This is the supplementary material for the paper: "Kernel regression in high dimensions: Refined analysis beyond double descent", by Fanghui Liu, Zhenyu Liao, and Johan A.K. Suykens. The supplementary material (Appendix) is organized as follows.

- Section A provides high dimensional linearizations of some typical smooth kernels as concrete examples of Table 2.
- In Section B, we demonstrate that, a kernel matrix in high dimensions admits the same eigenvalue decay as $\widetilde{\boldsymbol{X}}$ and $\boldsymbol{X}\boldsymbol{X}^\top/d$.
- Our proof framework includes the error decomposition in Section C, the error bound for the bias in Section D and for the variance in Section E, respectively.
- Section F discusses the quantity function $\mathcal{N}_{\widetilde{\boldsymbol{X}}}^{n\lambda+\gamma}$ based on three eigenvalue decays: *harmonic decay*, *polynomial decay*, and *exponential decay* in the $n < d$ and $n > d$ regimes.
- Some additional experiments are presented in Section G to further validate our theoretical results.

A Examples of kernels and their linearizations

In this section, we present linearization of some typical kernels by Eq. (5). Here we assume that $\alpha, \beta, \gamma \geq 0$ to ensure the positive definiteness of the approximated kernel matrix $\widetilde{\boldsymbol{K}}^{\text{lin}}$. Table 4 reports the results of three inner-product kernels including polynomial kernel, linear kernel, exponential kernel; as well as a radial kernel: the common-used Gaussian kernel. We can find that $\alpha, \gamma \geq 0$. Specifically, $\beta > 0$ avoids a trivial solution.

Table 4: Linearizations of typical kernels in high dimensions.

kernel	formulation	α	β	γ
polynomial kernels	$k(\boldsymbol{x}, \boldsymbol{x}') := \left(1 + \frac{1}{d}\langle \boldsymbol{x}, \boldsymbol{x}'\rangle\right)^p$	$1 + p(p-1)\frac{\text{tr}\left(\boldsymbol{\Sigma}_d^2\right)}{2d^2}$	p	$(1+\tau)^p - 1 - p\tau$
linear kernel	$k(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{d}\langle \boldsymbol{x}, \boldsymbol{x}'\rangle$	0	1	0
exponential kernel	$k(\boldsymbol{x}, \boldsymbol{x}') = \exp(\frac{2}{d}\langle \boldsymbol{x}, \boldsymbol{x}'\rangle)$	$1 + 2\frac{\text{tr}\left(\boldsymbol{\Sigma}_d^2\right)}{d^2}$	2	$\exp(2\tau) - 1 - 2\tau$
Gaussian kernel	$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{1}{d}\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2\right)$	$\exp(-2\tau)\left[1 + 2\frac{\text{tr}\left(\boldsymbol{\Sigma}_d^2\right)}{d^2}\right]$	$2\exp(-2\tau)$	$1 - 2\tau\exp(-2\tau) - \exp(-2\tau)$

B Eigenvalue decay equivalence

In this section, we demonstrate that, in high dimensions, a kernel matrix induced by inner-product kernels or radial kernels admits the same eigenvalue decay as $\widetilde{\boldsymbol{X}} = \beta\boldsymbol{X}\boldsymbol{X}^\top/d + \alpha\mathbf{1}\mathbf{1}^\top$ and $\boldsymbol{X}\boldsymbol{X}^\top/d$.

For notational simplicity, denote the inner-product kernel matrix $\boldsymbol{K}_{\text{inner}}$ and its linearization $\widetilde{\boldsymbol{K}_{\text{inner}}^{\text{lin}}}$; the radial kernel matrix $\boldsymbol{K}_{\text{radial}}$ and its linearization $\widetilde{\boldsymbol{K}_{\text{radial}}^{\text{lin}}}$.

Proposition 2. *The inner-product kernel matrix $\boldsymbol{K}_{\text{inner}}$ admits the same eigenvalue decay as $\widetilde{\boldsymbol{X}}$ and $\boldsymbol{X}\boldsymbol{X}^\top/d$.*

Proof. According to Theorem 2.1 in [26], the inner-product kernel matrix $\boldsymbol{K}_{\text{inner}}$ can be well approximated by $\widetilde{\boldsymbol{K}_{\text{inner}}^{\text{lin}}}$ with

$$\widetilde{\boldsymbol{K}_{\text{inner}}^{\text{lin}}} := \beta\frac{\boldsymbol{X}\boldsymbol{X}^\top}{d} + \gamma\boldsymbol{I} + \alpha\mathbf{1}\mathbf{1}^\top,$$

in a spectral norm sense, where α, β, γ are given in Table 2. As a result, with high probability, the inner-product kernel matrix $\boldsymbol{K}_{\text{inner}}$ and its linearization $\widetilde{\boldsymbol{K}_{\text{inner}}^{\text{lin}}}$ has the same eigenvalue. That means, $\boldsymbol{K}_{\text{inner}}$ admits the same eigenvalue decay as $\widetilde{\boldsymbol{X}} := \beta\boldsymbol{X}\boldsymbol{X}^\top/d + \alpha\mathbf{1}\mathbf{1}^\top$ via a constant shift γ.

Next, we shall demonstrate that $\boldsymbol{K}_{\text{inner}}$ admits the same eigenvalue decay as $\boldsymbol{X}\boldsymbol{X}^\top/d$. Since $\mathbf{1}\mathbf{1}^\top$ is a rank-one matrix with $\lambda_1(\mathbf{1}\mathbf{1}^\top) = n$, with Weyl's inequality and $\lambda_n \leq \lambda_{n-1} \leq \ldots \leq \lambda_1$, we have

$$\beta\lambda_1\left(\frac{\boldsymbol{X}\boldsymbol{X}^\top}{d}\right) + \gamma \leq \lambda_1(\widetilde{\boldsymbol{K}_{\text{inner}}^{\text{lin}}}) \leq \beta\lambda_1\left(\frac{\boldsymbol{X}\boldsymbol{X}^\top}{d}\right) + \gamma + \alpha n,$$

and

$$\beta\lambda_i\left(\frac{\boldsymbol{X}\boldsymbol{X}^\top}{d}\right) + \gamma \leq \lambda_i(\widetilde{\boldsymbol{K}_{\text{inner}}^{\text{lin}}}) \leq \beta\lambda_{i-1}\left(\frac{\boldsymbol{X}\boldsymbol{X}^\top}{d}\right) + \gamma\,, \quad i = 2, 3, \ldots n\,,$$

so that the eigenvalue of $\widetilde{\boldsymbol{K}_{\text{inner}}^{\text{lin}}}$ *interlaced* with those of $\beta\boldsymbol{X}\boldsymbol{X}^\top/d + \gamma\boldsymbol{I}$. We can thus conclude that the eigenvalue decay of $\widetilde{\boldsymbol{K}_{\text{inner}}^{\text{lin}}}$ is the same as that of $\boldsymbol{X}\boldsymbol{X}^\top/d$ with a constant shift and scaling, which do not effect the trend of eigenvalue decay. Accordingly, the inner-product-type kernel matrix $\boldsymbol{K}_{\text{inner}}$ and its linearization $\widetilde{\boldsymbol{K}_{\text{inner}}^{\text{lin}}}$, $\widetilde{\boldsymbol{X}}$ admit the same eigenvalue decay as $\boldsymbol{X}\boldsymbol{X}^\top/d$, which concludes the proof. □

Proposition 2 also provides a justification to study the eigenvalue decay of a radial kernel matrix. According to Theorem 2.2 in [26], the radial kernel matrix $\boldsymbol{K}_{\text{radial}}$ can be well approximated by $\widetilde{\boldsymbol{K}_{\text{radial}}^{\text{lin}}}$ with

$$\widetilde{\boldsymbol{K}_{\text{radial}}^{\text{lin}}} := \beta\frac{\boldsymbol{X}\boldsymbol{X}^\top}{d} + \gamma\boldsymbol{I} + \alpha\mathbf{1}\mathbf{1}^\top + h'(2\tau)\boldsymbol{A} + \frac{1}{2}h''(2\tau)\boldsymbol{A}\odot\boldsymbol{A}\,,$$

in a spectral norm sense, where α, β, γ are given in Table 2. Recall $\boldsymbol{A} := \mathbf{1}\boldsymbol{\psi}^\top + \boldsymbol{\psi}\mathbf{1}^\top$, where $\boldsymbol{\psi} \in \mathbb{R}^n$ with $\psi_i := \|\boldsymbol{x}_i\|_2^2/d - \tau$, we find that $\boldsymbol{A}$ is a rank 2 matrix with its eigenvalues $\lambda(\boldsymbol{A}) = \mathbf{1}^\top\boldsymbol{\psi} \pm \sqrt{n}\|\boldsymbol{\psi}\|_2$, and thus we have $\text{rank}(\boldsymbol{A}\odot\boldsymbol{A}) = 3$.[3] Hence, by virtue of Proposition 2, apart from the top 5 eigenvalues of the radial kernel matrix $\boldsymbol{K}_{\text{radial}}$, its remaining eigenvalues follow with

$$\beta\lambda_i\left(\frac{\boldsymbol{X}\boldsymbol{X}^\top}{d}\right) + \gamma \leq \lambda_i(\widetilde{\boldsymbol{K}_{\text{radial}}^{\text{lin}}}) \leq \beta\lambda_{i-1}\left(\frac{\boldsymbol{X}\boldsymbol{X}^\top}{d}\right) + \gamma\,, \quad i = 6, 7, \ldots n\,.$$

Accordingly, $\boldsymbol{K}_{\text{radial}}$ admits the same eigenvalue decay as $\boldsymbol{X}\boldsymbol{X}^\top/d$.

C Proof of Lemma 1

Proof. By virtue of the closed form of the KRR estimator in Eq. (3) and $\boldsymbol{\epsilon} := \boldsymbol{y} - f_\rho(\boldsymbol{X})$, we have

$$f_{\boldsymbol{z},\lambda}(\boldsymbol{x}) - f_\rho(\boldsymbol{x}) = k(\boldsymbol{x},\boldsymbol{X})^\top(\boldsymbol{K}+n\lambda\boldsymbol{I})^{-1}\boldsymbol{\epsilon} + k(\boldsymbol{x},\boldsymbol{X})^\top(\boldsymbol{K}+n\lambda\boldsymbol{I})^{-1}f_\rho(\boldsymbol{X}) - f_\rho(\boldsymbol{x})\,,$$

where $f_\rho(\boldsymbol{X}) = [f_\rho(\boldsymbol{x}_1), f_\rho(\boldsymbol{x}_1), \cdots, f_\rho(\boldsymbol{x}_n)]^\top \in \mathbb{R}^n$. According to $\mathbb{E}_{y|\boldsymbol{x}}[\boldsymbol{\epsilon}] = 0$, we then have

$$\mathbb{E}_{y|\boldsymbol{x}}\big\|f_{\boldsymbol{z},\lambda} - f_\rho\big\|_{\mathcal{L}^2_{\rho_X}}^2 = \mathbb{E}_{\boldsymbol{x}}\big\|k(\boldsymbol{x},\cdot)^\top(\boldsymbol{K}+n\lambda\boldsymbol{I})^{-1}f_\rho(\boldsymbol{X}) - f_\rho\big\|_{\mathcal{L}^2_{\rho_X}}^2 + \mathbb{E}_{y,\boldsymbol{x}}\big\|k(\boldsymbol{x},\cdot)^\top(\boldsymbol{K}+n\lambda\boldsymbol{I})^{-1}\boldsymbol{\epsilon}\big\|_{\mathcal{L}^2_{\rho_X}}^2\,.$$

Based on the definition of $\mathtt{B}$, we decompose $\mathtt{B}$ as

$$\begin{aligned}\mathtt{B} &:= \mathbb{E}_{\boldsymbol{x}}\big\|k(\boldsymbol{x},\cdot)^\top(\boldsymbol{K}+n\lambda\boldsymbol{I})^{-1}f_\rho(\boldsymbol{X}) - f_\rho\big\|_{\mathcal{L}^2_{\rho_X}}^2 = \|f_{\boldsymbol{X},\lambda} - f_\rho\|_{\mathcal{L}^2_{\rho_X}}^2 \\ &\leq 2\|f_{\boldsymbol{X},\lambda} - f_\lambda\|_{\mathcal{L}^2_{\rho_X}}^2 + 2\|f_\lambda - f_\rho\|_{\mathcal{L}^2_{\rho_X}}^2\,,\end{aligned}$$

which concludes our proof. □

D Proof for the bias

The error bound for the bias is given by the following theorem.

[3] This can be proved using rank-one decomposition of $\boldsymbol{A}$.

Theorem 3. *(Bias) Under Assumption 4 (source condition with $0 < r \leq 1$), Assumption 5 (capacity condition with $0 \leq \eta \leq 1$), let $0 < \delta < 1/2$, taking the regularization parameter $\lambda := \bar{c}n^{-\vartheta}$ with $0 \leq \vartheta \leq \frac{1}{1+\eta}$, there holds with probability at least $1 - 2\delta$, we have*

$$\mathtt{B} \leq 2\left(\|f_{\boldsymbol{X},\lambda} - f_\lambda\|^2_{\mathcal{L}^2_{\rho_X}} + \|f_\lambda - f_\rho\|^2_{\mathcal{L}^2_{\rho_X}}\right) \lesssim n^{-2\vartheta r}\log^4\left(\frac{2}{\delta}\right).$$

In our error decomposition, $\|f_\lambda - f_\rho\|^2_{\mathcal{L}^2_{\rho_X}}$ is independent of data $\boldsymbol{X}$ that corresponds to the approximation error in learning theory [30]; while the first term $\|f_{\boldsymbol{X},\lambda} - f_\lambda\|^2_{\mathcal{L}^2_{\rho_X}}$ depends on $\boldsymbol{X}$, termed as bias-sample error. To prove Theorem 3, we need to bound the approximation error and the bias-sample error as follows.

D.1 Bound approximation error

In learning theory, the approximation error $\|f_\lambda - f_\rho\|_{\mathcal{L}^2_{\rho_X}}$ can be estimated by the source condition in Assumption 4.

Lemma 2. *(Lemma 3 in [61]) Under the source condition in Assumption 4 with $0 < r \leq 1$, the approximation error can be given by*

$$\|f_\lambda - f_\rho\|_{\mathcal{L}^2_{\rho_X}} = \|(L_K + \lambda I)^{-1}L_K f_\rho - f_\rho\|_{\mathcal{L}^2_{\rho_X}} \leq \lambda^r \|L_K^{-r} f_\rho\|_{\mathcal{L}^2_{\rho_X}} \leq R\lambda^r.$$

D.2 Bound bias-sample error

To bound the bias-sample error $\|f_{\boldsymbol{X},\lambda} - f_\lambda\|_{\mathcal{L}^2_{\rho_X}}$, we need the following lemma.

Lemma 3. *(Lemma 17 in [62]) For any $0 < \delta < 1$, it holds with probability at least $1 - \delta$ that*

$$\|(L_K + \lambda I)^{-1/2}(L_K - L_{K,\boldsymbol{X}})\| \leq \frac{2\kappa}{\sqrt{n}}\left\{\frac{\kappa}{\sqrt{n\lambda}} + \sqrt{\mathcal{N}(\lambda)}\right\}\log\left(\frac{2}{\delta}\right),$$

where $\kappa := \max\{1, \sup_{\boldsymbol{x}\in X}\sqrt{k(\boldsymbol{x},\boldsymbol{x})}\}$.

Then the bias-sample error can be decomposed into several parts.

Lemma 4. *Under Assumption 4, we have*

$$\begin{aligned}\|f_{\boldsymbol{X},\lambda} - f_\lambda\| \leq R\lambda^{1/2}\|(L_{K,\boldsymbol{X}} + \lambda I)^{-1/2}(L_K + \lambda I)^{1/2}\|\|(L_K + \lambda I)^{-1/2}(L_K - L_{K,\boldsymbol{X}})\|^r \\ \|(L_K + \lambda I)^{-1/2}(L_K - L_{K,\boldsymbol{X}})(L_K + \lambda I)^{-1}\|^{1-r}.\end{aligned}$$

Proof of Lemma 4. According to the definition of $f_{\boldsymbol{X},\lambda}$ and f_λ, we have

$$f_{\boldsymbol{X},\lambda} - f_\lambda = (L_{K,\boldsymbol{X}} + \lambda I)^{-1}L_{K,\boldsymbol{X}} f_\rho - (L_K + \lambda I)^{-1}L_K f_\rho.$$

Due to $(A + \lambda I)^{-1}A = I - \lambda(A + \lambda I)^{-1}$ for any bounded positive operator A, we have

$$(L_K + \lambda I)^{-1}L_K f_\rho - (L_{K,\boldsymbol{X}} + \lambda I)^{-1}L_{K,\boldsymbol{X}} f_\rho = \lambda\left[(L_{K,\boldsymbol{X}} + \lambda I)^{-1} - (L_K + \lambda I)^{-1}\right] f_\rho.$$

Further, by virtue of the first order decomposition of operator difference: $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ for any invertible bounded operator and using the source condition in Assumption 4, the above equation can be further expressed as

$$\begin{aligned}&(L_K + \lambda I)^{-1}L_K f_\rho - (L_{K,\boldsymbol{X}} + \lambda I)^{-1}L_{K,\boldsymbol{X}} f_\rho = \lambda(L_{K,\boldsymbol{X}} + \lambda I)^{-1}(L_K - L_{K,\boldsymbol{X}})(L_K + \lambda I)^{-1}L_K^r g_\rho \\ &= \lambda^{1/2}\left(\lambda^{1/2}(L_{K,\boldsymbol{X}} + \lambda I)^{-1/2}\right)\left((L_{K,\boldsymbol{X}} + \lambda I)^{-1/2}(L_K + \lambda I)^{1/2}\right) \\ &\left((L_K + \lambda I)^{-1/2}(L_K - L_{K,\boldsymbol{X}})(L_K + \lambda I)^{-(1-r)}\right)\left((L_K + \lambda I)^{-r}L_K^r\right) g_\rho.\end{aligned}$$

Besides, using $\|AB^t\| \leq \|A\|^{1-t}\|AB\|^t$ with $t \in [0,1]$ for any bounded linear operator A and positive semi-definite operator B in Proposition 9 in [37], we have

$$\begin{aligned}\|(L_K + \lambda I)^{-1/2}(L_K - L_{K,\boldsymbol{X}})(L_K + \lambda I)^{-(1-r)}\| \leq \|(L_K + \lambda I)^{-1/2}(L_K - L_{K,\boldsymbol{X}})\|^r \\ \|(L_K + \lambda I)^{-1/2}(L_K - L_{K,\boldsymbol{X}})(L_K + \lambda I)^{-1}\|^{1-r},\end{aligned}$$

where we choose $A := (L_K + \lambda I)^{-1/2}(L_K - L_{K,\boldsymbol{X}})$, $B := (L_K + \lambda I)^{-1}$, and $t := 1 - r \in [0, 1)$. Accordingly, we can conclude our proof due to $\|(L_{K,\boldsymbol{X}} + \lambda I)^{-1/2}\| \leq 1/\sqrt{\lambda}$ and $\|(L_K + \lambda I)^{-r} L_K^r\| \leq 1$. □

Remark: The proof framework of Lemma 4 is similar to Lemma 4 in [37] but we consider a more general case $0 < r \leq 1$ than $1/2 \leq r \leq 1$ in [37]. Although $0 < r < 1/2$ appears to be unattainable as claimed in [37], we follow with [62, 63] on a quite general case with $r > 0$.

To prove Theorem 3, we also need the following two lemmas.

Lemma 5. *(Proposition 6 in [37]) Let $\delta \in (0, 1/2]$, it holds with probability at least $1 - 2\delta$ that*

$$\begin{aligned}&\|(L_K + \lambda I)^{-1/2}(L_K - L_{K,\boldsymbol{X}})(L_K + \lambda I)^{-1}\| \\ &\leq \|(L_K + \lambda I)^{-1/2}(L_K - L_{K,\boldsymbol{X}})(L_K + \lambda I)^{-1/2}\|\|(L_K + \lambda I)^{-1/2}\| \leq \left(\frac{\kappa^2}{3n\lambda} + \sqrt{\frac{\kappa^2}{n\lambda}}\right)\frac{1}{\sqrt{\lambda}}\,.\end{aligned}$$

Lemma 6. *For any $0 < \delta < 1$, with probability at least $1 - \delta$, we have*

$$\|(L_{K,\boldsymbol{X}} + \lambda I)^{-1/2}(L_K + \lambda I)^{1/2}\| \leq 1 + \frac{2\kappa}{\sqrt{n\lambda}}\left\{\frac{\kappa}{\sqrt{n\lambda}} + \sqrt{\mathcal{N}(\lambda)}\right\}\log\left(\frac{2}{\delta}\right)\,.$$

Proof of Lemma 6. By virtue of a second order decomposition of operator difference in Lemma 16 [62], we have

$$A^{-1} - B^{-1} = B^{-1}(B - A)A^{-1}(B - A)B^{-1} + B^{-1}(B - A)B^{-1}\,,$$

which leads to

$$A^{-1}B = I + B^{-1}(B - A) + B^{-1}(B - A)A^{-1}(B - A)\,. \tag{11}$$

Accordingly, denote $A := L_{K,\boldsymbol{X}} + \lambda I$ and $B := L_K + \lambda I$, we can derive that

$$\begin{aligned}&\|(L_{K,\boldsymbol{X}} + \lambda I)^{-1/2}(L_K + \lambda I)^{1/2}\| \leq \|(L_{K,\boldsymbol{X}} + \lambda I)^{-1}(L_K + \lambda I)\|^{1/2} \\ &\leq \sqrt{1 + \lambda^{-1/2}\|(L_K + \lambda I)^{-1/2}(L_K - L_{K,\boldsymbol{X}})\| + \lambda^{-1}\|(L_K + \lambda I)^{-1/2}(L_K - L_{K,\boldsymbol{X}})\|^2} \\ &\leq \sqrt{1 + \mathcal{A} + \mathcal{A}^2} \leq 1 + \mathcal{A}\,,\end{aligned}$$

where $\mathcal{A} := \frac{2\kappa}{\sqrt{n\lambda}}\left\{\frac{\kappa}{\sqrt{n\lambda}} + \sqrt{\mathcal{N}(\lambda)}\right\}\log(2/\delta)$ by Lemma 3. The first inequality holds by $\|A^s B^s\| \leq \|AB\|^s$ with $0 \leq s \leq 1$ for positive operators A and B on Hilbert spaces [39]. The second inequality can be derived by Eq. (11), $\|(L_{K,\boldsymbol{X}} + \lambda I)^{-1}\| \leq 1/\lambda$ and $\|(L_K + \lambda I)^{-1/2}\| \leq 1/\sqrt{\lambda}$. □

Remark: Lemma 7.2 in [64] gives $\|(L_{K,\boldsymbol{X}} + \lambda I)^{-1/2}(L_K + \lambda I)^{1/2}\| \leq \sqrt{2}$ by assuming $\lambda > \frac{9}{n}\log\frac{n}{\delta}$; whereas our result does not require extra conditions on λ.

Based on the above lemmas, we are ready to prove Theorem 3.

Proof of Theorem 3. We first estimate $\|(L_{K,\boldsymbol{X}} + \lambda I)^{-1/2}(L_K + \lambda I)^{1/2}\|$ in Lemma 6 by taking $\lambda := \bar{c}n^{-\vartheta}$ and the capacity condition in Assumption 5: $\mathcal{N}(\lambda) \leq Q^2\lambda^{-\eta}$ with $\eta \in [0, 1]$. Accordingly, we have

$$\begin{aligned}\|(L_{K,\boldsymbol{X}} + \lambda I)^{-1/2}(L_K + \lambda I)^{1/2}\| &\leq 1 + \frac{2\kappa}{\sqrt{n\lambda}}\left\{\frac{\kappa}{\sqrt{n\lambda}} + \sqrt{\mathcal{N}(\lambda)}\right\}\log\left(\frac{2}{\delta}\right) \\ &\leq 1 + \left(\frac{2\kappa^2}{\bar{c}}n^{-(1-\vartheta)} + 2\kappa\bar{c}^{-(\frac{1}{2}+\frac{\eta}{2})}Qn^{-\frac{1-\vartheta-\vartheta\eta}{2}}\right)\log\left(\frac{2}{\delta}\right) \\ &\leq \left(1 + \frac{2\kappa(\kappa + Q)}{\bar{c}}n^{-\frac{1-\vartheta-\vartheta\eta}{2}}\right)\log\left(\frac{2}{\delta}\right)\,,\end{aligned}$$

where we use $\log^r(2/\delta) \leq \log(2/\delta)$ due to $\log(2/\delta) > 1$ in the last inequality. Since $\|(L_{K,\boldsymbol{X}} + \lambda I)^{-1/2}(L_K + \lambda I)^{1/2}\|$ converges to zero when n is large enough, we require $\vartheta < \frac{1}{1+\eta}$ to ensure a positive convergence rate, which implies

$\vartheta \le 1$. Then we bound $\|(L_K+\lambda I)^{-1/2}(L_K - L_{K,\boldsymbol{X}})\|^r$ by Lemma 3. By virtue of $(a+b)^r \le a^r + b^r$ for any $r \in (0,1]$ and $a, b \ge 0$, we have

$$\begin{aligned}\|(L_K+\lambda I)^{-1/2}(L_K - L_{K,\boldsymbol{X}})\|^r &\le \left(\frac{2\kappa}{\sqrt{n}}\right)^r \left\{\frac{\kappa}{(n\lambda)^{\frac{r}{2}}} + [\mathcal{N}(\lambda)]^{\frac{r}{2}}\right\} \log\left(\frac{2}{\delta}\right) \\ &\le (2\kappa)^r (n\bar{c})^{-\frac{r}{2}} \left[\kappa n^{-\frac{r(1-\vartheta)}{2}} + \frac{Q}{\bar{c}^{\frac{\eta r}{2}}} n^{\frac{\vartheta\eta r}{2}}\right] \log\left(\frac{2}{\delta}\right) \\ &\le \frac{2\kappa(Q+\kappa)}{\bar{c}} n^{-\frac{(1-\vartheta\eta)r}{2}} \log\left(\frac{2}{\delta}\right),\end{aligned}$$

where the second one admits by the capacity condition in Assumption 5. Similarly, to bound $\|(L_K+\lambda I)^{-1/2}(L_K - L_{K,\boldsymbol{X}})(L_K+\lambda I)^{-1}\|^{1-r}$ by Lemma 5, we can derive that

$$\begin{aligned}\|(L_K+\lambda I)^{-1/2}(L_K - L_{K,\boldsymbol{X}})(L_K+\lambda I)^{-1}\|^{1-r} &\le \lambda^{-\frac{1-r}{2}} \left(\kappa^2 (n\lambda)^{-(1-r)} + \kappa (n\lambda)^{-\frac{1-r}{2}}\right) \\ &\le \frac{\kappa^2}{\bar{c}} \left(n^{(\frac{3}{2}\vartheta - 1)(1-r)} + n^{(\vartheta - \frac{1}{2})(1-r)}\right) \\ &\le \frac{\kappa^2}{\bar{c}} n^{(\vartheta-\frac{1}{2})(1-r)}.\end{aligned}$$

Combining the above three inequalities, we have

$$\begin{aligned}\|f_{\boldsymbol{X},\lambda} - f_\lambda\| &\le \frac{4R\kappa^3(Q+\kappa)^2}{\bar{c}^3} n^{-\frac{(1-\vartheta\eta)r+\vartheta}{2}} n^{(\vartheta-\frac{1}{2})(1-r)} \log^2\left(\frac{2}{\delta}\right) \\ &\le \widetilde{C_{R,Q,\kappa,\bar{c}}} n^{-\frac{1-\vartheta(\eta r+1-2r)}{2}} \log^2\left(\frac{2}{\delta}\right),\end{aligned}$$

where $\widetilde{C_{R,Q,\kappa,\bar{c}}} := 4R\kappa^3(Q+\kappa)^2/\bar{c}^3$ is independent of n and d.

Finally, the bias can be bounded by

$$\begin{aligned}\mathtt{B} &\le 2\|f_{\boldsymbol{X},\lambda} - f_\lambda\|^2_{\mathcal{L}^2_{\rho_X}} + 2\|f_\lambda - f_\rho\|^2_{\mathcal{L}^2_{\rho_X}} \\ &\le 2R^2 n^{-2\vartheta r} + \widetilde{C_1} n^{-[1-\vartheta(\eta r+1-2r)]} \log^4\left(\frac{2}{\delta}\right) \\ &\le \widetilde{C} n^{-2\vartheta r} \log^4\left(\frac{2}{\delta}\right),\end{aligned}$$

where the third inequality holds by $2\vartheta r \le 1 - \vartheta(\eta r + 1 - 2r)$ due to $\vartheta \le \frac{1}{1+\eta}$, and $\widetilde{C}, \widetilde{C_1}$ are some constants independent of n and d. Accordingly, we can conclude the proof. □

E Proof for the variance

Formally, we have the following theorem to bound the variance.

Theorem 4. *(Variance) Under Assumptions 2, 3, then for $0 < \delta < 1$ with probability $1 - \delta - d^{-2}$, $\theta = \frac{1}{2} - \frac{2}{8+m}$, and d large enough, for any given $\varepsilon > 0$, we have*

$$\mathtt{V} \lesssim \mathtt{V}_1 + \mathtt{V}_2,$$

where $\mathtt{V}_1 := \frac{\sigma^2\beta}{d} \mathcal{N}^{n\lambda+\gamma}_{\widetilde{\boldsymbol{X}}}$ and $\mathtt{V}_2$ is the residual term with

$$\mathtt{V}_2 := \begin{cases} \dfrac{\sigma^2 \log^{2+4\varepsilon} d}{(n\lambda+\gamma)^2 d^{4\theta-1}}, & \textit{inner-product kernels} \\ \dfrac{\sigma^2}{(n\lambda+\gamma)^2} d^{-2\theta} \log^{1+\varepsilon} d, & \textit{radial kernels}. \end{cases}$$

For inner-product kernels, our proof framework follows [18], and is briefly discussed in Section E.1. Nevertheless, error bound on radial kernels has not been investigated in [18] and is more subtle to handle (than that of inner-product kernels) due to the additionally introduced $\boldsymbol{A}$ and $\boldsymbol{A} \odot \boldsymbol{A}$ in Table 2. Accordingly, we mainly focus on proofs for radial kernels.

E.1 Inner-product kernel matrices

In this subsection, we consider the inner-product kernel case with $k(\boldsymbol{x}, \boldsymbol{x}') = h\left(\langle \boldsymbol{x}, \boldsymbol{x}' \rangle / d\right)$. We briefly introduce our results that can be derived from proofs of Theorem 2 in [18] for completeness.

To prove Theorem 4, define

$$\widetilde{\boldsymbol{K}^{\text{lin}}}(\boldsymbol{X}, \boldsymbol{X}) := (n\lambda + \gamma)\boldsymbol{I} + \alpha \mathbf{1}\mathbf{1}^\top + \beta \frac{\boldsymbol{X}\boldsymbol{X}^\top}{d} \in \mathbb{R}^{n \times n}\,,\ k^{\text{lin}}(\boldsymbol{x}, \boldsymbol{X}) := h(0)\mathbf{1} + \beta \frac{\boldsymbol{X}\boldsymbol{x}^\top}{d} \in \mathbb{R}^{n \times 1}\,, \tag{12}$$

and $k^{\text{lin}}(\boldsymbol{X}, \boldsymbol{x})$ is the transpose of $k^{\text{lin}}(\boldsymbol{x}, \boldsymbol{X})$. Note that γ in $\widetilde{\boldsymbol{K}^{\text{lin}}}$ corresponds to the *implicit* regularization and $n\lambda$ corresponds to the *explicit* regularization. Now we prove Theorem 4 for inner-product kernels.

Proof of Theorem 4 for inner-product kernels. According to the definition of $\mathtt{V}$, we have

$$\begin{aligned}
\mathtt{V} &= \mathbb{E}_{\boldsymbol{x},y} \text{tr}\left[k(\boldsymbol{x}, \boldsymbol{X})^\top (\boldsymbol{K} + n\lambda \boldsymbol{I})^{-1} \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top (\boldsymbol{K} + n\lambda \boldsymbol{I})^{-1} k(\boldsymbol{x}, \boldsymbol{X})\right] = \mathbb{E}_{\boldsymbol{x}} \|(\boldsymbol{K} + n\lambda \boldsymbol{I})^{-1} k(\boldsymbol{x}, \boldsymbol{X})\|_2^2 \ \mathbb{E}_{y|\boldsymbol{x}} \|\boldsymbol{\epsilon}\|_2^2 \\
&\leq \sigma^2 \mathbb{E}_{\boldsymbol{x}} \|(\boldsymbol{K} + n\lambda \boldsymbol{I})^{-1} k(\boldsymbol{x}, \boldsymbol{X})\|_2^2 \\
&\leq \sigma^2 \|(\boldsymbol{K} + n\lambda \boldsymbol{I})^{-1} \widetilde{\boldsymbol{K}^{\text{lin}}}\|_2^2 \ \mathbb{E}_{\boldsymbol{x}} \|[\widetilde{\boldsymbol{K}^{\text{lin}}}]^{-1} k^{\text{lin}}(\boldsymbol{x}, \boldsymbol{X})\|_2^2 + \sigma^2 \|(\boldsymbol{K} + n\lambda \boldsymbol{I})^{-1}\|_2^2 \ \mathbb{E}_{\boldsymbol{x}} \|k(\boldsymbol{x}, \boldsymbol{X}) - k^{\text{lin}}(\boldsymbol{x}, \boldsymbol{X})\|_2^2\,,
\end{aligned} \tag{13}$$

where the first inequality comes from Assumption 2. To bound the terms in Eq. (13), we need

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}} \|[\widetilde{\boldsymbol{K}^{\text{lin}}}]^{-1} k^{\text{lin}}(\boldsymbol{x}, \boldsymbol{X})\|_2^2 = \mathbb{E}_{\boldsymbol{x}} \,\text{tr}\left[\left[\widetilde{\boldsymbol{K}^{\text{lin}}}\right]^{-1} \left(\beta \frac{\boldsymbol{X}\boldsymbol{x}}{d} + h(0)\mathbf{1}\right)\left(\beta \frac{\boldsymbol{x}^\top \boldsymbol{X}^\top}{d} + h(0)\mathbf{1}^\top\right) \left[\widetilde{\boldsymbol{K}^{\text{lin}}}\right]^{-1}\right] \\
&\leq \frac{1}{d} \|\boldsymbol{\Sigma}_d\|_2 \,\text{tr}\left(\left[\widetilde{\boldsymbol{K}^{\text{lin}}}\right]^{-1} \beta^2 \frac{\boldsymbol{X}\boldsymbol{X}^\top}{d} \left[\widetilde{\boldsymbol{K}^{\text{lin}}}\right]^{-1}\right) + \frac{1}{d} \,\text{tr}\left(\left[\widetilde{\boldsymbol{K}^{\text{lin}}}\right]^{-1} h(0)^2 \mathbf{1}\mathbf{1}^\top \left[\widetilde{\boldsymbol{K}^{\text{lin}}}\right]^{-1}\right) \\
&\leq \frac{\beta}{d} \|\boldsymbol{\Sigma}_d\|_2 \sum_{j=1}^{n} \frac{\lambda_j\left(\widetilde{\boldsymbol{X}}\right)}{\left[n\lambda + \gamma + \lambda_j\left(\widetilde{\boldsymbol{X}}\right)\right]^2} + \frac{1}{d} \frac{h(0)^2 n}{\left[n\lambda + \gamma + \lambda_1\left(\widetilde{\boldsymbol{X}}\right)\right]^2} \\
&\asymp \frac{\beta}{d} \mathcal{N}_{\widetilde{\boldsymbol{X}}}^{n\lambda+\gamma} + \mathcal{O}\left(\frac{1}{nd}\right)\,.
\end{aligned} \tag{14}$$

To bound the remaining terms in Eq. (13), we also need the following results that can be obtained from [18]:
(i) By Proposition A.2 in [18], with probability at least $1 - \delta - d^{-2}$, for $\theta = \frac{1}{2} - \frac{2}{8+m}$ and any given $\varepsilon > 0$, we have $\left\|\boldsymbol{K} + n\lambda \boldsymbol{I} - \widetilde{\boldsymbol{K}^{\text{lin}}}\right\|_2 \leq d^{-\theta}\left(\delta^{-1/2} + \log^{0.5+\varepsilon} d\right)$ and $\mathbb{E}_{\boldsymbol{x}} \left\|k(\boldsymbol{x}, \boldsymbol{X}) - k^{\text{lin}}(\boldsymbol{x}, \boldsymbol{X})\right\|_2^2 \leq \widetilde{C_1} d^{-(4\theta-1)} \log^{2+4\varepsilon} d$.
(ii) $\left\|(\boldsymbol{K} + n\lambda \boldsymbol{I})^{-1}\right\|_2 \leq \frac{2}{n\lambda+\gamma}$ and $\left\|(\boldsymbol{K} + n\lambda \boldsymbol{I})^{-1} \widetilde{\boldsymbol{K}^{\text{lin}}}\right\|_2 \leq 2$ provided d is large enough such that $d^{-\theta}\left(\delta^{-1/2} + \log^{0.5+\varepsilon} d\right) \leq \gamma/2$.

Combining the above results, with probability at least $1 - \delta - d^{-2}$, for any given $\varepsilon > 0$, The error bound for the variance in Eq. (13) can be further given by

$$\begin{aligned}
\mathtt{V} &\leq \sigma^2 \mathbb{E}_{\boldsymbol{x}} \|(\boldsymbol{K} + n\lambda \boldsymbol{I})^{-1} k(\boldsymbol{x}, \boldsymbol{X})\|_2^2 \\
&\leq 2\sigma^2 \left\|(\boldsymbol{K} + n\lambda \boldsymbol{I})^{-1} \widetilde{\boldsymbol{K}^{\text{lin}}}\right\|_2^2 \ \mathbb{E}_{\boldsymbol{x}} \|[\widetilde{\boldsymbol{K}^{\text{lin}}}]^{-1} k^{\text{lin}}(\boldsymbol{X}, \boldsymbol{x})\|_2^2 + 2\sigma^2 \|\boldsymbol{K}^{-1}\|_2^2 \ \mathbb{E}_{\boldsymbol{x}} \left\|k(\boldsymbol{x}, \boldsymbol{X}) - k^{\text{lin}}(\boldsymbol{x}, \boldsymbol{X})\right\|_2^2 \\
&\asymp \frac{8\sigma^2 \beta}{d} \|\boldsymbol{\Sigma}_d\|_2 \sum_{j=1}^{n} \frac{\lambda_j(\widetilde{\boldsymbol{X}})}{\left[n\lambda + \gamma + \lambda_j(\widetilde{\boldsymbol{X}})\right]^2} + \frac{8\sigma^2}{(n\lambda+\gamma)^2} \widetilde{C_1} d^{-(4\theta-1)} \log^{2+4\varepsilon} d \\
&\asymp \frac{\sigma^2 \beta}{d} \mathcal{N}_{\widetilde{\boldsymbol{X}}}^{n\lambda+\gamma} + \frac{\sigma^2}{(n\lambda+\gamma)^2} d^{-(4\theta-1)} \log^{2+4\varepsilon} d\,,
\end{aligned}$$

which concludes the proof. □

E.2 Radial kernel matrices

In this subsection, we consider the radial kernel case with $k(\boldsymbol{x}, \boldsymbol{x}') = h\left(\frac{1}{d}\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2\right)$. Since the linearization of radial kernel matrices incurs in two additionally terms $\boldsymbol{A}$ and $\boldsymbol{A} \odot \boldsymbol{A}$, estimation for radial kernels is more technical than that of inner-product kernels. Accordingly, to prove Theorem 4 for radial kernels, we need to introduce the following notations and auxiliary results.

E.2.1 Auxiliary results

Recall $\tau := \mathrm{tr}(\boldsymbol{\Sigma}_d)/d$, define

$$\begin{aligned}\widetilde{\boldsymbol{K}^{\mathrm{lin}}}(\boldsymbol{X},\boldsymbol{X}) &:= (\gamma+n\lambda)\boldsymbol{I} + \alpha\mathbf{1}\mathbf{1}^\top + \beta\frac{\boldsymbol{X}\boldsymbol{X}^\top}{d} + h'(2\tau)\boldsymbol{A} + \frac{1}{2}h''(2\tau)\boldsymbol{A}\odot\boldsymbol{A}\\ k^{\mathrm{lin}}(\boldsymbol{x},\boldsymbol{X}) &:= h(2\tau)\mathbf{1} + \beta\frac{\boldsymbol{X}\boldsymbol{x}^\top}{d} - \frac{\beta}{2}\boldsymbol{A}(\boldsymbol{x},\boldsymbol{X}) \in \mathbb{R}^{n\times 1}\,,\end{aligned} \tag{15}$$

where $\boldsymbol{A}(\boldsymbol{x},\boldsymbol{X}) := \psi_{\boldsymbol{x}} + [\psi_1,\psi_2,\cdots,\psi_n]^\top$ with $\psi_{\boldsymbol{x}} = \|\boldsymbol{x}\|_2^2/d - \tau$ and $\psi_i = \|\boldsymbol{x}_i\|_2^2/d - \tau$ for $i = 1,2,\ldots,n$. As discussed in Appendix B, we conclude that $\widetilde{\boldsymbol{K}^{\mathrm{lin}}}$ admits the same eigenvalue decay as $\widetilde{\boldsymbol{X}}$ since $\boldsymbol{A}$ is a rank-2 matrix. Accordingly, we have the following results.

Proposition 3. *Given $\boldsymbol{A}(\boldsymbol{x},\boldsymbol{X})$ in Eq. (15), we have $\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{X}\boldsymbol{x}\boldsymbol{A}(\boldsymbol{X},\boldsymbol{x})] = \mu_3\boldsymbol{X}\boldsymbol{\Sigma}_d^{1/2}\,\mathrm{diag}(\boldsymbol{\Sigma}_d)\mathbf{1}_n^\top$, where $\mu_3 := \mathbb{E}[\boldsymbol{t}(j)^3]$ does not depend on j because each entry in $\boldsymbol{t}$ are independent for $j = 1,2,\ldots,d$. Further, $\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{X}\boldsymbol{x}\boldsymbol{A}(\boldsymbol{X},\boldsymbol{x})]$ is a rank-one matrix with its eigenvalue $\lambda_1(\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{X}\boldsymbol{x}\boldsymbol{A}(\boldsymbol{X},\boldsymbol{x})]) = \mathcal{O}(\sqrt{n/d})$.*

Proof of Proposition 3. According to the definition in Assumption 3, $\boldsymbol{x}_i = \boldsymbol{\Sigma}_d^{1/2}\boldsymbol{t}_i$ with $\mathbb{E}[\boldsymbol{t}_i(j)] = 0$ and $\mathbb{V}[\boldsymbol{t}_i(j)] = 1$, we have the following expression

$$\mathbb{E}_{\boldsymbol{t}}[\boldsymbol{t}\boldsymbol{t}^\top\boldsymbol{\Sigma}_d\boldsymbol{t}] = \mathbb{E}_{\boldsymbol{t}}\left[\boldsymbol{t}\sum_{i,j=1}^{d}\boldsymbol{t}(i)(\boldsymbol{\Sigma}_d)_{ij}\boldsymbol{t}(j)\right] = \mu_3[(\boldsymbol{\Sigma}_d)_{11},(\boldsymbol{\Sigma}_d)_{22},\cdots,(\boldsymbol{\Sigma}_d)_{dd}]^\top\,,$$

where $\mu_3 := \mathbb{E}(t_i^3)$. Accordingly, $\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{X}\boldsymbol{x}\boldsymbol{A}(\boldsymbol{X},\boldsymbol{x})]$ can be computed by

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{X}\boldsymbol{x}\boldsymbol{A}(\boldsymbol{X},\boldsymbol{x})] &= \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{X}\boldsymbol{x}(\psi_1+\psi_{\boldsymbol{x}}),\boldsymbol{X}\boldsymbol{x}(\psi_2+\psi_{\boldsymbol{x}})\cdots,\boldsymbol{X}\boldsymbol{x}(\psi_n+\psi_{\boldsymbol{x}})]\\ &= \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{X}\boldsymbol{x}\psi_{\boldsymbol{x}},\boldsymbol{X}\boldsymbol{x}\psi_{\boldsymbol{x}},\cdots,\boldsymbol{X}\boldsymbol{x}\psi_{\boldsymbol{x}}]\\ &= \boldsymbol{X}\boldsymbol{\Sigma}_d^{1/2}\left[\frac{\mathbb{E}_{\boldsymbol{t}}[\boldsymbol{t}\boldsymbol{t}^\top\boldsymbol{\Sigma}_d\boldsymbol{t}]}{d},\frac{\mathbb{E}_{\boldsymbol{t}}[\boldsymbol{t}\boldsymbol{t}^\top\boldsymbol{\Sigma}_d\boldsymbol{t}]}{d},\cdots,\frac{\mathbb{E}_{\boldsymbol{t}}[\boldsymbol{t}\boldsymbol{t}^\top\boldsymbol{\Sigma}_d\boldsymbol{t}]}{d}\right]\\ &= \mu_3\boldsymbol{X}\boldsymbol{\Sigma}_d^{1/2}\,\mathrm{diag}(\boldsymbol{\Sigma}_d)\mathbf{1}_n^\top\,.\end{aligned}$$

Note that, the matrix $\mathrm{diag}(\boldsymbol{\Sigma}_d)\mathbf{1}_n^\top$ is a rank-one matrix, which implies $\mathrm{rank}(\boldsymbol{X}\boldsymbol{\Sigma}_d^{1/2}\,\mathrm{diag}(\boldsymbol{\Sigma}_d)\mathbf{1}_n^\top) \le 1$. Accordingly, its non-zero eigenvalue $\lambda_1(\boldsymbol{X}\boldsymbol{\Sigma}_d^{1/2}\,\mathrm{diag}(\boldsymbol{\Sigma}_d)\mathbf{1}_n^\top)$ admits

$$\frac{1}{d}\lambda_1(\boldsymbol{X}\boldsymbol{\Sigma}_d^{1/2}\,\mathrm{diag}(\boldsymbol{\Sigma}_d)\mathbf{1}_n^\top) = \frac{1}{d}\sum_{i=1}^{n}\boldsymbol{x}_i^\top\boldsymbol{\Sigma}_d^{\frac{1}{2}}\,\mathrm{diag}(\boldsymbol{\Sigma}_d) = \frac{1}{d}\sum_{i=1}^{n}\boldsymbol{t}_i^\top\boldsymbol{\Sigma}_d\,\mathrm{diag}(\boldsymbol{\Sigma}_d)\,.$$

Due to $\mathbb{E}[\boldsymbol{t}_i^\top\boldsymbol{\Sigma}_d\,\mathrm{diag}(\boldsymbol{\Sigma}_d)] = 0$ and $\mathbb{V}[\boldsymbol{t}_i^\top\boldsymbol{\Sigma}_d\,\mathrm{diag}(\boldsymbol{\Sigma}_d)] = \|\boldsymbol{\Sigma}_d\,\mathrm{diag}(\boldsymbol{\Sigma}_d)\|_2^2$, which, with a central limit theorem argument, implies $\sum_{i=1}^n\boldsymbol{t}_i^\top\boldsymbol{\Sigma}_d\,\mathrm{diag}(\boldsymbol{\Sigma}_d) = \mathcal{O}(\sqrt{nd})$ due to $\|\boldsymbol{\Sigma}_d\,\mathrm{diag}(\boldsymbol{\Sigma}_d)\|_2 \le \|\boldsymbol{\Sigma}_d\|_2\|\,\mathrm{diag}(\boldsymbol{\Sigma}_d)\|_2 \le \widetilde{C}\|\,\mathrm{diag}(\boldsymbol{\Sigma}_d)\|_2$. Accordingly, we can conclude that $\frac{1}{d}\lambda_1(\boldsymbol{X}\boldsymbol{\Sigma}_d^{1/2}\,\mathrm{diag}(\boldsymbol{\Sigma}_d)\mathbf{1}_n^\top) = \mathcal{O}(\sqrt{n/d})$. □

Proposition 4. *Given $\boldsymbol{A}(\boldsymbol{x},\boldsymbol{X})$ in Eq. (15), we have $\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{A}(\boldsymbol{x},\boldsymbol{X})\boldsymbol{A}(\boldsymbol{X},\boldsymbol{x})] = \boldsymbol{\psi}\boldsymbol{\psi}^\top + \mathcal{O}(1/d)$. Further, it has only one non-zero eigenvalue that admits $\lambda_1(\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{A}(\boldsymbol{x},\boldsymbol{X})\boldsymbol{A}(\boldsymbol{X},\boldsymbol{x})]) = \mathcal{O}(n)$.*

Proof of Proposition 4. By virtue of the following results [26]

$$\begin{aligned}\frac{1}{d}\mathbb{E}_{\boldsymbol{x}}\|\boldsymbol{x}\|_2^2 &= \frac{1}{d}\mathbb{E}_{\boldsymbol{t}}[\boldsymbol{t}^\top\boldsymbol{\Sigma}_d\boldsymbol{t}] = \tau\\ \mathbb{V}_{\boldsymbol{x}}\left[\frac{\|\boldsymbol{x}\|_2^2}{d}\right] &= \frac{1}{d^2}\left((\mu_4-3)\sum_{i=1}^{d}((\boldsymbol{\Sigma}_d)_{ii})^2 + 2\,\mathrm{tr}(\boldsymbol{\Sigma}_d^2)\right) = \mathcal{O}\left(\frac{1}{d}\right)\,,\end{aligned}$$

where $\mu_4 := \mathbb{E}[\boldsymbol{t}(i)^4]$ does not depend on i. Accordingly, each entry in $\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{A}(\boldsymbol{x},\boldsymbol{X})\boldsymbol{A}(\boldsymbol{X},\boldsymbol{x})]$ can be computed as

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{A}(\boldsymbol{x},\boldsymbol{X})\boldsymbol{A}(\boldsymbol{X},\boldsymbol{x})]_{ij} &= \mathbb{E}_{\boldsymbol{x}}[(\psi_i+\psi_{\boldsymbol{x}})(\psi_j+\psi_{\boldsymbol{x}})]\\ &= \psi_i\psi_j + (\psi_i+\psi_j)\mathbb{E}_{\boldsymbol{x}}\psi_{\boldsymbol{x}} + \mathbb{E}_{\boldsymbol{x}}[\psi_{\boldsymbol{x}}^2]\\ &= \psi_i\psi_j + \mathbb{V}_{\boldsymbol{x}}\left[\frac{\|\boldsymbol{x}\|_2^2}{d}\right]\\ &= \psi_i\psi_j + \frac{\mu_4-3}{d^2}\operatorname{tr}(\boldsymbol{\Sigma}_d\odot\boldsymbol{\Sigma}_d) + \frac{2\operatorname{tr}(\boldsymbol{\Sigma}_d^2)}{d^2}\,.\end{aligned}$$

Then we have

$$\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{A}(\boldsymbol{x},\boldsymbol{X})\boldsymbol{A}(\boldsymbol{X},\boldsymbol{x})] = \boldsymbol{\psi}\boldsymbol{\psi}^\top + \mathcal{O}(1/d)\,.$$

Therefore, $\boldsymbol{\psi}\boldsymbol{\psi}^\top$ is a rank-one matrix with $\lambda_1(\boldsymbol{\psi}\boldsymbol{\psi}^\top) = \|\boldsymbol{\psi}\|_2^2 = \mathcal{O}(n)$. Then $\lambda_1(\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{A}(\boldsymbol{x},\boldsymbol{X})\boldsymbol{A}(\boldsymbol{X},\boldsymbol{x})])$ can be estimated by

$$\|\boldsymbol{\psi}\|_2^2 \le \lambda_1(\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{A}(\boldsymbol{x},\boldsymbol{X})\boldsymbol{A}(\boldsymbol{X},\boldsymbol{x})]) \le \|\boldsymbol{\psi}\|_2^2 + n\underbrace{\left[\frac{\mu_4-3}{d^2}\operatorname{tr}(\boldsymbol{\Sigma}_d\odot\boldsymbol{\Sigma}_d) + \frac{2\operatorname{tr}(\boldsymbol{\Sigma}_d^2)}{d^2}\right]}_{=\mathcal{O}(1/d)},$$

which implies $\lambda_1(\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{A}(\boldsymbol{x},\boldsymbol{X})\boldsymbol{A}(\boldsymbol{X},\boldsymbol{x})]) = \mathcal{O}(n)$. □

Lemma 7. *Given a radial kernel, under Assumption 3, for $\theta = \frac{1}{2} - \frac{2}{8+m}$, we have with probability at least $1-d^{-2}$ with respect to the draw of $\boldsymbol{X}$, for d large enough, for any given $\varepsilon > 0$, we have*

$$\mathbb{E}_{\boldsymbol{x}}\left\|k(\boldsymbol{x},\boldsymbol{X}) - k^{\mathrm{lin}}(\boldsymbol{x},\boldsymbol{X})\right\|_2^2 \le \widetilde{C_1} d^{-2\theta}\log^{1+\varepsilon} d\,,$$

where $\widetilde{C_1}$ is some constant independent of n and d.

Remark: In fact, we only need the $(5+m)$-moment in Assumption 3 but we still follow with it for simplicity.

Proof of Lemma 7. We start with the entry-wise Taylor expansion for the smooth kernel at 2τ with $\tau := \operatorname{tr}(\boldsymbol{\Sigma}_d)/d$

$$\begin{aligned}k(\boldsymbol{x},\boldsymbol{x}_j) = h(\frac{1}{d}\|\boldsymbol{x}-\boldsymbol{x}_j\|_2^2) &= h(2\tau) + h'(2\tau)(\frac{1}{d}\|\boldsymbol{x}-\boldsymbol{x}_j\|^2 - 2\tau) + \frac{h''(2\tau)}{2}\left(\frac{1}{d}\|\boldsymbol{x}-\boldsymbol{x}_j\|^2 - 2\tau\right)^2 + \mathcal{O}(d^{-3/2})\\ &= h(2\tau) + h'(2\tau)(\psi_{\boldsymbol{x}} + \psi_j - \frac{2\boldsymbol{x}^\top\boldsymbol{x}_j}{d}) + \frac{h''(2\tau)}{2}\left(\psi_{\boldsymbol{x}} + \psi_j - \frac{2\boldsymbol{x}^\top\boldsymbol{x}_j}{d}\right)^2 + \mathcal{O}(d^{-3/2})\,,\end{aligned}$$

where $\psi_j = \|\boldsymbol{x}_j\|_2^2/d - \tau$ for $j = 1, 2, \ldots, n$ as defined before. Accordingly, by virtue of $k^{\mathrm{lin}}(\boldsymbol{x},\boldsymbol{x}_j) = \frac{\beta\boldsymbol{x}^\top\boldsymbol{x}_j}{d} - \frac{\beta}{2}(\psi_{\boldsymbol{x}}+\psi_j)$ and Corollary 2 in [26], with probability at least $1-d^{-2}$, for any $\varepsilon > 0$, we have

$$k(\boldsymbol{x},\boldsymbol{x}_j) - k^{\mathrm{lin}}(\boldsymbol{x},\boldsymbol{x}_j) = \frac{h''(2\tau)}{2}\left(\frac{1}{d}\|\boldsymbol{x}-\boldsymbol{x}_j\|_2^2 - 2\tau\right)^2 \le \widetilde{C} d^{-1+\frac{4}{m}}(\log d)^{\frac{1+\varepsilon}{2}}\,,$$

where we only need $(5+m)$-moment. Therefore, with probability at least $1-d^{-2}$, for any given $\varepsilon > 0$, we have

$$\left\|k(\boldsymbol{x},\boldsymbol{X}) - k^{\mathrm{lin}}(\boldsymbol{x},\boldsymbol{X})\right\|_2 \le C_1 d^{-1/2+\frac{4}{m}}(\log d)^{\frac{1+\varepsilon}{2}} \le \widetilde{C_1} d^{-\theta}(\log d)^{\frac{1+\varepsilon}{2}}\,,$$

which implies

$$\mathbb{E}_{\boldsymbol{x}}\left\|k(\boldsymbol{x},\boldsymbol{X}) - k^{\mathrm{lin}}(\boldsymbol{x},\boldsymbol{X})\right\|_2^2 \le \widetilde{C_2} d^{-2\theta}\log^{1+\varepsilon} d\,,$$

where $\widetilde{C_1}$ and $\widetilde{C_2}$ are some constant independent of n and d. □

E.2.2 Proofs of Theorem 4 for radial kernels

Now we are ready to prove Theorem 4 for radial kernels.

Proof of Theorem 4 for radial kernels. Similar to Eq. (13), to estimate $\mathtt{V} \leq \sigma^2 \mathbb{E}_{\boldsymbol{x}} \|(\boldsymbol{K}+n\lambda \boldsymbol{I})^{-1} k(\boldsymbol{x}, \boldsymbol{X})\|_2^2$, we need to bound subsequently the following terms: $\left\|\boldsymbol{K}+n\lambda \boldsymbol{I}-\widetilde{\boldsymbol{K}^{\mathrm{lin}}}(\boldsymbol{X}, \boldsymbol{X})\right\|_2$, $\left\|(\boldsymbol{K}+n\lambda \boldsymbol{I})^{-1}\right\|_2$, $\left\|(\boldsymbol{K}+n\lambda \boldsymbol{I})^{-1} \widetilde{\boldsymbol{K}^{\mathrm{lin}}}(\boldsymbol{X}, \boldsymbol{X})\right\|_2$, $\mathbb{E}_{\boldsymbol{x}} \|[\widetilde{\boldsymbol{K}^{\mathrm{lin}}}(\boldsymbol{X}, \boldsymbol{X})]^{-1} k^{\mathrm{lin}}(\boldsymbol{x}, \boldsymbol{X})\|_2^2$, and $\mathbb{E}_{\boldsymbol{x}} \|k(\boldsymbol{x}, \boldsymbol{X})-k^{\mathrm{lin}}(\boldsymbol{x}, \boldsymbol{X})\|_2^2$.

In [26], the approximation error between radial kernel matrices and their linearization can be decomposed into three parts: the first-order term A_1, the second-order term A_2, and the third-order term A_3

$$\left\|\boldsymbol{K}+n\lambda \boldsymbol{I}-\widetilde{\boldsymbol{K}^{\mathrm{lin}}}(\boldsymbol{X}, \boldsymbol{X})\right\|_2 := A_1+A_2+A_3\,,$$

where A_1 and A_3 admit $\|A_1\|_2 \leq d^{-\theta} \log^{2+4\varepsilon} d$ and $\|A_3\|_2 \leq d^{-\theta} \log^{2+4\varepsilon} d$. The second-order term A_2 admits $\Pr(\|A_2\| \leq d^{-\theta} \delta^{-1/2}) \leq \delta$ by Proposition A.2 in [18] and [26]. Accordingly, with probability at least $1-\delta-d^{-2}$, for $\theta=\frac{1}{2}-\frac{2}{8+m}$ and any given $\varepsilon>0$, we have

$$\left\|\boldsymbol{K}+n\lambda \boldsymbol{I}-\widetilde{\boldsymbol{K}^{\mathrm{lin}}}(\boldsymbol{X}, \boldsymbol{X})\right\|_2 \leq d^{-\theta}\left(\delta^{-1/2}+\log^{2+4\varepsilon} d\right)\,.$$

According to Proposition 3 and 4, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}} \|[\widetilde{\boldsymbol{K}^{\mathrm{lin}}}(\boldsymbol{X}, \boldsymbol{X})]^{-1} k^{\mathrm{lin}}(\boldsymbol{X}, \boldsymbol{x})\|_2^2 \\
&= \beta^2 \mathbb{E}_{\boldsymbol{x}} \operatorname{tr}\left[[\widetilde{\boldsymbol{K}^{\mathrm{lin}}}]^{-1}\Big(\frac{\boldsymbol{X}\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{X}^\top}{d^2}-\frac{\boldsymbol{X}\boldsymbol{x}\boldsymbol{A}(\boldsymbol{X}, \boldsymbol{x})}{d}+\frac{1}{4}\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{X})\boldsymbol{A}(\boldsymbol{X}, \boldsymbol{x})+h(2\tau)^2 \mathbf{1}\mathbf{1}^\top\Big)[\widetilde{\boldsymbol{K}^{\mathrm{lin}}}]^{-1}\right] \\
&\leq \beta^2 \operatorname{tr}\left([\widetilde{\boldsymbol{K}^{\mathrm{lin}}}]^{-1}\Big(\frac{\boldsymbol{X}\boldsymbol{X}^\top \|\boldsymbol{\Sigma}_d\|_2}{d^2}-\frac{\mu_3 \boldsymbol{X}\boldsymbol{\Sigma}_d^{1/2} \operatorname{diag}(\boldsymbol{\Sigma}_d)\mathbf{1}_n^\top}{d}+\frac{1}{4}\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{X})\boldsymbol{A}(\boldsymbol{X}, \boldsymbol{x})+h(2\tau)^2 \mathbf{1}\mathbf{1}^\top\Big)[\widetilde{\boldsymbol{K}^{\mathrm{lin}}}]^{-1}\right) \\
&= \frac{\beta^2 \|\boldsymbol{\Sigma}_d\|_2}{d} \sum_{i=1}^n \frac{\lambda_i(\boldsymbol{X}\boldsymbol{X}^\top/d)}{[\lambda_i(\widetilde{\boldsymbol{K}^{\mathrm{lin}}})]^2}-\frac{\beta^2 \mu_3}{d} \frac{\lambda_1(\boldsymbol{X}\boldsymbol{\Sigma}_d^{1/2} \operatorname{diag}(\boldsymbol{\Sigma}_d)\mathbf{1}_n^\top)}{[\lambda_1(\widetilde{\boldsymbol{K}^{\mathrm{lin}}})]^2}+\beta^2 \frac{4\lambda_1(\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{X})\boldsymbol{A}(\boldsymbol{X}, \boldsymbol{x})])+h(2\tau)^2 n}{[\lambda_1(\widetilde{\boldsymbol{K}^{\mathrm{lin}}})]^2} \\
&\asymp \frac{\beta^2 \|\boldsymbol{\Sigma}_d\|_2}{d} \sum_{i=1}^n \frac{\lambda_i(\boldsymbol{X}\boldsymbol{X}^\top/d)}{[\lambda_1(\widetilde{\boldsymbol{K}^{\mathrm{lin}}})]^2}+\frac{\mathcal{O}(\sqrt{n/d})}{[\lambda_1(\widetilde{\boldsymbol{K}^{\mathrm{lin}}})]^2}+\frac{\mathcal{O}(n)}{[\lambda_1(\widetilde{\boldsymbol{K}^{\mathrm{lin}}})]^2}+\frac{\mathcal{O}(n)}{[\lambda_1(\widetilde{\boldsymbol{K}^{\mathrm{lin}}})]^2} \\
&\asymp \frac{\beta}{d} \mathcal{N}_{\widetilde{\boldsymbol{X}}}^{n\lambda+\gamma}+\mathcal{O}\left(\frac{1}{n}\right)\,.
\end{aligned} \tag{16}$$

It can be found that, the above error bounds are the same as that of inner-product kernels, except two additional terms due to the considered $\boldsymbol{A}$ and $\boldsymbol{A} \odot \boldsymbol{A}$ in the linearization, which can be shown small in the large n, d regime.

By virtue of $\left\|(\boldsymbol{K}+n\lambda \boldsymbol{I})^{-1}\right\|_2 \leq \frac{2}{n\lambda+\gamma}$ and $\left\|(\boldsymbol{K}+n\lambda \boldsymbol{I})^{-1} \widetilde{\boldsymbol{K}^{\mathrm{lin}}}\right\|_2 \leq 2$ in [18], Lemma 7, and the above equations, with probability at least $1-\delta-d^{-2}$, for any given $\varepsilon>0$, we have

$$\begin{aligned}
\mathtt{V} &\leq \sigma^2 \mathbb{E}_{\boldsymbol{x}} \|(\boldsymbol{K}+n\lambda \boldsymbol{I})^{-1} k(\boldsymbol{x}, \boldsymbol{X})\|_2^2 \\
&\leq 2\sigma^2 \left\|(\boldsymbol{K}+n\lambda \boldsymbol{I})^{-1} \widetilde{\boldsymbol{K}^{\mathrm{lin}}}\right\|_2^2 \mathbb{E}_{\boldsymbol{x}} \|[\widetilde{\boldsymbol{K}^{\mathrm{lin}}}]^{-1} k^{\mathrm{lin}}(\boldsymbol{X}, \boldsymbol{x})\|_2^2+2\sigma^2 \|\boldsymbol{K}^{-1}\|_2^2 \mathbb{E}_{\boldsymbol{x}} \left\|k(\boldsymbol{x}, \boldsymbol{X})-k^{\mathrm{lin}}(\boldsymbol{x}, \boldsymbol{X})\right\|_2^2 \\
&\leq 8\sigma^2 \mathbb{E}_{\boldsymbol{x}} \|[\widetilde{\boldsymbol{K}^{\mathrm{lin}}}]^{-1} k^{\mathrm{lin}}(\boldsymbol{X}, \boldsymbol{x})\|_2^2+\frac{8\sigma^2}{(n\lambda+\gamma)^2} \widetilde{C_1} d^{-2\theta} \log^{1+\varepsilon} d \\
&\asymp \frac{\sigma^2 \beta}{d} \mathcal{N}_{\widetilde{\boldsymbol{X}}}^{n\lambda+\gamma}+\frac{\sigma^2}{(n\lambda+\gamma)^2} d^{-2\theta} \log^{1+\varepsilon} d\,,
\end{aligned} \tag{17}$$

where the second inequality admits by Lemma 7, and the last inequality follows by Eq. (16). Finally, we conclude the proof. □

F Proof of Proposition 1

In this section, we discuss $\mathcal{N}_{\widetilde{\boldsymbol{X}}}^{n\lambda+\gamma}$ based on three eigenvalue decays: *harmonic decay*, *polynomial decay*, and *exponential decay* under two regimes $n<d$ and $n>d$.

F.1 $n < d$ case

Recall $b := n\lambda + \gamma > 0$, and $\mathcal{N}_{\widetilde{\boldsymbol{X}}}^b := \sum_{i=1}^n \frac{\lambda_i(\widetilde{\boldsymbol{X}})}{\left[b+\lambda_i(\widetilde{\boldsymbol{X}})\right]^2}$, define $F(\lambda_i) := \frac{\lambda_i}{(b+\lambda_i)^2}$ where λ_i is short for $\lambda_i(\widetilde{\boldsymbol{X}})$. We notice that, when $\lambda_i \leq b$, $F(\lambda_i)$ is an increasing function of λ_i, and thus a decreasing function of i when the above three eigenvalue decays are considered. Likewise, when $\lambda_i \geq b$, $F(\lambda_i)$ is a decreasing function of λ_i, and thus an increasing function of i. Without loss of generality, we assume that the first q eigenvalues satisfy $\lambda_i \geq b$ with $i = 1, 2, \cdots, q$ and the remaining $n-q$ eigenvalues satisfy $\lambda_i \leq b$ with $i = m+1, m+2 \cdots, n$. Clearly, the integer q can be chosen from 0 to n. Accordingly, denote $r_* := \operatorname{rank}(\widetilde{\boldsymbol{X}})$ which includes the rank-deficient case, $\mathcal{N}_{\widetilde{\boldsymbol{X}}}^b$ can be upper bounded by the Riemann sum as follows.

Harmonic decay $\lambda_i(\widetilde{\boldsymbol{X}}) \propto n/i$ for $i \in \{1, 2, \ldots, r_*\}$ and $\lambda_i(\widetilde{\boldsymbol{X}}) = 0$ for $i \in \{r_*+1, \ldots, n\}$

$$
\begin{aligned}
\frac{1}{d}\mathcal{N}_{\widetilde{\boldsymbol{X}}}^b = \frac{1}{d}\sum_{i=1}^{r_*} \frac{n/i}{(b+n/i)^2} &= \frac{1}{d}\sum_{i=1}^{q} \frac{n/i}{(b+n/i)^2} + \frac{1}{d}\sum_{i=q+1}^{r_*} \frac{n/i}{(b+n/i)^2} \\
&\leqslant \frac{1}{nd}\int_1^{q+1} \frac{t}{\left(1+\frac{bt}{n}\right)^2}\mathrm{d}t + \frac{1}{nd}\int_{q+1}^{r_*+1} \frac{t}{\left(1+\frac{bt}{n}\right)^2}\mathrm{d}t \\
&= \frac{n}{b^2 d}\int_{\frac{b}{n}}^{\frac{(r_*+1)b}{n}} \frac{u}{(1+u)^2}\mathrm{d}u \text{ with the change of variable } u = tb/n \\
&= \frac{n}{b^2 d}\left[\ln\frac{n+(r_*+1)b}{n+b} + \frac{n}{n+b+r_*b} - \frac{n}{n+b}\right] \\
&\leqslant \frac{n}{b^2 d}\ln\frac{n+(r_*+1)b}{n+b} = \mathcal{O}(\frac{n}{b^2 d})\,.
\end{aligned}
$$

Polynomial decay: $\lambda_i(\widetilde{\boldsymbol{X}}) \propto ni^{-2a}$ with $a > 1/2$ for $i \in \{1, 2, \ldots, r_*\}$ and $\lambda_i(\widetilde{\boldsymbol{X}}) = 0$ for $i \in \{r_*+1, \ldots, n\}$. Hence, we actually aim to bound

$$
\begin{aligned}
\frac{1}{d}\mathcal{N}_{\widetilde{\boldsymbol{X}}}^b = \frac{1}{d}\sum_{i=1}^{r_*} \frac{ni^{-2a}}{(b+ni^{-2a})^2} &= \frac{1}{d}\sum_{i=1}^{q} \frac{ni^{-2a}}{(b+ni^{-2a})^2} + \frac{1}{d}\sum_{i=q+1}^{r_*+1} \frac{ni^{-2a}}{(b+ni^{-2a})^2} \\
&\leqslant \frac{1}{nd}\int_1^{r_*+1} \frac{t^{2a}}{\left(1+\frac{t^{2a}b}{n}\right)^2}\mathrm{d}t \\
&= \frac{1}{2abd}\left(\frac{n}{b}\right)^{\frac{1}{2a}}\int_{b/n}^{(r_*+1)^{2a}b/n} \frac{u^{\frac{1}{2a}}}{(1+u)^2}\mathrm{d}u \text{ with the change of variable } u = t^{2a}b/n \\
&\leqslant \widetilde{C}\frac{1}{2abd}\left(\frac{n}{b}\right)^{\frac{1}{2a}} \quad \text{since the integral is finite due to } 2a > 1
\end{aligned}
$$

Exponential decay: $\lambda_i(\widetilde{\boldsymbol{X}}) \propto ne^{-ai}$ with $a > 0$ for $i \in \{1, 2, \ldots, r_*\}$ and $\lambda_i(\widetilde{\boldsymbol{X}}) = 0$ for $i \in \{r_*+1, \ldots, n\}$.

We aim to bound the sum as

$$
\begin{aligned}
\frac{1}{d}\mathcal{N}_{\widetilde{\boldsymbol{X}}}^b = \frac{1}{d}\sum_{i=1}^{r_*} \frac{ne^{-ai}}{(b+ne^{-ai})^2} &= \frac{1}{d}\sum_{i=1}^{q} \frac{ne^{-ai}}{(b+ne^{-ai})^2} + \frac{1}{d}\sum_{i=q+1}^{r_*} \frac{ne^{-ai}}{(b+ne^{-ai})^2} \\
&\leqslant \frac{1}{d}\int_1^{r_*+1} \frac{ne^{-at}}{(b+ne^{-at})^2}\mathrm{d}t \\
&= \frac{1}{ad}\int_{ne^{-a(r_*+1)}}^{ne^{-a}} \frac{1}{(b+u)^2}\mathrm{d}u \text{ with the change of variable } u = ne^{-at} \\
&= \frac{1}{ad}\left(\frac{1}{b+ne^{-a(r_*+1)}} - \frac{1}{b+ne^{-a}}\right)\,.
\end{aligned}
$$

Note that, the monotonicity of $\mathcal{N}_{\widetilde{\boldsymbol{X}}}^b$ (also $\mathtt{V}_1$) with respect to n is relatively clear for *harmonic decay* and *polynomial decay* but is unclear in the case of *exponential decay*. Here we study the monotonicity in the *exponential decay*.

Denote the function $G(n) := \left(\frac{1}{b+ne^{-a(r_*+1)}} - \frac{1}{b+ne^{-a}}\right)$ with $b := n\lambda + \gamma$, taking $\lambda := \bar{c}n^{-\vartheta}$, its derivation is

$$G'(n) = \frac{-\bar{c}(1-\vartheta)n^{-\vartheta} - e^{-a(r_*+1)}}{\left[cn^{1-\vartheta} + \gamma + ne^{-a(r_*+1)}\right]^2} + \frac{\bar{c}(1-\vartheta)n^{-\vartheta} + e^{-a}}{\left[cn^{(1-\vartheta)} + \gamma + ne^{-a}\right]^2}, \tag{18}$$

which can be rewritten as

$$G'(n) = \frac{\bar{c}(1-\vartheta)n^{-\vartheta} + e^{-a}}{\left[cn^{1-\vartheta} + \gamma + ne^{-a(r_*+1)}\right]^2} \left(\underbrace{\frac{\left[\bar{c}n^{1-\vartheta} + \gamma + ne^{-a(r_*+1)}\right]^2}{\left[\bar{c}n^{(1-\vartheta)} + \gamma + ne^{-a}\right]^2}}_{\triangleq H_1(n)} - \underbrace{\frac{\bar{c}(1-\vartheta)n^{-\vartheta} + e^{-a(r_*+1)}}{\bar{c}(1-\vartheta)n^{-\vartheta} + e^{-a}}}_{\triangleq H_2(n)} \right).$$

It can be found that both $H_1(n)$ and $H_2(n)$ are decreasing functions with n. More specifically, their maximum and minimum can be achieved with

$$\max_n H_1(n) = H_1(1) = \left(\frac{\bar{c}+\gamma+e^{-a(r_*+1)}}{\bar{c}+\gamma+e^{-a}}\right)^2, \quad \min_n H_1(n) = \lim_{n\to\infty} H_1(n) = \left(\frac{e^{-a(r_*+1)}}{e^{-a}}\right)^2,$$

and

$$\max_n H_2(n) = H_2(1) = \frac{\bar{c}(1-\vartheta)+e^{-a(r_*+1)}}{\bar{c}(1-\vartheta)+e^{-a}}, \quad \min_n H_2(n) = \lim_{n\to\infty} H_2(n) = \frac{e^{-a(r_*+1)}}{e^{-a}}.$$

Accordingly, if $H_1(1) < H_2(1)$, we obtain a decreasing function $G(n)$ of n, which implies that $\mathcal{N}^b_{\widetilde{\boldsymbol{X}}}$ will decrease with n. Here the condition $H_1(1) < H_2(1)$ indicates

$$\left(\frac{\bar{c}+\gamma+e^{-a(r_*+1)}}{\bar{c}+\gamma+e^{-a}}\right)^2 \leq \frac{\bar{c}(1-\vartheta)+e^{-a(r_*+1)}}{\bar{c}(1-\vartheta)+e^{-a}},$$

which is equivalent to

$$(\vartheta\bar{c}+\gamma)^2 \leq \left[e^{-a} + (1-\vartheta)\bar{c}\right]\left[e^{-a(r_*+1)} + (1-\vartheta)\bar{c}\right]. \tag{19}$$

Accordingly, if the above inequality holds, $\mathcal{N}^b_{\widetilde{\boldsymbol{X}}}$ will decrease with n. In Section G.2, we will experimentally check whether this condition holds or not.

F.2 $n > d$ case and the large n limit

In this section, we consider the $n > d$ case, and further study the trend of $\mathtt{V}_1$ as $n \to \infty$. Note that, in this case, $\boldsymbol{X}\boldsymbol{X}^\top/d$ has at most $r_* \leq d$ non-zero eigenvalues. Accordingly, the Riemann sum is counted to r_* instead of n. Similar to the above description, we also consider the following three eigenvalue decays.

Harmonic decay $\lambda_i(\widetilde{\boldsymbol{X}}) \propto n/i$, $i \in \{1, 2, \cdots, d\}$

$$\begin{aligned}
\frac{1}{d}\mathcal{N}^b_{\widetilde{\boldsymbol{X}}} = \frac{1}{d}\sum_{i=1}^{r_*}\frac{n/i}{(b+n/i)^2} &= \frac{1}{d}\sum_{i=1}^{q}\frac{n/i}{(b+n/i)^2} + \frac{1}{d}\sum_{i=q+1}^{r_*}\frac{n/i}{(b+n/i)^2} \\
&\leqslant \frac{n}{b^2 d}\int_{\frac{b}{n}}^{\frac{(r_*+1)b}{n}} \frac{u}{(1+u)^2}\mathrm{d}u \\
&= \frac{n}{b^2 d}\left[\ln\frac{n+(r_*+1)b}{n+b} + \frac{n}{n+b+r_*b} - \frac{n}{n+b}\right].
\end{aligned}$$

In particular, taking the limit of $n \to \infty$, we have

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{d}\mathcal{N}_{\widetilde{\boldsymbol{X}}}^{b} &= \lim_{n\to\infty} \frac{n}{b^2 d}\left[\ln \frac{n+(r_*+1)b}{n+b} + \frac{n}{n+b+r_* b} - \frac{n}{n+b}\right] \\
&= \lim_{n\to\infty} \frac{n}{b^2 d} \ln \frac{n+(r_*+1)b}{n+b} + \lim_{n\to\infty} \frac{n}{b^2 d}\left(\frac{n}{n+b+r_* b} - \frac{n}{n+b}\right) \\
&= \frac{r_*}{d}\left(\lim_{n\to\infty} \frac{1}{b}\frac{n}{n+b} - \lim_{n\to\infty} \frac{n^2}{b(n+b+r_*)(n+b)}\right) \\
&\leq \lim_{n\to\infty} \frac{1}{b}\frac{n}{n+b} - \lim_{n\to\infty} \frac{n^2}{b(n+b+r_*)(n+b)} \\
&= 0\,.
\end{aligned}$$

Accordingly, by the squeeze theorem, we can conclude, given d, $\mathcal{N}_{\widetilde{\boldsymbol{X}}}^{b}$ tends to zero when $n \to \infty$.

Polynomial decay: $\lambda_i(\widetilde{\boldsymbol{X}}) \propto n i^{-2a}$ with $a > 1/2$, $i \in \{1, 2, \cdots, d\}$

$$\begin{aligned}
\frac{1}{d}\mathcal{N}_{\widetilde{\boldsymbol{X}}}^{b} = \frac{1}{d}\sum_{i=1}^{r_*} \frac{n i^{-2a}}{\left(b + n i^{-2a}\right)^2} &\leqslant \frac{1}{2abd}\left(\frac{n}{b}\right)^{\frac{1}{2a}} \int_{b/n}^{(r_*+1)^{2a} b/n} \frac{u^{\frac{1}{2a}}}{(1+u)^2}\mathrm{d}u \\
&\leqslant \frac{1}{2abd}\left(\frac{n}{b}\right)^{\frac{1}{2a}} \int_{0}^{\infty} \frac{u^{\frac{1}{2a}}}{(1+u)^2}\mathrm{d}u \\
&\leqslant \widetilde{C}\frac{1}{2abd}\left(\frac{n}{b}\right)^{\frac{1}{2a}} \qquad \text{since the integral is finite due to } 2a > 1
\end{aligned}$$

Since the integral $\int \frac{u^{\frac{1}{2a}}}{(1+u)^2}\mathrm{d}u$ can behave rather differently for different choices of a, here we take $a = 1$ as an example. Taking the limit of $n \to \infty$, we have

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{d}\mathcal{N}_{\widetilde{\boldsymbol{X}}}^{b} &= \lim_{n\to\infty} \frac{1}{2bd}\left(\frac{n}{b}\right)^{\frac{1}{2}} \int_{b/n}^{(r_*+1)^2 b/n} \frac{u^{\frac{1}{2}}}{(1+u)^2}\mathrm{d}u \\
&= \frac{1}{2bd}\lim_{n\to\infty} \left(\frac{n}{b}\right)^{\frac{1}{2}} \left(\arctan(\sqrt{u}) - \frac{\sqrt{u}}{u+1}\right)\Bigg|_{b/n}^{(r_*+1)^2 b/n} \\
&= \frac{1}{2bd}\lim_{n\to\infty} \sqrt{\frac{n}{b}}\left((r_*+1)\sqrt{b/n} - \frac{(r_*+1)\sqrt{b/n}}{(r_*+1)^2 b/n} - \sqrt{b/n} + \frac{\sqrt{b/n}}{b/n+1}\right) \text{ using } \lim_{x\to 0}\tfrac{\arctan x}{x} = 1. \\
&= 0\,.
\end{aligned}$$

Exponential decay: $\lambda_i(\widetilde{\boldsymbol{X}}) \propto n e^{-ai}$ with $a > 0$, $i \in \{1, 2, \cdots, d\}$

$$\begin{aligned}
\frac{1}{d}\mathcal{N}_{\widetilde{\boldsymbol{X}}}^{b} = \frac{1}{d}\sum_{i=1}^{r_*} \frac{n e^{-ai}}{\left(b + n e^{-ai}\right)^2} &= \frac{1}{d}\sum_{i=1}^{q} \frac{n e^{-ai}}{\left(b + n e^{-ai}\right)^2} + \frac{1}{d}\sum_{i=q+1}^{r_*} \frac{n e^{-ai}}{\left(b + n e^{-ai}\right)^2} \\
&\leqslant \frac{1}{ad}\int_{n e^{-a(r_*+1)}}^{n e^{-a}} \frac{1}{(b+u)^2}\mathrm{d}u \\
&= \frac{1}{ad}\left(\frac{1}{b + n e^{-a(r_*+1)}} - \frac{1}{b + n e^{-a}}\right).
\end{aligned}$$

Taking the limit of $n \to \infty$, we can directly have $\lim_{n\to\infty} \frac{1}{d}\mathcal{N}_{\widetilde{\boldsymbol{X}}}^{b} = 0$.

G Additional Experiments

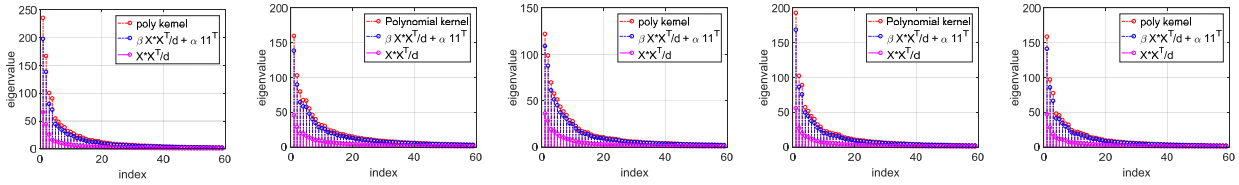
In this section, we present additional experiments including the following parts:

- In Section G.1, we add the *MNIST* dataset [59] to verify the eigenvalue decay equivalence, and evaluate the effect by different orders in polynomial kernel.
- In Section G.2, our model works in a polynomial kernel setting under the *polynomial decay* and *exponential decay* of $\widetilde{\boldsymbol{X}}$ on the synthetic dataset.

G.1 Eigenvalue decay equivalence

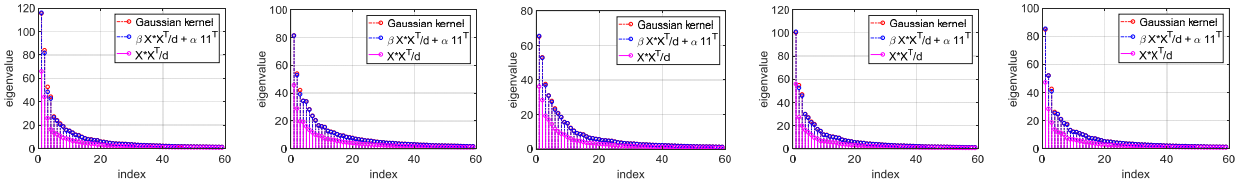
Apart from the *YearPredictionMSD* dataset in the main text, we add the *MNIST* dataset [59] to verify the eigenvalue decay equivalence. We also compute eigenvalues of $\widetilde{\boldsymbol{X}} := \beta \boldsymbol{X}\boldsymbol{X}^\top/d + \alpha \mathbf{1}\mathbf{1}^\top$ for validation. Here the parameters α depends on the covariate $\boldsymbol{\Sigma}_d$, which can be empirically estimated by the sample covariance $\frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i - \frac{1}{n}\sum_{j=1}^n \boldsymbol{x}_j)(\boldsymbol{x}_i - \frac{1}{n}\sum_{j=1}^n \boldsymbol{x}_j)^\top$.

Results on the polynomial kernel with order 3 and the Gaussian kernel are presented in Figure 6 and 7, respectively. It can be observed that, the nonlinear kernel matrix $\boldsymbol{K}$ admits almost the same eigenvalue as $\widetilde{\boldsymbol{X}} := \beta \boldsymbol{X}\boldsymbol{X}^\top/d + \alpha \mathbf{1}\mathbf{1}^\top$ with a constant shift γ, and accordingly exhibits the same eigenvalue decay with $\widetilde{\boldsymbol{X}}$ and $\boldsymbol{X}\boldsymbol{X}^\top/d$.



(a) digit 1 (b) digit 3 (c) digit 5 (d) digit 7 (e) digit 10

Figure 6: Top 60 eigenvalues of Polynomial kernel with order 3 and its linearization on the MNIST dataset. Note that the largest eigenvalue λ_1 is not plotted for better display.



(a) digit 1 (b) digit 3 (c) digit 5 (d) digit 7 (e) digit 10

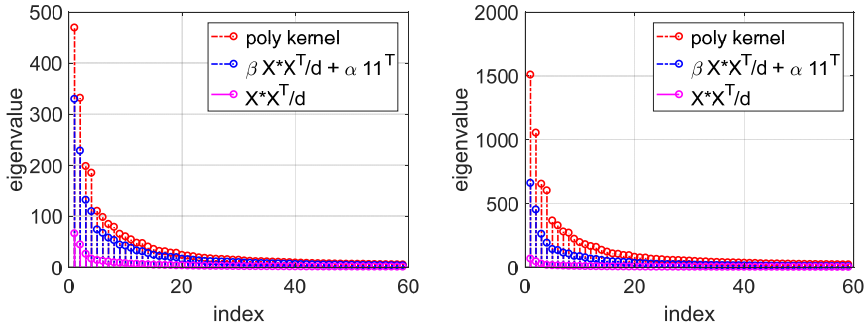
Figure 7: Top 60 eigenvalues of Gaussian kernel and its linearization on the MNIST dataset. Note that the largest eigenvalue λ_1 is not plotted for better display.

Besides, to study eigenvalue decay effected by the order in polynomial kernels, we present results of the order $p = 5$ and $p = 10$ in Figure 8. Experimental results show that, there is some gap between the original kernel and its linearization in higher orders. This is because, nonlinear kernel approximated by linear model here is based on Taylor expansion, which would incur in some residual errors as higher order in polynomial kernels brings in stronger non-linearity.

G.2 Results on the synthetic dataset

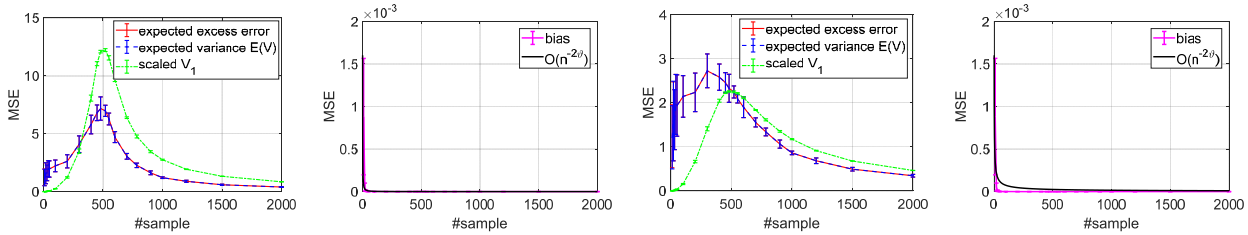
Here we evaluate our model with the polynomial kernel on the synthetic dataset under the polynomial/exponential decay of $\boldsymbol{\Sigma}_d$. The data generation process follows with our experiments part in the main text such that $\widetilde{\boldsymbol{X}}$ admits the polynomial/exponential decay.

Results on the *polynomial decay* and the *exponential decay* are shown in Figure 9 and Figure 10, respectively. We find that, the bias achieves the certain $\mathcal{O}(n^{-2\vartheta r})$ convergence rate on both decays; while the variance shows different configurations on these two decays. To be specific, the tend of $\mathtt{V}_1$ on the *polynomial decay* is unimodal,



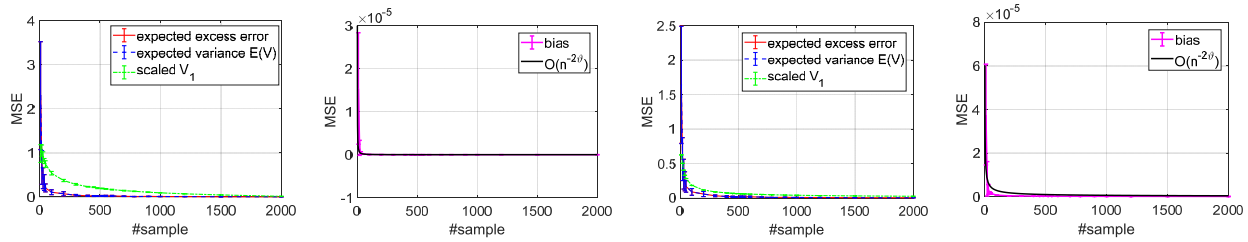
(a) *order 5* (b) *order 10*

Figure 8: Top 60 eigenvalues of polynomial kernel matrices and their linearizations on the MNIST dataset (digit 1). Note that the largest eigenvalue λ_1 is not plotted for better display.



(a) $\vartheta = 2/3$ (b) $\vartheta = 2/3$ (c) $\vartheta = 1/3$ (d) $\vartheta = 1/3$

Figure 9: Polynomial decay of $\widetilde{\boldsymbol{X}}$ in the polynomial kernel case: MSE of the expected excess risk, the variance in Eq. (10), our derived $\mathtt{V}_1$, the bias in Eq. (9), and our derived convergence rate $\mathcal{O}(n^{-2\vartheta r})$ with $r = 1$ in Theorem 2 under different ϑ.



(a) $\vartheta = 2/3$ (b) $\vartheta = 2/3$ (c) $\vartheta = 1/3$ (d) $\vartheta = 1/3$

Figure 10: Exponential decay of $\widetilde{\boldsymbol{X}}$ in the polynomial kernel case: MSE of the expected excess risk, the variance in Eq. (10), our derived $\mathtt{V}_1$, the bias in Eq. (9), and our derived convergence rate $\mathcal{O}(n^{-2\vartheta r})$ with $r = 1$ in Theorem 2 under different ϑ.

and thus the risk curve is bell-shaped. However, in Figure 10, $\mathtt{V}_1$ on the *exponential decay* monotonically decreases with n even if we set $\bar{c}$ to 10^{-5}, 10^{-8} for a small regularization scheme.

Here we attempt to explain this phenomenon. In our setting, γ is set to zero. The condition in Eq. (19) can be reformulated as

$$(2\vartheta - 1)\bar{c} \leq e^{-a}(1 - \vartheta)\,.$$

Clearly, if we choose $0 < \vartheta < 1/2$, the condition in Eq. (19) always holds. Hence, $\mathtt{V}_1$ will monotonically decreases with n. If $1/2 < \vartheta < 1$, we examine our result with $a = 1$ and $\vartheta = 2/3$. We conclude that the used $\bar{c} = 0.01 < e^{-1}$, so the tend of $\mathtt{V}_1$ is monotonically decreasing with n.