high probability over the distribution of z.

To conclude, we remark that, differently from previous work (Bubeck & Sellke, 2021; Dohmatob, 2022), we opt for a *scale invariant* definition of sensitivity. This derives from the term $\|z\|_2$ appearing in the RHS of (6), and it allows us to apply the definition to data with arbitrary scaling. In particular, if we set $\|z\|_2 = 1$, then (4) would yield $\|\Delta_{\text{adv}}\|_2 \leq \delta$. Both Bubeck & Sellke (2021) and Dohmatob (2022) consider this scaling of the data: the former paper uses $\mathcal{O}(1)$ Lipschitz constant as an indication of a robust model, and the latter considers a term similar to $\|\nabla_z f(z,\theta)\|_2$.

Robustness of generalized linear regression. For generalized linear models with feature map Φ, plugging (1) in (6) gives

$$\mathcal{S}_{f_\theta}(z) = \|z\|_2 \left\|\nabla_z \Phi(z)^\top \theta\right\|_2. \tag{7}$$

This expression can also provide the sensitivity of the model trained with gradient descent. Let us assume that $f(x,\theta_0) = 0$ for all $x \in \mathbb{R}^d$ and, as before, that the kernel of the feature map is invertible. Then, by plugging the value of θ^* from (3) in the previous equation, we get

$$\mathcal{S}_{f_{\theta^*}}(z) = \|z\|_2 \left\|\nabla_z \Phi(z)^\top \Phi^\top K^{-1} Y\right\|_2. \tag{8}$$

4. Main Result for Random Features

In this section, we provide our law of robustness for the *random features* model. In particular, we will show that, for a wide class of activations, the sensitivity of random features is $\gg 1$ and, therefore, the *model is not robust*.

We assume the data labels $Y \in \mathbb{R}^N$ to be given by $Y = g(X) + \epsilon$. Here, $g : \mathbb{R}^d \to \mathbb{R}$ is the ground-truth function applied row-wise to $X \in \mathbb{R}^{N\times d}$ and $\epsilon \in \mathbb{R}^N$ is a noise vector independent of X, having independent sub-Gaussian entries with zero mean and variance $\mathbb{E}[\epsilon_i^2] = \varepsilon^2$ for all i. The *random features (RF) model* takes the form

$$f_{\text{RF}}(x,\theta) = \Phi_{\text{RF}}(x)^\top \theta, \qquad \Phi_{\text{RF}}(x) = \phi(Vx), \tag{9}$$

where V is a $k \times d$ matrix, such that $V_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$, and ϕ is an activation function applied component-wise. The number of parameters of this model is k, as V is fixed and $\theta \in \mathbb{R}^k$ contains the trainable parameters. We initialize the model at $\theta_0 = 0$ so that $f_{\text{RF}}(x,\theta_0) = 0$ for all x. After training, we can write the sensitivity using (8), which gives

$$\mathcal{S}_{\text{RF}}(z) = \|z\|_2 \left\|\nabla_z \Phi_{\text{RF}}(z)^\top \Phi_{\text{RF}}^\top K_{\text{RF}}^{-1} Y\right\|_2, \tag{10}$$

where we use the shorthands $\Phi_{\text{RF}} := \Phi_{\text{RF}}(X) = \phi(XV^\top) \in \mathbb{R}^{N\times k}$ for the RF map and $K_{\text{RF}} := \Phi_{\text{RF}}\Phi_{\text{RF}}^\top \in \mathbb{R}^{N\times N}$ for the RF kernel. We will show that the kernel is in fact invertible in the argument of our main result, Theorem 4.5. Throughout this section, we make the following assumptions.

Assumption 4.1 (Data distribution)**.** The input data $(x_1, \ldots, x_N)$ are N i.i.d. samples from the distribution P_X, which satisfies the following properties:

1. $\|x\|_2 = \sqrt{d}$, i.e., the data are distributed on the sphere of radius $\sqrt{d}$.
2. $\mathbb{E}[x] = 0$, i.e., the data are centered.
3. P_X satisfies the *Lipschitz concentration property*. Namely, there exists an absolute constant $c > 0$ such that, for every Lipschitz continuous function $\varphi : \mathbb{R}^d \to \mathbb{R}$, we havefor all $t > 0$,

$$\mathbb{P}\left(|\varphi(x) - \mathbb{E}_X[\varphi(x)]| > t\right) \leq 2e^{-ct^2/\|\varphi\|_{\text{Lip}}^2}. \tag{11}$$

The first two assumptions can be achieved by simply preprocessing the raw data. For the third assumption, we remark that the family of Lipschitz concentrated distributions covers a number of important cases, e.g., standard Gaussian (Vershynin, 2018), uniform on the sphere and on the unit (binary or continuous) hypercube (Vershynin, 2018), or data obtained via a Generative Adversarial Network (GAN) (Seddik et al., 2020). The third requirement is common in the related literature (Nguyen et al., 2021; Bombari et al., 2022) and it is equivalent to the isoperimetry condition required by Bubeck & Sellke (2021). We remark that the three conditions of Assumption 4.1 are often replaced by a stronger requirement (e.g., data uniform on the sphere), see (Montanari & Zhong, 2022; Dohmatob, 2022).

Finally, we note that our choice of data scaling is different from the related work (Bubeck & Sellke, 2021; Dohmatob, 2022), as they both consider $\|z\|_2 = 1$. We choose $\|z\|_2 = \sqrt{d}$, as we believe it to be more natural in practice. Consider, for example, the case in which z is an image and d is the number of pixels. If the pixel values are renormalized in a fixed range, then $\|z\|_2$ scales as $\sqrt{d}$. This choice shouldn't worry the reader, as our definition of sensitivity (6) is scale invariant, facilitating the comparison between previous results and ours.

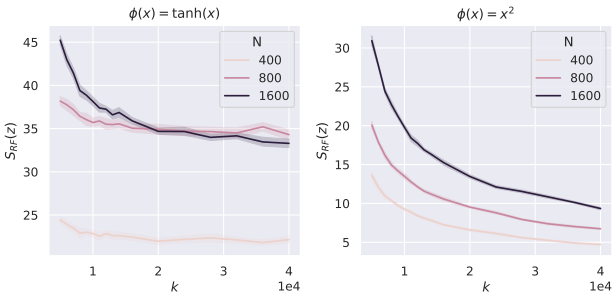

Figure 1. Sensitivity as a function of the number of parameters k, for an RF model with synthetic data sampled from a Gaussian distribution with input dimension $d = 1000$. Different curves correspond to different values of the sample size $N \in \{400, 800, 1600\}$. We plot the average over 10 independent trials and the confidence interval at 1 standard deviation. The values of the sensitivity are significantly larger for $\phi(x) = \tanh(x)$ (left) than for $\phi(x) = x^2$ (right).

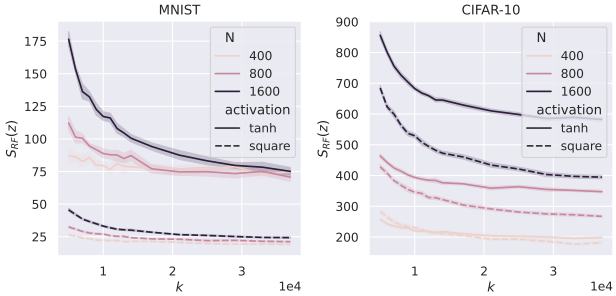

Figure 2. Sensitivity as a function of the number of parameters k, for an RF model with inputs taken from two classes of the MNIST (left) and CIFAR-10 (right) datasets. We pre-process the data normalizing the pixel values from 0 to 1. Different curves correspond to different values of the sample size $N \in \{400, 800, 1600\}$ and to different activations ($\phi(x) = \tanh(x)$ or $\phi(x) = x^2$). We plot an average over 10 independent trials and the confidence interval at 1 standard deviation. Again, $\phi(x) = x^2$ leads to a model which is more robust (or, equivalently, less sensitive) than $\phi(x) = \tanh(x)$.

Assumption 4.2 (Activation function)**.** The activation function ϕ satisfies the following properties:

1. ϕ is a non-linear L-Lipschitz function.
2. Its first and second order derivatives ϕ' and ϕ'' are respectively L_1 and L_2-Lipschitz functions.

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

Assumption 4.3 (Over-parameterization)**.**

$$N \log^3 N = o(k), \tag{12}$$

$$d \log^4 d = o(k). \tag{13}$$

In words, the number of neurons k scales faster than the input dimension d and the number of data points N. Such an over-parameterized setting allows to *(i)* perfectly interpolate the data (this follows directly from the invertibility of the kernel, which can be readily deduced from the proof of Theorem 4.5), and *(ii)* achieve minimum test error, see (Mei & Montanari, 2022). This is also in line with the current trend of heavily over-parametrized deep-learning models (Nakkiran et al., 2020).

Assumption 4.4 (High-dimensional data)**.**

$$N \log^3 N = o\left(d^{3/2}\right), \tag{14}$$

$$\log^8 k = o(d). \tag{15}$$

While the second condition is rather mild, the first one lower bounds the input dimension d in a way that appears to be crucial to prove our main Theorem 4.5. Understanding whether (14) can be relaxed is left as an open question.

At this point, we are ready to state our main result for random features.

Theorem 4.5. *Let Assumptions 4.1, 4.2, 4.3 and 4.4 hold, and let $Y = g(X) + \epsilon$, such that the entries of ϵ are sub-Gaussian with zero mean and variance $\mathbb{E}[\epsilon_i^2] = \varepsilon^2$ for all i, independent between each other and from everything else. Let $z \sim P_X$ be a test sample independent from the training set (X, Y), and let the derivative of the activation function satisfy $\mathbb{E}_{\rho\sim\mathcal{N}(0,1)}[\phi'(\rho)] \neq 0$. Define $\mathcal{S}_{\mathrm{RF}}(z)$ as in (10). Then,*

$$\mathcal{S}_{\mathrm{RF}}(z) = \Omega\left(\sqrt[6]{N}\right) \gg 1, \tag{16}$$

with probability at least $1 - \exp(-c \log^2 N)$ over X, z, V and ϵ.

In words, Theorem 4.5 shows that a vast class of over-parameterized RF models is not robust against adversarial perturbations. A few remarks are now in order.

Beyond the universal law of robustness. We recall that the results by Bubeck & Sellke (2021) give a lower bound on the sensitivity of order $\sqrt{Nd/k}$. A lower bound of the same order is obtained by Dohmatob (2022), whose results in the finite-width setting require N, d, k to follow a proportional scaling (i.e., $N = \Theta(d) = \Theta(k)$). This leaves as an open question the characterization of the robustness for sufficiently over-parameterized random features (i.e., when $k \gg Nd$). Our Theorem 4.5 resolves this question by showing that the RF model is in fact not robust for *any* degree of over-parameterization (as long as $k \gg N, d$ and Assumption 4.4 is satisfied).

Impact of the activation function and numerical results. For the result of Theorem 4.5 to hold, we require the additional assumption $\mathbb{E}_{\rho\sim\mathcal{N}(0,1)}[\phi'(\rho)] \neq 0$. This may not be just a technical requirement. In fact, if $\mathbb{E}_{\rho\sim\mathcal{N}(0,1)}[\phi'(\rho)] = 0$, then a key term appearing in the lower bound for the sensitivity (i.e., the Frobenius norm of the *interaction matrix* defined in (19)) has a drastically different scaling, see Theorem 4.8 in the proof outline below. The importance of the condition $\mathbb{E}_{\rho\sim\mathcal{N}(0,1)}[\phi'(\rho)] \neq 0$ is also confirmed by our numerical simulations in the synthetic setting considered in Figure 1. The left plot (corresponding to $\phi(x) = \tanh(x)$ s.t. $\mathbb{E}_{\rho\sim\mathcal{N}(0,1)}[\phi'(\rho)] \neq 0$) displays values of the sensitivity which are significantly larger than those in the right plot (corresponding to $\phi(x) = x^2$ s.t. $\mathbb{E}_{\rho\sim\mathcal{N}(0,1)}[\phi'(\rho)] = 0$). Furthermore, the sensitivity for $\phi(x) = \tanh(x)$ appears to reach a plateau for large values of k, while it keeps decreasing in k when $\phi(x) = x^2$. A similar impact of the activation function can be observed when considering standard datasets (MNIST and CIFAR-10), see Figure 2.[3] Additional experiments with different activation functions can be found in Appendix E.

4.1. Outline of the Argument of Theorem 4.5

The proof of Theorem 4.5 can be divided into three steps. It will be convenient to define the shorthand

$$A(z) := \nabla_z \Phi_{\text{RF}}(z)^\top \Phi_{\text{RF}}^\top K_{\text{RF}}^{-1} \in \mathbb{R}^{d\times N}. \tag{17}$$

Step 1. Fitting the noise lower bounds the sensitivity. First, we lower bound the sensitivity of our trained model with a quantity which does not depend on the labels and grows proportionally with the noise.

Theorem 4.6. *Let Assumptions 4.1, 4.2, 4.3 and 4.4 hold. Define $A(z)$ as in (17). Then, we have*

$$\mathcal{S}_{\text{RF}}(z) \geq \frac{\varepsilon}{2}\|z\|_2 \left\|A(z)\right\|_F, \tag{18}$$

with probability at least $1-\exp\left(-c\left\|A(z)\right\|_F^2 / \left\|A(z)\right\|_{\text{op}}^2\right)$ over ϵ, where c is a numerical constant.

The proof exploits the independence between $g(X)$ and ϵ and the Hanson-Wright inequality. The details are contained in Appendix C.1.

Theorem 4.6 removes the labels from the expression, and reduces the problem of estimating the robustness to characterizing the Frobenius norm of $A(z)$ (as long as $\|A(z)\|_F \gg \|A(z)\|_{\text{op}}$, so that (18) holds with high probability). This is in line with (Bubeck & Sellke, 2021; Dohmatob, 2022), which provide upper bounds on the robustness of models fitting the data below noise level.

[3]The code used to obtain the results in Figures 1-4 is available at the GitHub repository `simone-bombari/beyond-universal-robustness`.

Step 2. Splitting between interaction and kernel components. Next, we split the matrix $A(z)$ into two separate objects. The first is the *interaction matrix* given by

$$\mathcal{I}_{\text{RF}}(z) := \nabla_z \Phi_{\text{RF}}(z)^\top \tilde{\Phi}_{\text{RF}}^\top, \tag{19}$$

where we use the shorthand $\tilde{\Phi}_{\text{RF}} := \Phi_{\text{RF}} - \mathbb{E}_X[\Phi_{\text{RF}}] \in \mathbb{R}^{N\times k}$. The second is the *centered kernel* $\tilde{K}_{\text{RF}} := \tilde{\Phi}_{\text{RF}}\tilde{\Phi}_{\text{RF}}^\top$, whose spectrum can be studied separately.

Theorem 4.7. *Let Assumptions 4.1, 4.2, 4.3 and 4.4 hold. Define $A(z)$ as in (17). Then, we have*

$$\begin{aligned}\|A(z)\|_F &\geq \lambda_{\max}^{-1}\left(\tilde{K}_{\text{RF}}\right)\|\mathcal{I}_{\text{RF}}(z)\|_F - C\sqrt{N+d}/d,\\ \|A(z)\|_F &\leq \lambda_{\min}^{-1}(\tilde{K}_{\text{RF}})\|\mathcal{I}_{\text{RF}}(z)\|_F + C\sqrt{N+d}/d,\end{aligned} \tag{20}$$

which implies

$$\begin{aligned}\|A(z)\|_F &\geq C_1\frac{d}{kN}\|\mathcal{I}_{\text{RF}}(z)\|_F - C\sqrt{N+d}/d,\\ \|A(z)\|_F &\leq C_2 k^{-1}\|\mathcal{I}_{\text{RF}}(z)\|_F + C\sqrt{N+d}/d,\end{aligned} \tag{21}$$

with probability at least $1-\exp(-c\log^2 N)$ over X and V, where c, C, C_1 and C_2 are absolute constants.

Proof sketch. To prove the claim, first we characterize the extremal eigenvalues of the kernel $K_{\text{RF}} := \Phi_{\text{RF}}\Phi_{\text{RF}}^\top$ and of its centered counterpart $\tilde{K}_{\text{RF}} := \tilde{\Phi}_{\text{RF}}\tilde{\Phi}_{\text{RF}}^\top$, with $\tilde{\Phi}_{\text{RF}} = \Phi_{\text{RF}} - \mathbb{E}_X\Phi_{\text{RF}}$, see Appendix C.2. Then, we perform a delicate centering step on the matrix $A(z)$ defined in (17). Informally, we show that

$$\|\nabla_z\Phi_{\text{RF}}(z)^\top\Phi_{\text{RF}}^\top K_{\text{RF}}^{-1}\|_F \approx \|\nabla_z\Phi_{\text{RF}}(z)^\top\tilde{\Phi}_{\text{RF}}^\top\tilde{K}_{\text{RF}}^{-1}\|_F, \tag{22}$$

see Lemma C.13 for a precise statement. This is the key technical ingredient, since the rank-one components coming from the average feature matrix have a large operator norm, which would trivialize the lower bound on the sensitivity. More specifically, the centering leads to two rank-one terms which are successively removed via the Sherman-Morrison formula and a number of ad-hoc estimates, see Appendix C.3. Finally, the proof of (20)-(21) follows by combining (22) with the bounds on the smallest/largest eigenvalues of $\tilde{K}_{\text{RF}}$, and it appears at the end of Appendix C.3. □

Theorem 4.7 unveils the crucial role played by the interaction matrix $\mathcal{I}_{\text{RF}}(z)$ on the robustness of the model. This term is the only one containing the test sample z, and it depends on how the term $\nabla_z\Phi_{\text{RF}}(z)$ *aligns* (or "interacts") with the centered feature map of the training data $\tilde{\Phi}_{\text{RF}}$.

Step 3. Estimating $\|\mathcal{I}_{\text{RF}}(z)\|_F$. Finally, we provide a precise estimate on the norm of the interaction matrix $\|\mathcal{I}_{\text{RF}}(z)\|_F$. To do so, we assume that the test point z is independently sampled from the data distribution P_X.

Theorem 4.8. *Let Assumptions 4.1, 4.2, 4.3 and 4.4 hold. Let $z \sim P_X$ be sampled independently from the training set (X, Y), and $\mathcal{I}_{\text{RF}}(z)$ be defined as in (19). Then,*

$$\left| \|\mathcal{I}_{\text{RF}}(z)\|_F - \mathbb{E}_\rho^2[\phi'(\rho)] \frac{k\sqrt{N}}{\sqrt{d}} \right| = o\left(\frac{k\sqrt{N}}{\sqrt{d}} \right), \quad (23)$$

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

Proof sketch. A direct calculation gives that

$$\|\mathcal{I}_{\text{RF}}(z)\|_F^2 = \sum_{i=1}^N \left\| V^\top \text{diag}\left(\phi'(Vz)\right) \tilde{\phi}(Vx_i) \right\|_2^2,$$

and we bound separately each term of the sum. Note that the three terms $V^\top$, $\text{diag}\left(\phi'(Vz)\right)$ and $\tilde{\phi}(Vx_i)$ are correlated by the presence of V. However, V contributes to the last two terms only via a single projection (along the direction of z and x_i, respectively). Hence, the key idea is to use a Taylor expansion to split $\text{diag}\left(\phi'(Vz)\right)$ and $\tilde{\phi}(Vx_i)$ into a component correlated with V and an independent one. The correlated components are computed exactly and the remaining independent term is shown to have a negligible effect. The details are in Appendix C.4. □

At this point, the proof of Theorem 4.5 follows by combining the results of Theorems 4.6-4.8. The details are in Appendix C.5.

We note that the results of Theorems 4.6-4.8 do not require the assumption $\mathbb{E}_{\rho\sim\mathcal{N}(0,1)}[\phi'(\rho)] \neq 0$. In fact, the role of this additional condition is clarified by the statement of Theorem 4.8: if $\mathbb{E}_{\rho\sim\mathcal{N}(0,1)}[\phi'(\rho)] \neq 0$, then $\|\mathcal{I}_{\text{RF}}(z)\|_F$ is of order $\Theta(k\sqrt{N/d})$; otherwise, $\|\mathcal{I}_{\text{RF}}(z)\|_F$ is drastically smaller. This provides an explanation to the qualitatively different behavior of the sensitivity for $\phi(x) = \tanh(x)$ and $\phi(x) = x^2$ displayed in Figures 1-2. To conclude, the interaction matrix appears to be the key quantity capturing the impact of the activation function on the robustness of the ERM solution.

5. Main Result for NTK Regression

In this section, we provide our law of robustness for the *NTK* model. In particular, we will show that, for a class of *even* activations, the sensitivity of NTK can be $\mathcal{O}(1)$ and, therefore, the *model is robust* for some scaling of the parameters.

We consider the following two-layer neural network

$$f_{\text{NN}}(x, w) = \sum_{i=1}^k \phi\left(W_{i:}^1 x\right) - \sum_{i=1}^k \phi\left(W_{i:}^2 x\right). \quad (24)$$

Here, the hidden layer contains $2k$ neurons; ϕ is an activation function applied component-wise; $W^1, W^2 \in \mathbb{R}^{k\times d}$ denote the first and second half of the weights of the hidden layer, respectively; for $j \in \{1, 2\}$, $W_{i:}^j$ denotes the i-th row of W^j; the first k weights of the second layer are set to 1, and the last k weights to -1. We indicate with w the vector containing the parameters of this model, i.e., $w = [\text{vec}(W^1), \text{vec}(W^2)] \in \mathbb{R}^p$, with $p = 2kd$. For convenience, we initialize the network so that its output is 0. This has been shown to be necessary to have a robust model after lazy training (Dohmatob & Bietti, 2022; Wang et al., 2022). Specifically, we let $w_0 = [\text{vec}(W_0^1), \text{vec}(W_0^2)]$ be the vector of the parameters at initialization, where we take $[W_0^1]_{i,j} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$ and $W_0^2 = W_0^1$. Here, with a slight abuse of notation, we use the subscript 0 to refer to the initialization, and not to indicate matrix rows. This readily implies that $f_{\text{NN}}(x, w_0) = 0$, for all x.

Now, the *NTK regression model* takes the form

$$f_{\text{NTK}}(x, \theta) = \Phi_{\text{NTK}}(x)^\top \theta, \;\; \Phi_{\text{NTK}}(x) = \nabla_w f_{\text{NN}}(x, w)|_{w=w_0}. \quad (25)$$

Here, the vector trainable parameters is $\theta \in \mathbb{R}^p$, again with $p = 2kd$, which is initialized with $\theta_0 = w_0$. This is the same model considered by Dohmatob & Bietti (2022); Montanari & Zhong (2022). We remark that $f_{\text{NTK}}(x, \theta)$ is equivalent to the linearization of $f_{\text{NN}}(x, w)$ around the initial point w_0, see e.g. (Jacot et al., 2018; Bartlett et al., 2021).

An application of the chain rule gives

$$\begin{aligned} \Phi_{\text{NTK}}(x) &= \left[x \otimes \phi'(W_0^1 x), -x \otimes \phi'(W_0^2 x)\right] \\ &=: \left[\Phi'_{\text{NTK}}(x), -\Phi'_{\text{NTK}}(x)\right], \end{aligned} \quad (26)$$

where the last equality follows from the fact that $W_0^1 = W_0^2$, and the definition $\Phi'_{\text{NTK}}(x) := x \otimes \phi'(W_0^1 x)$. Thus, since $\theta_0 = w_0 = [\text{vec}(W_0^1), \text{vec}(W_0^2)] = [\text{vec}(W_0^1), \text{vec}(W_0^1)]$, we have

$$\begin{aligned} f_{\text{NTK}}(x, \theta_0) &= \Phi_{\text{NTK}}(x)^\top \theta_0 \\ &= \left[\Phi'_{\text{NTK}}(x), -\Phi'_{\text{NTK}}(x)\right]^\top \left[\text{vec}(W_0^1), \text{vec}(W_0^1)\right] = 0. \end{aligned} \quad (27)$$

This means that our model's output is 0 at initialization. Then, we can use (8) to express the sensitivity of the trained NTK regression model as

$$\mathcal{S}_{\text{NTK}}(z) = \|z\|_2 \left\| \nabla_z \Phi_{\text{NTK}}(z)^\top \Phi_{\text{NTK}}^\top K_{\text{NTK}}^{-1} Y \right\|_2, \quad (28)$$

where $\Phi_{\text{NTK}} \in \mathbb{R}^{N\times p}$ contains in its i-th row the feature map of the i-th training sample $\Phi_{\text{NTK}}(x_i)$ and $K_{\text{NTK}} = \Phi_{\text{NTK}}\Phi_{\text{NTK}}^\top$. The invertibility of the kernel will again follow from the proof of our main result, Theorem 5.4. Throughout this section, we make the following assumptions.

Assumption 5.1 (Activation function)**.** The activation function ϕ satisfies the following properties:

1. ϕ is a non-linear, even function.
2. Its first order derivative ϕ' is an L-Lipschitz function.

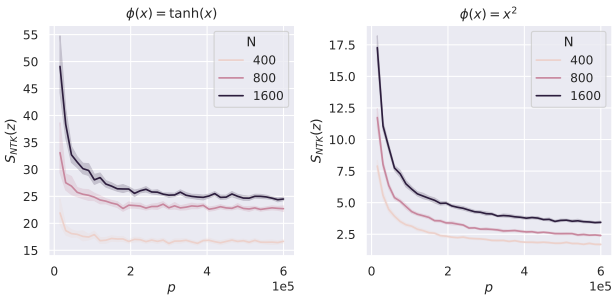

Figure 3. Sensitivity as a function of the number of parameters $p = dk$, for an NTK model with synthetic data sampled from a Gaussian distribution with fixed input dimension $d = 1000$. Different curves correspond to different values of the sample size $N \in \{400, 800, 1600\}$. We plot the average over 10 independent trials and the confidence interval at 1 standard deviation. The values of the sensitivity are significantly larger for $\phi(x) = \tanh(x)$ (left) than for $\phi(x) = x^2$ (right).

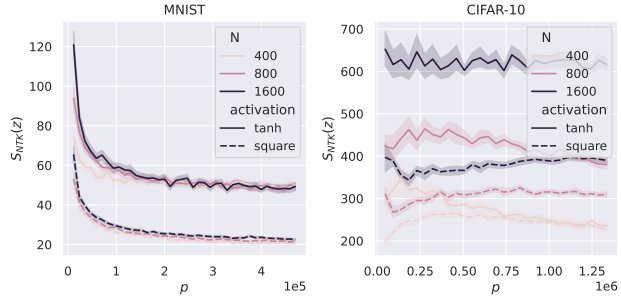

Figure 4. Sensitivity as a function of the number of parameters $p = dk$, for an NTK model with data taken from two classes of the MNIST (left) and CIFAR-10 (right) datasets. We pre-process the data normalizing the pixel values from 0 to 1. Different curves correspond to different values of the sample size $N \in \{400, 800, 1600\}$ and to different activations ($\phi(x) = \tanh(x)$ or $\phi(x) = x^2$). We plot an average over 10 independent trials and the confidence interval at 1 standard deviation. Again, $\phi(x) = x^2$ leads to a model which is more robust (or, equivalently, less sensitive) than $\phi(x) = \tanh(x)$.

We restrict our analysis to an even activation function, as this requirement significantly simplifies the derivations. The impact of the activation on the robustness will also be discussed at the end of this section.

Assumption 5.2 (Minimum over-parameterization)**.**

$$N \log^8 N = o(kd). \tag{29}$$

Eq. (29) provides the weakest possible requirement on the number of parameters of the model capable to guarantee interpolation for generic data points, as it leads to a lower bound on the smallest eigenvalue of K_{NTK} (Bombari et al., 2022).

Assumption 5.3 (High-dimensional data)**.**

$$k = \mathcal{O}(d). \tag{30}$$

This requirement is purely technical, and we leave as a future work the interesting problem of characterizing the sensitivity of the NTK regression model when $d = o(k)$.

We will still use Assumption 4.1 on the data distribution P_X, and the only assumption on the labels is that $\|Y\|_2 = \Theta(\sqrt{N})$, which corresponds to the natural requirement that the output is $\Theta(1)$. At this point, we are ready to state our main result for NTK regression.

Theorem 5.4. *Let Assumptions 4.1, 5.1, 5.2 and 5.3 hold. Then, we have*

$$\mathcal{S}_{\text{NTK}}(z) = \mathcal{O}\left(\log k \left(1 + \sqrt{\frac{N}{k}}\right) \sqrt{\frac{N}{k}}\right), \tag{31}$$

with probability at least $1 - Ne^{-c\log^2 k} - e^{-c\log^2 N}$ *over* X *and* w_0.

Furthermore, when $N = \mathcal{O}(k)$, (31) *simplifies to*

$$\mathcal{S}_{\text{NTK}}(z) = \mathcal{O}\left(\log k \sqrt{\frac{Nd}{p}}\right), \tag{32}$$

where $p = 2dk$ *denotes the number of parameters of the model.*

Proof sketch. As for the RF model, we crucially separate the contribution of the interaction matrix and of the kernel of the model. Specifically, we upper bound the RHS of (28) with

$$\|z\|_2 \left\|\nabla_z \Phi_{\text{NTK}}(z)^\top \Phi_{\text{NTK}}^\top\right\|_{\text{op}} \left\|K_{\text{NTK}}^{-1}\right\|_{\text{op}} \|Y\|_2 . \tag{33}$$

Now, each term in (33) is treated separately. First, by assumption, $\|Y\|_2 = \Theta(\sqrt{N})$. Next, we have that $\left\|K_{\text{NTK}}^{-1}\right\|_{\text{op}} = \lambda_{\min}^{-1}(K_{\text{NTK}}) = \mathcal{O}\left((dk)^{-1}\right)$, which can be deduced from (Bombari et al., 2022), see Lemma D.1. Finally, the bound on $\left\|\nabla_z \Phi_{\text{NTK}}(z)^\top \Phi_{\text{NTK}}^\top\right\|_{\text{op}}$ is computed explicitly through an application of the chain-rule (see Lemma D.3), and it critically depends on the data being 0-mean and on the activation being even. □

In words, Theorem 5.4 shows that, for even activations and k scaling super-linearly in N, $\mathcal{S}_{\text{NTK}}(z) = \mathcal{O}(1)$, namely, the NTK model is robust against adversarial perturbations. A few remarks are now in order.

Saturating the lower bound from the universal law of robustness. We highlight that our upper bound on the sensitivity in (32) exhibits the same scaling as the lower bound by Bubeck & Sellke (2021), thus saturating the universal law

of robustness. A lower bound on the sensitivity of the same order is also provided by Dohmatob (2022), albeit restricted to the setting in which $k = \Theta(d)$ and $k = \mathcal{O}(N)$. Note that Theorem 5.4 upper bounds the sensitivity of the model obtained by running gradient descent (over θ) on $f_{\text{NTK}}(x, \theta)$. It is well known, see e.g. (Chizat et al., 2019), that a similar solution is obtained by running gradient descent (over w) on $f_{\text{NN}}(x, w)$. This result holds after suitably rescaling the initialization of $f_{\text{NN}}(x, w)$, and the rescaling does not affect our Theorem 5.4. As a result, Theorem 5.4 proves that a class of two-layer neural networks has sensitivity of order $\sqrt{Nd/p}$ (up to logarithmic factors), thus resolving in the affirmative Conjecture 2 of (Bubeck et al., 2021).

Impact of the activation function and numerical results. Theorem 5.4 applies to even activation functions. At the technical level, the symmetry in ϕ directly "centers" the kernel in the expression (28) of the sensitivity, which largely simplifies our argument. We conjecture that this symmetry may in fact be fundamental to guarantee the robustness of the model. In fact, the right plot of Figure 3 shows that, for $\phi(x) = x^2$, the sensitivity keeps decreasing, as the number of neurons k (and, therefore, the number of parameters p) grows. In contrast, for $\phi(x) = \tanh(x)$, the sensitivity plateaus at a value which is an order of magnitude larger than for $\phi(x) = x^2$ (see the left plot of Figure 3). This difference is still visible even for real-world datasets (MNIST, CIFAR-10), see Figure 4.

6. Conclusions

Our paper provides a precise and quantitative characterization of how the robustness of the solution obtained via empirical risk minimization depends on the model at hand: for random features and non-even activations, over-parameterization does not help, and we provide bounds tighter than the universal law of robustness proposed by Bubeck & Sellke (2021); for NTK regression and even activations, the model is robust as soon as $p > dN$, i.e., the universal law of robustness is saturated. Numerical results on synthetic and standard datasets confirm the impact of the model on the robustness of the solution. The present contribution focuses on empirical risk minimization, which represents the cornerstone of deep learning training algorithms. In contrast, a number of recent papers has focused on adversarial training, considering the trade-off between generalization and robust error, see e.g. (Raghunathan et al., 2019; Zhang et al., 2019; Dobriban et al., 2020; Carmon et al., 2019; Min et al., 2021; Hassani & Javanmard, 2022) and references therein. We conclude by mentioning that the technical tools developed in this work could be applied also to the models deriving from adversarial training, and we leave such an analysis as future work.

Acknowledgements

Simone Bombari and Marco Mondelli were partially supported by the 2019 Lopez-Loreta prize, and the authors would like to thank Hamed Hassani for helpful discussions.

References

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

Adamczak, R., Litvak, A. E., Pajor, A., and Tomczak-Jaegermann, N. Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. *Constructive Approximation*, 34(1): 61–88, 2011.

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.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning (ICML)*, 2019.

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

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

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases*, pp. 387–402, 2013.

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)*, 2022.

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

Bubeck, S., Li, Y., and Nagaraj, D. M. A law of robustness for two-layers neural networks. In *Conference on Learning Theory (COLT)*, pp. 804–820, 2021.

Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J. C., and Liang, P. S. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

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

Dobriban, E., Hassani, H., Hong, D., and Robey, A. Provable tradeoffs in adversarially robust classification. *arXiv preprint arXiv:2006.05161*, 2020.

Dohmatob, E. Fundamental tradeoffs between memorization and robustness in random features and neural tangent regimes. *arXiv preprint arXiv:2106.02630*, 2022.

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

Donhauser, K., Tifrea, A., Aerni, M., Heckel, R., and Yang, F. Interpolation can hurt robust generalization even when there is no noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Du, S. S., Lee, J. D., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

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.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Hassani, H. and Javanmard, A. The curse of overparametrization in adversarial training: Precise analysis of robust generalization for random features regression. *arXiv preprint arXiv:2201.05149*, 2022.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

Huang, H., Wang, Y., Erfani, S., Gu, Q., Bailey, J., and Ma, X. Exploring architectural ingredients of adversarially robust deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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.

Javanmard, A. and Soltanolkotabi, M. Precise statistical analysis of classification accuracies for adversarial training. *The Annals of Statistics*, 50(4):2127–2156, 2022.

Javanmard, A., Soltanolkotabi, M., and Hassani, H. Precise tradeoffs in adversarial training for linear regression. In *Conference on Learning Theory (COLT)*, 2020.

Kurakin, A., Goodfellow, I. J., and Bengio, S. Adversarial machine learning at scale. In *International Conference on Learning Representations (ICLR)*, 2017.

Liao, Z. and Couillet, R. On the spectrum of random features maps of high dimensional data. In *International Conference on Machine Learning (ICML)*, 2018.

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.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICML)*, 2018.

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.

Min, Y., Chen, L., and Karbasi, A. The curious case of adversarially robust models: More data can help, double descend, or hurt generalization. In *Uncertainty in Artificial Intelligence (UAI)*, 2021.

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.

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., 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.

Raghunathan, A., Steinhardt, J., and Liang, P. Certified defenses against adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018.

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J. C., and Liang, P. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

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

Sambale, H. Some notes on concentration for α-subexponential random variables. *arXiv preprint arXiv:2002.10761*, 2020.

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.

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.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Taheri, H., Pedarsani, R., and Thrampoulidis, C. Asymptotic behavior of adversarial training in binary classification. *arXiv preprint arXiv:2010.13275*, 2020.

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.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. In *International Conference on Learning Representations (ICLR)*, 2019.

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

Wang, Y., Ullah, E., Mianjy, P., and Arora, R. Adversarial robustness is at odds with lazy training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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.

Wong, E. and Kolter, Z. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning (ICML)*, 2018.

Wu, B., Chen, J., Cai, D., He, X., and Gu, Q. Do wider neural networks really help adversarial robustness? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2019.

Zhu, Z., Liu, F., Chrysos, G., and Cevher, V. Robustness in deep learning: The good (width), the bad (depth), and the ugly (initialization). In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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 its 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)). It is useful to also define the generic Orlicz norm of order α of a real random variable X as

$$\|X\|_{\psi_\alpha} := \inf\{t > 0 \ : \ \mathbb{E}[\exp(|X|^\alpha/t^\alpha)] \leq 2\}. \tag{34}$$

From this definition, it follows that $\||X|^\gamma\|_{\psi_{\alpha/\gamma}} = \|X\|_{\psi_\alpha}^\gamma$. Finally, we recall the property that if X and Y are scalar random variables, we have $\|XY\|_{\psi_1} \leq \|X\|_{\psi_2} \|Y\|_{\psi_2}$.

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

$$\mathbb{P}\left(|\varphi(x) - \mathbb{E}_X[\varphi(x)]| > t\right) \leq 2e^{-ct^2/\|\varphi\|_{\text{Lip}}^2}. \tag{35}$$

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 $\varphi : \mathbb{R} \to \mathbb{R}$ is Lipschitz, we have that $\varphi(X)$ 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 two matrices $A, B \in \mathbb{R}^{m\times n}$, we denote by $A \circ B$ their Hadamard product, and by $A * B = [(A_{1:} \otimes B_{1:}), \ldots, (A_{m:} \otimes B_{m:})]^\top \in \mathbb{R}^{m\times n^2}$ their row-wise Kronecker product (also known as Khatri-Rao product). We denote $A^{*2} = A * A$. We 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 $v^\top A v \geq 0$.

B. Useful Lemmas

Lemma B.1. *Let Z be a matrix with rows $z_i \in \mathbb{R}^n$, for $i \in [N]$, such that $N = o(n/\log^4 n)$. Assume that $\{z_i\}_{i\in[N]}$ are independent sub-exponential random vectors with $\psi = \max_i \|z_i\|_{\psi_1}$. Let $\eta_{\min} = \min_i \|z_i\|_2$ and $\eta_{\max} = \max_i \|z_i\|_2$. Set $\xi = \psi K + K'$, and $\Delta := C\xi^2 N^{1/4} n^{3/4}$. Then, we have that*

$$\lambda_{\min}\left(ZZ^\top\right) \geq \eta_{\min}^2 - \Delta, \qquad \lambda_{\max}\left(ZZ^\top\right) \leq \eta_{\max}^2 + \Delta \tag{36}$$

holds with probability at least

$$1 - \exp\left(-cK\sqrt{N}\log\left(\frac{2n}{N}\right)\right) - \mathbb{P}\left(\eta_{max} \geq K'\sqrt{n}\right), \tag{37}$$

where c and C are numerical constants.

Proof. Following the notation in (Adamczak et al., 2011), we define

$$B := \sup_{u\in\mathbb{R}^N : \|u\|_2=1} \left| \left\| \sum_{i=1}^N u_i z_i \right\|_2^2 - \sum_{i=1}^N u_i^2 \|z_i\|_2^2 \right|^{\frac{1}{2}}. \tag{38}$$

Then, for any $u \in \mathbb{R}^N$ with unit norm, we have that

$$\|Zu\|_2^2 = \left\| \sum_{i=1}^N u_i z_i \right\|_2^2 - \sum_{i=1}^N u_i^2 \|z_i\|_2^2 + \sum_{i=1}^N u_i^2 \|z_i\|_2^2 \geq \min_i \|z_i\|_2^2 - B^2, \tag{39}$$

which implies that

$$\lambda_{\min}\left(ZZ^\top\right) = \inf_{u\in\mathbb{R}^N:\|u\|_2=1} \|Zu\|_2^2 \geq \min_i \|z_i\|_2^2 - B^2. \tag{40}$$

In the same way, it can also be proven that

$$\lambda_{\max}\left(ZZ^\top\right) \leq \max_i \|z_i\|_2^2 + B^2. \tag{41}$$

In our case, $z_i \in \mathbb{R}^n$. In the statement of Theorem 3.2 of (Adamczak et al., 2011), let's fix $r = 1$, $m = N$, and $\theta = (N/n)^{1/4} < 1/4$. Then, we have that the condition required to apply Theorem 3.2 is satisfied, *i.e.* ,

$$N \log^2\left(2\sqrt[4]{\frac{n}{N}}\right) \leq \sqrt{Nn}. \tag{42}$$

By combining (40) and (41) with the upper bound on B given by Theorem 3.2 of (Adamczak et al., 2011), the desired result readily follows. □

Lemma B.2. *Let $\{x_i\}_{i=1}^N$ be random variables in $\mathbb{R}$ such that they all respect*

$$\mathbb{P}(|x_i| > t) < 2e^{-ct^\gamma}. \tag{43}$$

Then, we have

$$\max_i |x_i| \leq \log^{2/\gamma} N, \tag{44}$$

with probability at least $1 - 2\exp(-c\log^2 N)$, where c is a numerical constant.

Proof. For any $i \in [N]$,

$$\mathbb{P}(|x_i| > \log^{2/\gamma} N) < 2\exp(-c\log^2 N), \tag{45}$$

which gives

$$\mathbb{P}(\max_i |x_i| > \log^{2/\gamma} N) < N\mathbb{P}(|x_1| > \log^{2/\gamma} N) < 2\exp(\log N - c\log^2 N) < 2\exp(-c_1\log^2 N), \tag{46}$$

where the second step is obtained through a union bound. This gives the desired result. □

Lemma B.3. *Let $z \sim P_Z$ be a mean-0 and Lipschitz concentrated random vector (see (35)). Consider*

$$\Gamma(z) = z \otimes z - \mathbb{E}_z\left[z \otimes z\right]. \tag{47}$$

Then,

$$\|\Gamma(z)\|_{\psi_1} < C. \tag{48}$$

where C is a numerical constant.

Proof. We have that

$$\|\Gamma(z)\|_{\psi_1} = \sup_{\|u\|_2=1} \left\|u^\top\Gamma(z)\right\|_{\psi_1} = \sup_{\|U\|_F=1} \left\|z^\top Uz - \mathbb{E}_x\left[z^\top Uz\right]\right\|_{\psi_1}. \tag{49}$$

Since z is mean-0 and Lipschitz concentrated, we can apply the version of the Hanson-Wright inequality given by Theorem 2.3 in (Adamczak, 2015):

$$\mathbb{P}\left(|z^\top Uz - \mathbb{E}_x\left[z^\top Uz\right]| > t\right) < 2\exp\left(-\frac{1}{C_1}\min\left(\frac{t^2}{\|U\|_F^2}, \frac{t}{\|U\|_{\text{op}}}\right)\right), \tag{50}$$

where C_1 is a numerical constant. Thus, by Lemma 5.5 of (Soltanolkotabi et al., 2018), and using $\|U\|_F \geq \|U\|_{\text{op}}$, we conclude that

$$\|\Gamma(z)\|_{\psi_1} < C\|U\|_F = C, \tag{51}$$

for some numerical constant C, which gives the desired result. □

Lemma B.4. *Let $z \in \mathbb{R}^d$ be a sub-Gaussian vector such that $\|z\|_{\psi_2} = K$. Set $\Sigma = \mathbb{E}\left[zz^\top\right]$. Then,*

$$\|\Sigma\|_{\text{op}} \leq CK^2, \tag{52}$$

where C is a numerical constant.

Proof. Let w be the unitary eigenvector associated to the maximum eigenvalue of Σ. Then,

$$\|\Sigma\|_{\text{op}} = w^\top \Sigma w = \mathbb{E}\left[w^\top z z^\top w\right] = \mathbb{E}\left[(w^\top z)^2\right]. \tag{53}$$

Furthermore, we have that

$$\|z\|_{\psi_2} := \sup_{w' \text{s.t.} \|w'\|_2 = 1} \left\|(w')^\top z\right\|_{\psi_2} \geq \left\|w^\top z\right\|_{\psi_2} \geq \frac{1}{C}\sqrt{\mathbb{E}\left[(w^\top z)^2\right]}, \tag{54}$$

where C is a numerical constant, and the last inequality comes from Eq. (2.15) of (Vershynin, 2018). By combining (53) and (54) we get our desired result. □

Lemma B.5. *Let Z be an $N \times n$ matrix whose rows z_i are i.i.d. mean-0 sub-Gaussian random vectors in $\mathbb{R}^n$. Let $K = \|z_i\|_{\psi_2}$ the sub-Gaussian norm of each row. Then, we have*

$$\left\|ZZ^\top\right\|_{\text{op}} = K^2 \mathcal{O}\left(N + n\right), \tag{55}$$

with probability at least $1 - 2\exp(-cn)$, for some numerical constant c.

Proof. The result is equivalent to Lemma B.7 of (Bombari et al., 2022). □

Lemma B.6. *Let $\{Y_i\}_{i=1}^n$ be independent real random variables such that $\|Y_i\|_{\psi_\alpha} \leq K$ for every i (see (34)), with $\alpha \in (0, 2]$. Then, for every $t > 0$ we have,*

$$\mathbb{P}\left(\sum_{i=1}^n (Y_i - \mathbb{E}[Y_i]) > t\right) < 2\exp\left(-c\left(\frac{t}{K\sqrt{n}\log^{1/\alpha} n}\right)^\alpha\right), \tag{56}$$

where c is an absolute constant.

Proof. Let Y be the vector containing the Y_i's in its entries. Then, an application of Cauchy–Schwarz inequality gives that the function

$$f(Y) := \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i \tag{57}$$

is 1-Lipschitz.

By Lemma 5.6 of (Sambale, 2020), we also have that

$$\|\max_i |Y_i|\|_{\psi_\alpha} \leq CK \log^{1/\alpha} n, \tag{58}$$

where C is a numerical constant (that might depend on α). Thus, as f is also convex, we can apply Proposition 3.1 of (Sambale, 2020), obtaining

$$\mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n (Y_i - \mathbb{E}[Y_i]) > t\right) < 2\exp\left(-c\frac{t^\alpha}{K^\alpha \log n}\right). \tag{59}$$

Rescaling t gives the thesis. □

C. Proofs for Random Features

In this section, we will indicate with $X \in \mathbb{R}^{N\times d}$ the data matrix, such that its rows are sampled independently from P_X (see Assumption 4.1). We will indicate with $V \in \mathbb{R}^{k\times d}$ the random features matrix, such that $V_{ij} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$. The feature vectors will be therefore indicated by $\phi(XV^\top)$, where ϕ is the activation function, applied component-wise to the pre-activations $XV^\top$. We will also use the shorthands $\Phi := \phi(XV^\top) \in \mathbb{R}^{N\times k}$ and $K := \Phi\Phi^\top \in \mathbb{R}^{N\times N}$. We also define $\tilde{\Phi} := \Phi - \mathbb{E}_X[\Phi]$ and $\tilde{K} := \tilde{\Phi}\tilde{\Phi}^\top$. Notice that, for compactness, we dropped the subscripts "RF" from these quantities, as this section will only treat the proofs related to Section 4. Again for the sake of compactness, we will not re-introduce such quantities in the statements or the proofs of the following lemmas.

C.1. Proof of Theorem 4.6

Proof of Theorem 4.6. We have

$$\|AY\|_2^2 = \|A\left(g(X) + \epsilon\right)\|_2^2 = \epsilon^\top A^\top A\epsilon + g(X)^\top A^\top Ag(X) + 2g(X)^\top A^\top A\epsilon. \tag{60}$$

As ϵ is sub-Gaussian, we have $\|A\epsilon\|_{\psi_2} \leq C\|A\|_{\text{op}}$, where C is a numerical constant. Indicating with $M := \|Ag(X)\|_2$, we can write

$$\mathbb{P}(|g(X)^\top A^\top A\epsilon| > t) < 2\exp\left(-c\frac{t^2}{M^2\|A\|_{\text{op}}^2}\right). \tag{61}$$

Setting $t = y\|A\|_{\text{op}} M$ in the previous equation, we readily get

$$\mathbb{P}(|g(X)^\top A^\top A\epsilon| < y\|A\|_{\text{op}} M) > 1 - 2\exp\left(-cy^2\right). \tag{62}$$

We will condition on this event for the rest of the proof.

On the other hand, a direct application of Theorem 6.3.2 in (Vershynin, 2018), gives

$$\|\|A\epsilon\|_2 - \varepsilon\|A\|_F\|_{\psi_2} \leq C_1\|A\|_{\text{op}}, \tag{63}$$

where C_1 is an absolute constant (that depends on ε). This means that

$$\mathbb{P}\left(|\|A\epsilon\|_2 - \varepsilon\|A\|_F| > z\|A\|_{\text{op}}\right) < 2\exp\left(-c_1 z^2\right). \tag{64}$$

This also implies

$$\mathbb{P}\left(\|A\epsilon\|_2 \geq \varepsilon\|A\|_F/\sqrt{2}\right) > 1 - 2\exp\left(-c_2\frac{\|A\|_F^2}{\|A\|_{\text{op}}^2}\right). \tag{65}$$

We will condition also on this other event for the rest of the proof.

Using the triangle inequality, (60) gives

$$\|AY\|_2^2 \geq \epsilon^\top A^\top A\epsilon + M^2 - 2y\|A\|_{\text{op}} M \geq \epsilon^\top A^\top A\epsilon - y^2\|A\|_{\text{op}}^2 \geq \frac{\varepsilon^2}{2}\|A\|_F^2 - y^2\|A\|_{\text{op}}^2, \tag{66}$$

where the second inequality is obtained through minimization over M. Setting $y = \|A\|_F/(2\|A\|_{\text{op}})$ we get

$$\|AY\|_2 \geq \frac{\varepsilon}{2}\|A\|_F, \tag{67}$$

with probability at least $1 - \exp\left(-c_3\|A\|_F^2/\|A\|_{\text{op}}^2\right)$, which concludes the proof. □

C.2. Bounds on the Extremal Eigenvalues of the Kernel

Lemma C.1. *Let $v \in \mathbb{R}^d$ be distributed as $V_{j:}$, for some $j \in [k]$, independent from X. Then, we have*

$$\|\phi(Xv)\|_{\psi_2} = \mathcal{O}\left(\sqrt{N}\right), \tag{68}$$

where the sub-Gaussian norm is intended over the probability space of v. This result holds with probability 1 over X.

Proof. By triangular inequality, we have

$$\|\phi(Xv)\|_{\psi_2} \leq \|\phi(Xv) - \mathbb{E}_v\left[\phi(Xv)\right]\|_{\psi_2} + \|\mathbb{E}_v\left[\phi(Xv)\right]\|_{\psi_2}. \tag{69}$$

Let's bound the two terms separately. As ϕ is an L-Lipschitz function, and $\sqrt{d}v$ is a Lipschitz concentrated random vector, we have

$$\|\phi(Xv) - \mathbb{E}_v\left[\phi(Xv)\right]\|_{\psi_2} \leq CL\frac{\|X\|_{\text{op}}}{\sqrt{d}} \leq CL\frac{\|X\|_F}{\sqrt{d}} = CL\sqrt{N}, \tag{70}$$

where in the last equality we used the fact that $\|x_i\|_2 = \sqrt{d}$.

For the second term of (69), we have

$$\|\mathbb{E}_v\left[\phi(Xv)\right]\|_{\psi_2} \leq C\,\|\mathbb{E}_v\left[\phi(Xv)\right]\|_2. \tag{71}$$

Let $(\phi(Xv))_i$ denote the i-th component of the vector $\phi(Xv)$. As v is a Gaussian vector with covariance I/d and since $\|x_i\| = \sqrt{d}$ for all $i \in [N]$, we have $\mathbb{E}_v\left[(\phi(Xv))_i\right] = \mathbb{E}_\rho[\phi(\rho)] = \mathcal{O}(1)$, where ρ is a standard Gaussian random variable, and the last equality is true since ϕ is Lipschitz. Thus, we have

$$\|\mathbb{E}_v\left[\phi(Xv)\right]\|_{\psi_2} = \mathcal{O}\left(\sqrt{N}\right). \tag{72}$$

Putting together (69), (70), and (72), we get the thesis. □

Lemma C.2. *Let $v \in \mathbb{R}^d$ be distributed as $V_{j:}$, for some $j \in [k]$, independent from X. Then, we have*

$$\|\phi(Xv) - \mathbb{E}_X\left[\phi(Xv)\right]\|_{\psi_2} = \mathcal{O}\left(\sqrt{N}\right), \tag{73}$$

where the sub-Gaussian norm is intended over the probability space of v. This result holds with probability 1 over X.

Proof. By triangle inequality, we have

$$\|\phi(Xv) - \mathbb{E}_X\left[\phi(Xv)\right]\|_{\psi_2} \leq \|\phi(Xv)\|_{\psi_2} + \|\mathbb{E}_X\left[\phi(Xv)\right]\|_{\psi_2}. \tag{74}$$

The first term is $\mathcal{O}\left(\sqrt{N}\right)$, due to Lemma C.1.

For the second term, by Jensen inequality, we have

$$\|\mathbb{E}_X\left[\phi(Xv)\right]\|_{\psi_2} \leq \mathbb{E}_X\left[\|\phi(Xv)\|_{\psi_2}\right] \leq C\mathbb{E}_X\left[\sqrt{N}\right] = \mathcal{O}\left(\sqrt{N}\right). \tag{75}$$

Thus, the thesis readily follows. □

Lemma C.3. *Let $Y := \left(X^{*2} + (\alpha - 1)\mathbb{E}_X\left[X^{*2}\right]\right)$ for some $\alpha \in \mathbb{R}$, where $*$ refers to the Khatri-Rao product, defined in Section 3. Then, we have*

$$\lambda_{\min}\left(YY^\top\right) = \Omega(d^2), \tag{76}$$

with probability at least $1 - \exp(-c\log^2 N)$ over X.

Proof. First, we want to compare $YY^\top$ with $\tilde{Y}\tilde{Y}^\top$, where

$$\tilde{Y} := X^{*2} - \mathbb{E}_X\left[X^{*2}\right]. \tag{77}$$

Let's also define $\bar{Y} := \mathbb{E}_X \left[X^{*2}\right] := \mathbf{1}\mu^\top$, where $\mathbf{1} \in \mathbb{R}^N$ denotes a vector full of ones, and $\mu := \mathbb{E}\left[x_1 \otimes x_1\right]$. We will refer to the i-th row of X with x_i. It is straightforward to verify that $Y = \tilde{Y} + \alpha\bar{Y}$. Thus, we can write

$$\begin{aligned} YY^\top &= \tilde{Y}\tilde{Y}^\top + \alpha^2\bar{Y}\bar{Y}^\top + \alpha\tilde{Y}\bar{Y}^\top + \alpha\bar{Y}\tilde{Y}^\top \\ &= \tilde{Y}\tilde{Y}^\top + \alpha^2 \left\|\mu\right\|_2^2 \mathbf{1}\mathbf{1}^\top + \alpha\Lambda\mathbf{1}\mathbf{1}^\top + \alpha\mathbf{1}\mathbf{1}^\top\Lambda \\ &= \tilde{Y}\tilde{Y}^\top + \left(\alpha\left\|\mu\right\|_2 \mathbf{1} + \frac{\Lambda\mathbf{1}}{\left\|\mu\right\|_2}\right)\left(\alpha\left\|\mu\right\|_2 \mathbf{1} + \frac{\Lambda\mathbf{1}}{\left\|\mu\right\|_2}\right)^\top - \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\left\|\mu\right\|_2^2} \\ &\succeq \tilde{Y}\tilde{Y}^\top - \frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\left\|\mu\right\|_2^2}, \end{aligned} \tag{78}$$

where we define Λ as a diagonal matrix such that $\Lambda_{ii} = \mu^\top\tilde{Y}_{i:}$.

Since $\tilde{Y}_{i:} = x_i \otimes x_i - \mathbb{E}[x_i \otimes x_i] = x_i \otimes x_i - \mu$, where x_i is mean-0 and Lipschitz concentrated, by Lemma B.3, we have that $\left\|\tilde{Y}_{i:}\right\|_{\psi_1} \leq C$, for some numerical constant C. Furthermore, we have that

$$\left|\left\|\tilde{Y}_{i:}\right\|_2 - d\right| = \left|\left\|\tilde{Y}_{i:}\right\|_2 - \left\|x_i \otimes x_i\right\|_2\right| \leq \left\|\tilde{Y}_{i:} - x_i \otimes x_i\right\|_2 = \left\|\mu\right\|_2. \tag{79}$$

where the second step follows from triangle inequality. We can upper-bound the RHS with the following argument:

$$\left\|\mu\right\|_2 = \left\|\mathbb{E}_{x_1}\left[x_1 x_1^\top\right]\right\|_F \leq \sqrt{d}\left\|\mathbb{E}_{x_1}\left[x_1 x_1^\top\right]\right\|_{\text{op}} = \mathcal{O}\left(\sqrt{d}\right), \tag{80}$$

where the last equality is justified by Lemma B.4, since x_i is sub-Gaussian. Then, plugging this last relation in (79), we have that, for all $i \in [N]$,

$$\left\|\tilde{Y}_{i:}\right\|_2 = \Theta(d) < C_y d, \tag{81}$$

for some numerical constant C_y.

Note that $\tilde{Y}$ has N independent rows in $\mathbb{R}^{d^2}$ such that $\left\|\tilde{Y}_{i:}\right\|_{\psi_1} \leq C$ for all i and $\min_i \left\|\tilde{Y}_{i:}\right\|_2 = \Omega(d)$. Assumption 4.4 directly implies $N = o(d^2/\log^4 d^2)$. Thus, we can apply Lemma B.1, setting $K = 1$ and $K' = C_y$ in the statement, and we get

$$\lambda_{\min}\left(\tilde{Y}\tilde{Y}^\top\right) = \Omega(d^2) - cN^{1/4}d^{3/2} = \Omega(d^2), \tag{82}$$

with probability at least $1 - \exp(-c\sqrt{N})$ over X, where c is a numerical constant.

To conclude the proof, we have to control the last term of (78). As $\Lambda_{ii} = \mu^\top\tilde{Y}_{i:}$ and $\left\|\tilde{Y}_{i:}\right\|_{\psi_1} \leq C$, we can write

$$\mathbb{P}(|\Lambda_{ii}|/\left\|\mu\right\|_2 > t) < 2\exp(-c_1 t), \tag{83}$$

where c_1 is a numerical constant, and the probability is intended over X. After applying Lemma B.2 with $\gamma = 1$, the last relation implies that

$$\left\|\Lambda/\left\|\mu\right\|_2\right\|_{\text{op}} = \mathcal{O}\left(\log^2 N\right), \tag{84}$$

with probability at least $1 - 2\exp(-c_2\log^2 N)$ over X, where c_2 is a numerical constant.

Plugging (82) and (84) in (78), we get that, with probability at least $1 - \exp(-c_3\log^2 N)$,

$$\lambda_{\min}\left(YY^\top\right) \geq \lambda_{\min}\left(\tilde{Y}\tilde{Y}^\top\right) - \left\|\frac{\Lambda\mathbf{1}\mathbf{1}^\top\Lambda}{\left\|\mu\right\|_2^2}\right\|_{\text{op}} = \Omega(d^2) - N\mathcal{O}\left(\log^4 N\right) = \Omega(d^2). \tag{85}$$

where the last equality is again consequence of Assumption 4.4. □

Lemma C.4. *We have that*

$$\lambda_{\min}\left(\mathbb{E}_V\left[K\right]\right) = \Omega(k), \tag{86}$$

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

Proof. We have

$$\mathbb{E}_V\left[K\right] = \mathbb{E}_V\left[\sum_{i=1}^{k}\phi(XV_{i:}^\top)\phi(XV_{i:}^\top)^\top\right] = k\mathbb{E}_v\left[\phi(Xv)\phi(Xv)^\top\right] := kM, \tag{87}$$

where we used the shorthand v to indicate a random variable distributed as $V_{1:}$. We will also indicate with x_i the i-th row of X.

Exploiting the Hermite expansion of ϕ, we can write

$$M_{ij} = \mathbb{E}_v\left[\phi(x_i^\top v)\phi(x_j^\top v)\right] = \sum_{n=0}^{+\infty}\mu_n^2\frac{\left(x_i^\top x_j\right)^n}{d^n} = \sum_{n=0}^{+\infty}\mu_n^2\frac{\left[\left(X^{*n}\right)\left(X^{*n}\right)^\top\right]_{ij}}{d^n}, \tag{88}$$

where μ_n is the n-th Hermite coefficient of ϕ. Notice that the previous expansion was possible since $\|x_i\| = \sqrt{d}$ 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 $M \succeq \mu_m^2 M_m$ in a p.s.d. sense, where we define

$$M_m := \frac{1}{d^m}\left(X^{*m}\right)\left(X^{*m}\right). \tag{89}$$

Let's consider the case $m = 2$. Then, since all the rows of X are mean-0, independent and Lipschitz concentrated, such that $\|x_i\| = \sqrt{d}$, we can apply Lemma C.3 with $\alpha = 1$ to readily get

$$\lambda_{\min}\left(M_m\right) = \Omega(1), \tag{90}$$

with probability at least $1 - \exp(-c\log^2 N)$ over X.

Let's now consider the case $m \geq 3$. For $i \neq j$ we can exploit again the fact that x_i is sub-Gaussian and independent from x_j, obtaining

$$\mathbb{P}\left(\left|x_i^\top x_j\right| > t\sqrt{d}\right) < \exp\left(-c_1 t^2\right). \tag{91}$$

Performing a union bound we also get

$$\mathbb{P}\left(\max_{i,j}\left|x_i^\top x_j\right| > t\sqrt{d}\right) < N^2\exp\left(-c_1 t^2\right). \tag{92}$$

Then, by Gershgorin circle theorem, we have that

$$\lambda_{\min}\left(M_m\right) \geq 1 - \max_i\sum_{j\neq i}\frac{\left|\left(x_i^\top x_j\right)^m\right|}{d^m} \geq 1 - \frac{N}{d^m}\max_{i,j}\left|(x_i^\top x_j)^m\right| \geq 1 - \frac{Nt^m}{d^{m/2}}, \tag{93}$$

where the last inequality holds with probability $1 - N^2\exp\left(-c_1 t^2\right)$. Setting $t = \log N$, we get

$$\lambda_{\min}\left(M_m\right) \geq 1 - \frac{N\log^m N}{d^{m/2}} = 1 - \frac{N\log^3 N}{d^{3/2}}\frac{\log^{m-3} N}{d^{m-3/2}} = 1 - o\left(\frac{N\log^3 N}{d^{3/2}}\right) = 1 - o(1), \tag{94}$$

where the last inequality is a consequence of Assumption 4.4. This result holds with probability at least $1 - N^2\exp\left(-c_1\log^2 N\right) \geq 1 - \exp\left(-c_2\log^2 N\right)$, and the thesis readily follows. □

Lemma C.5. *We have that*

$$\lambda_{\min}\left(K\right) = \Omega(k). \tag{95}$$

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

Proof. First we define a truncated version of Φ, which we call $\bar{\Phi}$. Over such truncated matrix, we will be able to use known concentrations results, to prove that $\lambda_{\min}\left(\bar{\Phi}\bar{\Phi}^\top\right) = \Omega(\lambda_{\min}\left(\bar{G}\right))$, where we define $\bar{G} := \mathbb{E}_V\left[\bar{\Phi}\bar{\Phi}^\top\right]$. Finally, we will prove closeness between $\bar{G}$ and G, which is analogously defined as $G := \mathbb{E}_V\left[\Phi\Phi^\top\right]$. Let's introduce the shorthand $v_i := V_{i:}$. To be precise, we define

$$\bar{\Phi}_{:j} = \phi(Xv_j)\chi\left(\|\phi(Xv_j)\|_2^2 \leq R\right), \tag{96}$$

where χ is the indicator function. In this case, $\chi = 1$ if $\|\phi(Xv_j)\|_2^2 \leq R$, and $\chi = 0$ otherwise. As this is a column-wise truncation, it's easy to verify that

$$\Phi\Phi^\top \succeq \bar{\Phi}\bar{\Phi}^\top. \tag{97}$$

Denoting with v a generic weight v_i, we can also write

$$G = k\mathbb{E}_v \left[\phi\left(Xv\right)\phi\left(Xv\right)^\top\right], \tag{98}$$

$$\bar{G} = k\mathbb{E}_v \left[\phi\left(Xv\right)\phi\left(Xv\right)^\top \chi\left(\|\phi\left(Xv\right)\|_2^2 \leq R\right)\right]. \tag{99}$$

Since all the columns of $\bar{\Phi}$ have limited norm with probability one, we can use Matrix Chernoff Inequality (see Theorem 1.1 of (Tropp, 2012)). In particular, for $\delta \in [0, 1]$, we have,

$$\mathbb{P}\left(\lambda_{\min}\left(\bar{\Phi}\bar{\Phi}^\top\right) \leq (1-\delta)\lambda_{\min}\left(\bar{G}\right)\right) \leq N\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\lambda_{\min}\left(\bar{G}\right)/R}, \tag{100}$$

where we used the fact that $\lambda_{\max}\left(\bar{\Phi}_{:j}\bar{\Phi}_{:j}^\top\right) \leq R$ and $\bar{\Phi}\bar{\Phi}^\top = \sum_j \bar{\Phi}_{:j}\bar{\Phi}_{:j}^\top$. Setting $\delta = 1/2$ in the previous equation gives

$$\mathbb{P}\left(\lambda_{\min}\left(\bar{\Phi}\bar{\Phi}^\top\right) \leq \lambda_{\min}\left(\bar{G}\right)/2\right) \leq \exp\left(-c\lambda_{\min}\left(\bar{G}\right)/R + c_1 \log N\right), \tag{101}$$

where this probability is over V.

Finally, we prove that the matrices G and $\bar{G}$ are "close". In particular, we have

$$\begin{aligned}
\left\|G - \bar{G}\right\|_{\text{op}} &= k\left\|\mathbb{E}_v\left[\phi\left(Xv\right)\phi\left(Xv\right)^\top \chi\left(\|\phi\left(Xv\right)\|_2^2 > R\right)\right]\right\|_{\text{op}} \\
&\leq k\mathbb{E}_v\left[\left\|\phi\left(Xv\right)\phi\left(Xv\right)^\top \chi\left(\|\phi\left(Xv\right)\|_2^2 > R\right)\right\|_{\text{op}}\right] \\
&= k\mathbb{E}_v\left[\|\phi\left(Xv\right)\|_2^2 \chi\left(\|\phi\left(Xv\right)\|_2^2 > R\right)\right] \\
&= k\int_{s=0}^{\infty} \mathbb{P}_v\left(\|\phi\left(Xv\right)\|_2 \chi\left(\|\phi\left(Xv\right)\|_2^2 > R\right) > \sqrt{s}\right) \mathrm{d}s \\
&= k\int_{s=0}^{\infty} \mathbb{P}_v\left(\|\phi\left(Xv\right)\|_2 > \max(\sqrt{R}, \sqrt{s})\right) \mathrm{d}s,
\end{aligned} \tag{102}$$

where we applied Jensen inequality in the second line. Let's define $\psi := \|\phi\left(Xv\right)\|_{\psi_2}$, where the sub-Gaussian norm is intended over the probability space of v. We have

$$\begin{aligned}
\left\|G - \bar{G}\right\|_{\text{op}} &\leq k\int_{s=0}^{\infty} \exp\left(-c_2\frac{\max(R, s)}{\psi^2}\right) \mathrm{d}s \\
&\leq k\int_{s=0}^{\infty} \exp\left(-c_2\frac{R+s}{2\psi^2}\right) \mathrm{d}s \\
&= Ck\psi^2 \exp\left(-c_2\frac{R}{2\psi^2}\right),
\end{aligned} \tag{103}$$

where the first inequality follows from the definition of sub-Gaussian random variable, and the second one from $\max(R, s) \geq (R+s)/2$.

By Lemma C.1, we have $\psi = \mathcal{O}\left(\sqrt{N}\right)$. Then, setting $R = k/\log^2 N$, we get

$$\left\|G - \bar{G}\right\|_{\text{op}} \leq Ck\psi^2 \exp\left(-c\frac{R}{2\psi^2}\right) \leq C_1 k \exp\left(-c_2\frac{k}{N\log^2 N} + \log N\right) = o(k), \tag{104}$$

where the last equality is a consequence of Assumption 4.3. Notice that this happens with probability 1 over v.

Due to Lemma C.4, we have that $\lambda_{\min}(G) = \Omega(k)$, with probability at least $1 - \exp\left(-c_3 \log^2 N\right)$ over X. Conditioning on such event, we have

$$\lambda_{\min}\left(\bar{G}\right) \geq \lambda_{\min}(G) - \left\|G - \bar{G}\right\|_{\text{op}} = \Omega(k) - o(k) = \Omega(k). \tag{105}$$

To conclude, looking back at (101), we also get $\lambda_{\min}\left(\bar{\Phi}\bar{\Phi}^\top\right) = \Omega(k)$ with probability (over V) at least

$$1 - \exp\left(-c\lambda_{\min}\left(\bar{G}\right)/R + c_1 \log N\right) \geq 1 - \exp\left(-c_4 \frac{k \log^2 N}{k} + c_1 \log N\right) \geq 1 - \exp\left(-c_5 \log^2 N\right). \tag{106}$$

This concludes the proof. □

Lemma C.6. *We have that*

$$\lambda_{\min}\left(\mathbb{E}_V\left[\tilde{K}\right]\right) = \Omega(k), \tag{107}$$

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

Proof. Indicating with x_i the i-th row of X, and with $\tilde{\phi}(x_i^\top v) := \phi(x_i^\top v) - \mathbb{E}_X\left[\phi(x_1^\top v)\right]$, we can write

$$\mathbb{E}_V\left[\tilde{\Phi}\tilde{\Phi}^\top\right] = \mathbb{E}_V\left[\sum_{i=1}^{k} \tilde{\phi}(XV_{i:}^\top)\tilde{\phi}(XV_{i:}^\top)^\top\right] = k\mathbb{E}_v\left[\tilde{\phi}(Xv)\tilde{\phi}(Xv)^\top\right] := kM, \tag{108}$$

where we use the shorthand v to indicate a random variable distributed as $V_{1:}$.

Exploiting the Hermite expansion of ϕ, we can write

$$\begin{aligned}
M_{ij} &= \mathbb{E}_v\left[\tilde{\phi}(x_i^\top v)\tilde{\phi}(x_j^\top v)\right] \\
&= \mathbb{E}_v\left[\left(\phi(x_i^\top v) - \mathbb{E}_z\left[\phi(z^\top v)\right]\right)\left(\phi(x_j^\top v) - \mathbb{E}_y\left[\phi(y^\top v)\right]\right)\right] \\
&= \mathbb{E}_{vzy}\left[\phi(x_i^\top v)\phi(x_j^\top v) + \phi(z^\top v)\phi(y^\top v) - \phi(x_i^\top v)\phi(y^\top v) - \phi(x_j^\top v)\phi(z^\top v)\right] \\
&= \mathbb{E}_{zy}\left[\sum_{n=0}^{+\infty} \mu_n^2 \frac{\left(x_i^\top x_j\right)^n + \left(z^\top y\right)^n - \left(x_i^\top y\right)^n - \left(x_j^\top z\right)^n}{d^n}\right],
\end{aligned} \tag{109}$$

where μ_n is the n-th Hermite coefficient of ϕ, and $z \sim P_X$ and $y \sim P_X$ are two independent random variables, distributed as the data and independent from them. Notice that the previous expansion was possible since $\|x_i\| = \sqrt{d}$ for all $i \in [N]$. As ϕ is non-linear, there exists $n \geq 2$ such that $\mu_n^2 > 0$. Let m denote one of such n's (i.e, m is such that $\mu_m^2 > 0$), and define ϕ^* as the function that shares the same Hermite expansion as ϕ, but with its m-th coefficient being 0. We have

$$\begin{aligned}
M_{ij} &= \mathbb{E}_{zy}\left[\sum_{n=0}^{+\infty} \mu_n^2 \frac{\left(x_i^\top x_j\right)^n + \left(z^\top y\right)^n - \left(x_i^\top y\right)^n - \left(x_j^\top z\right)^n}{d^n}\right] \\
&= \mathbb{E}_{zy}\left[\sum_{n \neq m} \mu_n^2 \frac{\left(x_i^\top x_j\right)^n + \left(z^\top y\right)^n - \left(x_i^\top y\right)^n - \left(x_j^\top z\right)^n}{d^n}\right] \\
&\quad + \frac{\mu_m^2}{d^m}\left(\left(x_i^\top x_j\right)^m + \mathbb{E}_{zy}\left[\left(z^\top y\right)^m - \left(x_i^\top y\right)^m - \left(x_j^\top z\right)^m\right]\right) \\
&= \mathbb{E}_v\left[\tilde{\phi}^*(x_i^\top v)\tilde{\phi}^*(x_j^\top v)\right] + \frac{\mu_m^2}{d^m}\left(\left(x_i^\top x_j\right)^m + \mathbb{E}_{zy}\left[\left(z^\top y\right)^m - \left(x_i^\top y\right)^m - \left(x_j^\top z\right)^m\right]\right) \\
&=: \mathbb{E}_v\left[\tilde{\phi}^*(Xv)\tilde{\phi}^*(Xv)^\top\right]_{ij} + \mu_m^2\left[M_m\right]_{ij}.
\end{aligned} \tag{110}$$

From the last equation we deduce that $M \succeq \mu_m^2 M_m$, since $\mathbb{E}_v\left[\tilde{\phi}^*(Xv)\tilde{\phi}^*(X^\top v)\right]$ is PSD. Proving the thesis becomes therefore equivalent to prove that $\lambda_{\min}(M_m) = \Omega(1)$.

Let's consider the case $m = 2$. We can write

$$\begin{aligned}[M_m]_{ij} &= \frac{1}{d^2}\mathbb{E}_{zy}\left[\left(x_i^\top x_j - z^\top x_j + x_i^\top y - z^\top y\right)\left(x_i^\top x_j + z^\top x_j - x_i^\top y - z^\top y\right)\right] \\ &= \frac{1}{d^2}\mathbb{E}_{zy}\left[\left((x_i - z)^\top (x_j + y)\right)\left((x_i + z)^\top (x_j - y)^\top\right)\right],\end{aligned} \tag{111}$$

as the cross-terms will be linear in at least one between y and z, and $\mathbb{E}[x] = \mathbb{E}[z] = 0$.

Going back in matrix notation, we can write

$$M_m = \frac{1}{d^2}\mathbb{E}_{ZY}\left[\left((X+Z)(X-Y)^\top \circ (X-Z)(X+Y)^\top\right)\right], \tag{112}$$

where Y and Z are $N \times d$ matrices respectively containing N copies of y and z in their rows.

We can now expand this term as

$$\begin{aligned}M_m &= \frac{1}{d^2}\mathbb{E}_{ZY}\left[\left((X-Z)*(X+Z)\right)\left((X+Y)*(X-Y)\right)^\top\right] \\ &= \frac{1}{d^2}\mathbb{E}_Z\left[(X-Z)*(X+Z)\right]\mathbb{E}_Y\left[(X+Y)*(X-Y)\right]^\top.\end{aligned} \tag{113}$$

Since $\mathbb{E}[Y] = \mathbb{E}[Z] = 0$, the cross-terms cancel out again, giving

$$\mathbb{E}_Z\left[(X-Z)*(X+Z)\right] = X*X - \mathbb{E}_Z\left[Z*Z\right] = X*X - \mathbb{E}_X\left[X*X\right], \tag{114}$$

where the last equality is because all the rows of Z are distributed accordingly to P_X. The same equation holds for the term with Y in (113). Thus, we can conclude that

$$\lambda_{\min}\left(M_m\right) = \frac{1}{d^2}\lambda_{\min}\left(TT^\top\right), \tag{115}$$

where we defined $T = X*X - \mathbb{E}_X\left[X*X\right]$. As all the rows of X are mean-0, independent and Lipschitz concentrated, such that $\|x_i\| = \sqrt{d}$, we can apply Lemma C.3 with $\alpha = 0$ to readily get

$$\lambda_{\min}\left(M_m\right) = \Omega(1), \tag{116}$$

with probability at least $1 - \exp(-c\log^2 N)$ over X.

Let's now consider the case $m \geq 3$. Since x is a sub-Gaussian random variable, we have that $\left\|\frac{x^\top y}{\sqrt{d}}\right\|_{\psi_2} = C$ in the probability space of x. Notice that this holds for all y, as $\|y\|_2 = \sqrt{d}$. Then, its m-th momentum its bounded by $m^{m/2}$, up to a multiplicative constant. As m doesn't depend by the scalings of the problem, we can simply write

$$\mathbb{E}_x\left[\left(\frac{x^\top y}{\sqrt{d}}\right)^m\right] = \mathcal{O}\left(1\right). \tag{117}$$

In the same way, we can prove that

$$\mathbb{E}_x\left[\left(\frac{x^\top x_i}{\sqrt{d}}\right)^m\right] = \mathcal{O}\left(1\right), \qquad \mathbb{E}_y\left[\left(\frac{y^\top x_j}{\sqrt{d}}\right)^m\right] = \mathcal{O}\left(1\right). \tag{118}$$

Also, for $i \neq j$ we can exploit again the fact that x_i is sub-Gaussian and independent from x_j, obtaining

$$\mathbb{P}\left(\left|x_i^\top x_j\right| > t\sqrt{d}\right) < \exp\left(-ct^2\right). \tag{119}$$

Performing a union bound we also get

$$\mathbb{P}\left(\max_{i\neq j}\left|x_i^\top x_j\right| > t\sqrt{d}\right) < N^2\exp\left(-ct^2\right). \tag{120}$$

By Gershgorin circle theorem, and setting $t = \log N$, we have that

$$\begin{aligned}\lambda_{\min}\left(M_m\right) &\geq 1 - \mathcal{O}\left(d^{-m/2}\right) - \max_i \sum_{j\neq i} \frac{\left|\left(x_i^\top x_j\right)^m + \mathbb{E}_{xy}\left[\left(x^\top y\right)^m - \left(x_i^\top y\right)^m - \left(x_j^\top x\right)^m\right]\right|}{d^m} \\ &\geq 1 - \mathcal{O}\left(d^{-m/2}\right) - N \max_{i\neq j} \frac{\left|\left(x_i^\top x_j\right)^m + \mathbb{E}_{xy}\left[\left(x^\top y\right)^m - \left(x_i^\top y\right)^m - \left(x_j^\top x\right)^m\right]\right|}{d^m} \\ &\geq 1 - \mathcal{O}\left(N d^{-m/2}\right) - \frac{N \log^m N}{d^{m/2}} = 1 - o(1),\end{aligned} \tag{121}$$

where the third inequality holds with probability at least $1 - N^2 \exp\left(-c \log^2 N\right) = 1 - \exp\left(-c_1 \log^2 N\right)$, and the last equality is a consequence of Assumption 4.4. This concludes the proof. □

Lemma C.7. *We have that*

$$\lambda_{\min}\left(\tilde{K}\right) = \Omega(k), \tag{122}$$

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

Proof. The proof follows the same exact path as Lemma C.5. The only difference is that now ψ (defined right before (103)) will be $\psi := \|\phi\left(Xv\right) - \mathbb{E}_X\left[\phi\left(Xv\right)\right]\|_{\psi_2}$. To successfully conclude the proof, we have to show that $\psi = \mathcal{O}\left(\sqrt{N}\right)$, which is done in Lemma C.2. □

Lemma C.8. *We have that*

$$\left\|\tilde{K}\right\|_{\text{op}} = \mathcal{O}\left(\frac{k}{d}(N+d)\right), \tag{123}$$

with probability at least $1 - k\exp(-cd)$ *over* V *and* X.

Proof. We have that

$$\tilde{K} = \tilde{\Phi}\tilde{\Phi}^\top = \sum_{i=1}^{\lceil k/d \rceil} \tilde{\Phi}^i (\tilde{\Phi}^i)^\top, \tag{124}$$

where we define $\tilde{\Phi}^i$ as the $N \times d$ matrix containing the columns of $\tilde{\Phi}$ between $(i-1)d+1$ and id. If $i = \lceil k/d \rceil$, we have that the number of columns is $k - d(\lceil k/d \rceil - 1) \leq d$.

Every $\tilde{\Phi}^i$ has mean-0, i.i.d. rows $\tilde{\phi}(V^i x_j)$ (in the probability space of X), where V^i contains the rows of V between $(i-1)d+1$ and id. Notice that, besides corner cases, V^i is a $d \times d$ matrix.Since the x_j are Lipschitz concentrated, we also have

$$\left\|\tilde{\phi}(V^i x_j)\right\|_{\psi_2} \leq C \left\|V^i\right\|_{\text{op}} = \mathcal{O}\left(1\right), \tag{125}$$

where the last equality follows from Theorem 4.4.5 of (Vershynin, 2018) and from the fact that both dimensions of V^i are less or equal than d, and holds with probability $1 - \exp(-cd)$ over V^i. Conditioning on such event, a straightforward application of Lemma B.5 gives

$$\left\|\tilde{\Phi}^i (\tilde{\Phi}^i)^\top\right\|_{\text{op}} = \mathcal{O}\left(N+d\right), \tag{126}$$

with probability at least $1 - \exp(-c_1 d)$ over X.

Thus, performing a union bound over the i-s, we get

$$\max_i \left\|\Phi^i (\tilde{\Phi}^i)^\top\right\|_{\text{op}} = \mathcal{O}\left(N+d\right), \tag{127}$$

with probability at least $1 - k\exp(-c_2 d)$. Conditioning over such event, we have

$$\left\|\tilde{K}\right\|_{\text{op}} \leq \sum_{i=1}^{\lceil k/d \rceil} \left\|\tilde{\Phi}^i (\tilde{\Phi}^i)^\top\right\|_{\text{op}} = \mathcal{O}\left(\frac{k}{d}(N+d)\right), \tag{128}$$

which gives the thesis. □

C.3. Centering and Estimate of $\|A(z)\|_F$

In the following section it will be convenient to define $E(z) := \nabla_z \phi(Vz)^\top \phi(XV^\top)^\top$. Notice that this gives $A(z) = E(z)K^{-1}$. Furthermore, we will use the notation $\tilde{\cdot}$ to indicate quantities involving a centering with respect to the random variable X. In particular, we indicate

$$\tilde{\Phi} = \Phi - \mathbb{E}_X[\Phi], \qquad \tilde{K} = \tilde{\Phi}\tilde{\Phi}^\top, \qquad \tilde{E} := \nabla_z \phi(Vz)^\top \tilde{\Phi}^\top, \qquad \tilde{A} = \tilde{E}\tilde{K}^{-1}. \tag{129}$$

For the sake of compactness, we will not re-introduce any of these quantities in the statements or the proofs of the following lemmas, and we will drop the dependence on z of A, $\tilde{A}$, E and $\tilde{E}$.

Note the results of this section will be applied only when $\nu := \mathbb{E}_{x_i}[\phi(Vx_i)]$ is not the all-0 vector. Therefore, we will assume $\|\nu\|_2 > 0$ through all the proofs here. Along the section it will be convenient to recall that $\nabla_z \phi(Vz)^\top = V^\top \text{diag}(\phi'(Vz))$.

Lemma C.9. *Let $\mathbf{1} \in \mathbb{R}^N$ denote a vector full of ones. Then, the kernel matrix K can be written as*

$$K = \tilde{K} + \eta\eta^\top - \lambda\lambda^\top, \tag{130}$$

where

$$\eta := \|\nu\|_2 \mathbf{1} + \tilde{\Phi}\frac{\nu}{\|\nu\|_2} = \Phi\frac{\nu}{\|\nu\|_2}, \tag{131}$$

$$\lambda := \tilde{\Phi}\frac{\nu}{\|\nu\|_2}. \tag{132}$$

Proof. Note that $\Phi = \tilde{\Phi} + \mathbf{1}\nu^\top$. Then,

$$\begin{aligned} K &= \left(\tilde{\Phi} + \mathbf{1}\nu^\top\right)\left(\tilde{\Phi} + \mathbf{1}\nu^\top\right)^\top \\ &= \tilde{\Phi}\tilde{\Phi}^\top + \tilde{\Phi}\nu\mathbf{1}^\top + \mathbf{1}\nu^\top\tilde{\Phi}^\top + \|\nu\|_2^2 \mathbf{1}\mathbf{1}^\top \\ &= \tilde{\Phi}\tilde{\Phi}^\top + \left(\|\nu\|_2 \mathbf{1} + \frac{\tilde{\Phi}\nu}{\|\nu\|_2}\right)\left(\|\nu\|_2 \mathbf{1} + \frac{\tilde{\Phi}\nu}{\|\nu\|_2}\right)^\top - \frac{\tilde{\Phi}\nu\nu^\top\tilde{\Phi}^\top}{\|\nu\|_2^2} \\ &= \tilde{\Phi}\tilde{\Phi}^\top + \eta\eta^\top - \lambda\lambda^\top. \end{aligned} \tag{133}$$

□

Lemma C.10. *We have that*

$$\left| \left\|EK^{-1}\right\|_F - \left\|\tilde{E}K^{-1}\right\|_F \right| = \mathcal{O}\left(\frac{\sqrt{N+d}}{d}\right), \tag{134}$$

with probability at least $1 - \exp(-c\log^2 N) - k\exp(-cd)$ over X and V, where c is a numerical constant.

Proof. We have

$$E = \tilde{E} + \nabla_z \phi(Vz)^\top \nu \mathbf{1}^\top. \tag{135}$$

By triangle inequality, we can write

$$\left| \left\|EK^{-1}\right\|_F - \left\|\tilde{E}K^{-1}\right\|_F \right| \le \left\|\nabla_z \phi(Vz)^\top \nu \mathbf{1}^\top K^{-1}\right\|_F = \left\|\nabla_z \phi(Vz)^\top \nu\right\|_2 \left\|K^{-1}\mathbf{1}\right\|_2. \tag{136}$$

Let's look at the two factors separately. The first one can be upper bounded as follows

$$\left\|\nabla_z \phi(Vz)^\top \nu\right\|_2 = \left\|V^\top \text{diag}(\phi'(Vz))\,\nu\right\|_2 \le \|V\|_{\text{op}} L \|\nu\|_2 = \mathcal{O}\left(\sqrt{\frac{k}{d}}\|\nu\|_2\right), \tag{137}$$

where the first inequality holds as ϕ is L-Lipschitz, and the second equality holds with probability at least $1 - e^{-cd}$ due to (242). Notice that from (131) and (132), we can write

$$\mathbf{1} = \frac{\eta}{\|\nu\|_2} - \frac{\lambda}{\|\nu\|_2}. \tag{138}$$

By plugging this in the second factor of (136) and applying the triangle inequality, we have

$$\left\|K^{-1}\mathbf{1}\right\|_2 = \frac{1}{\|\nu\|_2}\left\|K^{-1}(\eta-\lambda)\right\|_2 \leq \frac{1}{\|\nu\|_2}\left(\left\|K^{-1}\eta\right\|_2 + \left\|K^{-1}\lambda\right\|_2\right). \tag{139}$$

Again, let's control these last two terms separately. Since we can write $\eta = \Phi\nu/\|\nu\|_2$, we get

$$\left\|K^{-1}\eta\right\|_2 = \left\|K^{-1}\Phi\nu/\|\nu\|_2\right\|_2 \leq \left\|K^{-1}\Phi\right\|_{\text{op}} = \left\|\Phi^+\right\|_{\text{op}} = \lambda_{\min}(K)^{-1/2} = \mathcal{O}\left(\sqrt{\frac{1}{k}}\right), \tag{140}$$

where the last equality holds with probability at $1-\exp(-c\log^2 N)$, due to Lemma C.5. We will condition on such event for the rest of the proof. For the second term, we have

$$\left\|K^{-1}\lambda\right\|_2 = \left\|K^{-1}\tilde{\Phi}\nu/\|\nu\|_2\right\|_2 \leq \left\|K^{-1}\right\|_{\text{op}}\left\|\tilde{\Phi}\right\|_{\text{op}} = \lambda_{\min}(K)^{-1}\,\mathcal{O}\left(\sqrt{\frac{k(N+d)}{d}}\right) = \mathcal{O}\left(\sqrt{\frac{N+d}{kd}}\right), \tag{141}$$

where the first equality holds due to Lemma C.8 with probability at least $1-k\exp(-c_1 d)$. Let's condition on this event too. Therefore, putting together (139), (140) and (141), we have

$$\left\|K^{-1}\mathbf{1}\right\|_2 = \frac{1}{\|\nu\|_2}\left(\mathcal{O}\left(\sqrt{\frac{1}{k}}\right) + \mathcal{O}\left(\sqrt{\frac{N+d}{kd}}\right)\right) = \frac{1}{\|\nu\|_2}\mathcal{O}\left(\sqrt{\frac{N+d}{kd}}\right). \tag{142}$$

By combining this last expression with (136)-(137), we conclude

$$\left|\left\|EK^{-1}\right\|_F - \left\|\tilde{E}K^{-1}\right\|_F\right| = \mathcal{O}\left(\sqrt{\frac{k}{d}}\|\nu\|_2\right)\frac{1}{\|\nu\|_2}\mathcal{O}\left(\sqrt{\frac{N+d}{kd}}\right) = \mathcal{O}\left(\frac{\sqrt{N+d}}{d}\right), \tag{143}$$

which gives the thesis. □

Lemma C.11. *We have that*

$$\left|\left\|\tilde{E}K^{-1}\right\|_F - \left\|\tilde{E}\left(K+\lambda\lambda^\top\right)^{-1}\right\|_F\right| = \mathcal{O}\left(\frac{\sqrt{N+d}}{d}\right), \tag{144}$$

with probability at least $1-\exp(-c\log^2 N) - k\exp(-cd)$ *over* X *and* V*, where* c *is a numerical constant.*

Proof. A direct application of the Sherman-Morrison formula gives

$$\left(K+\lambda\lambda^\top\right)^{-1} = K^{-1} - \frac{K^{-1}\lambda\lambda^\top K^{-1}}{1+\lambda^\top K^{-1}\lambda}. \tag{145}$$

By triangle inequality we know that $|\|a\|_2 - \|b\|_2| \leq \|a-b\|_2$. This implies

$$\left|\left\|\tilde{E}K^{-1}\right\|_F - \left\|\tilde{E}\left(K+\lambda\lambda^\top\right)^{-1}\right\|_F\right| \leq \left\|\tilde{E}K^{-1} - \tilde{E}\left(K+\lambda\lambda^\top\right)^{-1}\right\|_F = \left\|\tilde{E}\frac{K^{-1}\lambda\lambda^\top K^{-1}}{1+\lambda^\top K^{-1}\lambda}\right\|_F, \tag{146}$$

where the last equality is a consequence of (145). This term can be bounded as

$$\left\|\tilde{E}\frac{K^{-1}\lambda\lambda^\top K^{-1}}{1+\lambda^\top K^{-1}\lambda}\right\|_F \leq \left\|\tilde{E}\right\|_{\text{op}}\left\|\frac{K^{-1}\lambda\lambda^\top K^{-1}}{1+\lambda^\top K^{-1}\lambda}\right\|_F = \left\|\tilde{E}\right\|_{\text{op}}\frac{\lambda^\top K^{-2}\lambda}{1+\lambda^\top K^{-1}\lambda}. \tag{147}$$

The first factor can be bounded as follows

$$\left\|\tilde{E}\right\|_{\text{op}} = \left\|\nabla_z\phi(Vz)^\top\tilde{\Phi}^\top\right\|_{\text{op}} \leq \left\|V^\top\text{diag}\left(\phi'(Vz)\right)\right\|_{\text{op}}\left\|\tilde{\Phi}\right\|_{\text{op}} = \mathcal{O}\left(\sqrt{\frac{k}{d}}\right)\mathcal{O}\left(\sqrt{\frac{k(N+d)}{d}}\right) = \mathcal{O}\left(k\frac{\sqrt{N+d}}{d}\right), \tag{148}$$

with probability at least $1 - k\exp(-cd)$. Here, we used the fact that ϕ is Lipschitz, and we applied Lemma C.8 and Equation (242). Furthermore, since $\|M\|_{\text{op}} M - M^2$ is PSD for every symmetric PSD matrix M, the following relation holds

$$\lambda^\top K^{-2}\lambda \leq \left\|K^{-1}\right\|_{\text{op}} \lambda^\top K^{-1}\lambda. \tag{149}$$

Thus,

$$\frac{\lambda^\top K^{-2}\lambda}{1+\lambda^\top K^{-1}\lambda} \leq \left\|K^{-1}\right\|_{\text{op}} \frac{\lambda^\top K^{-1}\lambda}{1+\lambda^\top K^{-1}\lambda} \leq \left\|K^{-1}\right\|_{\text{op}} = \frac{1}{\lambda_{\min}(K)} = \mathcal{O}\left(\frac{1}{k}\right), \tag{150}$$

where the last equality holds with probability at least $1 - \exp(-c\log^2 N)$, because of Lemma C.5. Plugging the last equation in (147) together with (148), and comparing with (146), gives the desired result. □

Lemma C.12. *We have that*

$$\left|\left\|\tilde{E}\left(\tilde{K}+\eta\eta^\top\right)^{-1}\right\|_F - \left\|\tilde{E}\tilde{K}^{-1}\right\|_F\right| = \mathcal{O}\left(\frac{\sqrt{N+d}}{d}\right), \tag{151}$$

with probability at least $1 - \exp(-c\log^2 N) - k\exp(-cd)$ *over* X *and* V*, where* c *is a numerical constant.*

Proof. An application of the Sherman-Morrison formula gives

$$\left(\tilde{K}+\eta\eta^\top\right)^{-1} = \tilde{K}^{-1} - \frac{\tilde{K}^{-1}\eta\eta^\top\tilde{K}^{-1}}{1+\eta^\top\tilde{K}^{-1}\eta}. \tag{152}$$

This implies

$$\left|\left\|\tilde{E}\left(\tilde{K}+\eta\eta^\top\right)^{-1}\right\|_F - \left\|\tilde{E}\tilde{K}^{-1}\right\|_F\right| \leq \left\|\tilde{E}\left(\tilde{K}+\eta\eta^\top\right)^{-1} - \tilde{E}\tilde{K}^{-1}\right\|_F = \left\|\tilde{E}\frac{\tilde{K}^{-1}\eta\eta^\top\tilde{K}^{-1}}{1+\eta^\top\tilde{K}^{-1}\eta}\right\|_F, \tag{153}$$

where the last equality is a consequence of (152). This term can be bounded as

$$\left\|\tilde{E}\frac{\tilde{K}^{-1}\eta\eta^\top\tilde{K}^{-1}}{1+\eta^\top\tilde{K}^{-1}\eta}\right\|_F \leq \left\|\tilde{E}\right\|_{\text{op}} \left\|\frac{\tilde{K}^{-1}\eta\eta^\top\tilde{K}^{-1}}{1+\eta^\top\tilde{K}^{-1}\eta}\right\|_F = \left\|\tilde{E}\right\|_{\text{op}} \frac{\eta^\top\tilde{K}^{-2}\eta}{1+\eta^\top\tilde{K}^{-1}\eta}. \tag{154}$$

The first factor can be bounded as in (148), in Lemma C.11, giving $\left\|\tilde{E}\right\|_{\text{op}} = \mathcal{O}\left(k\frac{\sqrt{N+d}}{d}\right)$ with probability at least $1 - k\exp(-cd)$. Similarly to (149), we can write

$$\eta^\top\tilde{K}^{-2}\eta \leq \left\|\tilde{K}^{-1}\right\|_{\text{op}} \eta^\top\tilde{K}^{-1}\eta. \tag{155}$$

Thus,

$$\frac{\eta^\top\tilde{K}^{-2}\eta}{1+\eta^\top\tilde{K}^{-1}\eta} \leq \left\|\tilde{K}^{-1}\right\|_{\text{op}} \frac{\eta^\top\tilde{K}^{-1}\eta}{1+\eta^\top\tilde{K}^{-1}\eta} \leq \left\|\tilde{K}^{-1}\right\|_{\text{op}} = \frac{1}{\lambda_{\min}\left(\tilde{K}\right)} = \mathcal{O}\left(\frac{1}{k}\right), \tag{156}$$

where the last equality holds with probability at least $1 - \exp(-c\log^2 N)$, because of Lemma C.7. Plugging the last equation in (154) together with the bound on $\left\|\tilde{E}\right\|_{\text{op}}$, and comparing with (153), gives the desired result. □

Lemma C.13. *We have that*

$$\left|\|A\|_F - \left\|\tilde{A}\right\|_F\right| = \mathcal{O}\left(\frac{\sqrt{N+d}}{d}\right), \tag{157}$$

with probability at least $1 - \exp(-c\log^2 N) - k\exp(-cd)$ *over* X *and* V*, where* c *is a numerical constant.*

Proof. Recalling that $A = EK^{-1}$ and $\tilde{A} = \tilde{E}\tilde{K}^{-1}$, by triangle inequality we have

$$\begin{aligned}\left|\|A\|_F - \left\|\tilde{A}\right\|_F\right| &\leq \left|\left\|EK^{-1}\right\|_F - \left\|\tilde{E}K^{-1}\right\|_F\right| \\ &\quad + \left|\left\|\tilde{E}K^{-1}\right\|_F - \left\|\tilde{E}\left(K+\lambda\lambda^\top\right)^{-1}\right\|_F\right| \\ &\quad + \left|\left\|\tilde{E}\left(\tilde{K}+\eta\eta^\top\right)^{-1}\right\|_F - \left\|\tilde{E}\tilde{K}^{-1}\right\|_F\right| \\ &= \mathcal{O}\left(\frac{\sqrt{N+d}}{d}\right)\end{aligned} \tag{158}$$

where the inequality holds because of Lemma C.10, Lemma C.11, and Lemma C.12, each of them with probability at least $1 - \exp(-c\log^2 N) - k\exp(-cd)$ over X and V. This gives the thesis. □

At this point, we are ready to prove Theorem 4.7.

Proof of Theorem 4.7. We have

$$\begin{aligned}\|A(z)\|_F &\leq \left\|\nabla_z\Phi(z)^\top\tilde{\Phi}^\top\tilde{K}^{-1}\right\|_F + \mathcal{O}\left(\sqrt{N+d}/d\right) \\ &\leq \left\|\nabla_z\Phi(z)^\top\tilde{\Phi}^\top\right\|_F\left\|\tilde{K}^{-1}\right\|_{\text{op}} + \mathcal{O}\left(\sqrt{N+d}/d\right) \\ &= \lambda_{\min}^{-1}(K)\mathcal{I}(z) + \mathcal{O}\left(\sqrt{N+d}/d\right),\end{aligned} \tag{159}$$

where the first inequality is justified by Lemma C.13 and holds with probability at least $1 - \exp(-c\log^2 N) - k\exp(-cd)$ over X and V. For the other bound of the first equation, we have

$$\begin{aligned}\|A(z)\|_F &\geq \left\|\nabla_z\Phi(z)^\top\tilde{\Phi}^\top\tilde{K}^{-1}\right\|_F - \mathcal{O}\left(\sqrt{N+d}/d\right) \\ &\geq \left\|\nabla_z\Phi(z)^\top\tilde{\Phi}^\top\right\|_F\lambda_{\min}\left(\tilde{K}^{-1}\right) - \mathcal{O}\left(\sqrt{N+d}/d\right) \\ &= \lambda_{\max}^{-1}(\tilde{K})\mathcal{I}(z) - \mathcal{O}\left(\sqrt{N+d}/d\right),\end{aligned} \tag{160}$$

where the first inequality is justified again by Lemma C.13, and holds with probability at least $1 - \exp(-c\log^2 N) - k\exp(-cd)$ over X and V.

The second equation directly follows from Lemma C.8 and Lemma C.6. □

C.4. Estimate of $\|\mathcal{I}_{\mathbf{RF}}(z)\|_F$

The goal of this sub-Section is to estimate the value of the Frobenius norm of the Interaction Matrix $\|\mathcal{I}_{\text{RF}}\|_F$, defined in (19). To do so, we will estimate the squared ℓ_2 norms of its columns. In particular, we define

$$\mathcal{I}_x := \left\|\nabla_z\phi(Vz)^\top\tilde{\phi}(Vx)\right\|_2. \tag{161}$$

Thus, we have $\|\mathcal{I}_{\text{RF}}\|_F^2 = \sum_i \mathcal{I}_{x_i}^2$. In this section, with $x \sim P_X$ we refer to a generic input data, where we drop the index for compactness. Also, we will assume $z \sim P_X$, independent from the training set.

Recalling that

$$\nabla_z\phi(Vz)^\top = V^\top\text{diag}\left(\phi'(Vz)\right), \tag{162}$$

we have

$$\mathcal{I}_x = \left\|V^\top\text{diag}\left(\phi'(Vz)\right)\tilde{\phi}(Vx)\right\|_2. \tag{163}$$

Along the proof, the following shorthands and notations will be useful

$$\mathcal{X}_{-jl} = \sum_{m\neq j} V_{lm}x_m, \qquad \mathcal{X}_l = \sum_{m=1}^{d} V_{lm}x_m, \tag{164}$$

$$\mathcal{Z}_{-jl} = \sum_{m \neq j} V_{lm} z_m, \qquad \mathcal{Z}_l = \sum_{m=1}^{d} V_{lm} z_m, \tag{165}$$

$$I_j = \sum_{l=1}^{k} V_{lj} \phi'\left(\mathcal{Z}_l\right) \tilde{\phi}\left(\mathcal{X}_l\right), \tag{166}$$

which imply

$$\mathcal{I}_x^2 = \sum_{j=1}^{d} I_j^2. \tag{167}$$

It will also be convenient to use the following notation

$$B_{-jl} := \left(\phi''\left(\mathcal{Z}_{-jl}\right) \tilde{\phi}\left(\mathcal{X}_{-jl}\right) z_j + \phi'\left(\mathcal{Z}_{-jl}\right) \phi'\left(\mathcal{X}_{-jl}\right) x_j\right), \tag{168}$$

$$B_{jl} := \left(\phi''\left(\mathcal{Z}_l\right) \tilde{\phi}\left(\mathcal{X}_l\right) z_j + \phi'\left(\mathcal{Z}_l\right) \phi'\left(\mathcal{X}_l\right) x_j\right). \tag{169}$$

For compactness, we will not necessarily re-introduce such quantities in the statements or the proofs of the lemmas in this section.

Lemma C.14. *Indicating with c a numerical constant, the following statements hold*

$$\max_l \left|\tilde{\phi}(\mathcal{X}_l)\right| = \mathcal{O}\left(\log k\right), \tag{170}$$

with probability at least $1 - \exp(-c \log^2 k)$, over x and V;

$$\max_l \left|\tilde{\phi}(\mathcal{X}_{-jl})\right| = \mathcal{O}\left(\log k\right), \tag{171}$$

with probability at least $1 - \exp(-c \log^2 k)$, over x and $\{V_{li}\}_{i \neq j, 1 \leq l \leq k}$;

$$\left|\sum_{l=1}^{k} V_{lj} \phi'\left(\mathcal{Z}_{-jl}\right) \tilde{\phi}\left(\mathcal{X}_{-jl}\right)\right| = o\left(\frac{k}{d}\right), \tag{172}$$

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

Proof. For the first result, as ϕ is Lipschitz and centered with respect to x, its sub-Gaussian norm (in the probability space of x) is $\mathcal{O}\left(\|V_{l:}\|_2\right) = \mathcal{O}\left(1\right)$, where the last equality holds with probability at least $1 - e^{-c_0 d}$ over $V_{l:}$, because of Theorem 3.1.1 in (Vershynin, 2018). In particular, with a union bound over $V_{l:}$, we have $\mathcal{O}\left(\max_l \|V_{l:}\|_2\right) = \mathcal{O}\left(1\right)$ with probability at least $1 - ke^{-c_0 d} \geq 1 - e^{-c_1 d}$, where the last inequality is justified by Assumption 4.4. Thus, by Lemma B.2, $\max_l \left|\tilde{\phi}(\mathcal{X}_l)\right| = \mathcal{O}\left(\log k\right)$, with probability at least $1 - \exp(-c_1 d) - \exp(-c_2 \log^2 k) \geq 1 - \exp(-c_3 \log^2 k)$, over x and V.

Notice that, with the same argument, we can prove that $\max_l \left|\tilde{\phi}(\mathcal{X}_{-jl})\right| = \mathcal{O}\left(\log k\right)$, with probability at least $1 - \exp(-c_4 \log^2 k)$, over x and $\{V_{li}\}_{i \neq j, 1 \leq l \leq k}$. We will condition on such event for the rest of the proof.

For the last statement, we have

$$\sum_{l=1}^{k} V_{lj} \phi'\left(\mathcal{Z}_{-jl}\right) \tilde{\phi}\left(\mathcal{X}_{-jl}\right) =: V_{:j}^{\top} u, \tag{173}$$

where u is a vector independent from $V_{:j}$ such that

$$\|u\|_2 \leq \sqrt{k} L \max_l \left|\tilde{\phi}\left(\mathcal{X}_{-jl}\right)\right| = \mathcal{O}\left(\log(k)\sqrt{k}\right). \tag{174}$$

Since $\|V_{:j}\|_{\psi_2} \leq C/\sqrt{d}$, we have

$$\mathbb{P}\left(\left|\sum_{l=1}^{k} V_{lj}\phi'\left(\mathcal{Z}_{-jl}\right)\tilde{\phi}\left(\mathcal{X}_{-jl}\right)\right| > \frac{\sqrt{d}}{\sqrt{k}/\log^2 k}\frac{k}{d}\right) < \exp(-c_5 \log^2 k), \tag{175}$$

where the probability is intended over $V_{:j}$. Thus, using Assumption 4.3, we readily get the desired result. □

Lemma C.15. *We have that*

$$\|\mathbb{E}_x\left[\phi'\left(\mathcal{X}_{-jl}\right)x_j\right]\|_{\psi_2} = \mathcal{O}\left(\frac{1}{\sqrt{d}}\right), \tag{176}$$

where the norm is intended over the probability space of $\{V_{li}\}_{i\neq j}$.

Proof. Let's define an auxiliary random variable $\rho \sim \mathcal{N}(0, 1/d)$, independent from any other random variable in the problem. Let's also define $V'_{l:}$ as the copy of $V_{l:}$ where we replaced the j-th component with ρ. Notice that $V'_{l:}$ is a Gaussian vector with covariance I/d. Let's also define $\mathcal{X}'_l$ in the same way, i.e., replacing V_{lj} with ρ in its expression. All the tail norms $\|\cdot\|_{\psi_r}$ in the proof will be referred to the probability space of $V'_{l:}$.

Since ϕ' is a Lipschitz function, we can write

$$|\mathbb{E}_x\left[\phi'\left(\mathcal{X}_{-jl}\right)x_j\right] - \mathbb{E}_x\left[\phi'\left(\mathcal{X}'_l\right)x_j\right]| \leq L_1\mathbb{E}_x\left[|\rho|x_j^2\right] \leq C|\rho|, \tag{177}$$

where we used the fact that x_j is sub-Gaussian in the last inequality, and where C is a numerical constant. Let's now look at the function

$$\varphi(v) := \mathbb{E}_x\left[\phi'\left(v^\top x\right)x_j\right]. \tag{178}$$

We have

$$\begin{aligned}|\varphi(v+v') - \varphi(v)| &= \left|\mathbb{E}_x\left[\left(\phi'\left((v+v')^\top x\right) - \phi'\left(v^\top x\right)\right)x_j\right]\right| \\ &\leq \mathbb{E}_x\left[\left|\left(\phi'\left((v+v')^\top x\right) - \phi'\left(v^\top x\right)\right)\right||x_j|\right] \\ &\leq \mathbb{E}_x\left[L_1|v'^\top x||x_j|\right].\end{aligned} \tag{179}$$

Since x is sub-Gaussian (and therefore also x_j is) we have that $|v'^\top x||x_j|$ is sub-exponential with norm (with respect to the probability space of x) upper bounded by $C_1\|v'\|_2$. Then, we have

$$|\varphi(v+v') - \varphi(v)| \leq \mathbb{E}_x\left[L_1|v'^\top x||x_j|\right] \leq C_2\|v'\|_2. \tag{180}$$

This implies that $\|\varphi\|_{\text{Lip}} \leq C_2$.

As $V'_{l:}$ is Gaussian (and hence Lipschitz concentrated) with covariance I/d, we can write

$$\left\|\varphi(V'_{l:}) - \mathbb{E}_{V'_{l:}}\left[\varphi(V'_{l:})\right]\right\|_{\psi_2} = \mathcal{O}\left(\frac{1}{\sqrt{d}}\right). \tag{181}$$

Furthermore,

$$\mathbb{E}_{V'_{l:}}\left[\varphi(V'_{l:})\right] = \mathbb{E}_x\left[x_j\mathbb{E}_{V'_{l:}}\left[\phi'\left(V'^\top_{l:}x\right)\right]\right] = \mathbb{E}_x\left[x_j\right]\mathbb{E}_{\rho'}\left[\phi'\left(\rho'\right)\right] = 0, \tag{182}$$

where ρ' is a standard Gaussian distribution independent of x. Note that the second step holds since $\|x\|_2 = \sqrt{d}$, and the last step is justified by $\mathbb{E}[x] = 0$.

This last equation implies

$$\begin{aligned}\|\mathbb{E}_x\left[\phi'\left(\mathcal{X}_{-jl}\right)x_j\right]\|_{\psi_2} &\leq \|\mathbb{E}_x\left[\phi'\left(\mathcal{X}_{-jl}\right)x_j\right] - \mathbb{E}_x\left[\phi'\left(\mathcal{X}'_l\right)x_j\right]\|_{\psi_2} + \|\mathbb{E}_x\left[\phi'\left(\mathcal{X}'_l\right)x_j\right]\|_{\psi_2} \\ &\leq C\|\rho\|_{\psi_2} + \|\varphi(V'_{l:})\|_{\psi_2} = \mathcal{O}\left(\frac{1}{\sqrt{d}}\right),\end{aligned} \tag{183}$$

where the first term in the last line is justified by (177). In fact, the random variable on the LHS is (with probability one) upper bounded by $C|\rho|$. Then, the bound subsists also between their tail norms.

Thus,

$$\mathbb{P}_{\{V_{li}\}_{i\neq j}}\left(\left|\mathbb{E}_x\left[\phi'\left(\mathcal{X}_{-jl}\right)x_j\right]\right|>t\right)=\mathbb{P}_{V'_{l:}}\left(\left|\mathbb{E}_x\left[\phi'\left(\mathcal{X}_{-jl}\right)x_j\right]\right|>t\right)<2\exp\left(-cd\,t^2\right), \tag{184}$$

where with $\mathbb{P}_{\{V_{li}\}_{i\neq j}}$ we refer to the probability space of $\{V_{li}\}_{i\neq j}$. Note that the first equality is true as the statement in the probability does not depend on the value of ρ. From this, we readily get the thesis. □

Lemma C.16. *We have that*

$$\left|I_j-\sum_l V_{lj}^2\left(\phi''\left(\mathcal{Z}_{-jl}\right)\tilde{\phi}\left(\mathcal{X}_{-jl}\right)z_j+\phi'\left(\mathcal{Z}_{-jl}\right)\phi'\left(\mathcal{X}_{-jl}\right)x_j\right)\right|=o\left(\frac{k}{d}\right). \tag{185}$$

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

Proof. Exploiting Taylor's Theorem, and the fact that both ϕ' and ϕ'' are Lipschitz, we can write

$$\phi'\left(\mathcal{Z}_l\right)=\phi'\left(\mathcal{Z}_{-jl}\right)+\phi''\left(\mathcal{Z}_{-jl}\right)V_{lj}z_j+a_{1lj}, \tag{186}$$

$$\tilde{\phi}\left(\mathcal{X}_l\right)=\phi\left(\mathcal{X}_{-jl}\right)+\phi'\left(\mathcal{X}_{-jl}\right)V_{lj}x_j+a_{2lj}-\mathbb{E}_x\left[\phi\left(\mathcal{X}_{-jl}\right)+\phi'\left(\mathcal{X}_{-jl}\right)V_{lj}x_j+a_{2lj}\right], \tag{187}$$

where $|a_{1lj}|=\mathcal{O}\left(V_{lj}^2z_j^2\right)$ and $|a_{2lj}|=\mathcal{O}\left(V_{lj}^2x_j^2\right)$. In the previous two equations, we expanded ϕ' around $\mathcal{Z}_{-jl}$ and ϕ around $\mathcal{X}_{-jl}$. For the rest of the proof (except from when expectations with respect to x are taken) we condition on the events that both $|x_j|$ and $|z_j|$ are $\mathcal{O}\left(\log k\right)$. As they are sub-Gaussian, this happens with probability at least $1-e^{-c_1\log^2 k}$ over x and z.

Invoking Lemma B.6, and setting $t=k/d^{1/\alpha}$, we can write (for $\alpha\in(0,2]$)

$$\sum_{l=1}^k|V_{lj}|^{2/\alpha}=\mathcal{O}\left(\frac{k}{d^{1/\alpha}}\right), \tag{188}$$

with probability at least $1-\exp(-c_2k^{\alpha/2}/\log k)$. This is true since $\mathbb{E}_\rho\left[\rho^{2/\alpha}\right]=C_\alpha$, where ρ is a standard Gaussian random variable. Thus, with probability at least $1-\exp(-c_3k^{1/5}/\log k)$ over V, we jointly have

$$\sum_{l=1}^k|V_{lj}|^2=\mathcal{O}\left(\frac{k}{d}\right),\qquad\sum_{l=1}^k|V_{lj}|^3=\mathcal{O}\left(\frac{k}{d^{3/2}}\right),\qquad\sum_{l=1}^k|V_{lj}|^5=\mathcal{O}\left(\frac{k}{d^{5/2}}\right). \tag{189}$$

We will condition on these events too. It will also be convenient to condition on the high probability events described by Lemma C.14. Thus, our discussion is now restricted to a probability space over x, z and V of measure at least $1-e^{-c_4\log^2 k}-e^{-c_5d}$.

We are now ready to estimate the cross terms of the product $V_{lj}\phi'\left(\mathcal{Z}_l\right)\tilde{\phi}\left(\mathcal{X}_l\right)$, exploiting the expansions in (186) and (187). Let's control the sums over l of all of them separately.

(a) $\phi'\left(\mathcal{Z}_l\right)\left(a_{2lj}-\mathbb{E}_x\left[a_{2lj}\right]\right)$

$$\begin{aligned}\left|\sum_{l=1}^k V_{lj}\phi'\left(\mathcal{Z}_l\right)\left(a_{2lj}-\mathbb{E}_x\left[a_{2lj}\right]\right)\right|&\leq C_1\sum_{l=1}^k|V_{lj}|L\left(|a_{2lj}|+\mathbb{E}_x\left[|a_{2lj}|\right]\right)\\&\leq C_2\left(x_j^2+1\right)\sum_{l=1}^k|V_{lj}|^3=\mathcal{O}\left(\frac{k\log^2 k}{d^{3/2}}\right)=o\left(\frac{k}{d}\right),\end{aligned} \tag{190}$$

where the second inequality is justified by the fact that x_j is sub-Gaussian, and the last equality by Assumption 4.4.

(b) $\tilde{\phi}\left(\mathcal{X}_l\right)a_{1jl}$

$$\begin{aligned}\left|\sum_{l=1}^k V_{lj}\tilde{\phi}\left(\mathcal{X}_l\right)a_{1jl}\right|&\leq C_3z_j^2\sum_{l=1}^k|V_{lj}|^3\left|\tilde{\phi}\left(\mathcal{X}_l\right)\right|\leq C_3z_j^2\max_l\left|\tilde{\phi}\left(\mathcal{X}_l\right)\right|\sum_{l=1}^k|V_{lj}|^3\\&=\mathcal{O}\left(\log^2(k)\log(k)\frac{k}{d^{3/2}}\right)=o\left(\frac{k}{d}\right),\end{aligned} \tag{191}$$

where the second to last equality follows from Lemma C.14, and the last from Assumption 4.4.

(c) $a_{1lj}\left(a_{2lj}-\mathbb{E}_x\left[a_{2lj}\right]\right)$ (as we counted it twice)

$$\left|\sum_{l=1}^{k} V_{lj} a_{1lj}\left(a_{2lj}-\mathbb{E}_x\left[a_{2lj}\right]\right)\right| \leq C_4\left(x_j^2+1\right) z_j^2 \sum_{l=1}^{k}\left|V_{lj}\right|^5=\mathcal{O}\left(\log^4(k)\frac{k}{d^{5/2}}\right)=o\left(\frac{k}{d}\right). \tag{192}$$

(d) $\phi'\left(\mathcal{Z}_{-jl}\right)\tilde{\phi}\left(\mathcal{X}_{-jl}\right)$

$$\left|\sum_{l=1}^{k} V_{lj}\phi'\left(\mathcal{Z}_{-jl}\right)\tilde{\phi}\left(\mathcal{X}_{-jl}\right)\right|=o\left(\frac{k}{d}\right), \tag{193}$$

as it directly follows from Lemma C.14.

(e) $\left(\phi'\left(\mathcal{Z}_{-jl}\right)+\phi''\left(\mathcal{Z}_{-jl}\right)V_{lj}z_j\right)\mathbb{E}_x\left[\phi'\left(\mathcal{X}_{-jl}\right)V_{lj}x_j\right]$

$$\begin{aligned}
&\left|\sum_{l=1}^{k} V_{lj}\left(\phi'\left(\mathcal{Z}_{-jl}\right)+\phi''\left(\mathcal{Z}_{-jl}\right)V_{lj}z_j\right)V_{lj}\mathbb{E}_x\left[\phi'\left(\mathcal{X}_{-jl}\right)x_j\right]\right| \\
&\quad\leq \max_l\left|\phi'\left(\mathcal{Z}_{-jl}\right)+\phi''\left(\mathcal{Z}_{-jl}\right)V_{lj}z_j\right|\max_l\left|\mathbb{E}_x\left[\phi'\left(\mathcal{X}_{-jl}\right)x_j\right]\right|\sum_{l=1}^{k}\left|V_{lj}\right|^2 \\
&\quad\leq \left(L+L_1\max_l\left|V_{lj}\right|\left|z_j\right|\right)\max_l\left|\mathbb{E}_x\left[\phi'\left(\mathcal{X}_{-jl}\right)x_j\right]\right|\sum_{l=1}^{k}\left|V_{lj}\right|^2 \\
&\quad=\mathcal{O}\left(\left(1+\frac{\log^2 k}{\sqrt{d}}\right)\frac{\log k}{\sqrt{d}}\right)\mathcal{O}\left(\frac{k}{d}\right)=o\left(\frac{k}{d}\right),
\end{aligned} \tag{194}$$

where the second to last equality comes from the fact that both $\left\|V_{lj}\right\|_{\psi_2}$ and $\left\|\mathbb{E}_x\left[\phi'\left(\mathcal{X}_{-jl}\right)x_j\right]\right\|_{\psi_2}$ are $\mathcal{O}\left(\frac{1}{\sqrt{d}}\right)$ (see Lemma C.15) and holds with probability at least $1-e^{-c_6\log^2 k}$ over V. The last equality is true for Assumption 4.4.

(f) $\phi''\left(\mathcal{Z}_{-jl}\right)V_{lj}z_j\phi'\left(\mathcal{X}_{-jl}\right)V_{lj}x_j$ As ϕ and ϕ' are Lipschitz

$$\left|\sum_{l=1}^{k} V_{lj}\phi''\left(\mathcal{Z}_{-jl}\right)V_{lj}z_j\phi'\left(\mathcal{X}_{-jl}\right)V_{lj}x_j\right|\leq LL_1|x_jz_j|\sum_{l=1}^{k}\left|V_{lj}\right|^3=\mathcal{O}\left(\log^2 k\frac{k}{d^{3/2}}\right)=o\left(\frac{k}{d}\right). \tag{195}$$

Then, we are left only with two cross terms, and we can write

$$I_j=\sum_l V_{lj}^2\left(\phi''\left(\mathcal{Z}_{-jl}\right)\tilde{\phi}\left(\mathcal{X}_{-jl}\right)z_j+\phi'\left(\mathcal{Z}_{-jl}\right)\phi'\left(\mathcal{X}_{-jl}\right)x_j\right)+\text{rest}, \tag{196}$$

where $|\text{rest}|$ can be upper-bounded by the sum of the terms (a)-(f), which gives

$$\left|I_j-\sum_l V_{lj}^2\left(\phi''\left(\mathcal{Z}_{-jl}\right)\tilde{\phi}\left(\mathcal{X}_{-jl}\right)z_j+\phi'\left(\mathcal{Z}_{-jl}\right)\phi'\left(\mathcal{X}_{-jl}\right)x_j\right)\right|=o\left(\frac{k}{d}\right) \tag{197}$$

with probability at least $1-\exp(-c_5d)-\exp(-c_7\log^2 k)$ over x, z and V. Hence, the thesis follows after using Assumption 4.4. □

Lemma C.17. *We have that*

$$\left|\sum_{l=1}^{k} V_{lj}^2B_{-jl}-\frac{1}{d}\sum_{l=1}^{k}B_{-jl}\right|=o\left(\frac{k}{d}\right), \tag{198}$$

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

Proof. We have

$$\sum_{l=1}^{k} V_{lj}^2 B_{-jl} = \sum_{l=1}^{k} \left(V_{lj}^2 - \mathbb{E}_V\left[V_{lj}^2\right]\right) B_{-jl} + \frac{1}{d}\sum_{l=1}^{k} B_{-jl}, \tag{199}$$

since $\mathbb{E}_V\left[V_{lj}^2\right] = 1/d$ for all l, j.

We call $v_l := V_{lj}^2 - \mathbb{E}_V\left[V_{lj}^2\right]$. Then, $v = (v_1, \ldots, v_k)$ is a mean-0 vector with independent and sub-exponential entries with norm $\mathcal{O}(1/d)$. This implies $\|v\|_{\psi_1} = \mathcal{O}(1/d)$. Thus,

$$\mathbb{P}_v\left(|v^\top u| > t\right) < 2\exp\left(-c\frac{dt}{\|u\|_2}\right), \tag{200}$$

for every vector u independent from v. If we define $u_l := B_{-jl}$, we have

$$\|u\|_2 \leq \sqrt{k}\left(\max_l |B_{-jl}|\right) \leq \sqrt{k}\left(L^2|x_j| + L_1|z_j|\max_l\left|\tilde{\phi}(\mathcal{X}_{-jl})\right|\right) = \mathcal{O}\left(\log^2(k)\sqrt{k}\right), \tag{201}$$

where the last equality comes from the fact that x_j and z_j are sub-Gaussian, and from the second statement in Lemma C.14. This holds with probability at least $1 - e^{-c_1 d} - e^{-c_2 \log^2 k}$ over x, z and $\{V_{li}\}_{i\neq j, 1\leq l\leq k}$, and we will condition on this event.

Then, merging this last result with (200) and setting $t = \log^4 k\sqrt{k}/d$ we have

$$\mathbb{P}_v\left(|v^\top u| > \frac{\log^4 k\sqrt{k}}{d}\right) < 2\exp\left(-c_3\log^2 k\right). \tag{202}$$

Performing a union bound gives the thesis. □

Lemma C.18. *We have that*

$$\left|\sum_{l=1}^{k} B_{-jl} - \sum_{l=1}^{k} B_{jl}\right| = o(k), \tag{203}$$

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

Proof. Let's condition on

$$|x_j| = \mathcal{O}(\log k), \qquad |z_j| = \mathcal{O}(\log k), \qquad \max_l |V_{lj}| = \mathcal{O}\left(\frac{\log k}{\sqrt{d}}\right). \tag{204}$$

These events happen with probability at least $1 - \exp(-c\log^2 k)$, as $\|x_j\|_{\psi_2} = \mathcal{O}(1)$, $\|z_j\|_{\psi_2} = \mathcal{O}(1)$ and $\|V_{jl}\|_{\psi_2} = \mathcal{O}\left(1/\sqrt{d}\right)$. The probability is over x, z and V. Also, we have $\max_l\left|\tilde{\phi}(\mathcal{X}_{-jl})\right| = \mathcal{O}(\log k)$ with probability at least $1 - e^{-c_1\log^2 k}$ over x, z and V, for the second statement in Lemma C.14. Let's condition on these high probability events for the rest of the proof, whenever we do not take expectations with respect to x.

Exploiting that ϕ, ϕ' and ϕ'' are all Lipschitz,

$$\max_l |\phi''(\mathcal{Z}_{-jl}) - \phi''(\mathcal{Z}_l)| = \mathcal{O}\left(\frac{\log^2 k}{\sqrt{d}}\right), \tag{205}$$

$$\max_l\left(|\phi(\mathcal{X}_{-jl}) - \phi(\mathcal{X}_{-jl})| + |\mathbb{E}_x\left[\phi(\mathcal{X}_{-jl}) - \phi(\mathcal{X}_{-jl})\right]|\right) \leq \mathcal{O}\left(\frac{\log^2 k}{\sqrt{d}}\right) + \max_l |V_{lj}|\mathbb{E}_x[|x_j|] = \mathcal{O}\left(\frac{\log^2 k}{\sqrt{d}}\right), \tag{206}$$

$$\max_l |\phi'(\mathcal{Z}_{-jl}) - \phi'(\mathcal{Z}_l)| = \mathcal{O}\left(\frac{\log^2 k}{\sqrt{d}}\right), \tag{207}$$

$$\max_l |\phi'(\mathcal{X}_{-jl}) - \phi'(\mathcal{X}_{-jl})| = \mathcal{O}\left(\frac{\log^2 k}{\sqrt{d}}\right). \tag{208}$$

Then, controlling all the cross terms, we get

$$\max_l |B_{-jl} - B_{jl}| = \mathcal{O}\left(\frac{\log^4 k}{\sqrt{d}}\right), \tag{209}$$

where this scaling is justified by the leading term

$$\max_l |\phi''(\mathcal{Z}_{-jl}) - \phi''(\mathcal{Z}_l)| \max_l \left|\tilde{\phi}(\mathcal{X}_{-jl})\right| |z_j| = \mathcal{O}\left(\frac{\log^2 k}{\sqrt{d}} \log k \log k\right). \tag{210}$$

Thus, we can conclude

$$\left|\sum_{l=1}^k B_{-jl} - \sum_{l=1}^k B_{jl}\right| \leq k \max_l |B_{-jl} - B_{jl}| = \mathcal{O}\left(\frac{k \log^4 k}{\sqrt{d}}\right) = o(k), \tag{211}$$

because of Assumption 4.4. This gives the thesis. □

Lemma C.19. *We have that*

$$\left|\sum_{l=1}^k B_{jl} - k\,\mathbb{E}_V[B_{j1}]\right| = o(k), \tag{212}$$

with probability at least $1 - \exp(-c \log^2 k)$ *over* x*,* z *and* V*.*

Proof. Since x_j and z_j are sub-Gaussian, we know that $|x_j|, |z_j| = \mathcal{O}(\log k)$ with probability at least $1 - \exp(-c \log^2 k)$ over x and z. We will condition on such event for the rest of the proof. Note that, once we fix the values of x and z, the B_{jl}'s are i.i.d. random variables in the probability space of V. Thus,

$$\sum_{l=1}^k B_{jl} = k\,\mathbb{E}_V[B_{j1}] + \sum_{l=1}^k (B_{jl} - \mathbb{E}_V[B_{jl}]). \tag{213}$$

Until the end of the proof, we only refer to the probability space of V.

We have

$$\left\|\tilde{\phi}(\mathcal{X}_l)\right\|_{\psi_2} \leq \|\phi(\mathcal{X}_l)\|_{\psi_2} + \|\mathbb{E}_x[\phi(\mathcal{X}_l)]\|_{\psi_2} \leq C\|x\|_2 \frac{1}{\sqrt{d}} + \mathbb{E}_x\left[\|\phi(\mathcal{X}_l)\|_{\psi_2}\right] \leq 2C, \tag{214}$$

where the second inequality is true because $\left\|x^\top V_{l:}\right\|_{\psi_2} \leq C_1 \|x\| \frac{1}{\sqrt{d}}$ and because ϕ is Lipschitz.

Since ϕ and ϕ' are both Lipschitz, we have that also $\|\phi''(\mathcal{Z}_l)\|_{\psi_2}$, $\|\phi'(\mathcal{Z}_l)\|_{\psi_2}$ and $\|\phi'(\mathcal{X}_l)\|_{\psi_2}$ are upper bounded by a numerical constant. Thus, we have that

$$\|B_{jl}\|_{\psi_1} \leq \log k. \tag{215}$$

An application of Bernstein inequality (cf. Theorem 2.8.1. in (Vershynin, 2018)) gives that

$$\mathbb{P}\left(\left|\sum_{l=1}^k (B_{jl} - \mathbb{E}_V[B_{jl}])\right| > \sqrt{k} \log^2 k\right) < \exp(-c_1 \log^2 k), \tag{216}$$

and the thesis readily follows. □

Lemma C.20. *Indicating with* ρ *a standard Gaussian random variable, we have that*

$$\left|\mathbb{E}_V[B_{j1}] - x_j \mathbb{E}_\rho^2[\phi'(\rho)]\right| = o(1), \tag{217}$$

with probability at least $1 - \exp(-c \log^2 k)$ *over* x *and* z*.*

Proof. Let's define the vector z' as follows

$$z' = \frac{\sqrt{d}\left(I - \frac{xx^\top}{d}\right)z}{\left\|\left(I - \frac{xx^\top}{d}\right)z\right\|_2}. \tag{218}$$

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

$$\frac{\left|z^\top z'\right|}{d} = \sqrt{1 - \left(\frac{x^\top z}{d}\right)^2} \geq 1 - \left(\frac{x^\top z}{d}\right)^2, \tag{219}$$

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

$$\|z - z'\|_2^2 = \|z\|_2^2 + \|z\|_2^2 - 2z^\top z' \leq 2d\left(1 - \left(1 - \left(\frac{x^\top z}{d}\right)^2\right)\right) = 2\frac{\left(x^\top z\right)^2}{d}. \tag{220}$$

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

$$\mathbb{P}\left(\|z - z'\|_2 > t\right) \leq \mathbb{P}\left(|x^\top z| > \sqrt{d}t/\sqrt{2}\right) < \exp(-ct^2), \tag{221}$$

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

Let's now introduce the shorthands $v := V_{1j}$ and $\mathcal{Z}_1' = v^\top z'$. We have

$$\left|\mathbb{E}_v\left[\phi''(\mathcal{Z}_1)\phi(\mathcal{X}_1)\right] - \mathbb{E}_v\left[\phi''(\mathcal{Z}_1')\phi(\mathcal{X}_1)\right]\right| \leq L_2\mathbb{E}_v\left[|v^\top(z - z')||\phi(\mathcal{X}_1)|\right]. \tag{222}$$

Since ϕ is Lipschitz, $\phi(\mathcal{X}_1)$ is sub-Gaussian with respect to the probability space of v. Since z and z' are independent from v, we deduce that

$$\left\|v^\top(z - z')\right\|_{\psi_2} \leq \|z - z'\|_2 \frac{C}{\sqrt{d}} = \mathcal{O}\left(\frac{\log k}{\sqrt{d}}\right), \tag{223}$$

with probability at least $1 - \exp(-c\log^2 k)$ over x. If we condition on such event, this implies that the expectation in (222) can be bounded as

$$\mathbb{E}_v\left[|v^\top(z - z')||\phi(\mathcal{X}_1)|\right] = \mathcal{O}\left(\frac{\log k}{\sqrt{d}}\right). \tag{224}$$

This is true since we are computing the first moment of a sub-exponential random variable.

A similar computation gives

$$\begin{aligned}&\left|\mathbb{E}_v\left[\phi''(\mathcal{Z}_1)\mathbb{E}_x[\phi(\mathcal{X}_1)]\right] - \mathbb{E}_{vx}\left[\phi''(\mathcal{Z}_1')\phi(\mathcal{X}_1)\right]\right| \\ &\quad \leq L_2\mathbb{E}_x\mathbb{E}_v\left[|v^\top(z - z')||\phi(\mathcal{X}_1)|\right] \leq C\frac{\mathbb{E}_x\left[\|z - z'\|_2\right]}{\sqrt{d}} = \mathcal{O}\left(\frac{1}{\sqrt{d}}\right),\end{aligned} \tag{225}$$

where the last step is a direct consequence of (221). Notice that this, together with the previous argument between (222) and (224), gives

$$\left|\mathbb{E}_v\left[\phi''(\mathcal{Z}_1)\tilde{\phi}(\mathcal{X}_1)\right] - \left(\mathbb{E}_v\left[\phi''(\mathcal{Z}_1')\phi(\mathcal{X}_1)\right] - \mathbb{E}_{vx}\left[\phi''(\mathcal{Z}_1')\phi(\mathcal{X}_1)\right]\right)\right| = \mathcal{O}\left(\frac{\log k}{d}\right). \tag{226}$$

With an equivalent argument of the one between (222) and (224), we can also show that

$$\left|\mathbb{E}_v\left[\phi'(\mathcal{Z}_1)\phi'(\mathcal{X}_1)\right] - \mathbb{E}_v\left[\phi'(\mathcal{Z}_1')\phi'(\mathcal{X}_1)\right]\right| = \mathcal{O}\left(\frac{\log k}{\sqrt{d}}\right), \tag{227}$$

with probability at least $1-\exp(-c\log^2 k)$ over x.

At this point, since $x^\top z' = 0$, we have have that (in the probability space of V), $\mathcal{X}_1$ and $\mathcal{Z}_1'$ can be treated as independent standard Gaussian random variables ρ_1 and ρ_2. We therefore have

$$\mathbb{E}_v\left[\phi''(\mathcal{Z}_1')\phi(\mathcal{X}_1)\right] = \mathbb{E}_{\rho_1\rho_2}\left[\phi''(\rho_1)\phi(\rho_2)\right] = \mathbb{E}_{\rho_1}\left[\phi''(\rho_1)\right]\mathbb{E}_{\rho_2}\left[\phi(\rho_2)\right], \tag{228}$$

$$\mathbb{E}_{vx}\left[\phi''(\mathcal{Z}_1')\phi(\mathcal{X}_1)\right] = \mathbb{E}_x\left[\mathbb{E}_{\rho_1\rho_2}\left[\phi''(\rho_1)\phi(\rho_2)\right]\right] = \mathbb{E}_{\rho_1}\left[\phi''(\rho_1)\right]\mathbb{E}_{\rho_2}\left[\phi(\rho_2)\right], \tag{229}$$

$$\mathbb{E}_v\left[\phi'(\mathcal{Z}_1')\phi'(\mathcal{X}_1)\right] = \mathbb{E}_{\rho_1\rho_2}\left[\phi'(\rho_1)\phi'(\rho_2)\right] = \mathbb{E}_{\rho_1}^2\left[\phi'(\rho_1)\right]. \tag{230}$$

Now, merging (226) with (228) and (229), we get

$$\left|\mathbb{E}_v\left[\phi''(\mathcal{Z}_1)\tilde{\phi}(\mathcal{X}_1)\right] - \left(\mathbb{E}_{\rho_1}\left[\phi''(\rho_1)\right]\mathbb{E}_{\rho_2}\left[\phi(\rho_2)\right] - \mathbb{E}_{\rho_1}\left[\phi''(\rho_1)\right]\mathbb{E}_{\rho_2}\left[\phi(\rho_2)\right]\right)\right| = \left|\mathbb{E}_v\left[\phi''(\mathcal{Z}_1)\phi(\mathcal{X}_1)\right]\right| = \mathcal{O}\left(\frac{\log k}{\sqrt{d}}\right). \tag{231}$$

While merging (227) with (230) returns

$$\left|\mathbb{E}_v\left[\phi'(\mathcal{Z}_1)\phi'(\mathcal{X}_1)\right] - \mathbb{E}_{\rho_1}^2\left[\phi'(\rho_1)\right]\right| = \mathcal{O}\left(\frac{\log k}{\sqrt{d}}\right). \tag{232}$$

The previous statements hold with probability at least $1-\exp(-c_1\log^2 k)$ over x.

Thus, conditioning on $|x_j| = \mathcal{O}(\log k)$ and $|z_j| = \mathcal{O}(\log k)$, we get

$$\left|\mathbb{E}_V[B_{j1}] - x_j\mathbb{E}_\rho^2[\phi'(\rho)]\right| = \mathcal{O}\left(\frac{\log^2 k}{\sqrt{d}}\right) = o(1), \tag{233}$$

with probability at least $1-\exp(-c_2\log^2 k)$ over x and z. This readily gives the thesis. □

At this point, we are ready to prove Theorem 4.8.

Proof of Theorem 4.8. Let's focus on $\mathcal{I}_{x_1}^2 = \sum_{j=1}^d I_j^2$, where with a slight abuse of notation we now refer to I_j as the quantity defined in (166), with the variable x_1 instead of x. All the results from the previous lemmas hold as $x_1 \sim P_X$. We have, by triangle inequality,

$$\begin{aligned}\left|I_j - x_j\mathbb{E}_\rho^2[\phi'(\rho)]\frac{k}{d}\right| &\le \left|I_j - \sum_{l=1}^k V_{lj}^2 B_{-jl}\right| + \left|\sum_{l=1}^k V_{lj}^2 B_{-jl} - \frac{1}{d}\sum_{l=1}^k B_{-jl}\right| \\ &+ \frac{1}{d}\left|\sum_{l=1}^k B_{-jl} - \sum_{l=1}^k B_{jl}\right| + \frac{1}{d}\left|\sum_{l=1}^k B_{jl} - k\,\mathbb{E}_V[B_{j1}]\right| + \frac{k}{d}\left|\mathbb{E}_V[B_{j1}] - x_j\mathbb{E}_\rho^2[\phi'(\rho)]\right| \\ &= o\left(\frac{k}{d}\right) + o\left(\frac{k}{d}\right) + \frac{1}{d}o(k) + \frac{1}{d}o(k) + \frac{k}{d}o(1) = o\left(\frac{k}{d}\right),\end{aligned} \tag{234}$$

where the five terms in the third line are justified respectively by Lemmas C.16, C.17, C.18, C.19 and C.20, and jointly hold with probability at least $1-\exp(-c_2\log^2 k)$ over x_1 and z and V.

This holds for every $1 \le j \le d$ with probability at least $1 - d\exp(-c_2\log^2 k) \ge 1-\exp(-c_3\log^2 k)$. Thus, we can write

$$I_{x_1} \ge \sqrt{\sum_{j=1}^d \left(|x_j|\mathbb{E}_\rho^2[\phi'(\rho)]\frac{k}{d} - o\left(\frac{k}{d}\right)\right)^2} \ge \sqrt{\mathbb{E}_\rho^4[\phi'(\rho)]\frac{k^2}{d^2}\sum_{j=1}^d x_j^2 - o\left(\frac{k^2}{d}\right) - o\left(\frac{k^2\sum_j|x_j|}{d^2}\right)}. \tag{235}$$

By Cauchy-Schwartz, we have that $\sum_j |x_j| \le \sqrt{d\sum_j x_j^2} = d$. If $\mathbb{E}_\rho[\phi'(\rho)] \ne 0$, we have

$$I_{x_1} \ge \sqrt{\mathbb{E}_\rho^4[\phi'(\rho)]\frac{k^2}{d} - o\left(\frac{k^2}{d}\right)} = \mathbb{E}_\rho^2[\phi'(\rho)]\frac{k}{\sqrt{d}}\sqrt{1-o(1)} = \mathbb{E}_\rho^2[\phi'(\rho)]\frac{k}{\sqrt{d}} - o\left(\frac{k}{\sqrt{d}}\right). \tag{236}$$

Following the same strategy, we can also prove that

$$I_{x_1} \leq \mathbb{E}_\rho^2[\phi'(\rho)]\frac{k}{\sqrt{d}} + o\left(\frac{k}{\sqrt{d}}\right), \tag{237}$$

which finally gives

$$\left| I_{x_1} - \mathbb{E}_\rho^2[\phi'(\rho)]\frac{k}{\sqrt{d}} \right| = o\left(\frac{k}{\sqrt{d}}\right). \tag{238}$$

Notice that we easily retrieve the same result also in the case $\mathbb{E}_\rho[\phi'(\rho)] = 0$.

Performing now a union bound over all the data, and recalling that $\|\mathcal{I}_{\text{RF}}\|_F = \sqrt{\sum_{i=1}^N I_{x_i}^2} \geq \sqrt{N}\min_i I_{x_i}$ we have

$$\|\mathcal{I}_{\text{RF}}\|_F \geq \mathbb{E}_\rho^2[\phi'(\rho)]\frac{k\sqrt{N}}{\sqrt{d}} - o\left(\frac{k\sqrt{N}}{\sqrt{d}}\right), \tag{239}$$

with probability at least $1 - N\exp(-c_3\log^2 k) \geq 1 - \exp(-c_4\log^2 k)$ over X and z and V. The same argument proves also the upper bound, and the thesis readily follows. □

C.5. Proof of Theorem 4.5

First, we state and prove an auxiliary result upper bounding $\|A\|_{\text{op}}$.

Lemma C.21. *Let $A = \nabla_z\phi(Vz)^\top\phi(XV^\top)^\top K^{-1}$. Then, we have*

$$\|A\|_{\text{op}} = \mathcal{O}\left(\frac{1}{\sqrt{d}}\right), \tag{240}$$

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

Proof. Recalling that $\nabla_z\phi(Vz)^\top = V^\top\text{diag}\left(\phi'(Vz)\right)$, we have

$$\begin{aligned}
\|A\|_{\text{op}} &= \left\|\nabla_z\phi(Vz)^\top\phi(XV^\top)^\top K^{-1}\right\|_{\text{op}} \\
&= \left\|V^\top\text{diag}\left(\phi'(Vz)\right)\phi(XV^\top)^\top K^{-1}\right\|_{\text{op}} \\
&\leq \|V\|_{\text{op}}\left\|\text{diag}\left(\phi'(Vz)\right)\right\|_{\text{op}}\left\|\Phi^+\right\|_{\text{op}} \\
&\leq \|V\|_{\text{op}} L\,\lambda_{\min}(K)^{-1/2}.
\end{aligned} \tag{241}$$

where in the last line we use that ϕ is an L-Lipschitz function and that $K^{-1} = (\Phi^+)^\top(\Phi^+)$. By Theorem 4.4.5 of (Vershynin, 2018), we have (as $d = o(k)$ by Assumption 4.3) that

$$\|V\|_{\text{op}} = \mathcal{O}\left(1 + \sqrt{\frac{k}{d}}\right) = \mathcal{O}\left(\sqrt{\frac{k}{d}}\right), \tag{242}$$

with probability at least $1 - \exp(-cd)$ over V. Lemma C.5 readily gives $\lambda_{\min}(K) = \Omega(k)$, with probability at least $1 - \exp(-c\log^2 N)$ over V and X. Taking a union bound on these two events gives the thesis. □

Proof of Theorem 4.5. By Theorem 4.8, we have $\|\mathcal{I}_{\text{RF}}\|_F \geq C_1\frac{k\sqrt{N}}{\sqrt{d}}$, with probability at least $1 - \exp(-c\log^2 k)$, which implies, by Theorem 4.7,

$$\|A(z)\|_F \geq C_2\frac{d}{kN}\frac{k\sqrt{N}}{\sqrt{d}} - C\sqrt{N+d}/d = C_2\frac{\sqrt{d}}{\sqrt{N}} - C\sqrt{N+d}/d \geq C_3\frac{\sqrt{d}}{\sqrt{N}}, \tag{243}$$

where the last inequality follows from Assumption 4.4, and the first holds with probability at least $1 - \exp(-c\log^2 N) - k\exp(-cd)$. This also means that

$$\mathcal{S}(z) \geq C_4\sqrt{d}\,\|A(z)\|_F \tag{244}$$

holds with probability at least $1 - \exp\left(-cd^2/N\right)$, because of Theorem 4.6 and Lemma C.21. Thus, we can conclude that

$$\mathcal{S}(z) \geq C_4 \|z\| \|A(z)\|_F \geq C_5 \frac{d}{\sqrt{N}} \geq C_6 \sqrt[6]{N}, \tag{245}$$

where the last inequality is a consequence of Assumption 4.4. □

D. Proofs for NTK Regression

In this section, we will indicate with $X \in \mathbb{R}^{N \times d}$ the data matrix, such that its rows are sampled independently from P_X (see Assumption 4.1). We will indicate with $W \in \mathbb{R}^{k \times d}$ the first half of the weight matrix at initialization, such that $W_{ij} \sim_{\text{i.i.d.}} \mathcal{N}(0, 1/d)$. The feature vector will be therefore be (see (26))

$$\Phi_{\text{NTK}}(x) = [x \otimes \phi'(Wx), -x \otimes \phi'(Wx)] \tag{246}$$

where ϕ is the activation function, applied component-wise to the pre-activations Wx.

Notice that, for compactness, we dropped the subscripts "NTK" from these quantities, as this section will only treat the proofs related to Section 5. Always for the sake of compactness, we will not re-introduce such quantities in the statements or the proofs of the following Lemmas.

Lemma D.1. *We have that*

$$\lambda_{\min}(K) = \Omega(kd), \tag{247}$$

with probability at least $1 - Ne^{-c\log^2 k} - e^{-c\log^2 N}$ *over* X *and* W.

Proof. The thesis follows from Theorem 3.1 of (Bombari et al., 2022). Notice that our assumptions on the data distribution P_X are stronger, and that our initialization of the very last layer (which differs from their Gaussian initialization) does not change the result. Our assumption $k = \mathcal{O}(d)$ respects the *loose pyramidal topology* condition, and we left the overparameterization assumption unchanged. A main difference is that we do not assume the activation function ϕ to be Lipschitz anymore. This, however, stop being a necessary assumption since we are working with a 2-layer neural network, and ϕ doesn't appear in the expression of Neural Tangent Kernel. □

Lemma D.2. *We have that*

$$\max_i \left|\phi'(Wx_i)^\top \phi'(Wz)\right| = \mathcal{O}\left(\sqrt{k}\log k\right), \tag{248}$$

with probability at least $1 - \exp(-c\log^2 k)$ *over* X *and* W.

Proof. Let's look at the term $i = 1$, and let's call $x := x_1$ for compactness. As in Lemma C.20, we define z' as follows

$$z' = \frac{\sqrt{d}\left(I - \frac{xx^\top}{d}\right) z}{\left\|\left(I - \frac{xx^\top}{d}\right) z\right\|_2}. \tag{249}$$

We have $x^\top z' = 0$, and $\|z'\|_2 = \sqrt{d}$. Furthermore, by (221), we also have that $\|z - z'\|_2$ is a sub-Gaussian random variable, in the probability space of x. It will also be convenient to characterize $\|\phi'(Wz')\|_2$. We have, indicating with w_j the j-th row of W,

$$\begin{aligned}\|\phi'(Wz')\|_2^2 &= \sum_{j=1}^k \phi'^2(w_j^\top z') = \sum_{j=1}^k \mathbb{E}_W\left[\phi'^2(w_j^\top z')\right] + \sum_{j=1}^k \left(\phi'^2(w_j^\top z') - \mathbb{E}_W\left[\phi'^2(w_j^\top z')\right]\right) \\ &= k\mathbb{E}_\rho\left[\phi'^2(\rho)\right] + \sum_{j=1}^k \left(\phi'^2(\rho_j) - \mathbb{E}_\rho\left[\phi'^2(\rho)\right]\right),\end{aligned} \tag{250}$$

where the ρ and ρ_j's indicate independent standard Gaussian random variables. Since ϕ' is Lipschitz and non-zero, we have that $\mathbb{E}_\rho\left[\phi'^2(\rho)\right] = \Theta(1)$. Furthermore, by Bernstein inequality (cf. Theorem 2.8.1 in (Vershynin, 2018)), we have

$\sum_{j=1}^{k}\left(\phi'^2(\rho_j)-\mathbb{E}_\rho\left[\phi'^2(\rho)\right]\right)=\mathcal{O}\left(k\right)$ with probability at least $1-2\exp(-ck)$. This means that, with this probability (which is over W), we have

$$\left\|\phi'(Wz')\right\|_2 \leq C\sqrt{k}, \tag{251}$$

where C is a numerical constant. Notice that the same result holds for $\left\|\phi'(Wx)\right\|_2$. We also have $\left\|W\right\|_{\text{op}}=\mathcal{O}\left(1\right)$, which happens with probability at least $1-\exp(-ck)$ over W, because of Theorem 4.4.5 of (Vershynin, 2018) and Assumption 5.3.

Since ϕ' is Lipschitz, we can write

$$\left|\phi'(Wx)^\top\phi'(Wz)\right| \leq \left|\phi'(Wx)^\top\phi'(Wz')\right| + L\left\|\phi'(Wx)\right\|_2\left\|W\right\|_{\text{op}}\left\|z-z'\right\|_2. \tag{252}$$

Let's look at the first term of the LHS. Since $x^\top z'=0$, and $\left\|x\right\|_2=\left\|z'\right\|_2=\sqrt{d}$, we have

$$\phi'(Wx)^\top\phi'(Wz')=\phi'(\hat{\rho}_a)^\top\phi'(\hat{\rho}_b), \tag{253}$$

where $\hat{\rho}_a$ and $\hat{\rho}_b$ are two independent Gaussian vectors, each of them with independent entries. As ϕ is an even function (see Assumption 5.1), we have that $\phi'(\hat{\rho}_a)$ and $\phi'(\hat{\rho}_b)$ are independent sub-Gaussian vectors, with norm upper bounded by a numerical constant. Thus,

$$\mathbb{P}\left(\left|\phi'(Wx)^\top\phi'(Wz')\right| > C\sqrt{k}\log k\right) < 2\exp(-c_1\log^2 k), \tag{254}$$

where this holds due to (251). Let's now look at the second term of the LHS of (252). We have

$$\left\|\phi'(Wx)\right\|_2\left\|W\right\|_{\text{op}}\left\|z-z'\right\|_2 \leq C\sqrt{k}C_1\left\|z-z'\right\|_2=\mathcal{O}\left(\log k\sqrt{k}\right), \tag{255}$$

where the inequality holds with probability at least $1-\exp(-ck)$ over W, and the equality holds with probability at least $1-2\exp(-c_2\log^2 k)$ over x. Taking the intersection between the previously described high probability events, from (252) we get

$$\left|\phi'(Wx)^\top\phi'(Wz)\right|=\mathcal{O}\left(\sqrt{k}\log k\right), \tag{256}$$

with probability at least $1-2\exp(-c\log^2 k)$ over x and W. Thus, performing a union bound over the different x_i's, we get the thesis. □

Lemma D.3.

$$\left\|\nabla_z\Phi(z)^\top\Phi(X)^\top\right\|_{\text{op}}=\mathcal{O}\left(\log k\left(\sqrt{k}+\sqrt{N}\right)\sqrt{d}\right), \tag{257}$$

with probability at least $1-\exp(-c\log^2 k)$ *over* X*,* W *and* z*.*

Proof. We have, by application of the chain rule,

$$\begin{aligned}
\nabla_z\Phi(z)^\top\Phi(X)^\top &= 2\,\nabla_z\Phi'(z)^\top\Phi'(X)^\top \\
&= 2\,\nabla_z\left(z\otimes\phi'(Wz)\right)^\top\left(X*\phi'(XW^\top)\right)^\top \\
&= 2\left(I*\mathbf{1}\phi'^\top(Wz)+\mathbf{1}z^\top*W^\top\text{diag}\left(\phi''(Wz)\right)\right)\left(X*\phi'(XW^\top)\right)^\top \\
&= 2\,X^\top\circ\mathbf{1}\left(\phi'(XW^\top)\phi'(Wz)\right)^\top+2\,\mathbf{1}\left(Xz\right)^\top\circ W^\top\text{diag}\left(\phi''(Wz)\right)\phi'(XW^\top)^\top \\
&= 2\,X^\top\text{diag}\left(\phi'(XW^\top)\phi'(Wz)\right)+2\,W^\top\text{diag}\left(\phi''(Wz)\right)\phi'(XW^\top)^\top\text{diag}\left(Xz\right).
\end{aligned} \tag{258}$$

Let's now condition on the event $\left\|W\right\|_{\text{op}}=\mathcal{O}\left(1\right)$, which happens with probability at least $1-\exp(-cd)$ over W, because of Theorem 4.4.5 of (Vershynin, 2018) and Assumption 5.3.

Let's now look at the two terms separately. For the first, we have

$$\begin{aligned}
\left\|X^\top\text{diag}\left(\phi'(XW^\top)\phi'(Wz)\right)\right\|_{\text{op}} &\leq \left\|X\right\|_{\text{op}}\max_i\left|\phi'(Wx_i)^\top\phi'(Wz)\right| \\
&= \mathcal{O}\left(\log k\sqrt{k}\left(\sqrt{N}+\sqrt{d}\right)\right),
\end{aligned} \tag{259}$$

where the last equality is true because of Lemmas B.5 and D.2 and holds with probability at least $1 - \exp(-cd) - \exp(-c\log^2 k)$ over X and W. For the other term, we have

$$\begin{aligned}\left\|W^\top \mathrm{diag}\left(\phi''(Wz)\right)\phi'(XW^\top)\mathrm{diag}\left(Xz\right)\right\|_{\mathrm{op}} &\leq \|W\|_{\mathrm{op}} L \left\|\phi'(XW^\top)\right\|_{\mathrm{op}} \max_i \left|x_i^\top z\right| \\ &= \mathcal{O}\left(\log k\sqrt{d}\left(\sqrt{N}+\sqrt{k}\right)\right).\end{aligned} \tag{260}$$

The last step uses Lemma B.5 and the fact that we are taking the maximum over k sub-Gaussian random variables such that $\left\|x_i^\top z\right\|_{\psi_2} \leq C\|z\|_2 = C\sqrt{d}$, and it holds with probability at least $1 - 2\exp(-ck) - \exp(-c\log^2 k)$ over X.

Putting everything together, and using Assumption 5.3, we get

$$\begin{aligned}\left\|\nabla_z \Phi(z)^\top \Phi(X)^\top\right\|_{\mathrm{op}} &= \mathcal{O}\left(\log k\sqrt{k}\left(\sqrt{N}+\sqrt{d}\right)\right) + \mathcal{O}\left(\log k\sqrt{d}\left(\sqrt{N}+\sqrt{k}\right)\right) \\ &= \mathcal{O}\left(\log k\left(\sqrt{k}+\sqrt{N}\right)\sqrt{d}\right),\end{aligned} \tag{261}$$

which gives the thesis. □

D.1. Proof of Theorem 5.4

Proof of Theorem 5.4. We have

$$\begin{aligned}\mathcal{S}(z) &\leq \sqrt{d}\left\|\nabla_z \Phi(z)^\top \Phi(X)^\top\right\|_{\mathrm{op}} \left\|K^{-1}\right\|_{\mathrm{op}} \|Y\|_2 \\ &= \sqrt{d}\mathcal{O}\left(\log k\left(\sqrt{k}+\sqrt{N}\right)\sqrt{d}\right)\mathcal{O}\left(\frac{1}{dk}\right)\mathcal{O}\left(\sqrt{N}\right) \\ &= \mathcal{O}\left(\log k\left(1+\sqrt{N/k}\right)\frac{\sqrt{N}}{\sqrt{k}}\right),\end{aligned} \tag{262}$$

where the second line is justified by Lemmas D.1 and D.3, and holds with probability at least $1 - Ne^{-c\log^2 k} - e^{-c\log^2 N}$ over X and W. The second statement is immediate. □

E. Additional Experiments

In this section, we provide additional experimental results for the activation functions $\phi(x) = \cos(x)$ (which is even) and $\phi(x) = \frac{e^x}{1+e^x}$ (namely, softplus or smooth ReLU, which is not even). Again, our theoretical findings are supported by numerical evidence both on the RF (see Figure 5) and the NTK regression model (see Figure 6).

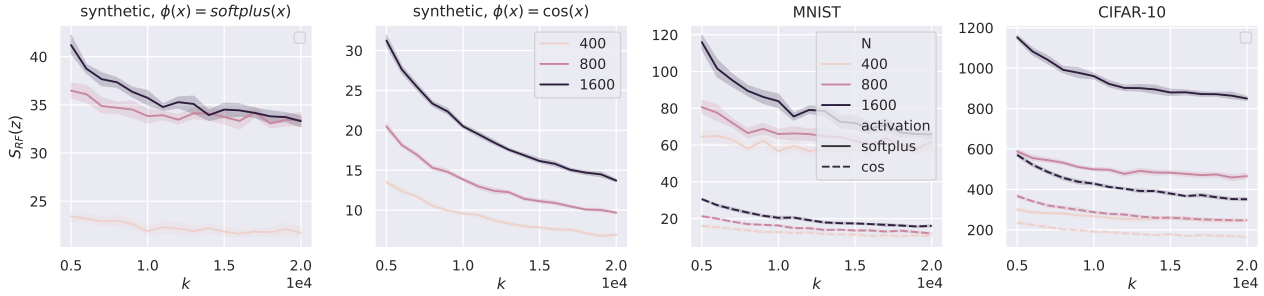

Figure 5. Sensitivity as a function of the number of parameters k, for an RF model with synthetic data sampled from a Gaussian distribution with input dimension $d = 1000$ (first and second plot), and with inputs taken from two classes of the MNIST and CIFAR-10 datasets (third and fourth plot respectively). Different curves correspond to different values of the sample size $N \in \{400, 800, 1600\}$ and to different activations ($\phi(x) = \mathrm{softplus}(x)$ or $\phi(x) = \cos(x)$). We plot the average over 10 independent trials and the confidence interval at 1 standard deviation. The values of the sensitivity are significantly larger for $\phi(x) = \mathrm{softplus}(x)$ than for $\phi(x) = \cos(x)$.

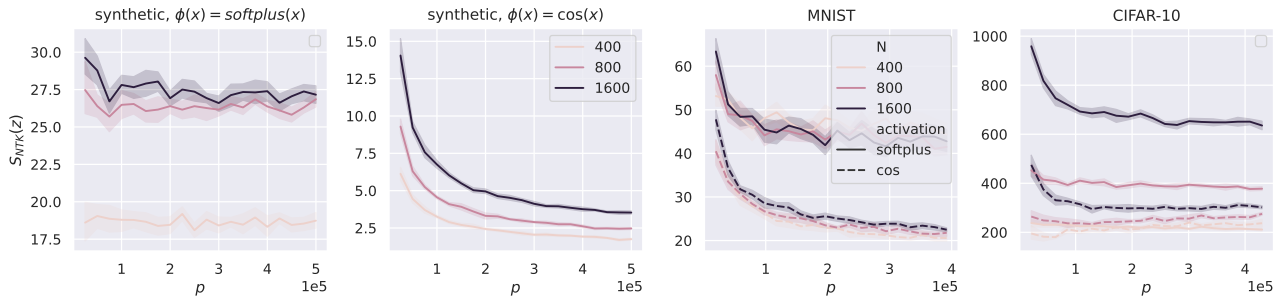

Figure 6. Sensitivity as a function of the number of parameters p, for an NTK model with synthetic data sampled from a Gaussian distribution with input dimension $d = 1000$ (first and second plot), and with inputs taken from two classes of the MNIST and CIFAR-10 datasets (third and fourth plot respectively). Different curves correspond to different values of the sample size $N \in \{400, 800, 1600\}$ and to different activations ($\phi(x) = \text{softplus}(x)$ or $\phi(x) = \cos(x)$). We plot the average over 10 independent trials and the confidence interval at 1 standard deviation. The values of the sensitivity are significantly larger for $\phi(x) = \text{softplus}(x)$ than for $\phi(x) = \cos(x)$.

The code to reproduce the experiments and all the plots in the paper is available at the public GitHub repository https://github.com/simone-bombari/beyond-universal-robustness.