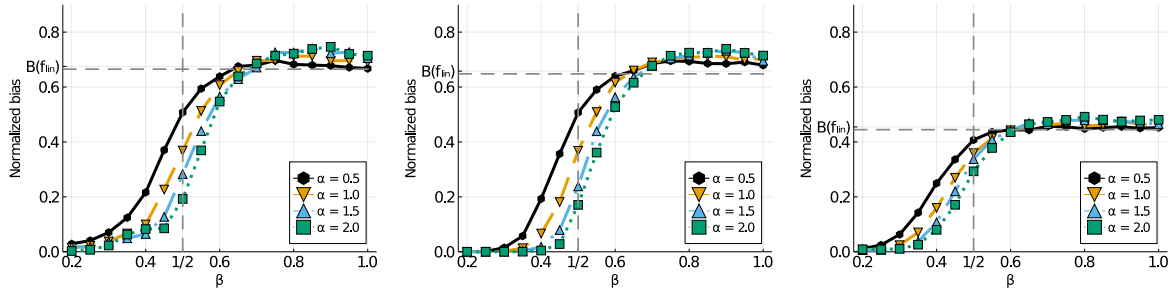
A. Experiments

This section contains additional experiments not shown in the main text. [4]

A.1. Polynomial Barrier

In this section, we provide additional experiments that discuss Theorem 3.1. In particular, we investigate kernels beyond the Laplace kernel and study the behaviour of the bias with respect to β when d is fixed and n varies. The experimental setting is the same as the one in Section 4.2.



(a) $\mathbb{P}_1$ (b) $\mathbb{P}_2$ (c) $\mathbb{P}_3$

Figure 5: The bias of the minimum norm interpolant $\mathbf{B}(\hat{f}_0)$ normalized by $\mathbf{B}(0)$ as a function of β for the α-exponential kernel with different choices of α and with $n = 4000$ i.i.d. samples drawn from (a) $\mathbb{P}_1$, (b) $\mathbb{P}_2$ and (c) $\mathbb{P}_3$.

Instead of comparing the bias curves for different input distribution as in Figure 2a, Figure 5 shows the bias with respect to β for the α-exponential kernel, i.e. $k(x, x') = \exp(-\|x - x'\|_2^\alpha)$, for different choices of α and hence, for kernels with distinct eigenvalue decays ($\alpha = 2$ results in an exponential eigenvalue decay while $\alpha < 2$ in a polynomial eigenvalue decay). Clearly, we can see that the curves transition at a similar value for β, which confirms the the discussion of Theorem 3.1 in Section 3.3 where we argue that the polynomial approximation barrier occurs independently of the eigenvalue decay.
Figure 6 shows the bias of the minimum norm interpolant $\mathbf{B}(\hat{f}_0)$ normalized by $\mathbf{B}(0)$ for the ground truth function $f^\star(x) = 2x_{(1)}^3$ and the Laplace kernel as in Section 4.2 with $\tau = d_{\text{eff}}$. We observe that the asymptotics already kick in for $d \approx 40$ since all curves for $d \geq 40$ resemble each other. This confirms the the trend in Figure 2b.

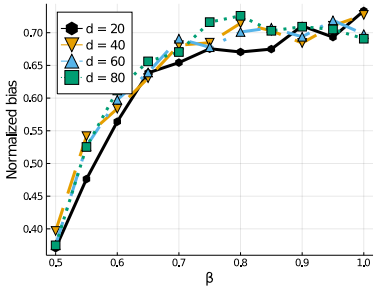
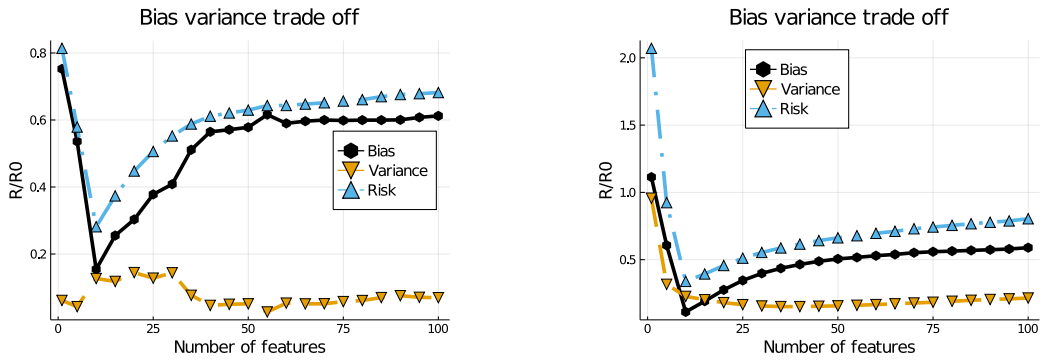


Figure 6: The bias of the minimum norm interpolant $\mathbf{B}(\hat{f}_0)$ normalized by $\mathbf{R}(0)$ as a function of β for different choices of d and samples generated from an isotropic Gaussian (as in Model $\mathbb{P}_1$) with $n = \lfloor d^{1/\beta} \rfloor$.

A.2. Feature selection - Synthetic

The goal of this experiment is to compare the bias variance trade-off of ridge regression and minimum norm interpolation. We use the same experimental setting as the ones used for Figure 4a (see Section 4.3). We set the bandwidth to $\tau = d_{\text{eff}}$ and choose the ridge parameter λ using 5-fold cross validation. While for small dimensions d, ridge regularization is crucial to achieve good performance, the bias becomes dominant as the dimension grows and the difference of the risks of both methods shrinks. This aligns well with Theorem 3.1, which predicts that the bias starts to increase with d for fixed n once we enter the asymptotic regime.

[4] Our code is publicly available at https://www.github.com/DonhauserK/High-dim-kernel-paper/



(a) Ridge regression

(b) Minimum norm interpolation

Figure 7: The bias-variance trade-off of the (a) ridge estimate and (b) minimum norm interpolant normalized by $\mathbf{B}(0)$ as a function of selected features for the synthetic experiment described in Section 4.3. Figure (b) is exactly the same as Figure 4a.

A.3. Feature selection - Real world

We now present details for our real world experiments to emphasize the relevance of feature selection when using kernel regression for practical applications, as discussed in Section 4.3.
We consider the following data sets:

1. The *residential housing* regression data set from the UCI website (Dua and Graff, 2017) where we predict the sales prices and construction costs given a variety of features including the floor area of the building or the duration of construction.

2. The *ALLAML* classification data set from the ASU feature selection website (Li et al., 2018) where we classify patients with acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) based on features gained from gene expression monitoring via DNA microarrays.

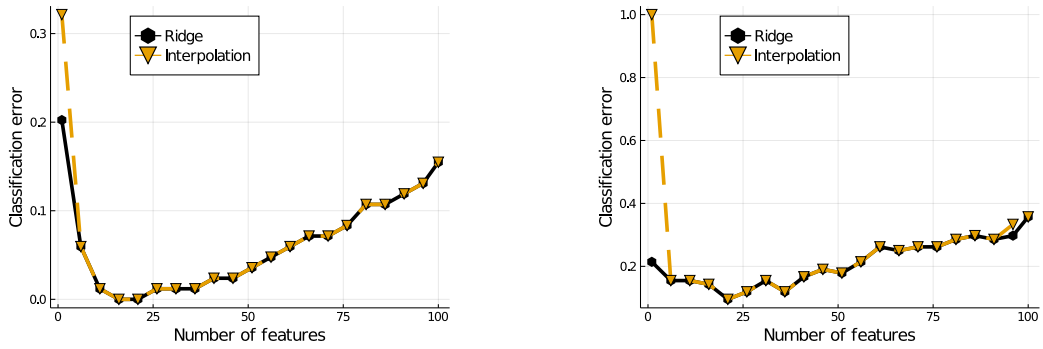
3. The *CLL_SUB_111* classification dataset from the ASU feature selection web-page (Li et al., 2018) where we classify genetically and clinically distinct subgroups of B-cell chronic lymphocytic leukemia (B-CLL) based on features consisting of gene expressions from high density oligonucleotide arrays. While the original dataset contains three different classes, we only use the classes 2 and 3 for our experiment to obtain a binary classification problem.

Because the number of features in the *ALLAML* and *CLL_SUB_111* datasets massively exceed the number of samples, we run the feature selection algorithm in (Chen et al., 2017) and pre-select the best 100 features chosen by the algorithm. In order to reduce the computational expenses, we run the algorithm in batches of 2000 features and iteratively remove all features except for the best 200 features chosen by the algorithm. We do this until we reduce the total number of features to 2000 and then select, in a last round, the final 100 features used for the further procedure. Reducing the amount of features to 100 is important for the computational feasibility of greedy forward features selection in our experiments. The properties of the datasets are summarized in Table 2.

Data set	*Binary CLL_SUB_111*	*ALLAML*	Residential Building Data Set
Features	11,340 (100)	7129 (100)	107
Samples	100	72	372
Type	Binary classification	Binary classification	Regression

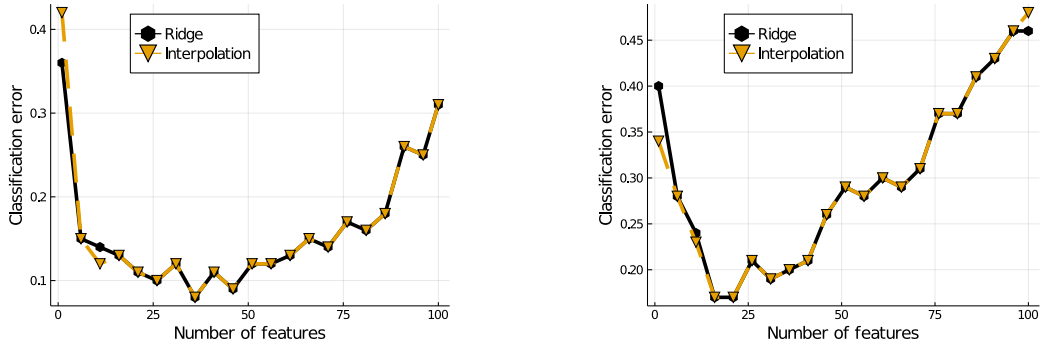
Table 2: Real world datasets used for the experiments. The value in the brackets shows the number of features after a pre-selection using the algorithm presented in (Chen et al., 2017).

Experimental setting: As a first step, we normalise both the vectors containing the single input features and the observations separately using ℓ_1 normalization. We use the Laplace kernel for computing the ridge and ridgeless estimate in all experiments. For each setting, we pick the bandwidth τ and the penalty coefficient λ (for the ridge estimator) using cross validation. We increase the number of features by greedily adding the feature that results in the lowest 5-fold cross validation risk. In addition, in order to study the effect of noise, we generate additional data sets where we add synthetic i.i.d. noise drawn from the uniform distribution on $[-1/2, 1/2]$ to the observations for the regression tasks and flip 20% of the label for



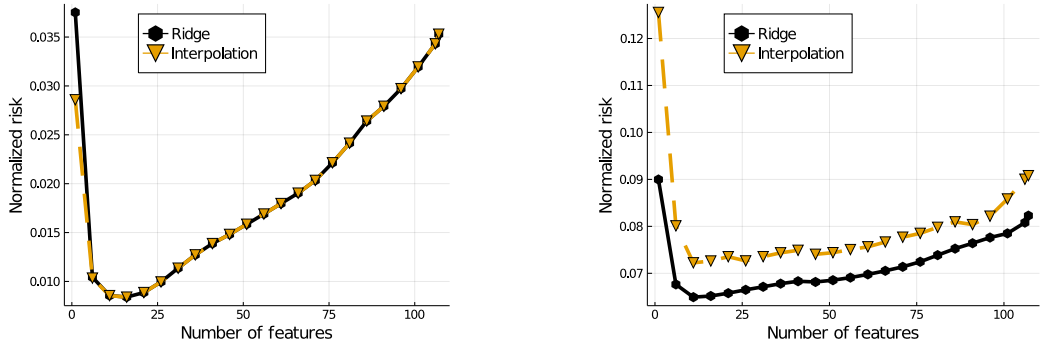
(a) *ALLAML* - original (b) *ALLAML* - 20% label noise

Figure 8: The classification error of the minimum norm interpolator and ridge estimator for the *ALLAML* dataset.



(a) *CLL_SUB_111* - original (b) *CLL_SUB_111* - 20% label noise

Figure 9: The classification error of the minimum norm interpolator and ridge estimator for the *CLL_SUB_111* dataset.



(a) Residential housing - original (b) Residential housing - add. noise

Figure 10: The risk $\mathbf{R}(\hat{f}_0)$ of the minimum norm interpolant respectively $\mathbf{R}(\hat{f}_\lambda)$ of the ridge estimate normalized by $\mathbf{R}(0)$ for the *residential housing* dataset with target construction costs.

the classification tasks.

Results of the experiments: The following figures present the results of our experiments on all datasets except for the ones predicting the sales prices in the *residential housing* dataset, which we presented in Figures 4b, 4c in the main text. Similar to the observations made in Section 4.3, Figures 8,9,10 show that the risk reaches its minimum around $d \approx 25$, with significant differences to the right at $d \approx 100$. In particular, this holds for both ridge regression and interpolation, which again shows that the bias becomes dominant as the dimension increases. Surprisingly, we also note that the relevance of ridge regularization seems to be much smaller for classification tasks than regression tasks.

B. Bounded Hilbert norm assumption

This section gives a formal statement of Lemma 2.1. We begin with the conditions under which the lemma holds. We consider tensor product kernels of the form

$$k(x, x') = \prod_{j=1}^{d} q(x_{(j)}, x'_{(j)})$$

with inputs $x, x' \in \mathcal{X}^{\otimes d} \subset \mathbb{R}^d$ with $\mathcal{X}$ compact and $\otimes d$ denotes the product space, for some kernel function q on $\mathcal{X}$ which may change with d (e.g. the scaling). In order to prevent the sequence of kernels k to diverge as $d \to \infty$, assume that there exists some probability measure on $\mathcal{X}$ with full support such that the trace of the kernel operator is bounded by 1, i.e. $\int q(x, x) d\mu(x) \leq 1$. Let $\|.\|_{\mathcal{H}_k}$ be the Hilbert norm induced by k. Then,

Lemma B.1 (Formal statement of Lemma 2.1)**.** *Let k satisfy the above conditions. Then, for any f that is a non-constant sparsely parameterized product function $f(x) = \prod_{j=1}^{m} f_j(x_{(j)})$ for some fixed $m \in \mathbb{N}$,*

$$\|f\|_{\mathcal{H}_k} \overset{d\to\infty}{\to} \infty.$$

Proof. For any $j > m$, define $f_j = 1$. First, we note that the proof follows trivially if any of the f_j is not contained in the RKHS induced by q since this implies that the Hilbert norm $\|f\|_{\mathcal{H}_k} = \infty$. Hence, we can assume that for all j, f_j is contained in the RKHS for all d. Furthermore, because k is a product kernel, we can write $\|f\|_{\mathcal{H}_k} = \prod_{j=1}^{d} \|f_j\|_{\mathcal{H}_q}$ where $\|.\|_{\mathcal{H}_q}$ is the Hilbert norm induced by q on $\mathcal{X}$. Because we are only interested to see whether the sequence of Hilbert norms diverge, without loss of generality we can assume that $m = 1$, and hence,

$$\|f\|_{\mathcal{H}_k} = \|f_1\|_{\mathcal{H}_q} (\|1\|_{\mathcal{H}_q})^{d-1}. \tag{9}$$

Next, by Mercer's theorem there exists an orthonormal eigenbasis $\{\phi_i\}_{i=1}^{\infty}$ in $\mathcal{L}_2(\mathcal{X}, \mu)$ with corresponding eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$ such that for any $g \in \mathcal{H}_q$, $\|g\|_{\mathcal{H}_q} = \sum_{i=1}^{\infty} \frac{(\langle f, \phi_i \rangle)^2}{\lambda_i}$, where $\langle f, \phi_i \rangle = \int \phi_i(x) f(x) d\mu(x)$. Note that because the kernel q depends on d, λ_i and ϕ_i also depend on d. Next, because by assumption $f(x) = 1$ is contained in the RKHS, there exists α_i such that for every $x \in \mathcal{X}$, $1 = \sum_{i=1}^{\infty} \alpha_i \phi_i(x)$ and $\sum_{i=1}^{\infty} \alpha_i^2 = 1$. Furthermore,

$$1 \geq \int q(x, x) d\mu(x) = \int \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(x) d\mu(x) = \sum_{i=1}^{\infty} \lambda_i.$$

Combining these results, we get that

$$\|1\|_{\mathcal{H}_q} = \sum_{i=1}^{\infty} \frac{\alpha_i^2}{\lambda_i} \geq 1.$$

Furthermore, there exists β_i such that $f_1(x) = \sum_{i=1}^{\infty} \beta_i \phi_i(x)$. Again, because we are only interested to see whether the sequence of Hilbert norms diverge, without loss of generality we can assume that $\sum_{i=1}^{\infty} \beta_i^2 = 1$ and hence also $\|f_j\|_{\mathcal{H}_q} \geq 1$. First, assume that there exists a subsequence such that $\|1\|_{\mathcal{H}_q} \to 1$. This implies that there exists a sequence $j_d \in \mathbb{N}$ such that $\alpha_{j_d}^2 \to 1$ and $\lambda_{j_d} \to 1$. Next, because by assumption $f_1 \neq 1$, there exists some constant $c_1 > 0$ such that for all d,

$$c_1 \leq \int (1 - f_1(x))^2 d\mu(x) = \sum_{i=1}^{\infty} (\alpha_i - \beta_i)^2.$$

Together with the fact that $\alpha_{j_d}^2 \to 1$ it then follows that $\sum_{i \neq j_d} \beta_i^2$ has to be asymptotically lower bounded by some positive non-zero constant c_2 and hence

$$\|f_1\|_{\mathcal{H}_q} \geq \frac{c_2}{(1 - \lambda_{j_d})} \to \infty.$$

This contradicts the assumption that $\|f_1\|_{\mathcal{H}_q}$ is upper bounded by some constant for every d. Hence, we are only left with the case where $\|1\|_{\mathcal{H}_q} \geq c > 1$, however, this case diverges due to Equation 9. Hence, the proof is complete. □

C. Proof of Theorem 3.1

Before presenting the proof of the (generalized) theorem, we first state the key concentration inequalities used throughout the proof. It is an extension of Lemma A.2 in the paper (El Karoui et al., 2010), which iteself is a consequence of the concentration of Lipschitz continuous functions of i.i.d random vectors.

Lemma C.1. *For any $\mathbb{P}_X \in \mathcal{Q}$ or $\mathcal{Q}_{\mathcal{S}^{d-1}}$, let $\boldsymbol{X} \in \mathbb{R}^{d\times n}$ consists of i.i.d. vectors $x_i \sim \mathbb{P}_X$ and $X \sim \mathbb{P}_X$ be independent of x_i. For any constants $\epsilon > 0$, define the events*

$$\mathcal{E}_{\boldsymbol{X}} := \left\{ \boldsymbol{X} \,|\, \max_{i,j} \left| x_i^\top x_j / \operatorname{tr}(\Sigma_d) - \delta_{i,j} \right| \leq n^{-\beta/2} (\log(n))^{(1+\epsilon)/2} \right\} \tag{10}$$

$$\mathcal{E}_{X|\boldsymbol{X}} := \left\{ X \mid \left| \, \|X\|_2^2 / \operatorname{tr}(\Sigma_d) - 1 \right| \leq n^{-\beta/2} (\log n)^{(1+\epsilon)/2} \text{ and } \max_i |x_i^\top X| / \operatorname{tr}(\Sigma_d) \leq n^{-\beta/2} (\log n)^{(1+\epsilon)/2} \right\} \tag{11}$$

Then, there exists some constant $C > 0$ such that for n sufficiently large,

$$\mathbb{P}(\mathcal{E}_{\boldsymbol{X}}) \geq 1 - n^2 \exp(-C(\log(n))^{(1+\epsilon)}) \tag{12}$$

$$\text{and} \quad \mathbb{P}(\mathcal{E}_{X|\boldsymbol{X}} | \mathcal{E}_{\boldsymbol{X}}) \geq 1 - (n+1)^2 \exp(-C(\log(n))^{(1+\epsilon)}) \tag{13}$$

In particular, the event $\mathcal{E}_{\boldsymbol{X}}$ holds almost surely with respect to the sequence of data sets $\boldsymbol{X}$ as $n \to \infty$, that is the probability that for infinitely many n, $\mathcal{E}_{\boldsymbol{X}}$ does not hold, is zero.

The proof of the lemma can be found in Section E.1.

Proof of Theorem 3.1. The proof of the theorem is primarily separated into two parts

- We first state Theorem C.2 which shows under the weaker Assumption C.1 that the results of 3.1 hold for the ridge estimate $\hat{f}_\lambda$ for non-vanishing $\lambda > 0$ or the ridgeless estimate whenever the eigenvalues of K are asymptotically lower bounded.
- We finish the proof for the ridgeless estimate by invoking Theorem C.2 and showing that K indeed has asymptotically lower bounded eigenvalues under the stricter assumptions A.1-A.3 imposed in Theorem 3.1.

For the clarity we denote with A.3 the β-dependent assumptions in Theorem 3.1

(A.3) *β-dependent assumptions:* g_i is $(\lfloor 2/\beta \rfloor + 1 - i)$-times continuously differentiable in a neighborhood of $(1,1)$ and there exists $j' > \lfloor 2/\beta \rfloor$ such that $g_{j'}(1,1) > 0$.

We start by introducing the following weaker assumptions that allows us to jointly treat α-exponential kernels and kernels satisfying Assumption A.1-A.3 when the kernel eigenvalues are lower bounded in Theorem C.2. Note that this assumption implies that the kernel is rotationally invariant.

(C.1) *Relaxation of Assumption A.1-A.3*: Define the neighborhood $N(\delta, \delta') \subset \mathbb{R}^d \times \mathbb{R}^d$ as

$$N(\delta, \delta') := \{(x, x') \in \mathbb{R}^d \times \mathbb{R}^d \mid (\|x\|_2^2, \|x'\|_2^2) \in [1-\delta, 1+\delta] \times [1-\delta, 1+\delta], x^\top x' \in [-\delta', \delta']\}.$$

The kernel function k is rotationally invariant and there exists a function g such that $k(x.x') = g(\|x\|_2^2, \|x'\|_2^2, x^\top x')$. Furthermore, g can be expanded as a power series of the form

$$k(x, x') = g(\|x\|_2^2, \|x'\|_2^2, x^\top x') = \sum_{j=0}^{m} g_j(\|x\|_2^2, \|x'\|_2^2)(x^\top x')^j + (x^\top x')^{m+1} r(\|x\|_2^2, \|x'\|_2^2, x^\top x') \tag{14}$$

with $m = \lfloor 2/\beta \rfloor$ that converges in a neighborhood $N(\delta, \delta')$ of the sphere for some $\delta, \delta' > 0$ and where g_i is $(\lfloor 2/\beta \rfloor + 1 - i)$-times continuously differentiable in an neighborhood of $(1,1)$ and the remainder term r is a continuous function in a neighborhood of $(1,1,0)$.

Theorem C.2 (Polynomial approximation barrier)**.** *Assume that the kernel k, respectively its restriction onto the unit sphere, satisfies Assumption C.1 and that the eigenvalues of $K + \lambda I_n$ are almost surely lower bounded by a positive constant with respect to the sequence of datasets $\boldsymbol{X}$ as $n \to \infty$. Furthermore, assume that the ground truth $f^\star$ is bounded and the input distribution satisfies B.1-B.2. Then, for $m = 2\lfloor 2/\beta \rfloor$ for $\mathbb{P}_X \in \mathcal{Q}$ and $m = \lfloor 2/\beta \rfloor$ for $\mathbb{P}_X \in \mathcal{Q}_{\mathcal{S}^{d-1}}$, the following results hold for both the ridge* (1) *and ridgeless estimator* (2) *$\hat{f}_\lambda$ wiht $\lambda \geq 0$.*

1. *The bias of the kernel estimators $\hat{f}_\lambda$ is asymptotically lower bounded, for any $\epsilon > 0$,*

$$\boldsymbol{B}(\hat{f}_\lambda) \geq \inf_{p \in \mathcal{P}_{\leq m}} \|f^\star - p\|_{\mathcal{L}_2(\mathbb{P}_X)} - \epsilon \;\; a.s. \text{ as } n \to \infty. \tag{15}$$

2. *We can find a polynomial p such that for any $\epsilon' > 0$, there exists $C > 0$, such that for any $\epsilon > 0$ asymptotically with probability $\geq 1 - n^2 \exp(-C(\log(n))^{1+\epsilon'})$ over the draws of X,*

$$\left| \mathbb{E}_Y \hat{f}_\lambda(X) - p(X) \right| \leq \epsilon \;\; a.s. \text{ as } n \to \infty. \tag{16}$$

Furthermore, for bounded kernel functions on the support of $\mathbb{P}_X$ the averaged estimator $\mathbb{E}_Y \hat{f}_\lambda$ converges in $\mathcal{L}_2(\mathbb{P}_X)$ to a polynomial $p \in \mathcal{P}_{\leq m}$,

$$\left\| \mathbb{E}_Y \hat{f}_\lambda - p \right\|_{\mathcal{L}_2(\mathbb{P}_X)} \to 0 \;\; a.s. \text{ as } n \to \infty. \tag{17}$$

The proof of this theorem can be found in Section C.1. Theorem C.2 states Theorem 3.1 under the assumption that $(K + \lambda I)$ has asymptotically lower bounded eigenvalues and the weaker Assumption C.1. For the proof of Theorem 3.1, it remains to show that Assumptions A.1-A.3 of Theorem 3.1 and the α-exponential kernel both

(a) satisfy Assumption C.1 and

(b) induce kernel matrices with almost surely asymptotically positive lower bounded eigenvalues

Point (a) is relatively simple to prove and deferred to Section E.6. The bulk of the work in fact lies in showing (b) separately for the case for A.1-A.3 and α-exponential kernels with $\alpha \in (0, 2)$ in the following two propositions, as these two cases require two different proof techniques.

Proposition C.3. *Assume that the kernel k, respectively its restriction onto the unit sphere, satisfies Assumption A.1-A.3 and the distribution $\mathbb{P}_X$ satisfies B.1-B.2. Then, for any $\gamma > 0$ and $m = \lfloor 2/\beta \rfloor$, conditioned on $\mathcal{E}_{\mathbf{X}}$,*

$$\lambda_{\min}(K) \geq g(1,1,1) - \sum_{i=0}^{m} g_i(1,1) - \gamma > 0 \tag{18}$$

where $\lambda_{\min}(K)$ is the minimum eigenvalue of the kernel matrix K.

Proposition C.4. *Assume that the Assumptions B.1-B.2 hold true. Then, the minimum eigenvalue of the kernel matrix of the α-exponential kernel with $\alpha \in (0, 2)$ is lower bounded by some positive constant almost surely as $n \to \infty$.*

The proof of the Propositions C.3 and C.4 can be found in the Sections C.2.1 and C.2.2 respectively which concludes the proof of the theorem.

Remark C.5. *The almost sure statement in Proposition C.4 can also be replaced with an in probability statement as in Lemma C.1. Hence the statements in Theorem 3.1 can also be replaced with an in probability statement.*

□

C.1. Proof of Theorem C.2

As a result of Lemma C.1 it is sufficient to condition throughout the rest of this proof on the intersection of the events $\mathcal{E}_{\mathbf{X}}$ and the event where the eigenvalues of the kernel matrix K are lower bounded by a positive constant.
For simplicity of notation, we define $z_i = \frac{x_i}{\sqrt{\tau}}$ and let $\mathbf{Z}$ be the $d \times n$ matrix with column vectors z_i. Define the random variable $Z = X/\sqrt{\tau}$ with $X \sim \mathbb{P}_X$ and denote with $\mathbb{P}_Z$ the probability distribution of Z. Define the event $\mathcal{E}_{Z|\mathbf{Z}}$ in the same way as $\mathcal{E}_{X|\mathbf{X}}$ for the normalised inputs z_i, Z and $\mathcal{E}_{\mathbf{Z}}$ like $\mathcal{E}_{\mathbf{X}}$. In the latter, we denote with $a \lesssim b$ that there exists a constant $C > 0$ such that $a \leq Cb$ with C independent of n, d. Furthermore, we make heavily use of the closed form solution for the estimator $\hat{f}_\lambda$,

$$\mathbb{E}_Y \hat{f}_\lambda(X) = f^\star(\mathbf{X})^\top (K + \lambda I_n)^{-1} k_Z$$

with $k_Z \in \mathbb{R}^n$ the vector with entries $(k_Z)_i = k_\tau(x_i, X) = k(z_i, Z)$ and $f^\star(\mathbf{X})$ the vector with entries $f^\star(\mathbf{X})_i = f^\star(x_i)$. This equation holds true for any $\lambda \geq 0$ and is a well known consequence of the representer theorem.
The idea of the proof is to decompose the analysis into the term emerging from the error in the high probability region $\mathcal{E}_{Z|\mathbf{Z}}$ and the error emerging from the low probability region $\mathcal{E}^{\mathsf{c}}_{Z|\mathbf{Z}}$. The proof essentially relies on the following lemma.

Lemma C.6. *We can construct a polynomial p of degree $\leq m$ such that for $n \to \infty$,*

1. $|p(Z) - \mathbb{E}_Y \hat{f}_\lambda(\sqrt{\tau}Z)| \to 0$*, uniformly for all* $Z \in \mathcal{E}_{Z|\mathbf{Z}}$
2. $\|\mathbb{1}_{Z \in \mathcal{E}^c_{Z|\mathbf{Z}}} p\|_{\mathcal{L}_2} \to 0$

The proof of the lemma can be found in Section E.2. As a result, Equation (16) follows immediately and Equation (17) is a consequence of

$$\left\| \mathbb{E}_Y \hat{f}_\lambda - p \right\|^2_{\mathcal{L}_2(\mathbb{P}_Z)}$$

The first two terms vanish due to Lemma C.6. To see that the third term vanishes, note that for n sufficiently large,

$$\begin{aligned}
&\mathbb{E}_Z \mathbb{1}_{Z \notin \mathcal{E}_{Z|\mathbf{Z}}} (\mathbb{E}_Y \hat{f}_\lambda(\sqrt{\tau}Z))^2 = \mathbb{E}_Z \mathbb{1}_{Z \notin \mathcal{E}_{Z|\mathbf{Z}}} (y^\top (K + \lambda I_n)^{-1} k_Z)^2 \\
\lesssim &\frac{n^2}{c^2_{\lambda_{\min}}} \mathbb{E}_Z \mathbb{1}_{Z \notin \mathcal{E}_{Z|\mathbf{Z}}} \max_i |k(z_i, Z)|^2 \lesssim n^2 P(\mathcal{E}^c_{Z|\mathbf{Z}}) \to 0
\end{aligned}$$

where we have used in the first inequality that by assumption $|f^\star|$ is bounded on the support of $\mathbb{P}_X$ and that $\lambda_{\min}(K + \lambda I_n) \geq c_{\min} > 0$ and in the second inequality that $|k|$ is bounded. Finally, the convergence to zero is due Lemma C.1.
Next, the lower bound for the bias. Due to Lemma C.8, we have that

$$\max_{Z \in \mathcal{E}_{Z|\mathbf{Z}}} \left| p(Z) - \mathbb{E}_Y \hat{f}_\lambda(\sqrt{\tau}Z) \right| = \max_{Z \in \mathcal{E}_{Z|\mathbf{Z}}} \left| p(Z) - f^\star(\sqrt{\tau}\mathbf{Z})^\top (K + \lambda \mathrm{I}_n)^{-1} k_Z \right| \to 0,$$

and hence, for any $\gamma_1 > 0$ and n sufficiently large,

$$\mathbb{E}_Z \mathbb{1}_{Z \in \mathcal{E}_{Z|\mathbf{Z}}} \left(f^\star(\sqrt{\tau}\mathbf{Z})^\top (K + \lambda \mathrm{I}_n)^{-1} k_Z - f^\star(\sqrt{\tau}Z) \right)^2 \geq \mathbb{E}_Z \mathbb{1}_{Z \in \mathcal{E}_{Z|\mathbf{Z}}} \left(p(Z) - f^\star(\sqrt{\tau}Z) \right)^2 - \gamma_1,$$

Furthermore, due to the second statement in Lemma C.6, we know that $\mathbb{E}_Z \mathbb{1}_{Z \notin \mathcal{E}_{Z|\mathbf{Z}}} (p(Z))^2 \to 0$ and because $f^\star$ is bounded by assumption, we can see that $\mathbb{E}_Z \mathbb{1}_{Z \notin \mathcal{E}_{Z|\mathbf{Z}}} (p(Z) - f^\star(\sqrt{\tau}Z))^2 \to 0$. Since $\hat{f}_\lambda$ only depends linearly on the observations y, we have $\mathbf{B}(\hat{f}) = \mathbb{E}_Z (f^\star(\sqrt{\tau}\mathbf{Z})^\top (K + \lambda \mathrm{I}_n)^{-1} k_Z - f^\star(\sqrt{\tau}Z))^2$. Thus, as a result, for any $\gamma_2 > 0$,

$$\begin{aligned}
\mathbf{B}(\hat{f}) &\geq \mathbb{E}_Z \mathbb{1}_{Z \in \mathcal{E}_{Z|\mathbf{Z}}} \left(p(Z) - f^\star(\sqrt{\tau}Z) \right)^2 - \gamma_1 \\
&\geq \mathbb{E}_Z \mathbb{1}_{Z \in \mathcal{E}_{Z|\mathbf{Z}}} \left(p(Z) - f^\star(\sqrt{\tau}Z) \right)^2 + \mathbb{E}_Z \mathbb{1}_{Z \notin \mathcal{E}_{Z|\mathbf{Z}}} \left(p(Z) - f^\star(\sqrt{\tau}Z) \right)^2 - \gamma_1 - \gamma_2 \\
&= \mathbb{E}_Z \left(p(Z) - f^\star(\sqrt{\tau}Z) \right)^2 - \gamma_1 - \gamma_2.
\end{aligned}$$

Thus, the result follows from the definition of the infimum. □

C.2. Proofs for the lower bound of the eigenvalues

C.2.1. Proof of Proposition C.3

We use the same notaiton as used in the proof of Theorem C.2. As a result of Lemma C.1 it is sufficent to condition on $\mathcal{E}_{\mathbf{X}}$ throughout the rest of this proof. The proof follows straight forwardly from the following Lemma C.7 which gives an asymptotic description of the kernel matrix K based on a similar analysis as the one used in the proof of Theorem 2.1 and 2.2 in the paper (El Karoui et al., 2010). In essence, it is again a consequence of the concentration inequality from Lemma C.1 and the stronger Assumption A.1-A.3 and in particular the power series expansion of g. We denote with $\circ i$ the i-times Hadamard product.

Lemma C.7. *Given that the assumption in Proposition C.3 hold. For $m = \lfloor 2/\beta \rfloor$,*

$$||K - M||_{op} \to 0$$

with

$$M = I \left(g(1,1,1) - \sum_{q=0}^{m} g_q(1,1) \right) + \sum_{q=0}^{m} (\mathbf{Z}^\top \mathbf{Z})^{\circ q} \circ G_{g_q}, \tag{19}$$

where G_{g_q} is the positive semi-definite matrix with entries $(G_{g_q})_{i,j} = g_q(\|z_i\|_2^2, \|z_j\|_2^2)$.

The proof of the lemma can be found in Section E.3. The proof of Proposition C.3 then follows straight forwardly when using Schur's product theorem which shows that

$$I\left(g(1,1,1)-\sum_{q=0}^{m} g_q(1,1)\right)+\sum_{q=0}^{m}(\mathbf{Z}^\top\mathbf{Z})^{\circ q}\circ G_{g_q}\succeq I\left(g(1,1,1)-\sum_{q=0}^{m} g_q(1,1)\right)$$

where we use that g_q are positive semi-definite by Assumption A.1. To see that the eigenvalues are lower bounded, we thus simply need to show that $g(1,1,1)-\sum_{q=0}^{m} g_q(1,1)>0$. This holds because the positive semi-definiteness of g_q implies that $g_q(1,1)\geq 0$ and hence $g(1,1,1)=\sum_{q=0}^{\infty} g_q(1,1)$ is a sum of positive coefficients and because by Assumption A.3 there exists $j'>\lfloor 2/\beta\rfloor$ such that $g_{j'}(1,1)>0$. Hence, there exists a positive constant $c>0$ such that $\lambda_{\min}(M)\geq c$. We can conclude the proof when applying Lemma C.7, which implies that $\lambda_{\min}(K)\to\lambda_{\min}(M)$ as $n\to\infty$. □

C.2.2. Proof of Proposition C.4

We use the same notaiton as used in the proof of Theorem C.2 and define D_α to be the $n\times n$ matrix with entries $(D_\alpha)_{i,j}=d_\alpha(z_i,z_j):=\|z_i-z_j\|_2^\alpha$. We separate the proof into two steps. In a first step, we decompose the $k(x,x')=d_\alpha(x,x')$ in the terms

$$\exp(-||x-x'||_2^\alpha)=\exp(\tilde{k}(x,x'))\exp(-||x-x'||_2^\alpha-\tilde{k}(x,x'))$$

such that $\exp(-||x-x'||_2^\alpha-\tilde{k}(x,x'))$ and $\exp(\tilde{k}(x,x'))$ are both positive semi-definite kernel functions. In particular, we construct $\tilde{k}$ such that the eigenvalues of the kernel matrix A of $\exp(\tilde{k}(x,x'))$ evaluated at $\mathbf{Z}$ are almost surely lower bounded by a positive constant. The proposition is then a straight forward consequence and shown in the last step.

Step 1: A matrix M is conditionally negative semi-definite if for every $v\in\mathbb{R}^n$ with $1^\top v=0$, $v^\top Mv\leq 0$. We can see from Chapter 3 Theorem 2.2 in (Berg et al., 1984) that d_α is a conditionally negative semi-definite function, that is that for any $m\in\mathbb{N}\setminus 0$ and any $\{x_1,...,x_m\}$, the corresponding kernel matrix M is conditionally negative semi-definite. As shown in Chapter 3 Lemma 2.1 in (Berg et al., 1984), a kernel function $\phi(x,x')$ is conditionally negative semi-definite, if and only if for any z_0, $(x,x')\to\phi(x,z_0)+\phi(z_0,x')-\phi(x,x')-\phi(z_0,z_0)$ is a positive semi-definite function. Hence, for any $z_0\in\mathbb{R}^n$, the kernel defined by

$$\tilde{k}(x,x')=d_\alpha(x,z_0)+d_\alpha(z_0,x')-d_\alpha(x,x')-d_\alpha(z_0,z_0) \tag{20}$$

is positive semi-definite.
The goal of the first step is now to show that we can find a vector z_0 such that the kernel matrix A of $\tilde{k}$ evaluated at $\mathbf{Z}$ has eigenvalues almost surely lower bounded by some positive constant. Essentially, the statement is a consequence of the following lemma, bounding the eigenvalues of D_α.

Lemma C.8. *Assume that $\mathbb{P}_X$ satisfies the Assumption B.1-B.2. Conditioned on $\mathcal{E}_{\boldsymbol{X}}$, for any n sufficiently large, all eigenvalues of the matrix D_α are bounded away from zero by a positive constant $c>0$, i.e. $\min_{i\leq n}|\lambda_i(D_\alpha)|\geq c$.*

The proof of the lemma can be found in Section E.4. We can see from Lemma C.1 that there exists almost surely over the draws of $\mathbf{Z}$ as $n\to\infty$ an additional vector z_0, such that for any two vectors $z,z'\in\mathbf{Z}\cup\{z_0\}$,

$$|z^\top z'-\delta_{z=z'}|\lesssim n^{-\beta/2}(\log(n))^{(1+\epsilon)/2}. \tag{21}$$

In particular, Lemma C.8 then also implies that the eigenvalues of the matrix

$$D_\alpha(\mathbf{Z},z_0)=\begin{pmatrix} D_\alpha & d_\alpha(\mathbf{Z},z_0)\\ d_\alpha(\mathbf{Z},z_0)^\top & d_\alpha(z_0,z_0),\end{pmatrix}$$

are bounded away from zero and since this matrix is conditionally negative semi-definite, we have that for any $v\in\mathbb{R}^n$,

$$\begin{pmatrix} v^T & -1^\top v\end{pmatrix} D_\alpha(\mathbf{Z},z_0)\begin{pmatrix} v^T\\ -1^\top v\end{pmatrix}\leq -\tilde{c}\|\begin{pmatrix} v^T\\ 1^\top v\end{pmatrix}\|_2^2, \tag{22}$$

where $\tilde{c}>0$ is a positive constant and we have used that $1^T\begin{pmatrix} v^T\\ -1^\top v\end{pmatrix}=0$. Throughout the rest of this proof, we conditioned on the event $\mathcal{E}_{\mathbf{X}}$ and the additional event that Equation (21) holds, and remark that the intersection of these two events holds true almost surely as $n\to\infty$.

As a result, we can see that

$$\begin{aligned}
&\begin{pmatrix} v^T & -1^\top v \end{pmatrix} D_\alpha(\mathbf{Z}, z_0) \begin{pmatrix} v^T \\ -1^\top v \end{pmatrix} \\
=&v^\top D_\alpha v - v^\top \left[\frac{1}{n} 1 1^\top d_\alpha(z_0, \mathbf{Z})\right] v - v^\top \left[\frac{1}{n} d_\alpha(z_0, \mathbf{Z})^\top 1 1^\top\right] v + v^\top \left[\frac{1}{n^2} 1 1^\top d_\alpha(z_0, z_0) 1 1^\top\right] v. \\
=&v^\top \underbrace{\left[D_\alpha - \frac{1}{n} 1 1^\top d_\alpha(z_0, \mathbf{Z}) - \frac{1}{n} d_\alpha(z_0, \mathbf{Z})^\top 1 1^\top + \frac{1}{n^2} 1 1^\top d_\alpha(z_0, z_0) 1 1^\top\right]}_{=-A} v,
\end{aligned} \tag{23}$$

where A is exactly the kernel matrix of $\tilde{k}$ evaluated at $\mathbf{Z}$. Hence, combining Equation (22) and (23) gives

$$v^T A v \geq \tilde{c}\,(v^T v + v^T 1 1^T v) \geq \tilde{c}\, v^T v.$$

We can conclude the first step of the proof when applying the Courant–Fischer–Weyl min-max principle which shows that A has lower bounded eigenvalues $\geq \tilde{c}$.

Step 2: We can write $\exp(-d_\alpha(x, x')) = \exp(\tilde{k}(x, x')) \exp(\phi(x, x'))$ with $\phi(x, x') := -d_\alpha(x, x') - \tilde{k}(x, x') = d_\alpha(z_0, z_0) - d_\alpha(x, z_0) - d_\alpha(z_0, x')$. It is straight forward to verify that $\exp(\phi(x, x'))$ is a positive semi-definite function. Hence, due to Schur's product theorem, the following sum is a sum of positive semi-definite functions

$$\exp(-d_\alpha(x, x')) = \exp(\tilde{k}(x, x')) \exp(\phi(x, x')) = \sum_{l=0}^{\infty} \frac{1}{l!} \tilde{k}(x, x')^l \exp(\phi(x, x')).$$

It is sufficient to show that the eigenvalues of the kernel matrix M of $\tilde{k}(x, x') \exp(\phi(x, x'))$, evaluated at $\mathbf{Z}$, are lower bounded. Let B be the kernel matrix of $\exp(\phi(x, x'))$ evaluated at $\mathbf{Z}$, we have that $M = A \circ B$, where $\circ$ is the Hadamard product and A the kernel matrix of $\tilde{k}$ from the previous step. We make the following claim from which the proof follows trivially using that the eigenvalues of A are lower bounded by a positive constant.

Claim: $B = \frac{1}{2e^4} 1 1^\top + \tilde{B}$ *with* $\tilde{B}$ *a positive semi-definite matrix.*

Proof of the claim: Let ψ be the vector with entries $\psi_i = \exp(-d_\alpha(z_i, z_0))$. Furthermore, let $\gamma = \exp(d_\alpha(z_0, z_0))$. We can write

$$B = \gamma \left(1 \psi^\top\right) \circ \left(\psi 1^\top\right) = \gamma \psi \psi^\top.$$

Next, using Lemma C.1 and the fact that $d_\alpha(x, x') = 2^{\alpha/2} + O(\frac{||x - x'||_2^2}{2} - 1)$, we can see that $\gamma \geq \exp(2^{\alpha/2}/2) > 1$. Hence, it is sufficient to show that $\psi \psi^\top - \frac{1}{2e^4} 1 1^\top$ is positive semi-definite. This is true if and only if $1^\top \psi \psi^\top 1 \geq \frac{1}{2e^4} 1^\top 1 1^\top 1$, which is equivalent to saying that $\left(\sum_{i=1}^n \exp(-d_\alpha(z_i, z_0))\right)^2 \geq \frac{n^2}{2e^4}$. Using again the same argument as for γ, we can see that $\max_i \left|2^{\alpha/2} - d_\alpha(z_i, z_0)\right| \to 0$ for any i, which completes the proof. □

C.3. Proof of Corollary 3.2

First, note that the Assumption A.1-A.3 straight forwardly hold true for the exponential inner product kernel with $k(x, x') = \exp(x^\top x') = \sum_{j=0}^{\infty} \frac{1}{j!} (x^\top x')^j$ and for the Gaussian kernel with

$$k(x, x') = \exp(-\|x - x'\|_2^2) = \sum_{j=0}^{\infty} \frac{2^j}{j!} (x^\top x')^j \exp(-\|x\|_2^2) \exp(-\|x'\|_2^2).$$

Next, note that the α-exponential kernel with $\alpha < 2$ is already explicitly covered in Theorem 3.1. Hence, the only thing left to show is that Theorem 3.1 also applies to ReLU-NTK.

We use the definition of the Neural Tangent Kernel presented in (Arora et al., 2019; Lee et al., 2018). Let L be the depth of the NTK and $\sigma : \mathbb{R} \to \mathbb{R}$ the activation function which is assumed to be almost everywhere differentiable. For any $i > 0$,

define the recursion

$$\Sigma^{(0)}(x,x') := x^\top x'$$
$$\Lambda^{(i)}(x,x') := \begin{pmatrix} \Sigma^{(i-1)}(x,x) & \Sigma^{(i-1)}(x,x') \\ \Sigma^{(i-1)}(x,x') & \Sigma^{(i-1)}(x',x') \end{pmatrix}$$
$$\Sigma^{(i)}(x,x') := c_\sigma \mathop{\mathbb{E}}_{(u,v)\sim\mathcal{N}(0,\Lambda^{(i)})} [\sigma(u)\sigma(v)]$$

with $c_\sigma := \left[\mathop{\mathbb{E}}_{v\sim\mathcal{N}(0,1)} \left[\sigma(v)^2\right]\right]^{-1}$. Furthermore, define

$$\dot{\Sigma}^{(i)} := c_{\dot\sigma} \mathop{\mathbb{E}}_{(u,v)\sim\mathcal{N}(0,\Lambda^{(i)})} [\dot\sigma(u)\dot\sigma(v)]$$

with $c_{\dot\sigma} := \left[\mathop{\mathbb{E}}_{v\sim\mathcal{N}(0,1)} \left[\dot\sigma(v)^2\right]\right]^{-1}$ where $\dot\sigma$ is the derivative of σ. The NTK k_{NTK} of depth $L \geq 1$ is then defined as

$$k_{\mathrm{NTK}}(x,x') := \sum_{i=1}^{L+1} \Sigma^{(i-1)}(x,x') \prod_{j=i}^{L+1} \dot\Sigma^{(j)}(x,x').$$

We call a function σ *k-homogeneous*, if for any $x \in \mathbb{R}$ and any $a > 0$, $\sigma(ax) = a^k\sigma(x)$. We can now show the following result from which the corollary follows.

Proposition C.9. *Assume that the activation function σ is k-homogeneous and both the activation function and its derivative possess a Hermite-polynomial series expansion (see (Daniely et al., 2016)) where there exits $j' \geq \lfloor 2/\beta \rfloor$ such that the j'-th coefficient $a_{j'} \neq 0$. Then, the NTK satisfies the Assumption A.1-A.3 and hence Theorem 3.1 applies.*

In fact, we can easily see that any non-polynomial activation function which is homogeneous and both the activation function and its derivative possesses a Hermite polynomial extension satisfies the assumptions in Proposition C.9. In particular, this includes the popular ReLU activation function $\sigma(x) = \max(x, 0)$. We refer to (Daniely et al., 2016) for the explicit expression of the Hermite polynomial extension. □

C.3.1. Proof of Proposition C.9

Essentially, the proof follows from the power series expression of the NTK k_{NTK} given in the following lemma.

Lemma C.10. *The NTK k_{NTK} possesses a power series expansion*

$$k_{\mathrm{NTK}}(x,x') = \sum_{j=0}^{\infty} (x,x')^j g_j(||x||_2^2, ||x'||_2^2),$$

which converges for any $x, x' \in \mathbb{R}^d$ with $x, x' \neq 0$. Furthermore, for any $u, u' \in \mathbb{R}_+$,

$$g_j(u,u') = \sum_{i=-\infty}^{\infty} \eta_{j,i} (uu')^{i/2}$$

and $\eta_{j,i} \geq 0$.

The proof of the lemma can be found in Section E.5. It is straight forward to verify from the proof of Lemma C.10 that $\Sigma^{(i)}$ and $\dot\Sigma^{(i)}$ are compositions of continuous functions and thus k_{NTK} is also continuous for any $x, x' \neq 0$. Next, note that the Lipschitz continuity (Assumption A.2) follows straight forwardly from Equation (31) in the proof of Lemma C.10. In order to show that any g_j from Lemma C.10 is smooth, recall that

$$g_j(u,u') = \sum_{l=-\infty}^{\infty} \eta_{j,l}^{(i+1)} (uu')^{l/2} =: h_j(xy).$$

Therefore, h_j is a Puiseux power series with divisor 2. Furthermore, the function $\tilde h_j(t) := h_j(t^2) = \sum_{l=-\infty}^{\infty} \eta_{j,l}^{(i+1)} (t)^l$ is a Laurent series which converges for every $t \neq 0$. Hence, we can conclude that $\tilde h_j$ is smooth for any $t \neq 0$ and thus

also h_j. Finally, because $(u, u') \to uu'$ is also a smooth function, we can conclude that any g_j is a smooth function for any $u, u' \neq 0$. Next, since for any $l \in \mathbb{Z}$, $(u, u') \to \alpha(uu')^{l/2}$ is trivially a positive semi-definite (PSD) function whenever $\alpha \geq 0$ and sums of PSD functions are again PSD, we can conclude that the $g_j(u, u') = \sum_{l=-\infty}^{\infty} \eta_{j,l}^{(i+1)}(uu')^{l/2}$ is PSD for any j. Therefore, we can conclude that Assumption A.1 holds as well.
The only thing left to show is Assumption A.3. While we have already shown that g_j are smooth in a neighborhood of $(1, 1)$, we still need to show that there exists $j' > \lfloor 2/\beta \rfloor$ such that $g_{j'}(1, 1) > 0$. However, this follows from the fact that by assumption there exists $j' > \lfloor 2/\beta \rfloor$ such that $a_{j'} \neq 0$ where a_j are the Hermite coefficients of the activation function σ. □

D. Different scalings τ

In this section, we present results for different choices of the scaling beyond the standard choice $\tau \asymp d_{\text{eff}}$. In Subsection D.1, we give a proof of Theorem 3.3 describing the *flat limit*, i.e. the limit of the interpolant where for any fixed n, d, $\tau \to \infty$. Furthermore, in order to get a more comprehensive picture, we additionally present straight forward results for other choices of τ in Section D.2.

D.1. Proof of Theorem 3.3

We use again the same notation as used for the proof of Theorem C.2 where we set $z_i = x_i/\sqrt{d_{\text{eff}}}$ and let $Z = X/\sqrt{d_{\text{eff}}}$ be the random variable with $X \sim \mathbb{P}_X$. We can again condition throughout the proof on the event $\mathcal{E}_{\mathbf{X}}$. In particular, we assume throughout the proof that n is sufficiently large since we are only interested in the asymptotic behaviour. Furthermore, recall the definition of D_α, which is the $n \times n$ matrix with entries $(D_\alpha)_{i,j} = \|z_i - z_j\|_2^\alpha$ and denote with D_α^{-1} its inverse. In addition, denote with d_Z^α the vector with entries $(d_Z^\alpha)_i = \|z_i - Z\|_2^\alpha$ and with d_α the function $d_\alpha(z, z') = \|z - z'\|_2^\alpha$.
First, although the limit $\lim_{\tau\to\infty} K^{-1}$ does not exists, we can apply Theorem 3.12 in (Lee et al., 2014) to show that the flat limit interpolator $f_{\text{FL}} := \lim_{\tau\to\infty} \hat{f}_0$ of any kernel satisfying the assumption in Theorem 3.3 exists and has the form

$$f_{\text{FL}}(Z) = \begin{pmatrix} y^\top & 0 \end{pmatrix} \begin{pmatrix} -D_\alpha & 1 \\ 1^T & 0 \end{pmatrix}^{-1} \begin{pmatrix} d_Z^\alpha \\ 1 \end{pmatrix}.$$

Furthermore, for the α-exponential kernel, we use Theorem 2.1 in (Blumenthal and Getoor, 1960) to show that it satisfies the assumptions imposed on the eigenvalue decay in Theorem 3.3.

Remark D.1. *The estimator f_{FL} is also called the polyharmonic spline interpolator. This estimator is invariant under rescalings of the input data which is also the reason why we can rescale the input data by $\sqrt{d_{\text{eff}}}$, i.e. consider $z_i = x_i/\sqrt{d_{\text{eff}}}$ as input data points.*

We already know from lemma C.8 that the matrix D_α has $n - 1$ negative eigenvalues and one positive eigenvalue. In particular, we have shown that $|\lambda_i(D_\alpha)| \geq \tilde{c} > 0$, for every i, where $\tilde{c}$ is some positive constant. Next, note that because D_α has full rank, we can conclude from Theorem 3.1 in (Ikramov and Savel'eva, 2000) that the matrix $\begin{pmatrix} -D_\alpha & 1 \\ 1^T & 0 \end{pmatrix}$ has n positive eigenvalues and one negative. Hence,

$$\det \begin{pmatrix} -D_\alpha & 1 \\ 1^T & 0 \end{pmatrix} = \det(-D_\alpha)(1^T D_\alpha^{-1} 1) \neq 0$$

and $1^T D_\alpha^{-1} 1 > 0$. In particular, this allows us to use the block matrix inverse to show that

$$\begin{pmatrix} -D_\alpha & 1 \\ 1^T & 0 \end{pmatrix}^{-1} = \begin{pmatrix} -D_\alpha^{-1} + \frac{D_\alpha^{-1} 1 1^T D_\alpha^{-1}}{1^T D_\alpha^{-1} 1} & D_\alpha^{-1} 1 \frac{1}{1^T D_\alpha^{-1} 1} \\ (D_\alpha^{-1} 1 \frac{1}{1^T D_\alpha^{-1} 1})^T & \frac{1}{1^T D_\alpha^{-1} 1} \end{pmatrix}$$

and therefore,

$$f_{\text{FL}}(Z) = y^T \underbrace{\left[-D_\alpha^{-1} + \frac{D_\alpha^{-1} 1 1^T D_\alpha^{-1}}{1^T D_\alpha^{-1} 1} \right]}_{=:A} d_Z^\alpha + \frac{y^T D_\alpha^{-1} 1}{1^T D_\alpha^{-1} 1}.$$

Next, using the Binomial series expansion, we can see that for any $q \in \mathbb{N}$,

$$d_\alpha(z_i, Z) = \sum_{j=0}^{q} \binom{j}{\alpha/2} 2^{\alpha/2} \left(\frac{1}{2}\|z_i - Z\|_2^2 - 1 \right)^j + O\left(\left(\frac{1}{2}\|z_i - Z\|_2^2 - 1 \right)^{q+1} \right).$$

Furthermore, by Lemma C.1 and the fact that $\|z_i - Z\|_2^2 = z_i^\top z_i + Z^\top Z - 2z_i^\top Z$, we can see that for $q = \lfloor 2/\beta \rfloor$, $nO\left(\left(\frac{1}{2}\|z_i - Z\|_2^2 - 1\right)^{q+1}\right) \to 0$. Hence, assuming that the absolute eigenvalues $|\lambda_i(A)|$ of A are all upper bounded by a non-zero positive constant, we can use exactly the same argument as used in the proof of Theorem C.2 to conclude the proof. We already know from Lemma C.8 that there exists some constant $c > 0$ independent of n, such that $\left|\left|D_\alpha^{-1}\right|\right|_{\text{op}} \leq c$. Thus, we only need to show that $\left|\left|\frac{D_\alpha^{-1}11^TD_\alpha^{-1}}{1^TD_\alpha^{-1}1}\right|\right|_{\text{op}}$ is almost surely upper bounded. Because $\frac{D_\alpha^{-1}11^TD_\alpha^{-1}}{1^TD_\alpha^{-1}1}$ is a rank one matrix, we know that

$$\left|\left|\frac{D_\alpha^{-1}11^TD_\alpha^{-1}}{1^TD_\alpha^{-1}1}\right|\right|_{\text{op}} = \frac{1^TD_\alpha^{-2}1}{1^TD_\alpha^{-1}1}.$$

Step 1: We show that

$$1^TD_\alpha^{-2}1 = O(1)$$

Let $\lambda_2, ..., \lambda_n$ be the $n-1$ negative eigenvalues of D_α^{-1} and λ_1 the only positive eigenvalue. Furthermore, let v_i be the corresponding orthonormal eigenvectors. Next, let $\alpha_i \in \mathbb{R}$ be such that $1 = \sum_{i=1}^n \alpha_i v_i$. Since $\|1\|_2 = \sqrt{n}$, we know that $\alpha_i \leq \sqrt{n}$. We have

$$1^TD_\alpha^{-1}1 = \alpha_1^2\frac{1}{\lambda_1} - \sum_{i=2}^n \alpha_i^2\frac{1}{|\lambda_i|} > 0$$

$$\text{and } 1^TD_\alpha^{-2}1 = \alpha_1^2\frac{1}{\lambda_1^2} + \sum_{i=2}^n \alpha_i^2\frac{1}{\lambda_i^2},$$

where we use that $1^TD_\alpha^{-1}1 > 0$ which we already know from the discussion above. Because by the binomial expansion, $d_\alpha(z_i, z_j) = 2^{\alpha/2} + O(\frac{1}{2}\|z_i - z_j\|_2^2 - 1)$, Lemma C.1 implies that $\max_{i\neq j} d_\alpha(z_i, z_j) \to 2^{\alpha/2}$ and hence, $\frac{1}{n}1^TD_\alpha 1 \gtrsim n$, for any n sufficiently large. Therefore, $\lambda_1 \gtrsim n$. Hence, there exists some constant $c > 0$ independent of n, such that $\alpha_1^2\frac{1}{\lambda_1} \leq c$. As a consequence,

$$c \geq \alpha_1^2\frac{1}{\lambda_1} \geq \sum_{i=1}^{n-1}\alpha_i^2\frac{1}{|\lambda_i|} \geq \tilde{c}\sum_{i=1}^{n-1}\alpha_i^2\frac{1}{\lambda_i^2},$$

where we use in the last inequality that $|\lambda_i| \geq \tilde{c}$. Furthermore, $\alpha_1^2\frac{1}{\lambda_1^2} = O(\frac{1}{n})$. Thus, we conclude that $1^TD_\alpha^{-2}1 = O(1)$.

Step 2: In order to prove the result, we are only left to study the case where $1^TD_\alpha^{-2}1 \geq 1^TD_\alpha^{-1}1 \to 0$. We prove by contradiction and assume that $\frac{1^TD_\alpha^{-2}1}{1^TD_\alpha^{-1}1} \to \infty$ and $1^TD_\alpha^{-1}1 \to 0$. Let $\gamma = 1^TD_\alpha^{-1}1$ and $v = D_\alpha^{-1}1$. Furthermore, let $\tilde{v} = (-\gamma, 0, \cdots, 0)^\top$. We know that $1^T(v + \tilde{v}) = 0$ and hence,

$$(v+\tilde{v})^TD_\alpha(v+\tilde{v}) \leq -\tilde{c}(v+\tilde{v})^T(v+\tilde{v})$$

Next, note that our assumption imply that for n sufficiently large, $v^\top v - \gamma^2 \geq 1/2 v^\top v$. Therefore,

$$-\frac{\tilde{c}}{2}v^Tv \geq (v+\tilde{v})^TD_\alpha(v+\tilde{v}) = v^TD_\alpha v + 2\tilde{v}^TD_\alpha v + \tilde{v}^TD_\alpha\tilde{v} = 1^TD_\alpha^{-1}1 + 2\tilde{v}^T1 + \gamma^2 d_\alpha(X_1, X_1) = \gamma - 2\gamma,$$

and using the fact that $\gamma = 1^TD_\alpha^{-1}1$ is positive, we get that $1 \geq \frac{\tilde{c}}{2}\frac{v^Tv}{\gamma}$. However, this contradicts the assumption that $\frac{v^Tv}{\gamma} \to \infty$ and hence we can conclude the proof. □

D.2. Additional results

In this section, we present some additional results for different choices of the scaling. The results presented in this section are straight forward but provide a more complete picture for different choices of the scaling τ. We use again the same notation as used in Appendix C.1.

First, we show the case where $\tau \to 0$. We assume that k is the α-exponential kernel with $\alpha \in (0, 2]$, i.e. $k(x, x') = \exp(-\|x - x'\|_2^\alpha)$.

Lemma D.2. *Let $\mathbb{P}_X$ satisfy Assumption B.1-B.2 and assume that the bandwidth $\tau/d_{eff} = O(n^{-\theta})$ with $\theta > 0$. Furthermore, assume that the ground truth function $f^\star$ is bounded. Then, conditioned on the event $\mathcal{E}_X$, for any $\lambda \geq 0$, with probability $\geq 1 - (n+1)^2\exp(-C(\log(n))^{1+\epsilon})$ over the draws of $X \sim \mathbb{P}_X$,*

$$\mathbb{E}_Y\hat{f}_\lambda(X) \to 0\,.$$

Proof. Let $\tilde{\tau} = \tau/d_{\text{eff}}$ and define $z_i = x_i/\sqrt{d_{\text{eff}}}$ and $Z = X/\sqrt{d_{\text{eff}}}$. Lemma C.1 shows that $\|z_i - z_j\|_2^2$ concentrates around $2(1-\delta_{i,j})$. Hence, $k_\tau(x_i, x_j) = \exp(-\tilde{\tau}^{-\alpha/2}\|z_i - z_j\|_2^\alpha) \to \delta_{i,j}$ because $\tilde{\tau} \to 0$. In fact, due to the assumption that $\tilde{\tau} = O(n^{-\theta})$ with $\theta > 0$, we can see that

$$||K - \mathrm{I}_n||_{\text{op}} \leq n \max_{i \neq j} |\exp(-n^{\theta\alpha/2}\|z_i - z_j\|_2^\alpha)| \to 0 \ ,$$

and with probability $\geq 1 - (n+1)^2 \exp(-C(\log(n))^{1+\epsilon})$ over the draws of X,

$$\|k_Z\|_1 \leq n \max_i |\exp(-n^{\theta\alpha/2}\|z_i - Z\|_2^\alpha)| \to 0.$$

Hence, the result follows immediately from $\hat{f}_\lambda(X) = y^T(K + \lambda I)^{-1}k_\tau(\mathbf{X}, X)$. □

We can also show a similar result for the case where $\tau \to \infty$ and λ does not vanish.

Lemma D.3. *Let $\mathbb{P}_X$ satisfy Assumption B.1-B.2 and assume that the bandwidth $\tau/d_{\text{eff}} = O(n^\theta)$ with $\theta > \frac{2}{\alpha}$. Furthermore, assume that $\lambda = \Omega(1)$ and that the ground truth function $f^\star$ is bounded. Then, conditioned on the event $\mathcal{E}_{\boldsymbol{X}}$, for any $\lambda \geq 0$, with probability $\geq 1 - (n+1)^2\exp(-C(\log(n))^{1+\epsilon})$ over the draws of $X \sim \mathbb{P}_X$,*

$$\mathbb{E}_Y \hat{f}_\lambda(X) \to c,$$

with $c = f(\boldsymbol{X})^T(11^T + \lambda I_n)^{-1}1$

Proof. We use the same notation as in Lemma D.3. Again due to Lemma C.1, we find that

$$\left|\left|K - 11^T\right|\right|_{\text{op}} \lesssim n \max_{i \neq j} |\exp(-n^{-\theta\alpha/2}\|z_i - z_j\|) - 1| \to 0 \ .$$

As a result, we observe that the kernel matrix K converges asymptotically to the rank one matrix 11^T. We remark that such a phenomenon can also be observed in Theorem 2.1 and 2.2 in (El Karoui et al., 2010) when $\operatorname{tr}(\Sigma_d)/d \to 0$. As a consequence, we observe that the eigenvalues of K^{-1} diverge as $n \to \infty$. However, because by assumption, $\lambda = \Omega(1)$, we can still conclude that the eigenvalues of $(K + \lambda \mathrm{I}_n)^{-1}$ are upper bounded by a constant independent of n. Hence, with probability $\geq 1 - (n+1)^2\exp(-C(\log(n))^{1+\epsilon})$ over the draws of $X \sim \mathbb{P}_X$,

$$\begin{aligned}\|(K + \lambda \mathrm{I}_n)^{-1}k_Z - (11^T + \lambda \mathrm{I}_n)^{-1}1\|_1 &\lesssim n \max_i |\exp(-n^{-\theta\alpha/2}c) - 1| \\ &\lesssim n^{-\theta\alpha/2+1} + O(n^{-\theta\alpha+1}) \to 0,\end{aligned}$$

where we have used that $\theta\alpha/2 > 1$.

The only thing left to show is that $f(\mathbf{X})^T(11^T + \lambda I_n)^{-1}1$ does not diverge. For this, let $a\mathrm{I}_n + b11^T$ be the inverse of $11^T + \lambda \mathrm{I}_n$. As a result of a simple computation we find that $a = \frac{\lambda(\lambda+1+(n-1))+(n-1)}{\lambda(\lambda+1+(n-1))}$ and $b = \frac{-1}{\lambda(\lambda+1+(n-1))}$. Hence,

$$\|(11^T + \lambda \mathrm{I}_n)^{-1}1\|_1 = n|a + (n-1)b| = \sqrt{n}|\lambda b| = \frac{n}{(\lambda + 1 + (n-1))} \to 1,$$

which completes the proof.

□

E. Technical lemmas

E.1. Proof of Lemma C.1

We begin with the following Lemma which is a direct consequence of the results in Appendix A in (El Karoui et al., 2010).

Lemma E.1 (Concentration of quadratic forms)**.** *Suppose the vector $x \in \mathbb{R}^d$, of some dimension $d \in \mathbb{N}_+$, is a random vector with either*

1. *i.i.d entries $x_{(i)}$ almost surely bounded $|x_{(i)}| \leq c$ by some constant $c > 0$ and with zero mean and unit variance.*
2. *standard normal distributed i.i.d entries.*

Let M be any symmetric matrix with $||M||_{op} = 1$ and let $M = M_+ - M_-$ be the decomposition of M into two positive semi-definite matrices M_+ and M_- with $||M_+||_{op}, ||M_-||_{op} \leq 1$. Then, there exists some positive constants C_1, C_2, C_3 independent of M such that for any $r > \zeta = C_1/\operatorname{tr}(M_+)$,

$$P(|x^T Mx/\operatorname{tr}(M_+) - \operatorname{tr}(M)/\operatorname{tr}(M_+)| > r) \lesssim \exp(-C_2 \operatorname{tr}(M_+)(r/2 - \zeta)^2) + \exp(-C_3 \operatorname{tr}(M_+)).$$

Case 1: Distribution $\mathbb{P}_X \in \mathcal{Q}$ Following the same argument as the one used in Corollary A.2 in (El Karoui et al., 2010), we can use Lemma E.1 to show that there exists constants $C_2, C_3 > 0$ such that for $n, d_{\text{eff}} \to \infty$,

$$P(|x_i^T x_j / \operatorname{tr}(\Sigma_d) - \delta_{i,j}| > r) \leq C_3[\exp(-C_2 \operatorname{tr}(\Sigma_d)(r/2 - \zeta)^2) + \exp(-C_2 \operatorname{tr}(\Sigma_d)].$$

We now make us of the Borel-Cantelli Lemma. For any $\epsilon > 0$, let $r(n) = \frac{1}{\sqrt{2}} n^{-\beta/2} (\log(n))^{(1+\epsilon)/2}$ and note that because $\operatorname{tr}(\Sigma_d) \asymp n^\beta$, ζ decays at rate $n^{-\beta}$ and in particular, for any n sufficiently large, $r(n)/2 > \zeta \to 0$. Hence, we can see that there exists some constant $C > 0$ such that for any n sufficiently large

$$P(|x_i^T x_j / \operatorname{tr}(\Sigma_d) - \delta_{i,j}| > r(n)) \leq \exp(-C (\log(n))^{1+\epsilon}).$$

Next, using the union bound, we get

$$P(\max_{i,j} |x_i^T x_j / \operatorname{tr}(\Sigma_d) - \delta_{i,j}| > r(n)) \leq n^2 \exp(-C (\log(n))^{1+\epsilon}).$$

And because $\epsilon > 0$, for any $N \in \mathbb{N}_+$, we have that

$$\sum_{n=N}^{\infty} n^2 \exp(-C (\log(n))^{1+\epsilon}) < \infty.$$

which allows us to apply the Borel-Cantelli Lemma. Hence,

$$\max_{i,j} \left| x_i^T x_j / \operatorname{tr}(\Sigma_d) - \delta_{i,j} \right| \leq \frac{1}{\sqrt{2}} n^{-\beta/2} (\log(n))^{(1+\epsilon)/2} \ \ a.s. \text{ as } n \to \infty, \tag{24}$$

which concludes the first step of the proof.
Next, we already know from the previous discussion that for any n sufficiently large,
$P(\mathcal{E}_{\mathbf{X}}) \geq 1 - n^2 \exp(-C (\log(n))^{(1+\epsilon)/2})$. Furthermore, because X is independently drawn from the same distribution as x_i, $P(\mathcal{E}_{\mathbf{X}} \cup \mathcal{E}_{X|\mathbf{X}}) \geq 1 - (n+1)^2 \exp(-C (\log(n))^{(1+\epsilon)/2})$. Hence, for any n sufficiently large,

$$P(\mathcal{E}_{\mathbf{X},X} | \mathcal{E}_{\mathbf{X}}) = \frac{P(\mathcal{E}_{\mathbf{X}} \cup \mathcal{E}_{X|\mathbf{X}})}{P(\mathcal{E}_{\mathbf{X}})} \geq \left[1 - (n+1)^2 \exp(-C (\log(n))^{(1+\epsilon)})\right]$$

This completes the first case of the proof, i.e. where $\mathbb{P}_X \in \mathcal{Q}$.

Case 2: Distribution $\mathbb{P}_X \in \mathcal{Q}_{\mathcal{S}^{d-1}}$ First, note that the case where $i = j$ is clear. Let $s_i = \frac{x_i}{\|x_i\|_2}$ and $z_i = \frac{x_i}{\sqrt{\operatorname{tr}(\Sigma_d)}}$. Since we are in the Euclidean space, the inner product is given by

$$(s_i^T s_j)^2 = \frac{(z_i^T z_j)^2}{\|z_i\|_2^2 \|z_j\|_2^2}.$$

Due to Equation (24), we have that $|\ \|z_i\|_2^2 - 1| \leq n^{-\beta/2} (\log(n))^{(1+\epsilon)/2}\ a.s.$ as $n \to \infty$ and
$(z_i^T z_j)^2 \leq (n^{-\beta/2} (\log(n))^{(1+\epsilon)/2})^2 a.s.$ as $n \to \infty$. Therefore,

$$(s_i^T s_j)^2 \leq (n^{-\beta/2} (\log(n))^{(1+\epsilon)/2})^2 \ \ a.s. \text{ as } n \to \infty.$$

The rest of the proof then follows straight forwardly. □

Proof of Lemma E.1. Let

$$f : \mathbb{R}^d \to \mathbb{R}, x \to \sqrt{x^T M_+ x / \operatorname{tr}(M_+)} = \frac{1}{\sqrt{\operatorname{tr}(M_+)}} \|M_+^{1/2} x\|_2.$$

We know that f is $\lambda_{\max}(M_+)/\sqrt{\operatorname{tr}(M_+)}$-Lipschitz continuous and hence also a $\sqrt{||M||_{\text{op}}}/\sqrt{\operatorname{tr}(M_+)}$-Lipschitz continuous. For the case where the entries x are bounded i.i.d. random variables, we can use the simple fact that the norm is convex in order to apply Corollary 4.10 in (Ledoux, 2001) and Proposition 1.8 in in (Ledoux, 2001). For the case where the entries are normally distributed we can apply Theorem V.I in (Milman and Schechtman, 1986). As a result, we can see that there exists a constant $C_4 > 0$ independent of M, such that

$$P\left(\left|\sqrt{x^T M x / \operatorname{tr}(M_+)} - \sqrt{\operatorname{tr}(M) / \operatorname{tr}(M_+)}\right| > r\right) \leq 4 \exp(4\pi) \exp(-C_4 \operatorname{tr}(M_+) r^2).$$

The proof then follows straight forwardly following line by line the proof of Lemma A.2 in (El Karoui et al., 2010).

□

E.2. Proof of Lemma C.6

For any $j \leq m+1$, $\alpha = (i_1, i_2)$, let $g_j^{(\alpha)}$ denote the partial derivatives $g_j^{(\alpha)}(x,y) = \frac{\partial^{|\alpha|}}{\partial t_1^{i_1} t_2^{i_2}} g_j(t_1,t_2)|_{x,y}$. Define $s = \lfloor 2/\beta \rfloor$
First of all, note that due to Lemma C.1, for any $\delta, \delta' > 0$ and n sufficiently large, it holds that for any $Z \in \mathcal{E}_{Z|\mathbf{Z}}$,

$$\begin{aligned} \text{for all } i \neq j: \quad & (z_i, z_j) \in N(\delta, \delta'), \\ \text{for all } i: \quad & (z_i, Z) \in N(\delta, \delta'). \end{aligned}$$

As a result, we can make use of Assumption C.1. The proof is separated into two steps where we separately show the two statements in Lemma C.6.

Proof of the first statement We construct a polynomial $p(Z)$ using the power series expansion of k from Assumption A.1 and in addition the Taylor series approximation of g_l around the point $(1,1)$. For any n sufficiently large, we can write

$$\begin{aligned} \text{for all } i \;\; k(z_i, Z) &= \sum_{l=0}^{s} (z_i^\top Z)^l g_l(\|z_i\|_2^2, \|z_j\|_2^2) + (z_i^\top Z)^{s+1} r(\|z_i\|_2^2, \|Z\|_2^2, z_i^\top Z) \\ &= \underbrace{\sum_{l=0}^{s} (z_i^\top Z)^l \sum_{l_1+l_2 \leq s-l} \frac{g_l^{(l_1,l_2)}(1,1)}{l_1! l_2!} (z_i^\top z_i - 1)^{l_1} (Z^\top Z - 1)^{l_2}}_{=:(v_Z)_i} \\ &+ \sum_{l=0}^{s} (z_i^\top Z)^l \sum_{l_1+l_2 = s+1-l} \frac{g_l^{(l_1,l_2)}(\eta_{l_1,l_2}^{l,i})}{l_1! l_2!} (z_i^\top z_i - 1)^{l_1} (Z^\top Z - 1)^{l_2} \\ &+ (z_i^\top Z)^{s+1} r(\|z_i\|_2^2, \|Z\|_2^2, z_i^\top Z). \end{aligned} \tag{25}$$

where $\eta_{l_1,l_2}^{l,i} \in B_r(1,1)$ are points contained in the closed ball around the point $(1,1)$ with radius $r^2 = (\|z_i\|_2^2 - 1)^2 + (\|Z\|_2^2 - 1)^2 \to 0$. Hence, using the fact that g_i is $s+1-i$-times continuously differentiable, we can see that any $\|g_l^{(l_1,l_2)}(\eta_{l_1,l_2}^{l,i})\|$ is almost surely upper bounded by some constant.
Let v_Z be the vector defined in Equation (25) We define the polynomial p as

$$p(Z) := f^\star(\mathbf{X})^\top (K + \lambda I_n)^{-1} v_Z$$

Note that p is a linear combination of the terms $(Z^\top Z)^{p_1} (z_i^\top Z)^{p_2}$ with $p_1 + p_2 \leq s$, and hence a polynomial of Z of degree at most $2s$. If g_l are constant, i.e. the kernel is an inner product kernel, v_Z contains only the terms $(z_i^\top Z)^l$ and hence $p(Z)$ is a polynomial of Z of degree at most s.
Next, because by assumption $|f^\star| \leq C_{f^\star}$ is bounded on the support of $\mathbb{P}_X$ by some constant $C_{f^\star}$, all entries of $f^\star(\mathbf{X})$ are bounded, and hence,

$$\left| \mathbb{E}_Y \hat{f}_\lambda(\sqrt{\tau} Z) - p(Z) \right| = \left| f^\star(\mathbf{X})^\top (K + \lambda I_n)^{-1} (k_Z - v_Z) \right| \leq C_{f^\star} \|(K + \lambda I_n)^{-1} (k_Z - v_Z)\|_1, \tag{26}$$

where we have used that $\mathbb{E}_Y \hat{f}_\lambda(\sqrt{\tau} Z) = f^\star(\mathbf{X})^\top (K + \lambda I_n)^{-1} k_Z$. Further, by assumption $\lambda_{\min}(K + \lambda I_n) \geq c_{\min} > 0$, and hence,

$$\|(K + \lambda I_n)^{-1} (k_Z - v_Z)\|_1 \leq \sqrt{n} \|(K + \lambda I_n)^{-1} (k_Z - v_Z)\|_2 \leq \frac{\sqrt{n}}{c_{\min}} \|k_Z - v_Z\|_2 \leq \frac{n}{c_{\min}} \max_i |(v_Z)_i - (k_Z)_i|.$$

Equation (25) yields

$$\begin{aligned} n \max_i |(v_Z)_i - (k_Z)_i| \leq n \max_i & \underbrace{\left| \sum_{q=0}^{s} (z_i^\top Z)^q \sum_{l_1+l_2 = s+1-q} \frac{g_q^{(l_1,l_2)}(\eta_{l_1,l_2}^{q,i})}{l_1! l_2!} (z_i^\top z_i - 1)^{l_1} (Z^\top Z - 1)^{l_2} \right|}_{=:B_1^i} \\ + n \max_{i \neq j} & \underbrace{\left| (z_i^\top Z)^{s+1} r(\|z_i\|_2^2, \|Z\|_2^2, z_i^\top Z) \right|}_{=:B_2^i}. \end{aligned}$$

In order to conclude the first step of the proof, we only need to show that both terms go to zero. First, we show that the term $n \max_i B_1^i \to 0$. Recall that $\left| g_q^{(l_1,l_2)}(\eta_{l_1,l_2}^{q,i}) \right|$ is upper bounded as $n \to \infty$ independent of i. Hence, we can apply Lemma C.1 which shows that for any integers q, l_1 and l_2 such that $q + l_1 + l_2 = s + 1$,

$$\left|(z_i^\top z_i - 1)^{l_1} (Z^\top Z - 1)^{l_2} (z_i^\top Z)^q\right| \lesssim (n^{-\beta/2}(\log(n))^{(1+\epsilon)/2})^{s+1}$$

which holds true for any positive constant $\epsilon > 0$. Finally, because $s = \lfloor 2/\beta \rfloor$, $(\beta/2)(s+1) > 1$, and hence

$$n \max_i B_1^i \lesssim n(n^{-\beta/2}(\log(n))^{(1+\epsilon)/2})^{s+1} \to 0.$$

Furthermore, because r is a continuous function and z_i, Z are contained in a closed neighborhood around of $(1, 1, 0)$, $r(\|z_i\|, \|Z\|, z_i^\top Z)$ is upper bounded by some constant independent of i as $n \to \infty$. Therefore, we also have that

$$n \max_i B_2^i \lesssim n \max_i \left|(z_i^\top Z)^{q+1}\right| \lesssim n \left(n^{-\beta/2}(\log(n))^{(1+\epsilon)/2}\right)^{s+1} \to 0,$$

where we have used again Lemma C.1. Hence, we can conclude the first step of the proof when observing that we have only assumed that $Z \in \mathcal{E}_{Z|\mathbf{Z}}$ and hence the convergence is uniformly.

Proof of the second statement We can see from the definition of p and the subsequent discussion that

$$\left\| p \mathbb{1}_{Z \in \mathcal{E}_{Z|\mathbf{Z}}^{\mathsf{c}}} \right\|_{\mathcal{L}_2(\mathbb{P}_Z)} = \left\| f^\star(\mathbf{X})^\top (K + \lambda I_n)^{-1} v_Z \mathbb{1}_{Z \in \mathcal{E}_{Z|\mathbf{Z}}^{\mathsf{c}}} \right\|_{\mathcal{L}_2(\mathbb{P}_Z)} \lesssim n \max_i \left\| (v_Z)_i \mathbb{1}_{Z \in \mathcal{E}_{Z|\mathbf{Z}}^{\mathsf{c}}} \right\|_{\mathcal{L}_2(\mathbb{P}_Z)},$$

and furthermore,

$$\max_i \left\| (v_Z)_i \mathbb{1}_{Z \in \mathcal{E}_{Z|\mathbf{Z}}^{\mathsf{c}}} \right\|_{\mathcal{L}_2(\mathbb{P}_Z)} \lesssim \sum_{q+l_1+l_2 \leq s} \left\| (z_i^\top Z)^q (z_i^\top z_i - 1)^{l_1} (Z^\top Z - 1)^{l_2} \mathbb{1}_{Z \in \mathcal{E}_{Z|\mathbf{Z}}^{\mathsf{c}}} \right\|_{\mathcal{L}_2(\mathbb{P}_Z)}.$$

We can decompose for n sufficiently large,

$$\begin{aligned}
&\left\| (z_i^\top Z)^q (\|z_i\|_2^2 - 1)^{l_1} \left(\|Z\|_2^2 - 1 \right)^{l_2} \mathbb{1}_{Z \in \mathcal{E}_{Z|\mathbf{Z}}^{\mathsf{c}}} \right\|_{\mathcal{L}_2(\mathbb{P}_Z)} \\
&\leq \left\| ((\|z_i\|_2^2 - 1) + 1)^{\frac{q}{2}} \|Z\|_2^q (\|z_i\|_2^2 - 1)^{l_1} \left(\|Z\|^2 - 1 \right)^{l_2} \mathbb{1}_{Z \in \mathcal{E}_{Z|\mathbf{Z}}^{\mathsf{c}}} \right\|_{\mathcal{L}_2(\mathbb{P}_Z)} \\
&\lesssim \left\| \|Z\|_2^q \left(\|Z\|_2^2 - 1 \right)^{l_2} \mathbb{1}_{Z \in \mathcal{E}_{Z|\mathbf{Z}}^{\mathsf{c}}} \mathbb{1}_{\|Z\|_2^2 \leq 2} \right\|_{\mathcal{L}_2(\mathbb{P}_Z)} + \left\| \|Z\|_2^q \left(\|Z\|_2^2 - 1 \right)^{l_2} \mathbb{1}_{\|Z\|_2^2 > 2} \right\|_{\mathcal{L}_2(\mathbb{P}_Z)},
\end{aligned}$$

where we have used Lemma C.1 in the second inequality. The first term vanishes trivially from the concentration inequality. Indeed, for n sufficiently large, we have that

$$\left\| \|Z\|_2^q \left(\|Z\|_2^2 - 1 \right)^{l_2} \mathbb{1}_{Z \in \mathcal{E}_{Z|\mathbf{Z}}^{\mathsf{c}}} \mathbb{1}_{\left\{\|Z\|_2^2 \leq 2\right\}} \right\|_{\mathcal{L}_2(\mathbb{P}_Z)} \lesssim P(\mathcal{E}_{Z|\mathbf{Z}}^{\mathsf{c}} | \mathcal{E}_{\mathbf{Z}}) \lesssim (n+1)^2 \exp(-C(\log(n))^{1+\epsilon}).$$

For the second term, note that we can see from the proof of Lemma E.1 that there exists some constant $c > 0$, such that for n sufficiently large,

$$P\left(\|Z\|_2^2 > r \right) \leq \exp(-cn^\beta r).$$

We can now apply integration by parts to show that

$$\begin{aligned}
&\left\| \|Z\|_2^q \left(\|Z\|_2^2 - 1 \right)^{l_2} \mathbb{1}_{\left\{\|Z\|_2^2 > 2\right\}} \right\|_{\mathcal{L}_2(\mathbb{P}_Z)}^2 \leq \int \|Z\|_2^{q+2l_2} \mathbb{1}_{\left\{\|Z\|_2^2 > 2\right\}} dP \\
&\lesssim \left[r^{q+2l_2} P(\|Z\|_2^2 > r) \right]_2^\infty - \int_2^\infty r^{q+2l_2-1} P\left(\|Z\|_2^2 > r \right) dr \lesssim \exp(-2cn^\beta)
\end{aligned}$$

Hence, combining these terms, we get the desired result

$$n \max_i \left\| (v_Z)_i \mathbb{1}_{Z \in \mathcal{E}_{Z|\mathbf{Z}}^{\mathsf{c}}} \right\|_{\mathcal{L}_2(\mathbb{P}_Z)} \lesssim n(n+1)^2 \exp(-C(\log(n))^{1+\epsilon}) \to 0 \text{ as } n \to \infty. \tag{27}$$

□

E.3. Proof of Lemma C.7

As in (El Karoui et al., 2010), we separately analyze the off and on diagonal terms of K. Let A be the off-diagonal matrix of K, with diagonal entries $A_{i,j} = (1-\delta_{i,j})K_{i,j}$ and let D be the diagonal matrix of K with entries $D_{i,j} = \delta_{i,j}K_{i,j}$. We have

$$\begin{aligned} \text{for all } i: \quad & D_{i,i} := K_{i,i} = g(\|z_i\|_2^2, \|z_i\|_2^2, z_i^\top z_i), \\ \text{for all } i \neq j: \quad & A_{i,j} := K_{i,j} = g(\|z_i\|_2^2, \|z_j\|_2^2, z_i^\top z_j). \end{aligned}$$

Similarily, decompose M into its off-diagonal, M_A, and its diagonal M_D. We have

$$||K - M||_{\text{op}} \leq ||A - M_A||_{\text{op}} + ||D - M_D||_{\text{op}} \tag{28}$$

We begin with the first term. Note that M_A has off-diangoal entries $(M_A)_{i,j} := \sum_{q=0}^m (z_i^\top z_j)^q g_q(\|z_i\|_2^2, \|z_j\|_2^2)$, and hence,

$$||M_A - A||_{\text{op}} \leq \|M_A - A\|_F \leq n \max_{i,j} |(M_A)_{i,j} - A_{i,j}| \lesssim n \max_{i,j} |(z_j^\top z_j - 1)^{m+1}| \to 0,$$

where we have used the same argument as used in the proof of Lemma C.6 and the fact that the Assumptions A.1-A.3 imply Assumption C.1, as shown in Lemma E.2.
Next, note that we can write

$$M_D := \left[g(1,1,1) - \sum_{q=0}^m g_q(1,1) \ + \sum_{q=0}^m \|z_i\|_2^{2q} g_q(\|z_i\|_2^2, \|z_i\|_2^2) \right] \mathrm{I}_n$$

Because $D - M_D$ is a diagonal matrix, for n sufficiently large,

$$\begin{aligned} ||D - M_D||_{\text{op}} &= \max_i \left| g(\|z_i\|_2^2, \|z_i\|_2^2, z_i^\top z_i) - g(1,1,1) + \sum_{q=0}^m g_q(1,1) \ - \sum_{q=0}^m \|z_i\|_2^{2q} g_q(\|z_i\|_2^2, \|z_i\|_2^2) \right| \\ &\leq \underbrace{\delta_L \sqrt{3}(\|z_i\|_2^2 - 1)}_{T_1} + \sum_{q=0}^m \underbrace{[\|z_i\|_2^{2q} - 1] g_q(\|z_i\|_2^2, \|z_i\|_2^2)}_{T_2} + \underbrace{[g_q(\|z_i\|_2^2, \|z_i\|_2^2) - g_q(1,1)]}_{T_3} \end{aligned}$$

where we have used that by assumption g is δ_L-Lipschitz continuous on the restriction $\{(x,x,x) | x \in [1-\delta_L, 1+\delta_L]\} \subset \Omega$ for some $\delta_L > 0$. Clearly $T_1 \to 0$ due to Lemma C.1. Furthermore, by Assumption C.1, for any $q \leq m$, g_q is continuously differentiable and hence also Lipschitz continuous in a closed ball around $(1,1)$. Thus, $T_3 \to 0$. Hence, it is only left to show that $T_2 \to 0$, which is a consequence of the following claim.

Claim: *For any $\epsilon > 0$ and any $q > 0$,*

$$\max_i \left| \frac{(x_i^\top x_i)^q}{(\mathrm{tr}(\Sigma_d))^q} - 1 \right| \leq c_q \max[n^{-\beta/2} (\log(n))^{(1+\epsilon)/2}, n^{-q\beta/2} (\log(n))^{q((1+\epsilon)/2)}]$$

where c_q is a constant only depending on q.

Proof of the claim: In order to prove the claim, recall that due to Lemma C.1, for every $q > 0$

$$\max_i \left| \left[\frac{(x_i^\top x_i)}{\mathrm{tr}(\Sigma_d)} - 1 \right]^q \right| \leq (n^{-\beta/2} (\log(n))^{(1+\epsilon)/2})^q.$$

We prove the claim by induction. The case where $q = 1$ holds trivially with $c_1 = 1$. For $q > 1$,

$$\max_i \left| \left[\frac{(x_i^\top x_i)}{\mathrm{tr}(\Sigma_d)} - 1 \right]^q \right| = \max_i \left| \frac{(x_i^\top x_i)^q}{(\mathrm{tr}(\Sigma_d))^q} + \sum_{j=1}^q (-1)^j \binom{q}{j} \left(\frac{x_i^\top x_i}{\mathrm{tr}(\Sigma_d)} \right)^{q-j} \right| \leq (n^{-\beta/2} (\log(n))^{(1+\epsilon)/2})^q.$$

Next, by induction, $\max_i \left|\frac{(x_i^\top x_i)^j}{(\mathrm{tr}(\Sigma_d))^j} - 1\right| \leq c_1 \max\left[n^{-\beta/2}(\log(n))^{(1+\epsilon)/2}, n^{-j\beta/2}(\log(n))^{j((1+\epsilon)/2)}\right]$ for any $j < q$. Furthermore, $\sum_{j=1}^q (-1)^j \binom{q}{j} = -1$. Hence,

$$\begin{aligned}
\max_i \left|\frac{(x_i^\top x_i)^j}{(\mathrm{tr}(\Sigma_d))^q} - 1\right| &= \max_i \left|\frac{(x_i^\top x_i)^q}{(\mathrm{tr}(\Sigma_d))^q} + \sum_{j=1}^q (-1)^j \binom{q}{j} \left(\frac{(x_i^\top x_i)^q}{(\mathrm{tr}(\Sigma_d))^q}\right)^{q-j}\right| \\
= \max_i \left|\frac{(x_i^\top x_i)^q}{(\mathrm{tr}(\Sigma_d))^q} + \sum_{j=1}^q (-1)^j \binom{q}{j} \left[\left(\frac{(x_i^\top x_i)^q}{(\mathrm{tr}(\Sigma_d))^q}\right)^{q-j} - 1\right] + \sum_{j=1}^q (-1)^j \binom{q}{j}\right| \\
\geq \max_i \left|\frac{(x_i^\top x_i)^q}{(\mathrm{tr}(\Sigma_d))^q} - 1\right| - \sum_{j=1}^q \binom{q}{j} c_j \max\left[n^{-\beta/2}(\log(n))^{(1+\epsilon)/2}, n^{-j\beta/2}(\log(n))^{j((1+\epsilon)/2)}\right].
\end{aligned}$$

As a result, we have

$$\max_i \left|\frac{(x_i^\top x_i)^q}{(\mathrm{tr}(\Sigma_d))^q} - 1\right| \leq (n^{-\beta/2}(\log(n))^{(1+\epsilon)/2})^q + \sum_{j=1}^q \binom{q}{j} c_j \max\left[n^{-\beta/2}(\log(n))^{(1+\epsilon)/2}, n^{-j\beta/2}(\log(n))^{j((1+\epsilon)/2)}\right],$$

which completes the induction and thus the proof □

E.4. Proof of Lemma C.8

We use the well known formula

$$t^\alpha = c_\alpha \int_0^\infty (1 - e^{-t^2 x^2}) x^{-1-\alpha} \,\mathrm{d}\, x$$

with

$$c_\alpha := \left(\int_0^\infty (1 - e^{-x^2}) x^{-1-\alpha} \,\mathrm{d}\, x\right)^{-1} > 0,$$

which holds for all $t \geq 0$. Hence, for $t := \|z_i - z_j\|_2^\alpha \geqslant 0$, we can write

$$\|z_i - z_j\|_2^\alpha = c_\alpha \int_0^\infty (1 - e^{-x^2 \|z_i - z_j\|_2^2}) x^{-1-\alpha} \,\mathrm{d}\, x.$$

We first study $\mu \in \mathbb{R}^n$ with $\sum_{1 \leqslant i \leqslant n} \mu_i = 0$. We have

$$\begin{aligned}
\mu^\top D_\alpha \mu = \sum_{1 \leqslant i,j \leqslant n} \|z_i - z_j\|_2^\alpha \mu_i \mu_j &= \sum_{1 \leqslant i,j \leqslant n} \mu_i \mu_j c_\alpha \int_0^\infty (1 - e^{-x^2 \|z_i - z_j\|_2^2}) x^{-1-\alpha} \,\mathrm{d}\, x \\
&= -c_\alpha \int_0^\infty x^{-1-\alpha} \sum_{1 \leqslant i,j \leqslant n} \mu_i \mu_j e^{-x^2 \|z_i - z_j\|_2^2} \,\mathrm{d}\, x.
\end{aligned}$$

Next, note that the Gaussian kernel satisfies Assumptions A.1-A.3 since
$\exp(-\|x - x'\|_2^2) = \sum_{j=0}^\infty \frac{2^j}{j!} (x^\top x')^j \exp(-\|x\|_2^2 - \|x'\|_2^2)$. Hence, we can conclude from Proposition C.3 that for every $x \in \mathbb{R}_+$ there exists a constant $c_{G,x}$, such that $\sum_{1 \leqslant i,j \leqslant n} \mu_i \mu_j e^{-x^2 \|z_i - z_j\|_2^2} \geq \|\mu\|_2^2 c_{G,x} > 0$ almost surely as $n \to \infty$. Thus, we can conclude that there exists a constant $\tilde{c} > 0$ independent of n, such that

$$\mu^\top D_\alpha \mu \leq -\tilde{c} \|\mu\|_2^2 \;\; a.s. \text{ as } n \to \infty. \tag{29}$$

Because the nullspace of the vector 1 spans a $n-1$ dimensional subspace, we can apply the Courant–Fischer–Weyl min-max principle from which we can see that the second largest eigenvalue of the matrix D_α satisfies $\lambda_2 \leq -\tilde{c}$ almost surely as $n \to \infty$. Since the sum of the eigenvalues $\sum_{i=1}^n \lambda_i = \mathrm{tr}(D_\alpha) = 0$, $\lambda_1 > (n-1)\tilde{c}$ almost surely as $n \to \infty$, which concludes the proof. □

E.5. Proof of Lemma C.10

We start the proof with a discussion of existing results in the literature. As shown in Appendix E.1 in (Arora et al., 2019), the homogeneity of σ allows us to write

$$\Sigma^{(i)}(x,x') = c_\sigma \left(\Sigma^{(i-1)}(x,x)\Sigma^{(i-1)}(x',x')\right)^{k/2} t_\sigma\left(\frac{\Sigma^{(i-1)}(x,x')}{\sqrt{\Sigma^{(i-1)}(x,x)\Sigma^{(i-1)}(x',x')}}\right), \tag{30}$$

with

$$t_\sigma(\rho) = \mathop{\mathbb{E}}_{(u,v)\sim\mathcal{N}(0,\tilde{\Lambda}(\rho))}[\sigma(u)\sigma(v)]$$

and $\tilde{\Lambda}(\rho) = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Furthermore, because $\dot{\sigma}$ is a $k-1$-homogeneous, we can analogously write

$$\dot{\Sigma}^{(i)} = c_{\dot{\sigma}} \left(\Sigma^{(i-1)}(x,x)\Sigma^{(i-1)}(x',x')\right)^{(k-1)/2} t_{\dot{\sigma}}\left(\frac{\Sigma^{(i-1)}(x,x')}{\sqrt{\Sigma^{(i-1)}(x,x)\Sigma^{(i-1)}(x',x')}}\right),$$

with $t_{\dot{\sigma}}(\rho) = \mathop{\mathbb{E}}_{(u,v)\sim\mathcal{N}(0,\tilde{\Lambda}(\rho))}[\dot{\sigma}(u)\dot{\sigma}(v)]$. The function t_σ is called the dual of the activation function (Definition 4 in (Daniely et al., 2016)). In particular, since by assumption σ and $\dot{\sigma}$ have a Hermite polynomial extension, Lemma 11 in (Daniely et al., 2016) provides some useful properties which hold for both t_σ and $t_{\dot{\sigma}}$. We only state them for t_σ:

1. Let $a_i \in \mathbb{R}$ be the coefficients of the Hermite polynomial extension of σ, then $t_\sigma(\rho) = \sum_{i=0}^{\infty} a_i^2 \rho^i$
2. The function t_σ is continuous in $[-1,1]$ and smooth in $(-1,1)$
3. The image of t_σ is $[-\gamma,\gamma]$ with $\gamma = \mathop{\mathbb{E}}_{v\sim\mathcal{N}(0,1)}\left[\sigma(v)^2\right] = 1/c_\sigma$
4. We have that $t_\sigma(1) = \mathop{\mathbb{E}}_{v\sim\mathcal{N}(0,1)}\left[\sigma(v)^2\right] = c_\sigma^{-1}$

Based on this discussion, we now prove the lemma. In a first step, we derive a closed form expression for $\Sigma^{(i)}(x,x)$. Based on the discussion above and particularly Equation (30) we can see that

$$\Sigma^{(i)}(x,x) = \left(\Sigma^{(i-1)}(x,x)\right)^k,$$

and by induction, we get that

$$\Sigma^{(i)}(x,x) = (x^\top x)^{k^i}. \tag{31}$$

Therefore, Equation (30) becomes

$$\Sigma^{(i)}(x,x') = c_\sigma (x^\top x)^{\frac{k^i}{2}} (x'^\top x')^{\frac{k^i}{2}} t_\sigma\left(\frac{\Sigma^{(i-1)}(x,x')}{(x^\top x)^{\frac{k^{i-1}}{2}}(x'^\top x')^{\frac{k^{i-1}}{2}}}\right).$$

The goal is now to show that whenever $x, x' \neq 0$, $\Sigma^{(i)}(x,x')$ can be expressed as a sum of the form

$$\Sigma^{(i)}(x,x') = \sum_{j=0}^{\infty}(x^\top x')^j \sum_{l=-\infty}^{\infty} \eta_{j,l}^{(i)}(||x||_2^2||x'||_2^2)^{l/2} \tag{32}$$

with $\eta_{j,l}^{(i)} \geq 0$. We prove by induction. In a first step, note that the case where $i=0$ holds trivially true. Next, assume that Equation (32) holds true for $\Sigma^{(i-1)}$. Due to the above discussion, t_σ can be expressed as a Taylor series around 0 with

positive coefficients a_i^2. Thus,

$$\begin{aligned}
\Sigma^{(i)}(x,x') &= c_\sigma (x^\top x)^{\frac{k^i}{2}} (x'^\top x')^{\frac{k^i}{2}} \sum_{m=0}^{\infty} a_m^2 \left(\frac{\Sigma^{(i-1)}(x,x')}{(||x||_2^2||x'||_2^2)^{\frac{k^{i-1}}{2}}} \right)^m \\
&= c_\sigma (||x||_2^2)^{\frac{k^i}{2}} (||x'||_2^2)^{\frac{k^i}{2}} \sum_{m=0}^{\infty} a_m^2 \left(\frac{\sum_{j=0}^{\infty} (x^\top x')^j \sum_{l=-\infty}^{\infty} \eta_{j,l}^{(i-1)} (||x||_2^2||x'||_2^2)^{l/2}}{(||x||_2^2||x'||_2^2)^{\frac{k^{i-1}}{2}}} \right)^m \\
&= c_\sigma (||x||_2^2)^{\frac{k^i}{2}} (||x'||_2^2)^{\frac{k^i}{2}} \sum_{m=0}^{\infty} a_m^2 \left(\sum_{j=0}^{\infty} (x^\top x')^j \sum_{l=-\infty}^{\infty} \eta_{j,l}^{(i-1)} (||x||_2^2||x'||_2^2)^{l/2-k^{i-1}/2} \right)^m \\
&= \sum_{j=0}^{\infty} (x^\top x')^j \sum_{l=-\infty}^{\infty} \eta_{j,l}^{(i)} (||x||_2^2||x'||_2^2)^{l/2}.
\end{aligned}$$

In order to guarantee that the last equation holds true, we need to show that the above multi-sum converges absolutely. To see this, first of all note that $\eta_{j,l}^{(i)} \geq 0$ because by assumption $\eta_{j,l}^{(i-1)} \geq 0$. Furthermore, given that $d > 1$, for any x, x' we can find $\tilde{x}, \tilde{x}'$ such that $||x||_2^2 = ||\tilde{x}||_2^2$, $||x'||_2^2 = ||\tilde{x}'||_2^2$ and $\tilde{x}^\top \tilde{x}' = |x^\top x'|$. Hence, the above multi-sum only consists of positive coefficients when evaluating at $\tilde{x}, \tilde{x}'$ and thus converges absolutely which completes the induction.
Next, note that any of the properties 1-4 from the above discussion also hold true for $t_{\dot{\sigma}}$. Therefore, we can use exactly the same argument for $\dot{\Sigma}^{(i)}$ to show that for any $x, x' \neq 0$,

$$\dot{\Sigma}^{(i)}(x,x') = \sum_{j=0}^{\infty} (x^\top x')^j \sum_{l=-\infty}^{\infty} \dot{\eta}_{j,l}^{(i)} (||x||_2^2||x'||_2^2)^{l/2}.$$

with $\dot{\eta}_{j,l}^{(i+1)} \geq 0$. Finally, we can conclude the proof because

$$k_{\mathrm{NTK}}(x,x') := \sum_{i=1}^{L+1} \Sigma^{(i-1)}(x,x') \prod_{j=i}^{L+1} \dot{\Sigma}^{(j)}(x,x')$$

and when using the same argument as used in the induction step above to show that the resulting multi-sum converges absolutely. □

E.6. Additional lemmas

Lemma E.2. *Any kernel which satisfies Assumption A.1 and A.3 also satisfies Assumption C.1.*

Proof. The only point which does not follow immediately is to show that r is a continuous function. For this, write g as a function of the variables x, y, z, i.e.

$$g(x,y,z) = \sum_{j=0}^{\infty} g_j(x,y) z^j.$$

For every x, y, define the function $g_{x,y}(z) = g(x,y,z)$. Due to the series expansion, we can make use of the theory on the Taylor expansion which implies that for any x, y, $g_{x,y}$ is a smooth function in the interior of $N(\delta)$ (using the definition from Assumption A.1). Hence, we can conclude that there exists a function $r_{x,y}$ such that

$$g_{x,y}(z) = \sum_{j=0}^{m} g_j(x,y) z^j + (z)^{m+1} r_{x,y}(z)$$

In particular, the smoothness of $g_{x,y}(z)$ implies that $r_{x,y}$ is continuous in the interior of $\{z : (x,y,z) \in N(\delta)\}$. Next, define $r(x,y,z) = r_{x,y}(z)$ and note that that the continuity of g implies that r is continuous everywhere except for the plane $z = 0$. Finally, because $r_{x,y}(0)$ exists point wise, we conclude that r exists and is a continuous function in the interior of $N(\delta)$. □

Lemma E.3. *Any RBF kernel $k(x,x') = h(\|x - x'\|_2^2)$ with h locally analytic around 2 satisfies Assumption C.1.*

Proof. Because by assumption h has a local Taylor series around 2, we can write

$$h(\|x - x'\|_2^2) = \sum_{j=0}^{\infty} h_j \, (\|x - x'\|_2^2 - 2)^j,$$

which converges absolutely for any $\left|\|x - x'\|_2^2 - 2\right| \leq \tilde{\delta}$ where $\tilde{\delta} > 0$ is the convergence radius of the Taylor series approximation. Next, using $\|x - x'\|_2^2 - 2 = \|x\|_2^2 - 1 + \|x'\|_2^2 - 1 - 2x^\top x'$, we can make use of the Binomial series, which gives

$$h(\|x - x'\|_2^2) = \sum_{j=0}^{\infty} h_j \sum_{i=0}^{j} \sum_{l=0}^{i} \binom{j}{i} \binom{i}{l} (\|x\|_2^2 - 1)^l (\|x'\|_2^2 - 1)^{i-l} (-2x^\top x')^{j-i}. \tag{33}$$

The goal is now to show that this multi series converges absolutely. Whenever $d > 1$, we can choose $\tilde{x}, \tilde{x}'$ from the set of convergent points such that $\|\tilde{x}\|_2^2 - 1 > 0$, $\|\tilde{x}'\|_2^2 - 1 > 0$ and $\tilde{x}^\top \tilde{x}' < 0$. As a result, we can see that for any j, the sum

$$\sum_{i=0}^{j} \sum_{l=0}^{i} \binom{j}{i} \binom{i}{l} (\|\tilde{x}\|_2^2 - 1)^l (\|\tilde{x}'\|_2^2 - 1)^{i-l} (-2\tilde{x}^\top \tilde{x}')^{j-i}$$

is a sum of non negative summands. Hence, we get that the sum

$$\sum_{j=0}^{\infty} \sum_{i=0}^{j} \sum_{l=0}^{i} h_j \binom{j}{i} \binom{i}{l} (\|\tilde{x}\|_2^2 - 1)^l (\|\tilde{x}'\|_2^2 - 1)^{i-l} (-2\tilde{x}^\top \tilde{x}')^{j-i}$$

converges absolutely. Thus, we can arbitrarily reorder the summands:

$$\begin{aligned} h(\|\tilde{x} - \tilde{x}'\|_2^2) &= \sum_{j=0}^{\infty} (\tilde{x}^\top \tilde{x}')^j \sum_{i=j}^{\infty} \sum_{l=0}^{i-j} h_i (-2)^j \binom{i}{j} \binom{i-j}{l} (\|\tilde{x}\|_2^2 - 1)^l (\|\tilde{x}'\|_2^2 - 1)^{i-l-j} \\ &=: \sum_{j=0}^{\infty} (\tilde{x}^\top \tilde{x}')^j g_j(\|\tilde{x}\|_2^2, \|\tilde{x}'\|_2^2) \end{aligned} \tag{34}$$

In particular, we obtain that the sum from Equation (34) converges absolutely for any x, x' with $\|\tilde{x}\|_2^2 - 1 > \left|\|x\|_2^2 - 1\right|$, $\|\tilde{x}'\|_2^2 - 1 > \left|\|x'\|_2^2 - 1\right|$ and $-\tilde{x}^\top \tilde{x}' > \left|x^\top x'\right|$. In fact, we can always find $\delta, \delta' > 0$ such that any $(x, x') \in \mathbb{R}^d \times \mathbb{R}^d$, with $(\|x\|_2^2, \|x'\|_2^2) \in [1 - \delta, 1 + \delta] \times [1 - \delta, 1 + \delta]$ and $x^\top x' \in [-\delta', \delta']$ satisfies these constraints and hence the sum from Equation (33) converges absolutely. We can then conclude the proof when noting that the functions g_i are implicitly defined using the Taylor series expansion. Therefore, we can conclude that g_i are smooth functions in a neighborhood of $(1, 1)$. The rest of the proof follows then trivially. □