Appendix

A. Omitted Proofs

In this section we provide complete proofs of the results in the main text.

A.1. Proof of Theorem 1

Theorem. *Consider a fully-connected neural network with NTK parametrization as introduced in Section 4.3, equipped with a 1-homogeneous activation function (such as ReLU). Set every bias to zero ($\beta_i = 0$) except for the output bias, $b^{(L)} \sim \mathcal{N}(0, \beta^2)$. Then it holds that both $\Theta^{(L)}$ and $\Sigma^{(L)}$ are semi-homogeneous kernels with $\zeta = \beta$.*

Proof. We will prove this statement via induction over the depth of the network. We will first show that a network without any bias is semi-homogeneous with parameter $\zeta = 0$. Fix any $\alpha \in \mathbb{R}_+$. Let us first consider the base case $l = 1$.

Base case: $\Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x}') = \Theta^{(1)}(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{d_0}\boldsymbol{x}^T\boldsymbol{x}'$. We easily deduce that

$$\Sigma^{(1)}(\alpha\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{d_0}(\alpha\boldsymbol{x})^T\boldsymbol{x}' + \beta^2 = \alpha\Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x}')$$

The same thing holds for the NTK $\Theta^{(1)}$. The base case thus holds.

Induction step: Assume that $\Sigma^{(l)}$ and $\Theta^{(l)}$ are semi-homogeneous with $\zeta = 0$. Let us first analyze the NNGP.

$$\Sigma^{(l+1)}(\alpha\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{0}, \Sigma^{(l)}|_{\alpha\boldsymbol{x},\boldsymbol{x}'})}\big[\sigma(z_1)\sigma(z_2)\big] = \frac{1}{2\pi\sqrt{\det(\Sigma^{(l)}|_{\alpha\boldsymbol{x},\boldsymbol{x}'})}}\int_{\mathbb{R}^2}\sigma(z_1)\sigma(z_2)e^{-\frac{1}{2}\boldsymbol{z}^T\Sigma^{(l)}|^{-1}_{\alpha\boldsymbol{x},\boldsymbol{x}'}\boldsymbol{z}}d\boldsymbol{z}$$

Now observe that

$$\begin{aligned}\det\left(\boldsymbol{\Sigma}^{(l)}|_{\alpha\boldsymbol{x},\boldsymbol{x}'}\right) &= \Sigma^{(l)}(\alpha\boldsymbol{x}, \alpha\boldsymbol{x})\Sigma^{(l)}(\boldsymbol{x}', \boldsymbol{x}') - \Sigma^{(l)}(\alpha\boldsymbol{x}, \boldsymbol{x}')^2 \\ &= \alpha^2\left(\Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x})\Sigma^{(l)}(\boldsymbol{x}', \boldsymbol{x}') - \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}')^2\right) \\ &= \alpha^2\det\left(\boldsymbol{\Sigma}^{(l)}|_{\boldsymbol{x},\boldsymbol{x}'}\right)\end{aligned}$$

On the other hand we have that

$$\begin{aligned}\left(\boldsymbol{\Sigma}^{(l)}|_{\alpha\boldsymbol{x},\boldsymbol{x}'}\right)^{-1} &= \frac{1}{\det\left(\boldsymbol{\Sigma}^{(l)}|_{\alpha\boldsymbol{x},\boldsymbol{x}'}\right)}\begin{pmatrix}\Sigma^{(l)}(\boldsymbol{x}', \boldsymbol{x}') & -\Sigma^{(l)}(\alpha\boldsymbol{x}, \boldsymbol{x}') \\ -\Sigma^{(l)}(\alpha\boldsymbol{x}, \boldsymbol{x}') & \Sigma^{(l)}(\alpha\boldsymbol{x}, \alpha\boldsymbol{x})\end{pmatrix} \\ &= \frac{1}{\alpha^2}\frac{1}{\det\left(\boldsymbol{\Sigma}^{(l)}|_{\boldsymbol{x},\boldsymbol{x}'}\right)}\begin{pmatrix}\Sigma^{(l)}(\boldsymbol{x}', \boldsymbol{x}') & -\alpha\Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}') \\ -\alpha\Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}') & \alpha^2\Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x})\end{pmatrix}\end{aligned}$$

We can hence write that

$$\boldsymbol{z}^T\left(\boldsymbol{\Sigma}^{(l)}|_{\alpha\boldsymbol{x},\boldsymbol{x}'}\right)^{-1}\boldsymbol{z} = \frac{1}{\det\left(\boldsymbol{\Sigma}^{(l)}|_{\boldsymbol{x},\boldsymbol{x}'}\right)}\left(z_1^2\frac{1}{\alpha^2}\Sigma^{(l)}(\boldsymbol{x}', \boldsymbol{x}') - 2z_1z_2\frac{1}{\alpha}\Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}') + z_2^2\Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x})\right)$$

Let us perform the substitution $u_1 = \frac{1}{\alpha}z_1$ and $u_2 = z_2$ with area element $d\boldsymbol{u} = \frac{1}{\alpha}d\boldsymbol{z}$. Then we can write

$$\begin{aligned}\boldsymbol{z}^T\left(\boldsymbol{\Sigma}^{(l)}|_{\alpha\boldsymbol{x},\boldsymbol{x}'}\right)^{-1}\boldsymbol{z} &= \frac{1}{\det\left(\boldsymbol{\Sigma}^{(l)}|_{\boldsymbol{x},\boldsymbol{x}'}\right)}\left(u_1^2\Sigma^{(l)}(\boldsymbol{x}', \boldsymbol{x}') - 2u_1u_2\Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}') + u_2^2\Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x})\right) \\ &= \boldsymbol{u}^T\left(\boldsymbol{\Sigma}^{(l)}|_{\boldsymbol{x},\boldsymbol{x}'}\right)^{-1}\boldsymbol{u}\end{aligned}$$

We can thus rewrite the integral as

$$\begin{aligned}\Sigma^{(l+1)}(\alpha\boldsymbol{x},\boldsymbol{x}') &= \frac{1}{2\pi\sqrt{\det(\Sigma^{(l)}|_{\alpha\boldsymbol{x},\boldsymbol{x}'})}}\int_{\mathbb{R}^2}\sigma(z_1)\sigma(z_2)e^{-\frac{1}{2}\boldsymbol{z}^T\Sigma^{(l)}|^{-1}_{\alpha\boldsymbol{x},\boldsymbol{x}'}\boldsymbol{z}}d\boldsymbol{z}\\ &= \frac{1}{2\pi\alpha\sqrt{\det\left(\boldsymbol{\Sigma}^{(l)}|_{\boldsymbol{x},\boldsymbol{x}'}\right)}}\int_{\mathbb{R}^2}\sigma(\alpha u_1)\sigma(u_2)e^{-\frac{1}{2}\boldsymbol{u}^T\Sigma^{(l)}|^{-1}_{\alpha\boldsymbol{x},\boldsymbol{x}'}\boldsymbol{u}}\alpha d\boldsymbol{u}\\ &= \alpha\Sigma^{(l+1)}(\boldsymbol{x},\boldsymbol{x}')\end{aligned}$$

where we have used the 1-homogenity of σ. Next we analyze the NTK $\Theta^{(l+1)}$. Here we have to control the additional term

$$\dot{\Sigma}^{(l+1)}(\alpha\boldsymbol{x},\boldsymbol{x}') = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}\left(\boldsymbol{0},\Sigma^{(l)}|_{\alpha\boldsymbol{x},\boldsymbol{x}'}\right)}\left[\dot{\sigma}(z_1)\dot{\sigma}(z_2)\right]$$

Since σ is 1-homogeneous, we know that its derivative is 0-homogeneous. We can thus apply the exact same computation as for $\Sigma^{(l+1)}$ to arrive at

$$\dot{\Sigma}^{(l+1)}(\alpha\boldsymbol{x},\boldsymbol{x}') = \dot{\Sigma}^{(l+1)}(\boldsymbol{x},\boldsymbol{x}')$$

Using the previous result and the induction hypothesis, we obtain

$$\begin{aligned}\Theta^{(l+1)}(\alpha\boldsymbol{x},\boldsymbol{x}') &= \Theta^{(l)}(\alpha\boldsymbol{x},\boldsymbol{x}')\dot{\Sigma}^{(l+1)}(\alpha\boldsymbol{x},\boldsymbol{x}') + \Sigma^{(l+1)}(\alpha\boldsymbol{x},\boldsymbol{x}') = \alpha\Theta^{(l)}(\boldsymbol{x},\boldsymbol{x}')\dot{\Sigma}^{(l+1)}(\boldsymbol{x},\boldsymbol{x}') + \alpha\Sigma^{(l+1)}(\boldsymbol{x},\boldsymbol{x}')\\ &= \alpha\Theta^{(l+1)}(\boldsymbol{x},\boldsymbol{x}')\end{aligned}$$

Given this result, we can now consider the kernel with an output bias β added. Let K denote either the NTK or NNGP kernel with an output bias and C the corresponding kernel without output bias. Then we obtain

$$\begin{aligned}K(\alpha\boldsymbol{x},\boldsymbol{x}') &= C(\alpha\boldsymbol{x},\boldsymbol{x}') + \beta^2 = \alpha C(\boldsymbol{x},\boldsymbol{x}') + \beta^2 = \alpha\left(C(\boldsymbol{x},\boldsymbol{x}') + \beta^2 - \beta^2\right) + \beta^2\\ &= \alpha K(\boldsymbol{x},\boldsymbol{x}') + \beta^2(1-\alpha)\end{aligned}$$

This concludes the proof. □

A.2. Proof of Lemma 2

Lemma. *Fix a semi-homogeneous kernel K and two data points sampled according to the adversarial spheres measure, $\boldsymbol{x},\boldsymbol{z}\sim p$. Consider the projection $\mathcal{P}(\boldsymbol{x})$. Denote $r = ||\boldsymbol{x}||_2$ and $\tilde{r} = ||\mathcal{P}(\boldsymbol{x})||_2$. Then it holds that:*

$$K(\mathcal{P}(\boldsymbol{x}),\boldsymbol{z}) = \frac{\tilde{r}}{r}K(\boldsymbol{x},\boldsymbol{z}) + \zeta^2(1-\frac{\tilde{r}}{r})$$

Proof. Realize that we can write the projection as

$$\mathcal{P}(\boldsymbol{x}) = \frac{\tilde{r}}{r}\boldsymbol{x} = \alpha\boldsymbol{x}$$

Obviously, $\alpha > 0$, thus we can apply the defining property of semi-homogeneous kernels to conclude

$$K(\mathcal{P}(\boldsymbol{x}),\boldsymbol{z}) = K(\alpha\boldsymbol{x},\boldsymbol{z}) = \frac{\tilde{r}}{r}K(\boldsymbol{x},\boldsymbol{z}) + \zeta^2(1-\frac{\tilde{r}}{r})$$

□

A.3. Proof of Corollary 2.1

Corollary. *Fix a semi-homogeneous kernel K and a data point sampled according to the adversarial spheres measure, $\boldsymbol{x}\sim p$. Consider the projection $\mathcal{P}(\boldsymbol{x})$. Denote $r = ||\boldsymbol{x}||_2$ and $\tilde{r} = ||\mathcal{P}(\boldsymbol{x})||_2$. Then it holds that*

$$f_K\left(\mathcal{P}(\boldsymbol{x})\right) = \frac{\tilde{r}}{r}f_K(\boldsymbol{x}) + \zeta^2\left(1-\frac{\tilde{r}}{r}\right)\gamma_K(n)$$

where we define $\gamma_K(n) = \mathbf{1}_n^T K(\boldsymbol{X},\boldsymbol{X})^{-1}\boldsymbol{y}$ and $\mathbf{1}_n = (1,\ldots,1)^T \in \mathbb{R}^n$.

Proof. We just need to apply the Lemma 2:

$$\begin{aligned} f_K\left(\mathcal{P}(\boldsymbol{x})\right) &= K(\mathcal{P}(\boldsymbol{x}), \boldsymbol{X})\left(K(\boldsymbol{X}, \boldsymbol{X})\right)^{-1} \boldsymbol{y} \\ &= \left(\frac{\tilde{r}}{r} K(\boldsymbol{x}, \boldsymbol{X}) + \zeta^2 \left(1 - \frac{\tilde{r}}{r}\right) \mathbf{1}_n\right) \left(K(\boldsymbol{X}, \boldsymbol{X})\right)^{-1} \boldsymbol{y} \\ &= \frac{\tilde{r}}{r} f_K(\boldsymbol{x}) + \zeta^2 \left(1 - \frac{\tilde{r}}{r}\right) \mathbf{1}_n \left(K(\boldsymbol{X}, \boldsymbol{X})\right)^{-1} \boldsymbol{y} \\ &= \frac{\tilde{r}}{r} f_K(\boldsymbol{x}) + \zeta^2 \left(1 - \frac{\tilde{r}}{r}\right) \gamma_K(n) \end{aligned}$$

□

A.4. Proof of Theorem 3

Theorem. *Take a semi-homogeneous kernel K and consider a training set $\mathcal{S}_{train} \overset{i.i.d.}{\sim} \mathcal{D}^n$ along with the corresponding adversarial set $\mathcal{S}_{adv}$. Then it holds that a_{adv} is quantized to only three values:*

$$a_{adv} \in \left\{0, 1-q, 1\right\}$$

Moreover, we can characterize the phase transitions in sample size n as

$$a_{adv} = \begin{cases} 0 & \text{if} \quad \gamma_K(n) \leq \frac{r_1}{\zeta^2(r_2 - r_1)} \\ 1 - q & \text{if} \quad \frac{r_1}{\zeta^2(r_2 - r_1)} \leq \gamma_K(n) \leq \frac{r_2}{\zeta^2(r_2 - r_1)} \\ 1 & \text{if} \quad \gamma_K(n) \geq \frac{r_2}{\zeta^2(r_2 - r_1)} \end{cases}$$

Proof. By an extension of Proposition 2 in Jacot et al. (2018), we know that the inverse of kernel matrix $K(\boldsymbol{X}, \boldsymbol{X})$ is well-defined, implying that we have perfect training accuracy:

$$f_K(\boldsymbol{X}) = \boldsymbol{y}$$

$$\begin{aligned} f_K\left(\mathcal{P}(\boldsymbol{x})\right) &= \frac{\tilde{r}}{r} f_K(\boldsymbol{x}) + \zeta^2 \left(1 - \frac{\tilde{r}}{r}\right) \gamma_K(n) \\ &= \frac{\tilde{r}}{r} y + \zeta^2 \left(1 - \frac{\tilde{r}}{r}\right) \gamma_K(n) \end{aligned}$$

Assume first that $y = 1$, implying $||\boldsymbol{x}||_2 = r_1$. Thus we need that

$$\begin{aligned} & \frac{r_2}{r_1} + \zeta^2 \left(1 - \frac{r_2}{r_1}\right) \gamma_K(n) < 0 \\ \iff & \gamma_K(n) > \frac{r_2}{\zeta^2(r_2 - r_1)} \end{aligned}$$

The case $y = -1$ is similarly obtained. Notice that the inequality is entirely independent of the specific form of $\boldsymbol{x}$. Thus this inequality will hold for all $\boldsymbol{x}$ with label $y = 1$ simultaneously. Since $r_1 < r_2$, the part of the adversarial data with label $y_{\text{adv}} = 1$ (or clean label $y = -1$) will be learnt first. This corresponds to a fraction of $1 - q$ of the entire training set, leading to $1 - q$ correctly classified adversarial examples. □

A.5. Proof of Lemma 4

Lemma. *Assume that K is of the above form and denote by C the corresponding homogeneous kernel. Then it holds that*

$$\gamma_K(n) = \frac{1}{1 + \beta^2 s\left(C(\boldsymbol{X}, \boldsymbol{X})^{-1}\right)} \gamma_C(n)$$

where we define $s(\boldsymbol{A}) = \sum_{i,j} A_{ij}$.

Proof. Using the Sherman–Morrison formula, we expand the inverse as follows:

$$\begin{aligned}\gamma_K(n) &= \mathbf{1}_n^T K(\boldsymbol{X},\boldsymbol{X})^{-1}\boldsymbol{y} = \mathbf{1}_n^T\left(C(\boldsymbol{X},\boldsymbol{X}) + \beta^2\mathbf{1}_n\mathbf{1}_n^T\right)^{-1}\boldsymbol{y} = \mathbf{1}_n^T\left(C(\boldsymbol{X},\boldsymbol{X})^{-1}+\right)\\ &= \mathbf{1}_n^T C(\boldsymbol{X},\boldsymbol{X})^{-1}\boldsymbol{y} - \mathbf{1}_n\left(\frac{\beta^2 C(\boldsymbol{X},\boldsymbol{X})^{-1}\mathbf{1}_n\mathbf{1}_n^T C(\boldsymbol{X},\boldsymbol{X})^{-1}}{1+\beta^2\mathbf{1}_n^T C(\boldsymbol{X},\boldsymbol{X})^{-1}\mathbf{1}_n}\right)\boldsymbol{y}\\ &= \gamma_C(n) - \frac{\beta^2 s\left(C(\boldsymbol{X},\boldsymbol{X})^{-1}\right)\gamma_C(n)}{1+\beta^2 s\left(C(\boldsymbol{X},\boldsymbol{X})^{-1}\right)}\\ &= \frac{\gamma_C(n)}{1+\beta^2 s\left(C(\boldsymbol{X},\boldsymbol{X})^{-1}\right)}\end{aligned}$$

□

A.6. Proof of Theorem 5

Theorem. *Consider the expected kernel $\tilde{K} = \mathbb{E}_{\boldsymbol{X}\sim p^n}\left[K(\boldsymbol{X},\boldsymbol{X})\right]$. We have that $\gamma_{\tilde{K}}$ is asymptotically given by*

$$\gamma_{\tilde{K}}(n) \propto \frac{C_1+\eta n}{C_2-\beta^2 C_3 n}$$

for constants $C_1, C_2, C_3, \eta \in \mathbb{R}$ and the limit is given by

$$\gamma_{\tilde{K}}(n) \xrightarrow{n\to\infty} \frac{r_1+r_2}{\beta^2(r_2-r_1)}$$

Proof. Define the quantities $\alpha = K(\boldsymbol{e}_1,\boldsymbol{e}_1)$ where $\boldsymbol{e}_1$ denotes the first unit vector. Notice that since K is a dot-product kernel, it holds for any $\boldsymbol{x}\in\boldsymbol{S}_1^{d-1}$ that $K(\boldsymbol{x},\boldsymbol{x}) = \alpha$. Moreover, define $\rho = \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim p_1}\left[K(\boldsymbol{x},\boldsymbol{x}')\right]$ and consider the expected kernel $\tilde{\boldsymbol{K}} = \mathbb{E}_{\boldsymbol{X}\sim p^n}[K(\boldsymbol{X},\boldsymbol{X})]$. Recall that we assume a balanced dataset, where the upperhalf of $\boldsymbol{X}$ is sampled according to p_{r_1} and the second half according to p_{r_2}. Denote the respective samples by $\boldsymbol{X}_+$ and $\boldsymbol{X}_-$ Let us first focus on the bias-free part $\gamma_C(n)$. The bias-free kernel is given by the following block structure:

$$\begin{aligned}C(\boldsymbol{X},\boldsymbol{X}) &= \begin{pmatrix} C(\boldsymbol{X}_+,\boldsymbol{X}_+) & C(\boldsymbol{X}_+,\boldsymbol{X}_-)\\ C(\boldsymbol{X}_-,\boldsymbol{X}_+) & C(\boldsymbol{X}_-,\boldsymbol{X}_-)\end{pmatrix} = \begin{pmatrix} r_1^2 C(\tilde{\boldsymbol{X}}_+,\tilde{\boldsymbol{X}}_+) & r_1 r_2 C(\tilde{\boldsymbol{X}}_+,\tilde{\boldsymbol{X}}_-)\\ r_1 r_2 C(\tilde{\boldsymbol{X}}_-,\tilde{\boldsymbol{X}}_+) & r_2^2 C(\tilde{\boldsymbol{X}}_-,\tilde{\boldsymbol{X}}_-)\end{pmatrix}\\ &:= \begin{pmatrix} r_1^2\boldsymbol{C}_{++} & r_1 r_2\boldsymbol{C}_{+-}\\ r_1 r_2\boldsymbol{C}_{-+} & r_2^2\boldsymbol{C}_{--}\end{pmatrix}\end{aligned}$$

where $\tilde{\boldsymbol{X}}$ denotes the projected data to the unit sphere $\boldsymbol{S}_1^{d-1}$. We will be interested in the blocks of the inverse $C(\boldsymbol{X},\boldsymbol{X})^{-1}$:

$$C(\boldsymbol{X},\boldsymbol{X})^{-1} = \begin{pmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{B} & \boldsymbol{D}\end{pmatrix}$$

Due to the symmetry, we observe that

$$\mathbf{1}_n^T C(\boldsymbol{X},\boldsymbol{X})^{-1}\boldsymbol{y} = s(\boldsymbol{A}) - s(\boldsymbol{D})$$

By the inverse formula for block matrices, we can analyse the first inverse block to obtain that

$$\begin{aligned}\boldsymbol{A} &= \frac{1}{r_1^2}\boldsymbol{C}_{++}^{-1} + \frac{1}{r_1^2}\boldsymbol{C}_{++}^{-1} r_1 r_2\boldsymbol{C}_{+-}\left(r_2^2\boldsymbol{C}_{--} - r_1^2 r_2^2\boldsymbol{C}_{-+}\frac{1}{r_1^2}\boldsymbol{C}_{++}^{-1}\boldsymbol{C}_{+-}\right)^{-1} r_1 r_2\boldsymbol{C}_{-+}\frac{1}{r_1^2}\boldsymbol{C}_{++}^{-1}\\ &= \frac{1}{r_1^2}\left(\boldsymbol{C}_{++}^{-1} + \boldsymbol{C}_{++}^{-1}\boldsymbol{C}_{+-}\left(\boldsymbol{C}_{--} - \boldsymbol{C}_{-+}\boldsymbol{C}_{++}^{-1}\boldsymbol{C}_{+-}\right)^{-1}\boldsymbol{C}_{-+}\boldsymbol{C}_{++}^{-1}\right)\end{aligned}$$

Next we analyze the right bottom block of the inverse:

$$\boldsymbol{D} = \left(r_2^2\boldsymbol{C}_{--} - r_1^2 r_2^2\boldsymbol{C}_{-+}\frac{1}{r_1^2}\boldsymbol{C}_{++}^{-1}\boldsymbol{C}_{+-}\right)^{-1} = \frac{1}{r_2^2}\left(\boldsymbol{C}_{--} - \boldsymbol{C}_{-+}\boldsymbol{C}_{++}^{-1}\boldsymbol{C}_{+-}\right)^{-1}$$

Similarly, we simplify the off-diagonal blocks to

$$\begin{aligned}\boldsymbol{B} &= -\frac{1}{r_2^2}\left(\boldsymbol{C}_{--} - \boldsymbol{C}_{-+}\boldsymbol{C}_{++}^{-1}\boldsymbol{C}_{+-}\right)^{-1} r_1 r_2 \boldsymbol{C}_{-+}\frac{1}{r_1^2}\boldsymbol{C}_{++}^{-1} \\ &= \frac{1}{r_1 r_2}\left(\boldsymbol{C}_{--} - \boldsymbol{C}_{-+}\boldsymbol{C}_{++}^{-1}\boldsymbol{C}_{+-}\right)^{-1}\boldsymbol{C}_{-+}\boldsymbol{C}_{++}^{-1}\end{aligned}$$

By introducing the inverse of the projected kernel matrix

$$C(\bar{\boldsymbol{X}}, \bar{\boldsymbol{X}})^{-1} = \begin{pmatrix}\bar{\boldsymbol{A}} & \bar{\boldsymbol{B}} \\ \bar{\boldsymbol{B}} & \bar{\boldsymbol{D}}\end{pmatrix}$$

we quickly realize again through the inverse block matrix formula, that

$$C(\boldsymbol{X}, \boldsymbol{X})^{-1} = \begin{pmatrix}\frac{1}{r_1^2}\bar{\boldsymbol{A}} & \frac{1}{r_1 r_2}\bar{\boldsymbol{B}} \\ \frac{1}{r_1 r_2}\bar{\boldsymbol{B}} & \frac{1}{r_2^2}\bar{\boldsymbol{D}}\end{pmatrix}$$

Thus we can see that

$$\mathbf{1}_n^T C(\boldsymbol{X}, \boldsymbol{X})^{-1}\boldsymbol{y} = \frac{1}{r_1^2}s(\bar{\boldsymbol{A}}) - \frac{1}{r_2^2}s(\bar{\boldsymbol{D}})$$

Let us now consider the expected kernel. Since both $\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{D}}$ are the Gram matrix of the same kernel on the unit sphere, their expected kernel agrees and thus we only need to calculate $s(\tilde{\boldsymbol{D}})$. We define

$$\boldsymbol{G} = (\alpha - \rho)\mathbb{1}_{m\times m} + \rho\mathbf{1}_m\mathbf{1}_m^T$$

as well as the expected off-diagonal part

$$\boldsymbol{H} = \rho\mathbf{1}_m\mathbf{1}_m^T$$

Consider the matrix

$$\tilde{\boldsymbol{D}}^{-1} = \boldsymbol{C}_{--} - \boldsymbol{C}_{-+}\boldsymbol{C}_{++}^{-1}\boldsymbol{C}_{+-}$$

In expectation, this reduces to

$$\boldsymbol{S} := \mathbb{E}\left[\tilde{\boldsymbol{D}}^{-1}\right] = \boldsymbol{G} - \boldsymbol{H}\boldsymbol{G}^{-1}\boldsymbol{H}$$

Again, making use of Sherman-Morrison, we can find a closed form expression for $\boldsymbol{G}^{-1}$:

$$\boldsymbol{G}^{-1} = \left((\alpha - \rho)\mathbb{1}_{m\times m} + \rho\mathbf{1}_m\mathbf{1}_m^T\right)^{-1} = \frac{1}{\alpha - \rho}\mathbb{1}_{m\times m} - \frac{\rho}{(\alpha - \rho)\left(\alpha + \rho(m-1)\right)}\mathbf{1}_m\mathbf{1}_m^T$$

and thus we can simplify

$$\begin{aligned}\boldsymbol{H}\boldsymbol{G}^{-1}\boldsymbol{H} &= \frac{\rho^2 m}{\alpha - \rho}\mathbf{1}_m\mathbf{1}_m^T - \frac{m^2\rho^3}{(\alpha - \rho)\left(\alpha + \rho(m-1)\right)}\mathbf{1}_m\mathbf{1}_m^T = \frac{(\alpha + \rho(m-1))\rho^2 m - m^2\rho^3}{(\alpha - \rho)\left(\alpha + \rho(m-1)\right)}\mathbf{1}_m\mathbf{1}_m^T \\ &= \frac{m\rho^2}{\alpha + \rho(m-1)}\mathbf{1}_m\mathbf{1}_m^T\end{aligned}$$

We can now show that $\mathbf{1}_m$ is an eigenvector of $\boldsymbol{S}$:

$$\begin{aligned}\mathbb{E}\left[\tilde{\boldsymbol{D}}^{-1}\right]\mathbf{1}_m &= (\alpha - \rho)\mathbf{1}_m + m\rho\mathbf{1}_m - \frac{m^2\rho^2}{\alpha + \rho(m-1)}\mathbf{1}_m \\ &= \frac{\left(\alpha + \rho(m-1)\right)^2 - m^2\rho^2}{\left(\alpha + \rho(m-1)\right)}\mathbf{1}_m \\ &= \frac{(\alpha - \rho)^2 + 2m\rho(\alpha - \rho)}{\left(\alpha + \rho(m-1)\right)}\mathbf{1}_m\end{aligned}$$

Now we know that $\mathbf{1}_m$ is also an eigenvector of $\boldsymbol{S}^{-1}$ with inverse eigenvalue and thus

$$s\left(\boldsymbol{S}^{-1}\right) = \mathbf{1}_m^T \boldsymbol{S}^{-1} \mathbf{1}_m = \frac{(\alpha + \rho(m-1))}{(\alpha-\rho)^2 + 2m\rho(\alpha-\rho)} \mathbf{1}_m^T \mathbf{1}_m = \frac{\rho m^2 + m(\alpha-\rho)}{(\alpha-\rho)^2 + 2m\rho(\alpha-\rho)}$$

Using the symmetry, we thus proved that

$$\gamma_{\tilde{\boldsymbol{C}}}(m) = \frac{r_2^2 - r_1^2}{r_1^2 r_2^2} \frac{\rho m^2 + m(\alpha-\rho)}{(\alpha-\rho)^2 + 2m\rho(\alpha-\rho)} \propto \frac{r_2^2 - r_1^2}{r_1^2 r_2^2} \frac{\rho m^2}{(\alpha-\rho)^2 + 2m\rho(\alpha-\rho)}$$

To finish the proof, we need to finally calculate $s\left(\tilde{\boldsymbol{C}}^{-1}\right)$. Luckily, since we already calculated the sum of the block diagonal, we only need to find the sum of the off-diagonal blocks, in expectation:

$$\boldsymbol{W} = \boldsymbol{S}^{-1} \boldsymbol{H} \boldsymbol{G}^{-1}$$

Again we find the that $\mathbf{1}_m$ is an eigenvector:

$$\begin{aligned}
\boldsymbol{W}\mathbf{1}_m &= -\boldsymbol{S}^{-1}\boldsymbol{H}\frac{1}{\alpha+\rho(m-1)}\mathbf{1}_m = -\boldsymbol{S}^{-1}\frac{m\rho}{\alpha+\rho(m-1)}\mathbf{1}_m = -\frac{m\rho}{\alpha+\rho(m-1)}\frac{(\alpha+\rho(m-1))}{(\alpha-\rho)^2+2m\rho(\alpha-\rho)}\mathbf{1}_m \\
&= -\frac{m\rho}{(\alpha-\rho)^2+2m\rho(\alpha-\rho)}\mathbf{1}_m
\end{aligned}$$

where we used that $\mathbf{1}_m$ is also an eigenvector of $\boldsymbol{G}^{-1}$, as seen above. Thus the sum of the off-diagonal term is

$$s(\boldsymbol{W}) = -\frac{m^2\rho}{(\alpha-\rho)^2 + 2m\rho(\alpha-\rho)}$$

Thus we can finally obtain that the sum of the inverse expected kernel $\tilde{\boldsymbol{C}}^{-1} = C(\boldsymbol{X}, \boldsymbol{X})^{-1}$ is given by

$$\begin{aligned}
s\left(\tilde{\boldsymbol{C}}^{-1}\right) &= \left(\frac{1}{r_1} + \frac{1}{r_2}\right) s(\tilde{\boldsymbol{D}}) + \frac{2}{r_1 r_2} s(\boldsymbol{W}) = \frac{1}{(\alpha-\rho)^2 + 2m\rho(\alpha-\rho)} \left(m^2\rho \frac{(r_1-r_2)^2}{r_1^2 r_2^r} + \frac{r_1^2 + r_2^2}{r_1^2 r_2^2} m(\alpha-\rho) \right) \\
&\propto \frac{m^2\rho \frac{(r_1-r_2)^2}{r_1^2 r_2^r}}{(\alpha-\rho)^2 + 2m\rho(\alpha-\rho)}
\end{aligned}$$

Now, we can put all the pieces together to obtain

$$\begin{aligned}
\gamma_{\tilde{\boldsymbol{K}}}(m) &= \frac{\gamma_{\tilde{\boldsymbol{C}}}(m)}{1 + \beta^2 s\left(\tilde{\boldsymbol{C}}^{-1}\right)} = \frac{\frac{r_2^2 - r_1^2}{r_1^2 r_2^2}\left(\rho m^2 + m(\alpha-\rho)\right)}{(\alpha-\rho)^2 + 2m\rho(\alpha-\rho) + \beta^2 m^2 \rho \frac{(r_1-r_2)^2}{r_1^2 r_2^r} + \beta^2 \frac{r_1^2 + r_2^2}{r_1^2 r_2^2} m(\alpha-\rho)} \\
&\propto \frac{m\rho \frac{r_2^2 - r_1^2}{r_1^2 r_2^2} + (\alpha-\rho)\frac{r_2^2 - r_1^2}{r_1^2 r_2^2}}{m\beta^2\rho \frac{(r_1-r_2)^2}{r_1^2 r_2^r} + \beta^2 \frac{r_1^2 + r_2^2}{r_1^2 r_2^2}(\alpha-\rho)} \\
&= \frac{C_1 + \eta m}{C_2 - \beta^2 C_3 m}
\end{aligned}$$

Moreover, we can easily derive the limit

$$\gamma_{\tilde{\boldsymbol{K}}}(m) \xrightarrow{m\to\infty} \frac{r_1 + r_2}{\beta^2(r_2 - r_1)}$$

□

B. Additional Numerical Experiments

B.1. Adversarial Accuracy

Here we present more empirical evidence backing the theoretical findings in Theorem 3 for more kernels and architectures. As demonstrated in 8, the phase transitions in the adversarial accuracy hold across different underlying architectures. We

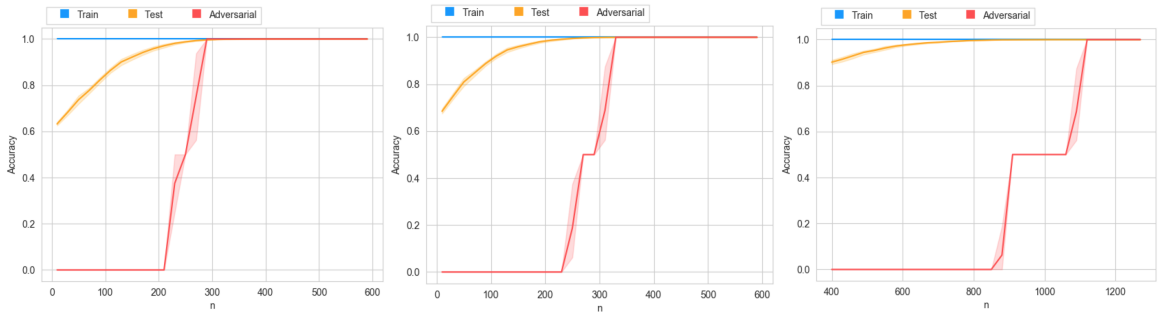

Figure 8. Train, test and adversarial accuracy for 5 layer NNGP (left), 9 layer NNGP (middle) and 5 layer NTK (right), plotted against sample size.

observe that the NTK in general is suffering more from the adversarial effect, compared to the NNGP. This is also visible in Figure 9 where we see that $\gamma_K(n)$ grows more slowly for the NTK, compared to the NNGP, making it hence also more slowly approach the phase transitions.

B.2. Behaviour of γ_K

We study the behavour of γ_K and the corresponding expected version $\gamma_{\tilde{\boldsymbol{K}}}$ for different architectures in Figure 9 We see that

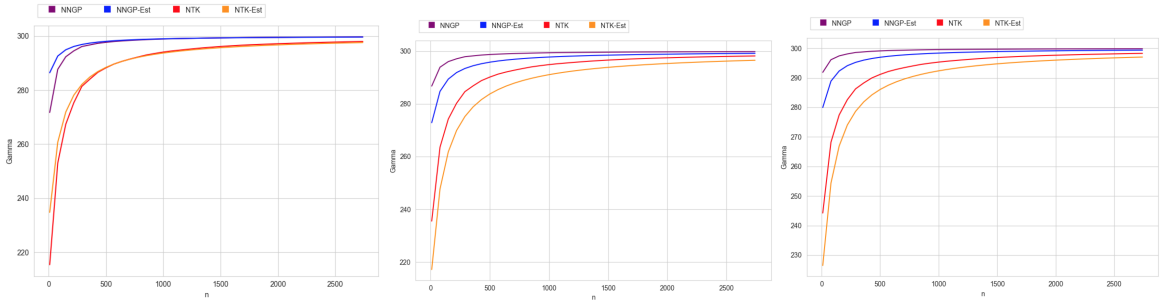

Figure 9. We plot γ_K and $\gamma_{\tilde{\boldsymbol{K}}}$ for a 3 layer NTK and NNGP (left), a 5 layer NTK and NNGP (middle) and a 7 layer NTK and NNGP (right)

the expected kernel induces a very good approximation $\gamma_{\tilde{\boldsymbol{K}}}$, especially for large sample sizes n. Moreover, the qualitative behaviour is also very well captured for smaller sample sizes. Transforming insights from $\gamma_{\tilde{\boldsymbol{K}}}$ to γ_K is thus sensible, especially for large sample sizes. Moreover, as anticipated in Theorem 5, all the kernels are converging to the same maximal capacity

$$\gamma_{\tilde{\boldsymbol{K}}}(m) \xrightarrow{m\to\infty} \frac{r_1 + r_2}{\beta^2(r_2 - r_1)}$$

B.3. Eigendecompositions

Here we study different decompositions, not just consisting of the dominant eigenfunction. In Figure 10 we verify that the dominant eigenfunction indeed captures all the signal in the data, leaving the ensemble of eigenfunctions consisting of all but the dominant one with no predictive power at all in terms of any accuracy. We then proceed to study if using the top 10 dominant eigenfunction brings any improvement in terms of the adversarial accuracy. Again this is not the case as visible in Figure 10. We tested more ensembles of eigenfunctions but none can improve over random guessing on the adversarial

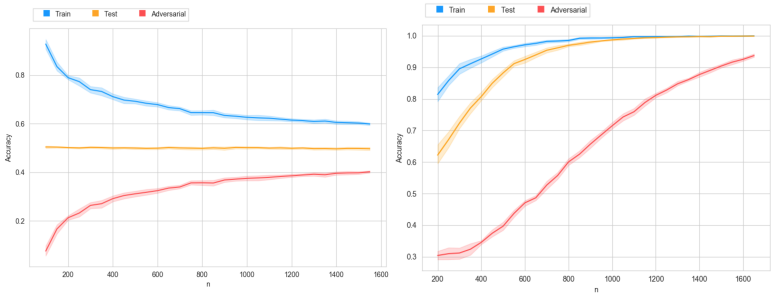

Figure 10. Train, test and adversarial accuracy for the ensemble of eigenfunctions consisting of all but the dominant one (left) and for the ensemble of the 10 most dominant eigenfunctions

dataset for small sample sizes. This renders any uniform convergence-based generalization bound still meaningless as it is lower-bounded by 0.5, which corresponds to random guessing for a binary task.