Appendix

A Glossary

Table 2: Glossary

Name	Notation	Expression	Dimension
sampling distribution	ρ	-	$\mathcal{X} \to \mathbb{R}^+$
sampling size	N	-	integer
input matrix	$\mathbf{X}$	$(x_i)_{i=1}^N \underset{iid}{\sim} \rho_{\mathcal{X}}$	$N \times d$
output vector	$\mathbf{y}$	$(y_i)_{i=1}^N$	$N \times 1$
sample	$\mathbf{Z}$	$(\mathbf{X}, \mathbf{y})$	$N \times (d+1)$
noise	ε	-	random scalar
noise variance	σ^2	$\mathbb{E}[\varepsilon^2]$	scalar
ridge	λ	-	scalar
finite-rank kernel	K	$\sum_{k=1}^M \lambda_k \psi_k(\cdot)\psi_k(\cdot)$	$\mathcal{X} \times \mathcal{X} \to \mathbb{R}$
kernel rank	M	-	integer
kth eigenfunction	ψ_k	-	$\mathcal{X} \to \mathbb{R}$
kth value	λ_k	-	scalar
-	$\boldsymbol{\psi}(x)$	$[\psi_k(x)]_{k=1}^M$	$M \times 1$
-	$\mathbf{\Psi}$	$[\psi_k(x_i)]_{k,i}$	$M \times N$
-	$\mathbf{\Lambda}$	$\mathrm{diag}\left[\lambda_k\right]$	$M \times M$
kernel matrix	$\mathbf{K}$	$[K(x_i, x_j)]_{i,j} = \mathbf{\Psi}^\top \mathbf{\Lambda} \mathbf{\Psi}$	$N \times N$
resolvent	$\mathbf{R}$	$(\mathbf{K} + \lambda N \mathbf{I}_N)^{-1}$	$N \times N$
target function	$\tilde{f}$	$\sum_{k=1}^M \tilde{\gamma}_k \psi_k + \tilde{\gamma}_{>M}\psi_{>M}$	$\mathcal{X} \to \mathbb{R}$
-	$\tilde{f}_{\leq M}$	$\sum_{k=1}^M \tilde{\gamma}_k \psi_k$	$\mathcal{X} \to \mathbb{R}$
kth target coefficient	$\tilde{\gamma}_k$	$\int_{\mathcal{X}} \tilde{f}(x)\psi_k(x) d\rho_{\mathcal{X}}(x)$	scalar
-	$\boldsymbol{\gamma}$	$[\gamma_k]$	$M \times 1$
orthonormal complement	$\psi_{>M}$	-	$\mathcal{X} \to \mathbb{R}$
complementary coefficient	$\tilde{\gamma}_{>M}$	-	scalar
-	$\mathbf{\Psi}_{>M}$	$[\psi_{>M}(x_i)]$	$1 \times N$
test error	$\mathcal{R}_{\mathbf{Z},\lambda}$	$\mathbb{E}_{x,\epsilon}\left[(f_{\mathbf{Z},\lambda}(x) - \tilde{f}(x))^2\right]$	scalar
bias	-	$\int_{\mathcal{X}} \left(f_{(\mathbf{X},\tilde{f}(\mathbf{X})),\lambda}(x) - \tilde{f}(x)\right)^2 d\rho(x)$	scalar
variance	-	$\mathcal{R}_{\mathbf{Z},\lambda} - \text{bias} = \mathbb{E}_{x,\varepsilon}\left(\mathbf{K}_x^\top \mathbf{R} \boldsymbol{\varepsilon}\right)^2$	scalar
fluctuation matrix	$\mathbf{\Delta}$	$\frac{1}{N}\mathbf{\Psi}\mathbf{\Psi}^\top - \mathbf{I}_M$	$M \times M$
fluctuation	δ	$\lVert\mathbf{\Delta}\rVert_{\text{op}}$	scalar
error vector	$\boldsymbol{E}$	$[\eta_k]$	$M \times 1$
-	η_k	$\frac{1}{N}\sum_{i=1}^N \psi_k(x_i)\psi_{>M}(x_i)$	scalar
-	$\mathbf{B}$	$(\mathbf{I}_M + \mathbf{\Delta} + \lambda \mathbf{\Lambda}^{-1})^{-1}$	$M \times M$
-	$\bar{\mathbf{P}}$	$\mathrm{diag}\left[\frac{\lambda_k}{\lambda_k + \lambda}\right]$	$M \times M$

B Classical KRR Theory

In an effort to keep our manuscript as self-contained as possible, we recall the Mercer decomposition, representer theorem for kernel ridge regression as well as the form of the bias-variance tradeoff in the KRR context.

B.1 Mercer Decomposition

We begin with a general kernel $K^{(\infty)} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

Proposition B.1. *[12] Fix a sample distribution ρ. Let $K^{(\infty)} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a reproducing kernel with corresponding RKHS $\mathcal{H}^{(\infty)}$. There exists a decreasing sequence of real numbers $\lambda_1 \geq \lambda_2 \geq \ldots$, called the eigenvalues of the kernel $K^{(\infty)}$; and a sequence of pairwise-orthonormal functions $\{\psi_k\}_{k=1}^{\infty} \subset L^2_\rho$, called the eigenfunctions of $K^{(\infty)}$, such that for all $x, x' \in \mathcal{X}$, we have*

$$K^{(\infty)}(x, x') = \sum_{k=1}^{\infty} \lambda_k \psi_k(x) \psi_k(x') \tag{14}$$

In particular, we assume $\lambda_k = 0$, $\forall k > M$. In this case, we say the kernel $K(x, x') = \sum_{k=1}^{M} \lambda_k \psi_k(x) \psi_k(x')$ is of finite rank M with corresponding (finite-dimensional) RKHS $\mathcal{H}$, recovering equation (2).

The first of these results, allows us to explicitly express the finite-rank kernel ridge regressor $f_{\mathbf{Z},\lambda}$.

Proposition B.2 (Representer Theorem - [38, Chapter 12]). *Let $\mathbf{R} \stackrel{\text{def.}}{=} (\mathbf{K} + \lambda N \mathbf{I}_N)^{-1} \in \mathbb{R}^{N \times N}$ be the resolvent matrix and recall the kernel ridge regressor $f_{\mathbf{Z},\lambda}$ given by equation (3):*

$$f_{\mathbf{Z},\lambda} \stackrel{\text{def.}}{=} \arg\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

Then, for every $x \in \mathcal{X}$, we have the expression

$$f_{\mathbf{Z},\lambda}(x) = \mathbf{y}^\top \mathbf{R} \mathbf{K}_x, \ \forall x \in \mathcal{X}, \tag{15}$$

where $\mathbf{K}_x \stackrel{\text{def.}}{=} [K(x_i, x)]_{i=1}^N \in \mathbb{R}^{N \times 1}$.

B.2 Compact Matrix Expression

First, let $\mathbf{\Psi} \stackrel{\text{def.}}{=} (\psi_k(x_i))_{k=1,i=1}^{M,N}$ be the random $M \times N$ matrix defined by evaluating the M eigenfunctions on all input training instances $\boldsymbol{X} \stackrel{\text{def.}}{=} (x_i)_{i=1}^N$, $\mathbf{\Lambda} \stackrel{\text{def.}}{=} \operatorname{diag}[\lambda_k] \in \mathbb{R}^{M \times M}$, and $\boldsymbol{\psi}(x) \stackrel{\text{def.}}{=} [\psi_k(x)]_{k=1}^M \in \mathbb{R}^{M \times 1}$. The advantage of this notation is that we can rewrite the equations in a more compact form. For equation (15):

$$f_{\mathbf{Z},\lambda}(x) = \mathbf{y}^\top \underbrace{(\mathbf{\Psi}^\top \mathbf{\Lambda} \mathbf{\Psi} + \lambda N \mathbf{I}_M)^{-1}}_{\mathbf{R}} \mathbf{\Psi}^\top \mathbf{\Lambda} \boldsymbol{\psi}(x); \tag{16}$$

for equation (4):

$$\tilde{f}(x) = \tilde{\boldsymbol{\gamma}}^\top \boldsymbol{\psi}(x) + \tilde{\gamma}_{>M} \psi_{>M}(x). \tag{17}$$

Last but not least, we define some important quantities for later analysis.

Definition B.3 (Fluctuation matrix). *The fluctuation matrix is the random $M \times M$-matrix given by $\mathbf{\Delta} \stackrel{\text{def.}}{=} \frac{1}{N} \mathbf{\Psi} \mathbf{\Psi}^\top - \mathbf{I}_M$. Our analysis will often involve the operator norm of $\mathbf{\Delta}$, which we denote by $\delta \stackrel{\text{def.}}{=} \|\mathbf{\Delta}\|_{op}$.*

The fluctuation matrix $\mathbf{\Delta}$ measures the first source of randomness in the KRR's test error. Namely it encodes the degree of non-orthonormality between the vectors obtained by evaluating of the M eigenfunctions $\psi_1, \ldots, \psi_M$ on the input $\mathbf{X}$.

The second source of randomness in the KRR's test error comes from the empirical evaluation of the dot product of the eigenfunction ψ_k's and the orthogonal complement $\psi_{>M}$:

Definition B.4 (Error Vector). $\boldsymbol{E} \stackrel{\text{def.}}{=} \frac{1}{N} \mathbf{\Psi} \psi_{>M}(\mathbf{X})$ *is called the error vector.*

The random matrix $\mathbf{\Delta}$ and the random vector $\boldsymbol{E}$ are centered; i.e. $\mathbb{E}_{\mathbf{X}}[\mathbf{\Delta}] = 0$ and $\mathbb{E}_{\mathbf{X}}[\boldsymbol{E}] = 0$.

B.3 Bias-Variance Decomposition

The derivation of several contemporary KRR generalization bounds [6, 26, 27] involves the classical Bias-Variance Trade-off:

Proposition B.5 (Bias-Variance Trade-off)**.** *Fix a sample* $\mathbf{Z}$*. Recall the definition 3.4 of test error* $\mathcal{R}_{\mathbf{Z},\lambda}$*, bias, and variance:*

$$\begin{aligned}
\mathcal{R}_{\mathbf{Z},\lambda} &\overset{\text{def.}}{=} \mathbb{E}_{x,\epsilon}\left[(f_{\mathbf{Z},\lambda}(x) - \tilde{f}(x))^2\right] = \mathbb{E}_{\epsilon}\left[\int_{\mathcal{X}} \left(f_{\mathbf{Z},\lambda}(x) - \tilde{f}(x)\right)^2 d\rho(x)\right];\\
\textit{bias} &\overset{\text{def.}}{=} \int_{\mathcal{X}} \left(f_{(\mathbf{X},\tilde{f}(\mathbf{X})),\lambda}(x) - \tilde{f}(x)\right)^2 d\rho(x);\\
\textit{variance} &\overset{\text{def.}}{=} \mathcal{R}_{\mathbf{Z},\lambda} - \textit{bias}.
\end{aligned}$$

Then, we can write $\textit{variance}_{\textit{test}} = \mathbb{E}_{x,\varepsilon}\left(\mathbf{K}_x^\top \mathbf{R}\boldsymbol{\varepsilon}\right)^2$ *and hence the test error* $\mathcal{R}_{\mathbf{Z},\lambda}$ *admits a decomposition:*

$$R_{\mathbf{Z},\lambda} = \textit{bias} + \mathbb{E}_{x,\varepsilon}\left(\mathbf{K}_x^\top \mathbf{R}\boldsymbol{\varepsilon}\right)^2.$$

Proof. See the proof of Theorem C.8. □

C Proofs

In this section, we will derive the essential lemmata and propositions for proving the main theorems.

C.1 Formula Derivation

We begin with writing the test error in convenient forms.

C.1.1 Bias

We first derive, from the definition of the bias, a convenient expression to proceed:

Proposition C.1 (Bias Expression)**.** *Let* $\boldsymbol{\Psi}_{>M} \overset{\text{def.}}{=} [\psi_{>M}(x_i)]_{i=1}^N$ *as an* $1 \times N$*- row vector,* $\binom{\boldsymbol{\Psi}}{\boldsymbol{\Psi}_{>M}}$ *as an* $(M+1) \times N$ *matrix. Denote* $\mathbf{P} \overset{\text{def.}}{=} \begin{pmatrix}\mathbf{P}_{\leq M} & \mathbf{P}_{>M}\end{pmatrix} = \boldsymbol{\Lambda}\boldsymbol{\Psi}\mathbf{R}\begin{pmatrix}\boldsymbol{\Psi}^\top & \boldsymbol{\Psi}_{>M}^\top\end{pmatrix} \in \mathbb{R}^{M\times(M+1)}$ *,* $\mathbf{P}_{\leq M} \in \mathbb{R}^{M\times M}$ *and* $\mathbf{P}_{>M} \in \mathbb{R}^{M\times 1}$*. Then the bias admits the following expression:*

$$\textit{bias} = \underbrace{\tilde{\gamma}_{>M}^2}_{\textit{Finite Rank Error}} + \underbrace{\|\tilde{\boldsymbol{\gamma}} - \mathbf{P}_{\leq M}\tilde{\boldsymbol{\gamma}} - \tilde{\gamma}_{>M}\mathbf{P}_{>M}\|_2^2}_{\textit{Fitting Error}}.$$

Proof. Recall that, by equations (16) and (17), we can write

$$\begin{aligned}
\tilde{f}(x) &= \tilde{\boldsymbol{\gamma}}^\top \boldsymbol{\psi}(x) + \tilde{\gamma}_{>M}\psi_{>M}(x),\\
f_{(\mathbf{X},\tilde{f}(\mathbf{X})),\lambda}(x) &= (\tilde{\boldsymbol{\gamma}}^\top \boldsymbol{\Psi} + \tilde{\gamma}_{>M}\boldsymbol{\Psi}_{>M}^\top)\mathbf{R}\boldsymbol{\Psi}^\top\boldsymbol{\Lambda}\boldsymbol{\Psi}(x).
\end{aligned}$$

Hence

$$\begin{aligned}
\text{bias} &= \mathbb{E}_x\left[\left(\tilde{\boldsymbol{\gamma}}^\top \boldsymbol{\psi}(x) + \tilde{\gamma}_{>M}\psi_{>M}(x) - (\tilde{\boldsymbol{\gamma}}^\top \boldsymbol{\Psi} + \tilde{\gamma}_{>M}\boldsymbol{\Psi}_{>M}^\top)\mathbf{R}\boldsymbol{\Psi}^\top\boldsymbol{\Lambda}\boldsymbol{\Psi}(x)\right)^2\right]\\
&= \left\|\begin{pmatrix}\tilde{\boldsymbol{\gamma}}\\ \tilde{\gamma}_{>M}\end{pmatrix} - \begin{pmatrix}\mathbf{P}\binom{\tilde{\boldsymbol{\gamma}}}{\tilde{\gamma}_{>M}}\\ 0\end{pmatrix}\right\|_2^2 \qquad (18)\\
&= \underbrace{\tilde{\gamma}_{>M}^2}_{\text{Finite Rank Error}} + \underbrace{\|\tilde{\boldsymbol{\gamma}} - \mathbf{P}_{\leq M}\tilde{\boldsymbol{\gamma}} - \tilde{\gamma}_{>M}\mathbf{P}_{>M}\|_2^2}_{\text{Fitting Error}},
\end{aligned}$$

in line (18), we use Parseval's identity. □

We proceed by reformulating the projection matrix $\mathbf{P}$, first with the left matrix $\mathbf{P}_{\leq M}$:

Lemma C.2. *Recall the following notations*

$$\mathbf{K} \stackrel{\text{def.}}{=} \mathbf{\Psi}^\top \mathbf{\Lambda} \mathbf{\Psi}, \quad \mathbf{R} \stackrel{\text{def.}}{=} (\mathbf{K} + \lambda N \mathbf{I}_M)^{-1}, \quad \mathbf{P}_{\leq M} \stackrel{\text{def.}}{=} \mathbf{\Lambda} \mathbf{\Psi} \mathbf{R} \mathbf{\Psi}^\top.$$

Define the symmetric random matrix $\mathbf{B} \stackrel{\text{def.}}{=} (\mathbf{I}_M + \mathbf{\Delta} + \lambda \mathbf{\Lambda}^{-1})^{-1}$. *It holds that*

$$\mathbf{P}_{\leq M} = \mathbf{I}_M - \lambda \mathbf{B} \mathbf{\Lambda}^{-1}.$$

Proof. We first observe that

$$\begin{aligned}
\mathbf{\Psi}\mathbf{\Psi}^\top \mathbf{P}_{\leq M} &= \mathbf{\Psi}\mathbf{\Psi}^\top \mathbf{\Lambda} \mathbf{\Psi} \mathbf{R} \mathbf{\Psi}^\top && (19)\\
&= \mathbf{\Psi} \mathbf{K} (\mathbf{K} + \lambda N \mathbf{I}_M)^{-1} \mathbf{\Psi}^\top \\
&= \mathbf{\Psi} \left(\mathbf{I}_M - \lambda N (\mathbf{K} + \lambda N \mathbf{I}_M)^{-1} \right) \mathbf{\Psi}^\top \\
&= \mathbf{\Psi}\mathbf{\Psi}^\top - \lambda N \mathbf{\Psi} (\mathbf{K} + \lambda N \mathbf{I}_M)^{-1} \mathbf{\Psi}^\top. && (20)
\end{aligned}$$

From lines (19)- (20) and the definition of the fluctuation matrix $\mathbf{\Delta}$ we deduce

$$\begin{aligned}
\frac{1}{N} \mathbf{\Psi}\mathbf{\Psi}^\top (\mathbf{I}_M - \mathbf{P}_{\leq M}) &= \lambda \mathbf{\Psi} (\mathbf{K} + \lambda N \mathbf{I}_M)^{-1} \mathbf{\Psi}^\top \\
(\mathbf{I}_M + \mathbf{\Delta})(\mathbf{I}_M - \mathbf{P}_{\leq M}) &= \lambda \mathbf{\Psi} \mathbf{R} \mathbf{\Psi}^\top \\
(\mathbf{I}_M + \mathbf{\Delta})(\mathbf{I}_M - \mathbf{P}_{\leq M}) &= \lambda \mathbf{\Lambda}^{-1} \mathbf{P}_{\leq M} \\
(\mathbf{\Lambda} + \mathbf{\Lambda}\mathbf{\Delta})(\mathbf{I}_M - \mathbf{P}_{\leq M}) &= \lambda \mathbf{P}_{\leq M} \\
\mathbf{\Lambda} + \mathbf{\Lambda}\mathbf{\Delta} &= (\mathbf{\Lambda} + \mathbf{\Lambda}\mathbf{\Delta} + \lambda \mathbf{I}_M) \mathbf{P}_{\leq M}. && (21)
\end{aligned}$$

Rearranging (21) and applying the definition of $\boldsymbol{B}$ we find that

$$\begin{aligned}
\mathbf{P}_{\leq M} &= (\mathbf{\Lambda} + \mathbf{\Lambda}\mathbf{\Delta} + \lambda \mathbf{I}_M)^{-1} (\mathbf{\Lambda} + \mathbf{\Lambda}\mathbf{\Delta}) && (22)\\
&= \mathbf{I}_M - \lambda (\mathbf{\Lambda} + \mathbf{\Lambda}\mathbf{\Delta} + \lambda \mathbf{I}_M)^{-1} \\
&= \mathbf{I}_M - \lambda (\mathbf{\Lambda} + \mathbf{\Lambda}\mathbf{\Delta} + \lambda \mathbf{I}_M)^{-1} \mathbf{\Lambda}\mathbf{\Lambda}^{-1} \\
&= \mathbf{I}_M - \lambda \mathbf{B} \mathbf{\Lambda}^{-1}. && (23)
\end{aligned}$$

□

Arguing analogously for the right matrix $\mathbf{P}_{>M}$, we draw the subsequent similar conclusion.

Lemma C.3. *Recall the following notations*

$$\mathbf{K} \stackrel{\text{def.}}{=} \mathbf{\Psi}^\top \mathbf{\Lambda} \mathbf{\Psi}, \quad \mathbf{R} \stackrel{\text{def.}}{=} (\mathbf{K} + \lambda N \mathbf{I}_M)^{-1}, \quad \mathbf{P}_{>M} \stackrel{\text{def.}}{=} \mathbf{\Lambda} \mathbf{\Psi} \mathbf{R} \mathbf{\Psi}_{>M}^\top,$$

$$\boldsymbol{E} \stackrel{\text{def.}}{=} \frac{1}{N} \mathbf{\Psi}\mathbf{\Psi}_{>M}^\top, \quad \mathbf{B} \stackrel{\text{def.}}{=} (\mathbf{I}_M + \mathbf{\Delta} + \lambda \mathbf{\Lambda}^{-1})^{-1}.$$

We have that $\mathbf{P}_{>M} = \mathbf{B}\boldsymbol{E}$.

Proof. Similarly to (19)- (20) we note that

$$\begin{aligned}
\mathbf{\Psi}\mathbf{\Psi}^\top \mathbf{P}_{>M} &= \mathbf{\Psi}\mathbf{\Psi}^\top \mathbf{\Lambda} \mathbf{\Psi} \mathbf{R} \mathbf{\Psi}_{>M}^\top \\
&= \mathbf{\Psi} \mathbf{K} (\mathbf{K} + \lambda N \mathbf{I}_M)^{-1} \mathbf{\Psi}_{>M}^\top \\
&= \mathbf{\Psi} \left(\mathbf{I}_M - \lambda N (\mathbf{K} + \lambda N \mathbf{I}_M)^{-1} \right) \mathbf{\Psi}_{>M}^\top \\
&= \mathbf{\Psi}\mathbf{\Psi}_{>M}^\top - \lambda N \mathbf{\Psi} (\mathbf{K} + \lambda N \mathbf{I}_M)^{-1} \mathbf{\Psi}_{>M}^\top.
\end{aligned}$$

Analogously to the computations in (22)-(23)

$$\begin{aligned}
(\mathbf{I}_M + \mathbf{\Delta}) \mathbf{P}_{>M} &= \boldsymbol{E} - \lambda \mathbf{\Psi} (\mathbf{K} + \lambda N \mathbf{I}_M)^{-1} \mathbf{\Psi}_{>M}^\top \\
(\mathbf{I}_M + \mathbf{\Delta}) \mathbf{P}_{>M} &= \boldsymbol{E} - \lambda \mathbf{\Lambda}^{-1} \mathbf{\Lambda} \mathbf{\Psi} (\mathbf{K} + \lambda N \mathbf{I}_M)^{-1} \mathbf{\Psi}_{>M}^\top \\
(\mathbf{I}_M + \mathbf{\Delta}) \mathbf{P}_{>M} &= \boldsymbol{E} - \lambda \mathbf{\Lambda}^{-1} \mathbf{P}_{>M} \\
(\mathbf{\Lambda} + \mathbf{\Lambda}\mathbf{\Delta}) \mathbf{P}_{>M} &= \mathbf{\Lambda} \boldsymbol{E} - \lambda \mathbf{P}_{>M} \\
(\mathbf{\Lambda} + \mathbf{\Lambda}\mathbf{\Delta} + \lambda \mathbf{I}_M) \mathbf{P}_{>M} &= \mathbf{\Lambda} \boldsymbol{E} \\
\mathbf{P}_{>M} &= (\mathbf{\Lambda} + \mathbf{\Lambda}\mathbf{\Delta} + \lambda \mathbf{I}_M)^{-1} \mathbf{\Lambda} \boldsymbol{E} \\
\mathbf{P}_{>M} &= \mathbf{B} \boldsymbol{E}.
\end{aligned}$$

□

Lemma C.4 (Fitting Error)**.** *Recall the notation*

$$\textit{fitting error} = \|\tilde{\gamma} - \mathbf{P}_{\leq M}\tilde{\gamma} - \tilde{\gamma}_{>M}\mathbf{P}_{>M}\|_2^2,$$
$$\mathbf{B} \overset{\text{def.}}{=} (\mathbf{I}_M + \mathbf{\Delta} + \lambda\mathbf{\Lambda}^{-1})^{-1}.$$

We have fitting error $= \left\|\mathbf{B}\left(\lambda\mathbf{\Lambda}^{-1}\tilde{\gamma} - \boldsymbol{E}\tilde{\gamma}_{>M}\right)\right\|_2^2.$

Proof. By lemmata C.2 and C.3,

$$\begin{aligned}\|\tilde{\gamma} - \mathbf{P}_{\leq M}\tilde{\gamma} - \tilde{\gamma}_{>M}\mathbf{P}_{>M}\|_2^2 &= \left\|\tilde{\gamma} - (\mathbf{I}_M - \lambda\mathbf{B}\mathbf{\Lambda}^{-1})\tilde{\gamma}\tilde{\gamma} - \mathbf{B}\boldsymbol{E}\tilde{\gamma}_{>M}\right\|_2^2 \\ &= \left\|\mathbf{B}\left(\lambda\mathbf{\Lambda}^{-1}\tilde{\gamma} - \boldsymbol{E}\tilde{\gamma}_{>M}\right)\right\|_2^2.\end{aligned}$$

□

Hence we come up with a new expression of the bias:

Proposition C.5 (Bias)**.** *Recall that* $\mathbf{B} \overset{\text{def.}}{=} (\mathbf{I}_M + \mathbf{\Delta} + \lambda\mathbf{\Lambda}^{-1})^{-1}$. *The bias* $\mathbb{E}_x\left(f_{\mathbf{X}}^{\lambda}(x) - \tilde{f}(x)\right)^2$ *has the following expression:*

$$\textit{bias} = \tilde{\gamma}_{>M}^2 + \left\|\mathbf{B}\left(\lambda\mathbf{\Lambda}^{-1}\tilde{\gamma} - \tilde{\gamma}_{>M}\boldsymbol{E}\right)\right\|_2^2.$$

Proof. We apply Proposition C.1 and Lemma C.4 to obtain the result. □

C.1.2 Variance

If we consider noise in the label, we have to compute the variance part of the test error.

Proposition C.6 (Variance Expression)**.** *Define*

$$\begin{aligned}\mathbf{M} &\overset{\text{def.}}{=} \mathbb{E}_x[\mathbf{K}_x\mathbf{K}_x^\top] \\ &= \mathbb{E}_x[\mathbf{\Psi}^\top\mathbf{\Lambda}\boldsymbol{\psi}(x)\boldsymbol{\psi}(x)^\top\mathbf{\Lambda}\mathbf{\Psi}] \\ &= \mathbf{\Psi}^\top\mathbf{\Lambda}\mathbb{E}_x[\boldsymbol{\psi}(x)\boldsymbol{\psi}(x)^\top]\mathbf{\Lambda}\mathbf{\Psi} \\ &= \mathbf{\Psi}^\top\mathbf{\Lambda}\mathbf{I}_M\mathbf{\Lambda}\mathbf{\Psi} \\ &= \mathbf{\Psi}^\top\mathbf{\Lambda}^2\mathbf{\Psi}.\end{aligned}$$

We can further simplify the variance part:

$$\begin{aligned}\textit{variance} &\overset{\text{def.}}{=} \mathbb{E}_{x,\varepsilon}\left[\left(\mathbf{K}_x^\top\mathbf{R}\boldsymbol{\varepsilon}\right)^2\right] \\ &= \mathbb{E}_{x,\varepsilon}\left[\boldsymbol{\varepsilon}^\top\mathbf{R}\mathbf{K}_x\mathbf{K}_x^\top\mathbf{R}\boldsymbol{\varepsilon}\right] \\ &= \mathbb{E}_{\varepsilon}\left[\boldsymbol{\varepsilon}^\top\mathbf{R}\mathbf{M}\mathbf{R}\boldsymbol{\varepsilon}\right] \\ &= \sigma^2 \cdot \mathrm{Tr}[\mathbf{R}\mathbf{M}\mathbf{R}].\end{aligned}$$

Theorem C.7 (Variance)**.** *Recall that* $\mathbf{B} \overset{\text{def.}}{=} (\mathbf{I}_M + \mathbf{\Delta} + \lambda\mathbf{\Lambda}^{-1})^{-1}$. *The variance part, variance, can be expressed as:*

$$\textit{variance} = \frac{\sigma^2}{N}\,\mathrm{Tr}\left[\mathbf{B}^2(\mathbf{I}_M + \mathbf{\Delta})\right].$$

Proof. We argue similarly as in lemma C.2. Since

$$\begin{aligned}\mathbf{\Psi}\mathbf{\Psi}^\top\mathbf{\Lambda}\mathbf{\Psi}\mathbf{R} &= \mathbf{\Psi}\mathbf{K}(\mathbf{K} + \lambda N\mathbf{I}_M)^{-1} \\ &= \mathbf{\Psi}(\mathbf{I}_M - \lambda N\mathbf{R}) \\ &= \mathbf{\Psi} - \lambda N\mathbf{\Psi}\mathbf{R},\end{aligned}$$

therefore, we deduce that

$$\begin{aligned}
(\mathbf{I}_M + \boldsymbol{\Delta})\boldsymbol{\Lambda}\boldsymbol{\Psi}\mathbf{R} &= \frac{1}{N}\boldsymbol{\Psi} - \lambda\boldsymbol{\Psi}\mathbf{R} \qquad (24)\\
(\mathbf{I}_M + \boldsymbol{\Delta})\boldsymbol{\Lambda}\boldsymbol{\Psi}\mathbf{R} &= \frac{1}{N}\boldsymbol{\Psi} - \lambda\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Psi}\mathbf{R}\\
(\mathbf{I}_M + \boldsymbol{\Delta} + \lambda\boldsymbol{\Lambda}^{-1})\boldsymbol{\Lambda}\boldsymbol{\Psi}\mathbf{R} &= \frac{1}{N}\boldsymbol{\Psi}\\
\boldsymbol{\Lambda}\boldsymbol{\Psi}\mathbf{R} &= \frac{1}{N}(\mathbf{I}_M + \boldsymbol{\Delta} + \lambda\boldsymbol{\Lambda}^{-1})^{-1}\boldsymbol{\Psi}\\
\boldsymbol{\Lambda}\boldsymbol{\Psi}\mathbf{R} &= \frac{1}{N}\mathbf{B}\boldsymbol{\Psi}. \qquad (25)
\end{aligned}$$

By leveraging the identity $\mathbf{M} = \boldsymbol{\Psi}^\top\boldsymbol{\Lambda}^2\boldsymbol{\Psi}$ and elementary properties of the trace map, the computations in (24)-(25) imply that

$$\begin{aligned}
\operatorname{Tr}[\mathbf{R}\mathbf{M}\mathbf{R}] &= \operatorname{Tr}[\mathbf{R}\boldsymbol{\Psi}^\top\boldsymbol{\Lambda}^2\boldsymbol{\Psi}\mathbf{R}] \qquad (26)\\
&= \operatorname{Tr}\left[(\boldsymbol{\Lambda}\boldsymbol{\Psi}\mathbf{R})^\top(\boldsymbol{\Lambda}\boldsymbol{\Psi}\mathbf{R})\right] \qquad (27)\\
&= \operatorname{Tr}\left[\left(\frac{1}{N}\mathbf{B}\boldsymbol{\Psi}\right)^\top\left(\frac{1}{N}\mathbf{B}\boldsymbol{\Psi}\right)\right] \qquad (28)\\
&= \frac{1}{N}\operatorname{Tr}\left[\frac{1}{N}\boldsymbol{\Psi}^\top\mathbf{B}^\top\mathbf{B}\boldsymbol{\Psi}\right]\\
&= \frac{1}{N}\operatorname{Tr}\left[\mathbf{B}^\top\mathbf{B}\cdot\frac{1}{N}\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\right] \qquad (29)\\
&= \frac{1}{N}\operatorname{Tr}\left[\mathbf{B}^\top\mathbf{B}(\mathbf{I}_M + \boldsymbol{\Delta})\right] \qquad (30)\\
&= \frac{1}{N}\operatorname{Tr}\left[\mathbf{B}^2(\mathbf{I}_M + \boldsymbol{\Delta})\right]; \qquad (31)
\end{aligned}$$

in more detail: in line (26), we use the definition of $\mathbf{M}$; in line (27), we use the fact that both $\boldsymbol{\Lambda}$ and $\mathbf{R}$ are symmetric; in line (28), we use line (25); in line (29), we use the cyclicity of the trace; in line (30), we use the definition of $\boldsymbol{\Delta}$; in line (31), we use the symmetry of $\mathbf{B}$. We obtain the result upon applying Lemma C.6. □

C.1.3 Test Error

The Bias-Variance trade-off (see Proposition B.5) decomposed the KRR's test error into two terms, the bias and variance. Since Propositions C.5 and C.7 give us exact expressions for the bias and variance, respectively, we deduce the following exact expression for the KRR's test error.

Theorem C.8 (Exact Formula for KRR's Test Error)**.** *The test error $\mathcal{R}_{\mathbf{Z},\lambda}$ of KRR equals*

$$\mathcal{R}_{\mathbf{Z},\lambda} = \underbrace{\overbrace{\left\|\mathbf{B}\left(\lambda\boldsymbol{\Lambda}^{-1}\tilde{\boldsymbol{\gamma}} - \tilde{\gamma}_{>M}\boldsymbol{E}_M\right)\right\|_2^2}^{\textit{fitting error}} + \overbrace{\tilde{\gamma}_{>M}^2}^{\textit{finite rank error}}}_{\textit{bias}} + \underbrace{\frac{\sigma^2_{\textit{noise}}}{N}\operatorname{Tr}\left[\mathbf{B}^2(\mathbf{I}_M + \boldsymbol{\Delta})\right]}_{\textit{variance}},$$

where $\mathbf{B} \overset{\text{def.}}{=} (\mathbf{I}_M + \boldsymbol{\Delta} + \lambda\boldsymbol{\Lambda}^{-1})^{-1}$.

Proof. We begin with the bias/variance decomposition:

$$\begin{aligned}
R^\lambda_{\mathbf{Z}} &\overset{\text{def.}}{=} \mathbb{E}_y\|f^\lambda_{\mathbf{Z}} - \tilde{f}\|^2_{L^2_{\rho_\mathcal{X}}}\\
&= \mathbb{E}_{x,y}\left(\mathbf{K}_x^\top\mathbf{R}\mathbf{y} - \tilde{f}(x)\right)^2\\
&= \mathbb{E}_{\varepsilon,x}\left(\mathbf{K}_x^\top\mathbf{R}(\tilde{f}(\mathbf{X}) + \boldsymbol{\varepsilon}) - \tilde{f}(x)\right)^2\\
&= \mathbb{E}_x\left(f^\lambda_{\mathbf{X}}(x) - \tilde{f}(x)\right)^2 + \mathbb{E}_{x,\varepsilon}\left[\left(\mathbf{K}_x^\top\mathbf{R}\boldsymbol{\varepsilon}\right)^2\right]\\
&= \text{bias} + \text{variance},
\end{aligned}$$

then we apply Propositions C.5 and C.7. $\square$

For the validation of the Theorem C.8, please see Appendix D for details.

The matrix $\mathbf{B}$ plays an important role in the expression since it encodes most information of the KRR. Therefore, the following subsection will discuss the approximation of the matrix $\mathbf{B}$.

C.2 Matrix Approximation

Recall that the matrix $\mathbf{B} \overset{\text{def.}}{=} (\mathbf{I}_M + \mathbf{\Delta} + \lambda\mathbf{\Lambda}^{-1})^{-1}$ is the inverse of a random matrix. The following lemma helps to approximate $\mathbf{B}$. Informally, it says that: given that $\delta \overset{\text{def.}}{=} \|\mathbf{\Delta}\|_{\text{op}} < 1$. We have

$$\mathbf{B} = \sum_{s=0}^{\infty} (-\bar{\mathbf{P}}\mathbf{\Delta})^s \bar{\mathbf{P}}$$

in operator norm $\|\cdot\|_{op}$ for an $M \times M$ matrix $\bar{P}$ depending only on the M eigenvalues $\{\lambda_k\}_{k=1}^M$ and on the ridge $\lambda > 0$. More precisely we have the following.

Lemma C.9 ($\mathbf{B}$-Expansion)**.** *Given that $\delta \overset{\text{def.}}{=} \|\mathbf{\Delta}\|_{op} < 1$. It holds that*

$$\lim_{n \uparrow \infty} \left\| \mathbf{B} - \sum_{s=0}^{n} (-\bar{\mathbf{P}}\mathbf{\Delta})^s \bar{\mathbf{P}} \right\|_{op} = 0$$

where $\bar{\mathbf{P}} \overset{\text{def.}}{=} \operatorname{diag}\left[\frac{\lambda_k}{\lambda_k + \lambda}\right]_k = \mathbf{\Lambda}(\mathbf{\Lambda} + \lambda \mathbf{I}_M)^{-1} \in \mathbb{R}^{M \times M}$.

Proof. Set $\mathbf{A} = \mathbf{I}_M + \lambda\mathbf{\Lambda}^{-1}$ and repeatedly use the formula $(\mathbf{A}+\mathbf{\Delta})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{\Delta}(\mathbf{A}+\mathbf{\Delta})^{-1}$ from [31], we have

$$\begin{aligned}
\mathbf{B} &\overset{\text{def.}}{=} (\mathbf{I}_M + \mathbf{\Delta} + \lambda\mathbf{\Lambda}^{-1})^{-1} \\
&= (\mathbf{A} + \mathbf{\Delta})^{-1} \\
&= \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{\Delta}(\mathbf{A} + \mathbf{\Delta})^{-1} \\
&= \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{\Delta}\left(\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{\Delta}(\mathbf{A} + \mathbf{\Delta})^{-1}\right) \\
&= \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{\Delta}\mathbf{A}^{-1} + (\mathbf{A}^{-1}\mathbf{\Delta})^2(\mathbf{A} + \mathbf{\Delta})^{-1} \\
&= \sum_{s=0}^{n} (-\mathbf{A}^{-1}\mathbf{\Delta})^s \mathbf{A}^{-1} + (-\mathbf{A}^{-1}\mathbf{\Delta})^{n+1}(\mathbf{A} + \mathbf{\Delta})^{-1}
\end{aligned}$$

Note that $\mathbf{A}^{-1} = (\mathbf{I}_M + \lambda\mathbf{\Lambda}^{-1})^{-1} = \mathbf{\Lambda}(\mathbf{\Lambda} + \lambda\mathbf{I}_M)^{-1} = \bar{\mathbf{P}}$ with operator norm $\frac{\lambda_1}{\lambda_1 + \lambda} < 1$, hence we have $(\mathbf{A}^{-1}\mathbf{\Delta})^{n+1} = (-\bar{\mathbf{P}}\mathbf{\Delta})^{n+1} \to 0$ in operator norm as $n \to \infty$. Hence

$$\begin{aligned}
\mathbf{B} &= \sum_{s=0}^{\infty} (-\mathbf{A}^{-1}\mathbf{\Delta})^s \mathbf{A}^{-1} \\
&= \sum_{s=0}^{\infty} (-\bar{\mathbf{P}}\mathbf{\Delta})^s \bar{\mathbf{P}}
\end{aligned}$$

in operator norm. $\square$

Due to the convergence result in lemma C.9, it is natural to define:

Definition C.10. *For any $n \in \mathbb{N} \cup \{\infty\}$, write $\mathbf{B}^{(n)} = \sum_{s=0}^{n} (-\bar{\mathbf{P}}\mathbf{\Delta})^s \bar{\mathbf{P}}$. For example, We have*

$$\begin{aligned}
\mathbf{B}^{(0)} &= \bar{\mathbf{P}} \\
\mathbf{B}^{(1)} &= \bar{\mathbf{P}} - \bar{\mathbf{P}}\mathbf{\Delta}\bar{\mathbf{P}} \\
\mathbf{B}^{(\infty)} &= \mathbf{B}
\end{aligned}$$

Although lemma C.9 is valid when $\delta < 1$, we need a slightly stronger condition that δ is upper bounded by an arbitrary constant strictly small than 1. For simplicity, we assume this constant to be $\frac{1}{2}$ in the following lemma:

Lemma C.11 (B-Approximation)**.** *Assume that $\delta \stackrel{\text{def.}}{=} \|\mathbf{\Delta}\|_{op} < \frac{1}{2}$. Let $\mathbf{B}^{(n)} = \sum_{s=0}^{n}(-\bar{\mathbf{P}}\mathbf{\Delta})^s\bar{\mathbf{P}}$ be the nth-order approximation of the matrix $\mathbf{B}$ as in definition C.10. Then we have*

$$\left\|\mathbf{B} - \mathbf{B}^{(n)}\right\|_{op} < 2\delta^{n+1}.$$

Proof. We first bound the operator norm of the matrix $\mathbf{B}$: since the minimum singular value of the matrix $\bar{\mathbf{P}}^{-1} + \mathbf{\Delta}$ is at least

$$\frac{\lambda_k + \lambda}{\lambda_k} - \|\Delta\|_{\text{op}} \geq 1 + \frac{\lambda}{\lambda_k} - \frac{1}{2} > \frac{1}{2},$$

and hence

$$\|\mathbf{B}\|_{\text{op}} = \left\|\left(\bar{\mathbf{P}}^{-1} + \mathbf{\Delta}\right)^{-1}\right\|_{\text{op}} < 2.$$

Also, we have

$$\begin{aligned}
\mathbf{B} - \mathbf{B}^{(n)} &= \sum_{s=n+1}^{\infty}(-\bar{\mathbf{P}}\mathbf{\Delta})^s\bar{\mathbf{P}} \\
&= (-\bar{\mathbf{P}}\mathbf{\Delta})^{n+1}\sum_{s=0}^{\infty}(-\bar{\mathbf{P}}\mathbf{\Delta})^s\bar{\mathbf{P}} \\
&= (-\bar{\mathbf{P}}\mathbf{\Delta})^{n+1}\mathbf{B}.
\end{aligned}$$

Hence $\left\|\mathbf{B} - \mathbf{B}^{(n)}\right\|_{\text{op}} \leq \left\|\bar{\mathbf{P}}\mathbf{\Delta}\right\|_{\text{op}}^{n+1}\|\mathbf{B}\|_{\text{op}} < \|\mathbf{\Delta}\|_{\text{op}}^{n+1} \cdot 2 = 2\delta^{n+1}$, since we have $\left\|\bar{\mathbf{P}}\right\|_{\text{op}} = \frac{\lambda_1}{\lambda_1+\lambda} < 1$. ☐

Note that the upper bound $\frac{1}{2}$ of δ can be replaced by any constant strictly small than 1 to get a similar conclusion.

Remark C.12. *Using the concentration result from random matrix theory, for $M < N$, one can show with high probability that the operator norm δ of the fluctuation matrix $\mathbf{\Delta}$ is less than 1.* [4]

See subsection C.3 for details. Then we can use the the above lemmata C.9 and C.11 to approximate the test error of KRR:

Proposition C.13 (Bias Approximation)**.** *Fix a sample $\mathbf{Z}$ of ρ such that $\delta \stackrel{\text{def.}}{=} \|\mathbf{\Delta}\|_{op} < \frac{1}{2}$. Then the bias$_{test}$ term is bounded above and below by*

$$\left|bias - \left(\left\|\bar{\mathbf{P}}w\right\|_2^2 + \tilde{\gamma}_{>M}^2\right)\right| \leq 2\delta\left\|\bar{\mathbf{P}}w\right\|_2^2 + \|w\|_2^2\,\delta^2 p(\delta),$$

where $\bar{\mathbf{P}} \stackrel{\text{def.}}{=} \mathbf{\Lambda}(\mathbf{\Lambda} + \lambda\mathbf{I}_M)^{-1}$, $w = \lambda\mathbf{\Lambda}^{-1}\tilde{\boldsymbol{\gamma}} - \tilde{\gamma}_{>M}\boldsymbol{E}$, and $p(\delta) \stackrel{\text{def.}}{=} 5 + 4\delta + 4\delta^2$. By writing $\boldsymbol{E} = (\eta_k)_{k=1}^M$, the upper-bound simplifies to

$$bias \leq (1 + 2\delta)\sum_{k=1}^{M}\frac{(\lambda\tilde{\gamma}_k - \tilde{\gamma}_{>M}\eta_k\lambda_k)^2}{(\lambda_k + \lambda)^2} + \tilde{\gamma}_{>M}^2 + \|w\|_2^2\,\delta^2 p(\delta).$$

Analogously, the lower bound can be derived.

[4] From there, we differentiate the approach from Bach [6]: From Propositions C.5 and C.7, it is inevitable to approximate the matrix $\mathbf{B}$, and we have $\mathbf{I}_M$ as support of the inverse. Bach instead uses RHKS basis to express the fluctuation matrix and is hence forced to use $\lambda\mathbf{I}_M$ as the support. As a result, he would need to require that the fluctuation is less than λ and hence his requirement on N is antiproportional to λ in Theorem 5.1.

Proof. Let $w = \lambda \mathbf{\Lambda}^{-1}\tilde{\boldsymbol{\gamma}} - \tilde{\gamma}_{>M}\boldsymbol{E}$. We apply lemma C.4 followed by the 1st-order approximation $\mathbf{B}^{(1)}$ of the matrix $\mathbf{B}$ in lemma C.11:

$$\begin{aligned}
\text{fitting error} &= \|\mathbf{B}w\|_2^2 = \left\|\mathbf{B}^{(1)}w + \left(\mathbf{B} - \mathbf{B}^{(1)}\right)w\right\|_2^2 \\
&= \left\|\mathbf{B}^{(1)}w\right\|_2^2 + w^\top\left(\mathbf{B} - \mathbf{B}^{(1)}\right)\mathbf{B}^{(1)}w + w^\top\mathbf{B}^{(1)}\left(\mathbf{B} - \mathbf{B}^{(1)}\right)w + \left\|\left(\mathbf{B} - \mathbf{B}^{(1)}\right)w\right\|_2^2 \\
&\le \left\|\mathbf{B}^{(1)}w\right\|_2^2 + 2\left\|\mathbf{B}^{(1)}\right\|_{\text{op}}\left\|\mathbf{B} - \mathbf{B}^{(1)}\right\|_{\text{op}}\|w\|_2^2 + \left\|\mathbf{B} - \mathbf{B}^{(1)}\right\|_{\text{op}}^2\|w\|_2^2 \\
&\le \left\|\mathbf{B}^{(1)}w\right\|_2^2 + 2\cdot(1+\delta)\cdot 2\delta^2\|w\|_2^2 + 4\delta^4\|w\|_2^2 \\
&\le \left\|\mathbf{B}^{(1)}w\right\|_2^2 + 4\|w\|_2^2\,\delta^2(1+\delta+\delta^2) \\
&\le \left\|\left(\bar{\mathbf{P}} - \bar{\mathbf{P}}\mathbf{\Delta}\bar{\mathbf{P}}\right)w\right\|_2^2 + 4\|w\|_2^2\,\delta^2(1+\delta+\delta^2) \\
&\le \left\|\mathbf{I}_M - \bar{\mathbf{P}}\mathbf{\Delta}\right\|_{\text{op}}^2\left\|\bar{\mathbf{P}}w\right\|_2^2 + 4\|w\|_2^2\,\delta^2(1+\delta+\delta^2) \\
&\le \left(1 + 2\left\|\bar{\mathbf{P}}\mathbf{\Delta}\right\|_{\text{op}} + \left\|\bar{\mathbf{P}}\mathbf{\Delta}\right\|_{\text{op}}^2\right)\left\|\bar{\mathbf{P}}w\right\|_2^2 + 4\|w\|_2^2\,\delta^2(1+\delta+\delta^2) \\
&\le \left(1 + 2\left\|\bar{\mathbf{P}}\mathbf{\Delta}\right\|_{\text{op}}\right)\left\|\bar{\mathbf{P}}w\right\|_2^2 + \|w\|_2^2\,\delta^2(5+4\delta+4\delta^2) \\
&\le (1+2\delta)\left\|\bar{\mathbf{P}}w\right\|_2^2 + \|w\|_2^2\,\delta^2(5+4\delta+4\delta^2).
\end{aligned}$$

Hence we have the upper bound:

$$\text{bias} \le \tilde{\gamma}_{>M}^2 + (1+2\delta)\left\|\bar{\mathbf{P}}w\right\|_2^2 + \|w\|_2^2\,\delta^2 p(\delta).$$

We argue similarly for the lower bound using: $\|\mathbf{A}\|_{\text{op}}\|v\|_2^2 \ge v^\top\mathbf{A}v \ge -\|\mathbf{A}\|_{\text{op}}\|v\|_2^2$ for any $\mathbf{A} \in \mathbb{R}^{M\times M},\ v \in \mathbb{R}^{M\times 1}$. ☐

We argue similarly for variance.

Proposition C.14 (Variance Approximation)**.** *Fix a sampling* $\mathbf{Z}$ *such that* $\delta \overset{\text{def.}}{=} \|\mathbf{\Delta}\|_{op} < \frac{1}{2}$. *Then we have*

$$\left|\textit{variance} - \frac{\sigma^2}{N}\sum_{k=1}^{M}\frac{\lambda_k^2}{(\lambda_k+\lambda)^2}\right| \le \delta\frac{\sigma^2}{N}\sum_{k=1}^{M}\frac{\lambda_k^2}{(\lambda_k+\lambda)^2} + M\frac{\sigma^2}{N}(1+\delta)\delta^2 p(\delta),$$

where $p(\delta) \overset{\text{def.}}{=} 5 + 4\delta + 4\delta^2$, *and* $\sigma^2 \overset{\text{def.}}{=} \mathbb{E}[\epsilon^2]$ *is the noise variance.*

Proof. Note that $\operatorname{Tr}\mathbf{A} \le M\|\mathbf{A}\|_{\text{op}}$ for any matrix $\mathbf{A} \in \mathbb{R}^{M\times M}$. Since $\mathbf{B}^2(\mathbf{I}_M + \mathbf{\Delta}) = (\mathbf{B}^{(1)})^2(\mathbf{I}_M + \mathbf{\Delta}) + 2\mathbf{B}^{(1)}\left(\mathbf{B} - \mathbf{B}^{(1)}\right)(\mathbf{I}_M + \mathbf{\Delta}) + \left(\mathbf{B} - \mathbf{B}^{(1)}\right)^2(\mathbf{I}_M + \mathbf{\Delta})$, we can bound the residue term by δ:

$$\begin{aligned}
&\operatorname{Tr}\left[2\mathbf{B}^{(1)}\left(\mathbf{B} - \mathbf{B}^{(1)}\right)(\mathbf{I}_M + \mathbf{\Delta}) + \left(\mathbf{B} - \mathbf{B}^{(1)}\right)^2(\mathbf{I}_M + \mathbf{\Delta})\right] \\
&\le M(1+\delta)\left\|\mathbf{B} - \mathbf{B}^{(1)}\right\|_{\text{op}}(2\left\|\mathbf{B}^{(1)}\right\|_{\text{op}} + \left\|\mathbf{B} - \mathbf{B}^{(1)}\right\|_{\text{op}}) \\
&\le M(1+\delta)\cdot 2\delta^2(2(1+\delta) + 2\delta^2) \\
&\le 4M\delta^2(1+\delta)(1+\delta+\delta^2),
\end{aligned}$$

For the main terms, we have

$$\begin{aligned}
\operatorname{Tr}[(\mathbf{B}^{(1)})^2(\mathbf{I}_M + \mathbf{\Delta})] &\le \operatorname{Tr}[\bar{\mathbf{P}}^2]\cdot\left\|(\mathbf{I}_M - \mathbf{\Delta}\bar{\mathbf{P}})^2(\mathbf{I}_M + \mathbf{\Delta})\right\|_{\text{op}} \\
&= \operatorname{Tr}[\bar{\mathbf{P}}^2]\left\|\mathbf{I}_M + \mathbf{\Delta}(\mathbf{I}_M - 2\bar{\mathbf{P}}) + (\mathbf{\Delta}\bar{\mathbf{P}})^2 - 2\mathbf{\Delta}\bar{\mathbf{P}}\mathbf{\Delta} + (\mathbf{\Delta}\bar{\mathbf{P}})^2\mathbf{\Delta}\right\|_{\text{op}} \\
&\le \operatorname{Tr}[\bar{\mathbf{P}}^2]\left\|\mathbf{I}_M + \mathbf{\Delta}(\mathbf{I}_M - 2\bar{\mathbf{P}})\right\|_{\text{op}} + M\left\|(\mathbf{\Delta}\bar{\mathbf{P}})^2 - 2\mathbf{\Delta}\bar{\mathbf{P}}\mathbf{\Delta} + (\mathbf{\Delta}\bar{\mathbf{P}})^2\mathbf{\Delta}\right\|_{\text{op}} \\
&\le \operatorname{Tr}[\bar{\mathbf{P}}^2](1+\delta) + M\delta^2(1+\delta).
\end{aligned}$$

We apply Theorem C.7 to yield a bound on variance:

$$\left|\text{variance} - \frac{\sigma^2}{N}\sum_{k=1}^{M}\frac{\lambda_k^2}{(\lambda_k+\lambda)^2}\right| \le \delta\frac{\sigma^2}{N}\sum_{k=1}^{M}\frac{\lambda_k^2}{(\lambda_k+\lambda)^2} + M\frac{\sigma^2}{N}(1+\delta)\delta^2 p(\delta).$$

□

Note that the above propositions C.13 and C.14 give absolute (non-probabilistic) bounds on the test error, once δ is controlled.

C.3 Concentration Results

In this subsection, we focus on bounding the operator norm δ of the fluctuation matrix $\mathbf{\Delta}$. We can assume the data-generating distribution ρ and eigenfunctions ψ_k are well-behaved in the sense that:

Assumption C.15 (Sub-Gaussian-ness)**.** *We assume probability distribution of the random variable $\psi_k(x)$, where $x \in \rho$, has sub-Gaussian norm bounded by a positive constant $G > 0$, for all $k \in \{1, ..., M\} \cup \{> M\}$*[5].

In particular, if the random variable $\psi_k(x)$ is bounded, the assumption C.15 is fulfilled. First, we establish some concentration results.

Lemma C.16 (Theorem 3.59 in [36])**.** *Let $\mathbf{A}$ be an $n \times N$ matrix with independent isotropic sub-Gaussian columns in $\mathbb{R}^n$ which sub-gaussian norm is bounded by a positive constant $G > 0$. Then for all $t \ge 0$, with probability at least $1 - 2\exp(-\frac{1}{3}t^2)$, we have*

$$\left\|\frac{1}{N}\mathbf{A}\mathbf{A}^\top - \mathbf{I}_n\right\|_{op} \le \max(a, a^2), \tag{32}$$

where $a \overset{\text{def.}}{=} C\sqrt{\frac{n}{N}} + \frac{t}{\sqrt{N}}$, for all constant $C \ge 12G^2$.

Proof. Let $a \overset{\text{def.}}{=} C\sqrt{\frac{n}{N}} + \frac{t}{\sqrt{N}}$ with $C > 0$ to be determined, and $\epsilon \overset{\text{def.}}{=} \max\{a, a^2\}$. The first step to show that :

$$\max_{x\in\mathcal{N}}\left|\frac{1}{N}\left\|\mathbf{A}^\top x\right\|_2^2 - 1\right| \le \epsilon$$

for some $\frac{1}{4}$-net $\mathcal{N}$ on the sphere $\mathbb{S}^{n-1} \subset \mathbb{R}^n$. Choose such a net $\mathcal{N}$ with $|\mathcal{N}| < \left(1 + \frac{2}{1/4}\right)^n = 9^n$. Let $\mathbf{A}_i$ be the ith column of the matrix $\mathbf{A}$ and let $Z_i \overset{\text{def.}}{=} \mathbf{A}_i^\top x$ be a random variable. By definition of $\mathbf{A}$, Z_i is centered with unit variance with sub-Gaussian norm upper bounded by G. Note that $G \ge \frac{1}{\sqrt{2}}\mathbb{E}[Z_i^2]^{1/2} = \frac{1}{\sqrt{2}}$, and the random variable $Z_i^2 - 1$ is centered and has sub-exponential norm upper bounded by $4G^2$. Hence by an exponential deviation inquality [6], we have, for any $x \in \mathbb{S}^{n-1}$:

$$\begin{aligned}\mathbb{P}\left\{\left|\frac{1}{N}\left\|\mathbf{A}^\top x\right\|_2^2 - 1\right| \ge \frac{\epsilon}{2}\right\} &= \mathbb{P}\left\{\left|\frac{1}{N}\sum_{i=1}^{N} Z_i^2 - 1\right| \ge \frac{\epsilon}{2}\right\} \\ &\le 2\exp\left(-\frac{1}{2}e^{-1}G^{-4}\min\{\epsilon, \epsilon^2\}\right) \\ &=\le 2\exp\left(-\frac{1}{2}e^{-1}G^{-4}a^2\right) \\ &\le 2\exp\left(-\frac{1}{2}e^{-1}G^{-4}(C^2 n + t^2)\right).\end{aligned}$$

[5] it means the orthonormal complement $\psi_{>M}$ is also mentioned in the assumption.

[6] This inequality is Corollary 5.17 from [36].

Then by union bound, we have

$$\mathbb{P}\left\{\max_{x\in\mathcal{N}}\left|\frac{1}{N}\left\|\mathbf{A}^\top x\right\|_2^2-1\right|\geq\frac{\epsilon}{2}\right\}\leq 9^n\cdot 2\exp\left(-\frac{1}{2}e^{-1}G^{-4}(C^2n+t^2)\right)$$
$$\leq 2\exp\left(-\frac{1}{2}e^{-1}G^{-4}t^2\right),$$

for $C\geq\sqrt{2e\log 9}G^2$. Since $12>\sqrt{2e\log 9}$, for simplicity, we assume $C>12G^2$. Moreover, since $G\geq\frac{1}{\sqrt{2}}$, we have $\frac{1}{2}e^{-1}G^{-4}\leq\frac{1}{3}$, we have

$$\mathbb{P}\left\{\max_{x\in\mathcal{N}}\left|\frac{1}{N}\left\|\mathbf{A}^\top x\right\|_2^2-1\right|\geq\frac{\epsilon}{2}\right\}\leq 2\exp\left(-\frac{1}{3}t^2\right).$$

Then by the $\frac{1}{4}$-net argument, with probability at least $1-2\exp\left(-\frac{1}{3}t^2\right)$, we have

$$\left\|\frac{1}{N}\mathbf{A}\mathbf{A}^\top-\mathbf{I}_n\right\|_{\mathrm{op}}\leq\frac{4}{2}\max_{x\in\mathcal{N}}\left|\frac{1}{N}\left\|\mathbf{A}^\top x\right\|_2^2-1\right|$$
$$\leq\epsilon=\max\{a,a^2\}.$$

□

Lemma C.17. *Assume Assumption C.15 holds and that $N>\exp(4(12G^2)^2(M+1))$. Then with a probability of at least $1-2/N$, we have*

$$\max\{\delta,\|\boldsymbol{E}_M\|_2\}\leq\sqrt{\frac{\log N}{N}}.$$

Proof. Set $n=M+1$, $\mathbf{A}=\binom{\boldsymbol{\Psi}_{\leq M}}{\psi_{>M}(\mathbf{X})^\top}\in\mathbb{R}^{(M+1)\times N}$. Then

$$\frac{1}{N}\mathbf{A}\mathbf{A}^\top-\mathbf{I}_n=\begin{pmatrix}\frac{1}{N}\boldsymbol{\Psi}_{\leq M}\boldsymbol{\Psi}_{\leq M}^\top & \boldsymbol{E}_M\\ \boldsymbol{E}_M^\top & \eta_{>M}+1\end{pmatrix}-\mathbf{I}_n=\begin{pmatrix}\boldsymbol{\Delta}_M & \boldsymbol{E}_M\\ \boldsymbol{E}_M^\top & \eta_{>M}\end{pmatrix}.$$

where $\eta_{>M}\overset{\text{def.}}{=}\frac{1}{N}\sum_{i=1}^N\psi_{>M}(x_i)^2-1$. On one hand, the operator norm of the above matrix bounds δ and $\|\boldsymbol{E}_M\|_2$ from above:

$$\left\|\begin{pmatrix}\boldsymbol{\Delta}_M & \boldsymbol{E}_M\\ \boldsymbol{E}_M^\top & \eta_{>M}\end{pmatrix}\right\|_{\mathrm{op}}=\max_{\|\mathbf{u}\|_2^2+v^2=1}\left\|\begin{pmatrix}\boldsymbol{\Delta}_M & \boldsymbol{E}_M\\ \boldsymbol{E}_M^\top & \eta_{>M}\end{pmatrix}\begin{pmatrix}\mathbf{u}\\ v\end{pmatrix}\right\|_2$$
$$=\max_{\|\mathbf{u}\|_2^2+v^2=1}\left\|\begin{pmatrix}\boldsymbol{\Delta}_M\mathbf{u}+v\boldsymbol{E}_M\\ \boldsymbol{E}_M^\top\mathbf{u}+\eta_{>M}v\end{pmatrix}\right\|_2$$
$$\geq\max_{\|\mathbf{u}\|_2^2+v^2=1}\|\boldsymbol{\Delta}_M\mathbf{u}+v\boldsymbol{E}_M\|_2$$
$$\geq\max_{\|\mathbf{u}\|_2^2=1,v=0}\|\boldsymbol{\Delta}_M\mathbf{u}+v\boldsymbol{E}_M\|_2$$
$$\geq\max_{\|\mathbf{u}\|_2^2=1}\|\boldsymbol{\Delta}_M\mathbf{u}\|_2=\delta,$$

and

$$\left\|\begin{pmatrix}\boldsymbol{\Delta}_M & \boldsymbol{E}_M\\ \boldsymbol{E}_M^\top & \eta_{>M}\end{pmatrix}\right\|_{\mathrm{op}}\geq\max_{\|\mathbf{u}\|_2^2+v^2=1}\|\boldsymbol{\Delta}_M\mathbf{u}+v\boldsymbol{E}_M\|_2\geq\max_{\|\mathbf{u}\|_2^2=0,|v|=1}\|\boldsymbol{\Delta}_M\mathbf{u}+v\boldsymbol{E}_M\|_2=\|\boldsymbol{E}_M\|_2.$$

On the other hand, set $t=\frac{1}{2}\sqrt{\log N},\ C=12G^2$, since $N>\exp(4C^2(M+1))$, we have

$$a=C\sqrt{\frac{n}{N}}+\frac{t}{\sqrt{N}}=12G^2\sqrt{\frac{M+1}{N}}+\frac{1}{2}\sqrt{\frac{\log N}{N}}\leq\sqrt{\frac{\log N}{N}}<1.$$

By Lemma C.17, then with probability of at least $1-2\exp(-\frac{1}{3}t^2)=1-2\exp(-\frac{1}{12})/N>1-2/N$, we have

$$\left\|\begin{pmatrix}\boldsymbol{\Delta}_M & \boldsymbol{E}_M\\ \boldsymbol{E}_M^\top & \eta_{>M}\end{pmatrix}\right\|_{\mathrm{op}}\leq\max\{a,a^2\}=a\leq\sqrt{\frac{\log N}{N}}.$$

Combine the both results and we conclude the upper bounds. □

In particular, as $N \to \infty$, δ vanishes almost surely. In empirical calculation, if the requirement $N > \exp(4(12G^2)^2(M+1))$ exponential in M is too demanding for a large integer M, we can take $t = N^s$ for any positive number $s \in \left(0, \frac{1}{2}\right)$ instead of $t = \frac{1}{2}\log N$. In this way, we decrease the requirement to N polynomial in M in sacrificing the decay from $\mathcal{O}\left(\sqrt{\frac{logN}{N}}\right)$ to $\mathcal{O}\left(N^{s-1/2}\right)$. For simplicity purpose, we do not list out the result with this decay in this paper.

C.4 Refined Test Error Analysis

We can apply the above concentration results to refine the following bounds on the finite-rank KRR test error.

Definition C.18. *To ease the notation, we denote* $\underline{r} \overset{\text{def.}}{=} \min_k\{|\tilde{\gamma}_k/\lambda_k|\}$ *and* $\overline{r} \overset{\text{def.}}{=} \max_k\{|\tilde{\gamma}_k/\lambda_k|\}$.

C.4.1 Refined Bounds on Bias

Recall that Proposition C.13 bounding the bias in terms of δ and η_k. For the former one, we can choose: for $N > \max\{\exp(4(12G^2)^2(M+1)), 9\}$, by Lemma C.17, with probability of at least $1 - 2/N$, we have $\delta \leq \sqrt{\frac{\log N}{N}} < \sqrt{\frac{\log 9}{9}} < \frac{1}{2}$. For the latter one, we have to control the vector w:

Lemma C.19. *Let* $w = \lambda \boldsymbol{\Lambda}^{-1}\tilde{\boldsymbol{\gamma}} - \tilde{\gamma}_{>M}\boldsymbol{E}$. *We have*

$$\lambda^2\underline{r}^2 M - 2\lambda\underline{r}|\tilde{\gamma}_{>M}|\sqrt{M}\,\|E\|_2 \leq \|w\|_2^2 \leq \left(\lambda\overline{r}\sqrt{M} + \tilde{\gamma}_{>M}\,\|E\|_2\right)^2;$$

$$\frac{\lambda^2\lambda_M}{(\lambda_M+\lambda)^2}\|\tilde{f}_{\leq M}\|_{\mathcal{H}}^2 - \frac{1}{2}|\tilde{\gamma}_{>M}|\|\tilde{f}_{\leq M}\|_{L^2_\rho}\|\boldsymbol{E}\|_2 \leq \left\|\bar{\mathbf{P}}w\right\|_2^2 \leq \lambda\|\tilde{f}_{\leq M}\|_{\mathcal{H}}^2 + \frac{1}{2}|\tilde{\gamma}_{>M}|\|\tilde{f}_{\leq M}\|_{L^2_\rho}\|\boldsymbol{E}\|_2 + \tilde{\gamma}_{>M}^2\|\boldsymbol{E}\|_2^2.$$

Proof. Since $\lambda^2\underline{r}^2M \leq \left\|\lambda\boldsymbol{\Lambda}^{-1}\tilde{\boldsymbol{\gamma}}\right\|_2^2 \leq \lambda^2\overline{r}^2M$ and $\|\tilde{\gamma}_{>M}\boldsymbol{E}\|_2^2 = \tilde{\gamma}_{>M}^2\|\boldsymbol{E}\|_2^2$, we have

$$\lambda^2\underline{r}^2 M - 2\lambda\underline{r}|\tilde{\gamma}_{>M}|\sqrt{M}\,\|E\|_2 \leq \|w\|_2^2 \leq \left(\lambda\overline{r}\sqrt{M} + \tilde{\gamma}_{>M}\,\|E\|_2\right)^2.$$

Similarly, we can bound $\left\|\bar{\mathbf{P}}w\right\|_2$. Observe that:

$$\left\|\bar{\mathbf{P}}w\right\|_2^2 = \underbrace{\lambda^2\sum_{k=1}^M\frac{\tilde{\gamma}_k^2}{(\lambda_k+\lambda)^2}}_{I}\underbrace{-2\lambda\tilde{\gamma}_{>M}\sum_{k=1}^M\frac{\tilde{\gamma}_k\lambda_k\eta_k}{(\lambda_k+\lambda)^2}}_{II}+\underbrace{\tilde{\gamma}_{>M}^2\sum_{k=1}^M\frac{\lambda_k^2\eta_k^2}{(\lambda_k+\lambda)^2}}_{III}.$$

Since $1 \geq \frac{\lambda}{\lambda_M+\lambda} \geq \frac{\lambda}{\lambda_k+\lambda}$, we have the upper bound:

$$I = \lambda^2\sum_{k=1}^M\frac{\tilde{\gamma}_k^2}{(\lambda_k+\lambda)^2} \leq \lambda\sum_{k=1}^M\frac{\lambda}{\lambda_k+\lambda}\frac{\tilde{\gamma}_k^2}{\lambda_k+\lambda} \leq \lambda\sum_{k=1}^M\frac{\tilde{\gamma}^2}{\lambda_k} = \lambda\|\tilde{f}_{\leq M}\|_{\mathcal{H}}^2. \tag{33}$$

where $\tilde{f}_{\leq M} \overset{\text{def.}}{=} \sum_{k=1}^M\tilde{\gamma}_k\psi_k = \tilde{f} - \tilde{\gamma}_{>M}\psi_{>M}$. For the lower bound, we have:

$$I = \lambda^2\sum_{k=1}^M\frac{\tilde{\gamma}_k^2}{(\lambda_k+\lambda)^2} \geq \lambda^2\sum_{k=1}^M\frac{\lambda_k}{(\lambda_k+\lambda)^2}\frac{\tilde{\gamma}_k^2}{\lambda_k} \geq \lambda^2\frac{\lambda_M}{(\lambda_M+\lambda)^2}\|\tilde{f}_{\leq M}\|_{\mathcal{H}}^2 \tag{34}$$

Similarly, since $4\lambda\lambda_k \leq (\lambda_k+\lambda)^2$,

$$|II| = 2\lambda|\tilde{\gamma}_{>M}|\sum_{k=1}^M\frac{|\tilde{\gamma}_k|\lambda_k|\eta_k|}{(\lambda_k+\lambda)^2} \leq \frac{1}{2}|\tilde{\gamma}_{>M}|\sum_{k+1}^M|\tilde{\gamma}_k\eta_k| \leq \frac{1}{2}|\tilde{\gamma}_{>M}|\sqrt{\sum_{k+1}^M\tilde{\gamma}_k^2\sum_{k=1}^M\eta_k^2} \leq \frac{1}{2}|\tilde{\gamma}_{>M}|\|\tilde{f}_{\leq M}\|_{L^2_\rho}\|\boldsymbol{E}\|_2.$$

And

$$III = \tilde{\gamma}_{>M}^2\sum_{k=1}^M\frac{\lambda_k^2\eta_k^2}{(\lambda_k+\lambda)^2} \leq \tilde{\gamma}_{>M}^2\sum_{k=1}^M\eta_k^2 = \tilde{\gamma}_{>M}^2\|\boldsymbol{E}\|_2^2.$$

$\square$

Combining the above result, we state the following theorem:

Theorem C.20. *For* $N > \max\left\{\exp(4(12G^2)^2(M+1)), 9\right\}$ *and for any constant* $C_1 > 8\left(\lambda\bar{r}\sqrt{M} + \frac{1}{2}|\tilde{\gamma}_{>M}|\right)^2 + \frac{5}{2}\|\tilde{f}\|^2_{L^2_\rho}$ *(independent to* N*), with a probability of at least* $1-2/N$*, we have the upper and lower bounds of bias:*

$$\textit{bias} \le \tilde{\gamma}^2_{>M} + \lambda\|\tilde{f}_{\le M}\|^2_{\mathcal{H}} + \left(\frac{1}{4}\|\tilde{f}\|^2_{L^2_\rho} + 2\lambda\|\tilde{f}_{\le M}\|^2_{\mathcal{H}}\right)\sqrt{\frac{\log N}{N}} + C_1\frac{\log N}{N};$$

$$\textit{bias} \ge \tilde{\gamma}^2_{>M} + \frac{\lambda^2\lambda_M}{(\lambda_M+\lambda)^2}\|\tilde{f}_{\le M}\|^2_{\mathcal{H}} - \left(\frac{1}{4}\|\tilde{f}\|^2_{L^2_\rho} + \frac{2\lambda^2}{\lambda_1+\lambda}\|\tilde{f}_{\le M}\|^2_{\mathcal{H}}\right)\sqrt{\frac{\log N}{N}} - C_1\frac{\log N}{N}.$$

For $\lambda \to 0$*, we have a simpler bound: with a probability of at least* $1-2/N$*, we have*

$$\begin{aligned}\lim_{\lambda\to 0}\textit{bias} &\le \left(1+\frac{\log N}{N}\right)\tilde{\gamma}^2_{>M} + 6\tilde{\gamma}^2_{>M}\left(\frac{\log N}{N}\right)^{\frac{3}{2}};\\ \lim_{\lambda\to 0}\textit{bias} &\ge \left(1-\frac{\log N}{N}\right)\tilde{\gamma}^2_{>M} - 6\tilde{\gamma}^2_{>M}\left(\frac{\log N}{N}\right)^{\frac{3}{2}}.\end{aligned} \tag{35}$$

For $\tilde{\gamma}^2_{>M} = 0$*, that is* $\tilde{f} \in \mathcal{H}$*, we have a simpler upper bound on bias: with a probability of at least* $1-2/N$*, we have*

$$\begin{aligned}\textit{bias} &\le \lambda\|\tilde{f}\|^2_{\mathcal{H}}\left(1+2\sqrt{\frac{\log N}{N}}\right) + C_1\frac{\log N}{N};\\ \textit{bias} &\ge \frac{\lambda^2\lambda_M}{(\lambda_M+\lambda)^2}\|\tilde{f}\|^2_{\mathcal{H}}\left(1-2\sqrt{\frac{\log N}{N}}\right) - C_1\frac{\log N}{N}.\end{aligned} \tag{36}$$

Proof. By Proposition C.13 and Lemma C.19,

$$\text{fitting error} \le (1+2\delta)\left\|\bar{\mathbf{P}}w\right\|_2^2 + \|w\|_2^2\,\delta^2 p(\delta) \tag{37}$$

$$\le (1+2\delta)(\lambda\|\tilde{f}_{\le M}\|^2_{\mathcal{H}} + \frac{1}{2}|\tilde{\gamma}_{>M}|\|\tilde{f}_{\le M}\|_{L^2_\rho}\|\boldsymbol{E}\|_2 + \tilde{\gamma}^2_{>M}\|\boldsymbol{E}\|_2^2) + \|w\|_2^2\,\delta^2 p(\delta) \tag{38}$$

$$\le (1+2\delta)(\lambda\|\tilde{f}_{\le M}\|^2_{\mathcal{H}} + \frac{1}{4}\|\tilde{f}\|^2_{L^2_\rho}\|\boldsymbol{E}\|_2 + \tilde{\gamma}^2_{>M}\|\boldsymbol{E}\|_2^2) + \|w\|_2^2\,\delta^2 p(\delta). \tag{39}$$

where in line (37), we use Proposition C.13; in line (38), we use Lemma C.19; in line (38), we use the fact that $2ab \le a^2+b^2$ where $a = |\tilde{\gamma}_{>M}|, b = \|\tilde{f}_{\le M}\|_{L^2_\rho}$.

Now we apply the concentration result in Lemma C.17: with a probability of at least $1-2/N$:

$$\begin{aligned}\text{fitting error} &\le \left(1+2\sqrt{\frac{\log N}{N}}\right)\lambda\|\tilde{f}_{\le M}\|^2_{\mathcal{H}} + \frac{1}{4}\|\tilde{f}\|^2_{L^2_\rho}\sqrt{\frac{\log N}{N}}\\ &\quad + \frac{\log N}{N}\left(\|w\|_2^2\,p(\delta) + (1+2\delta)\tilde{\gamma}^2_{>M} + \frac{1}{2}\|\tilde{f}\|^2_{L^2_\rho}\right)\\ &\le \lambda\|\tilde{f}_{\le M}\|^2_{\mathcal{H}} + \left(\frac{1}{4}\|\tilde{f}\|^2_{L^2_\rho} + 2\lambda\|\tilde{f}_{\le M}\|^2_{\mathcal{H}}\right)\sqrt{\frac{\log N}{N}} + C_1\frac{\log N}{N},\end{aligned}$$

where we choose $C_1 > 0$ to be such that:

$$\begin{aligned}\|w\|_2^2\,p(\delta) + (1+2\delta)\tilde{\gamma}^2_{>M} + \frac{1}{2}\|\tilde{f}\|^2_{L^2_\rho} &\le \|w\|_2^2\,p(\delta) + \left(1+2\delta+\frac{1}{2}\right)\|\tilde{f}\|^2_{L^2_\rho}\\ &\le \|w\|_2^2\,p\left(\frac{1}{2}\right) + \left(1+2\cdot\frac{1}{2}+\frac{1}{2}\right)\|\tilde{f}\|^2_{L^2_\rho}\\ &\le 8\left(\lambda\bar{r}\sqrt{M} + |\tilde{\gamma}_{>M}|\,\|E\|_2\right)^2 + \frac{5}{2}\|\tilde{f}\|^2_{L^2_\rho}\\ &\le 8\left(\lambda\bar{r}\sqrt{M} + \frac{1}{2}|\tilde{\gamma}_{>M}|\right)^2 + \frac{5}{2}\|\tilde{f}\|^2_{L^2_\rho} < C_1.\end{aligned}$$

Hence we have an upper bound for the bias. We argue similarly for the lower bound:

$$\begin{aligned}\text{fitting error} \geq &\left(1-2\sqrt{\frac{\log N}{N}}\right)\frac{\lambda^2\lambda_M}{(\lambda_M+\lambda)^2}\|\tilde{f}_{\leq M}\|_{\mathcal{H}}^2 - \frac{1}{4}\|\tilde{f}\|_{L^2_\rho}^2\sqrt{\frac{\log N}{N}} \\ &- \frac{\log N}{N}\left(\|w\|_2^2\, p(\delta) + (1+2\delta)\tilde{\gamma}_{>M}^2 + |\tilde{\gamma}_{>M}|\|\tilde{f}_{\leq M}\|_{L^2_\rho}\right) \\ \geq &\frac{\lambda^2\lambda_M}{(\lambda_M+\lambda)^2}\|\tilde{f}_{\leq M}\|_{\mathcal{H}}^2 - \left(\frac{1}{4}\|\tilde{f}\|_{L^2_\rho}^2 + \frac{2\lambda^2}{\lambda_1+\lambda}\|\tilde{f}_{\leq M}\|_{\mathcal{H}}^2\right)\sqrt{\frac{\log N}{N}} - C_1\frac{\log N}{N}.\end{aligned}$$

For $\lambda \to 0$, note that $w \to -\tilde{\gamma}_{>M}\boldsymbol{E}$. This yields

$$\begin{aligned}\lim_{\lambda\to 0}\text{fitting error} &\leq \lim_{\lambda\to 0}\left\{(1+2\delta)\left\|\bar{\mathbf{P}}w\right\|_2^2 + \|w\|_2^2\,\delta^2 p(\delta)\right\} \\ &= (1+2\delta)\left\|-\tilde{\gamma}_{>M}\boldsymbol{E}\right\|_2^2 + \left\|-\tilde{\gamma}_{>M}\boldsymbol{E}\right\|_2^2\delta^2 p(\delta) \\ &= \tilde{\gamma}_{>M}^2\left\|\boldsymbol{E}\right\|_2^2(1+\delta(2+\delta p(\delta))).\end{aligned}$$

Hence, by plugging in $\delta < \frac{1}{2}$, with probability of at least $1-2/N$,

$$\begin{aligned}\lim_{\lambda\to 0}\text{fitting error} &\leq \tilde{\gamma}_{>M}^2\frac{\log N}{N}\left(1+6\sqrt{\frac{\log N}{N}}\right) \\ \lim_{\lambda\to 0}\text{bias} &\leq \left(1+\frac{\log N}{N}\right)\tilde{\gamma}_{>M}^2 + 6\tilde{\gamma}_{>M}^2\left(\frac{\log N}{N}\right)^{\frac{3}{2}}.\end{aligned}$$

For lower bound, it follows similarly:

$$\lim_{\lambda\to 0}\text{bias} \geq \left(1-\frac{\log N}{N}\right)\tilde{\gamma}_{>M}^2 - 6\tilde{\gamma}_{>M}^2\left(\frac{\log N}{N}\right)^{\frac{3}{2}},$$

and we obtain line (35). For the case where $\tilde{\gamma}_{>M} = 0$, recalculate and simplify line (38) to obtain line (36). $\square$

C.4.2 Refined Bounds on Variance

Similarly, we can refine Theorem C.14 to get a bound on the variance:

Theorem C.21. *For $N > \max\left\{(12G)^4(M+1)^2, 9\right\}$, and set $C_2 = 12$ (independent to N), with a probability of at least $1-2/N$, we have the upper and lower bounds of variance:*

$$\begin{aligned}\textit{variance} &\leq \sigma^2\frac{M}{N}\left(1+\sqrt{\frac{\log N}{N}} + C_2\frac{\log N}{N}\right); \\ \textit{variance} &\geq \frac{\lambda_M^2}{(\lambda_M+\lambda)^2}\sigma^2\frac{M}{N}\left(1-\sqrt{\frac{\log N}{N}}\right) - C_2\sigma^2\frac{M}{N}\frac{\log N}{N}.\end{aligned}$$

Proof. We argue analogously as in Theorem C.20: by Proposition C.14 and Lemma C.17, we have

$$\begin{aligned}\text{variance} &\leq (1+\delta)\frac{\sigma^2}{N}\sum_{k=1}^{M}\frac{\lambda_k^2}{(\lambda_k+\lambda)^2} + M\frac{\sigma^2}{N}(1+\delta)\delta^2 p(\delta) \\ &\leq (1+\delta)\sigma^2\frac{M}{N} + \sigma^2\frac{M}{N}(1+\delta)\delta^2 p(\delta) \\ &\leq \left(1+\sqrt{\frac{\log N}{N}}\right)\sigma^2\frac{M}{N} + \sigma^2\frac{M}{N}\frac{\log N}{N}\left(1+\frac{1}{2}\right)p\left(\frac{1}{2}\right) \\ &\leq \left(1+\sqrt{\frac{\log N}{N}}\right)\sigma^2\frac{M}{N} + 12\sigma^2\frac{M}{N}\frac{\log N}{N}\end{aligned}$$

Hence we can choose $C_2 = 12$. For the lower bound, since $\frac{\lambda_k^2}{(\lambda_k+\lambda)^2} > \frac{\lambda_M^2}{(\lambda_M+\lambda)^2}$, we have

$$\begin{aligned}
\text{variance} &\geq (1-\delta)\frac{\sigma^2}{N}\sum_{k=1}^{M}\frac{\lambda_k^2}{(\lambda_k+\lambda)^2} - M\frac{\sigma^2}{N}(1+\delta)\delta^2 p(\delta) \\
&\geq \frac{\lambda_M^2}{(\lambda_M+\lambda)^2}\sigma^2\frac{M}{N}\left(1-\sqrt{\frac{\log N}{N}}\right) - 12\sigma^2\frac{M}{N}\frac{\log N}{N}.
\end{aligned}$$

□

Note that in both Theorems C.20 and C.21, the constants $C_1, C_2 > 0$ is not optimized.

D Numerical Validation

In this section, we illustrate our result for KRR with two different finite rank kernels.

D.1 Truncated NTK

First, we need to define a finite-rank kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We set $\mathcal{X} = \mathbb{S}^1 \subset \mathbb{R}^2$. By reparametrization, we write $\mathbb{S}^1 \cong [0, 2\pi]/_{0\sim 2\pi}$. We assume the data are drawn uniformly on the circle, that is $\rho_{\mathcal{X}} = \operatorname{unif}[\mathbb{S}^1]$. We can use the Fourier functions $\cos(k\cdot), \sin(k\cdot)$ as the orthogonal eigenfunctions of the kernel. We define the NTK

$$K^{(\infty)}(\theta, \theta') \stackrel{\text{def.}}{=} \frac{\cos(\theta-\theta')\,(\pi - |\theta-\theta'|)}{2\pi}$$

for all $\theta, \theta' \in [0, 2\pi]$. 3) We choose a rank-$M$ truncation $K(\theta, \theta') = \sum_{k=1}^{M} \lambda_k \psi_k(\theta)\psi_k(\theta')$ for all $\theta, \theta' \in [0, 2\pi]$. For the first few eigenvalues of the kernel, please see Table 3 for example. Before

k	1	2	3	4	5	6	7	∞
λ_k	$\frac{1}{\pi^2}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{5}{9\pi^2}$	$\frac{5}{9\pi^2}$	$\frac{17}{225\pi^2}$	$\frac{17}{225\pi^2}$	-
$\psi_k(\theta)$	1	$\sqrt{2}\cos(\theta)$	$\sqrt{2}\sin(\theta)$	$\sqrt{2}\cos(2\theta)$	$\sqrt{2}\sin(2\theta)$	$\sqrt{2}\cos(4\theta)$	$\sqrt{2}\sin(4\theta)$	-
$\sum_{k'=0}^{k}\lambda_{k'}$	0.1013	0.2263	0.3513	0.4076	0.4639	0.4716	0.4792	0.5

Table 3: The first few eigenvalues of the NTK

proceeding to test error computation, we present a training example, Figure 4, to give readers an intuition on the truncated NTK (tNTK).

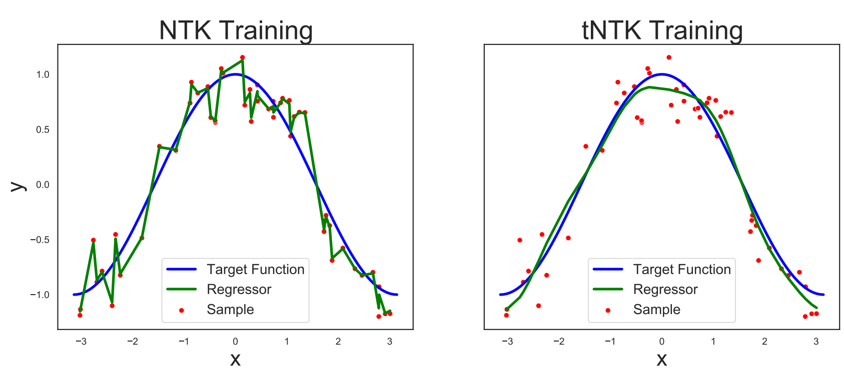

Figure 4: (left): NTK training; (right): tNTK training where $N = 50, M = 7$. $\sigma^2 = 0.05, \lambda = \sigma^2/N$.

D.2 Test Error Computations

In the following tNTK training, we set the hyperparameters as follows:

Target function We choose a simple target function $\tilde{f}(x) = \cos x = \frac{1}{\sqrt{2}}\psi_2(x)$. Throughout the experiment, we set the noise variance $\sigma^2 = 0.05$.

Ridge We choose $\lambda = \frac{\sigma^2}{N}$. In Figure 5 (left), we set $N = 50, \lambda = 0.05/50$ for tNTK training; (right) we set set $\lambda = 0.05/50$ for varying N from 10 to 200.

Error bars In Figure 7 (right), for each value of N, we run over 10 iterations of random samples and compute the test error. The error bars are shown as the difference between the upper and the lower quartiles.

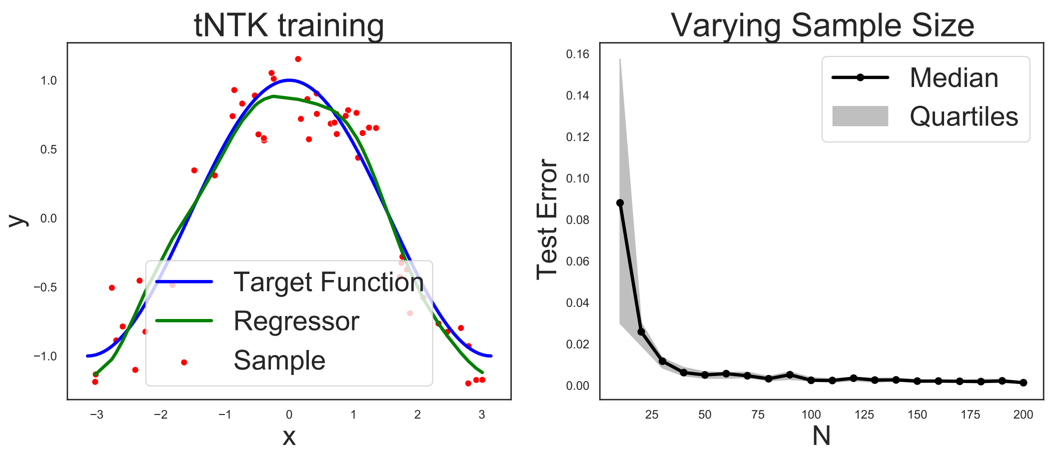

Figure 5: (left): tNTK training; (right): the decay of test error as N varies.

Lower bound See the subsection below.

D.3 Bound Comparison

We continue with the experiment on the tNTK this time with varying N and compare our upper bound with [6].

Upper bounds In Figure 6, the expression of Bach's and our upper bounds are directly computed:

$$\text{Bach's upper bound} = 4\lambda\|\tilde{f}\|_{\mathcal{H}}^2 + \frac{8\sigma^2 R^2}{\lambda N}(1 + 2\log N)$$

$$\text{Our upper bound without residue} = \lambda\|\tilde{f}\|_{\mathcal{H}}^2 \left(1 + 2\sqrt{\frac{\log N}{N}}\right),$$

where the constants $\|\tilde{f}\|_{\mathcal{H}}^2$ and R^2 can be computed directed from the choice of kernel and target function. For simplicity reason, we drop the residue term $C_1 \frac{\log N}{N}$ since it is overshadowed by the other terms and the constant C_1 is not optimized.

D.4 Legendre Kernel

To illustrate the bounds with another finite-rank, we choose a simple legendre kernel (LK):

$$K(x, z) = \sum_{k=0}^{M} \lambda_k P_k(x) P_k(z)$$

where P_k is the Legendre polynomial of degree k, and $\lambda_k > 0$ are the eigenvalues.

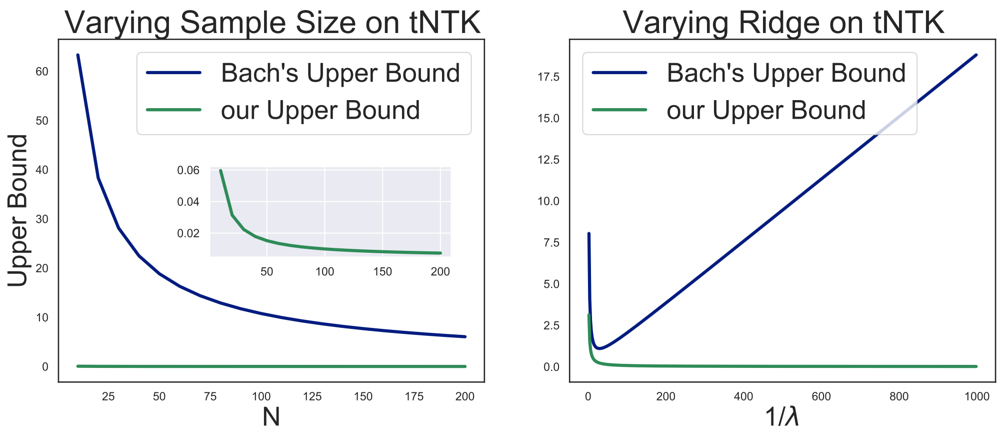

Figure 6: Bound improvement on tNTK. Residue term is dropped in our bound.

Eigenvalues To better compare the Legendre kernel K with the NTK, we choose $\lambda_k = C \cdot (k+1)^{-2}$ of quadratic decay such that the spectral sums are the same: $\sum_{k=0}^{\infty} \lambda_k = 0.5$. Hence we choose $C = 0.5 / \sum_{k=1}^{\infty} k^{-2} = \frac{3}{\pi^2}$.

Target function We choose a simple target function $\tilde{f}(x) = x^2 = \frac{1}{3}P_0(x) + \frac{2}{3}P_2(x)$. Throughout the experiment, we set the noise variance $\sigma^2 = 0.05$.

D.5 Test Error Computation

Ridge As before, our bound suggests that, to balance the bias and the variance with a fixed N, we can choose $\lambda = \frac{\sigma^2}{N}$. In Figure 7 (left), we set $N = 50$, $\lambda = 0.05/50$ for KRR training; (right) we set set $\lambda = 0.05/50$ for varying N from 10 to 200.

Error bars In Figure 7 (right), for each value of N, we run over 10 iterations of random samples and compute the test error. The error bars are shown as the different between the upper and the lower quartiles. The median is taken as average.

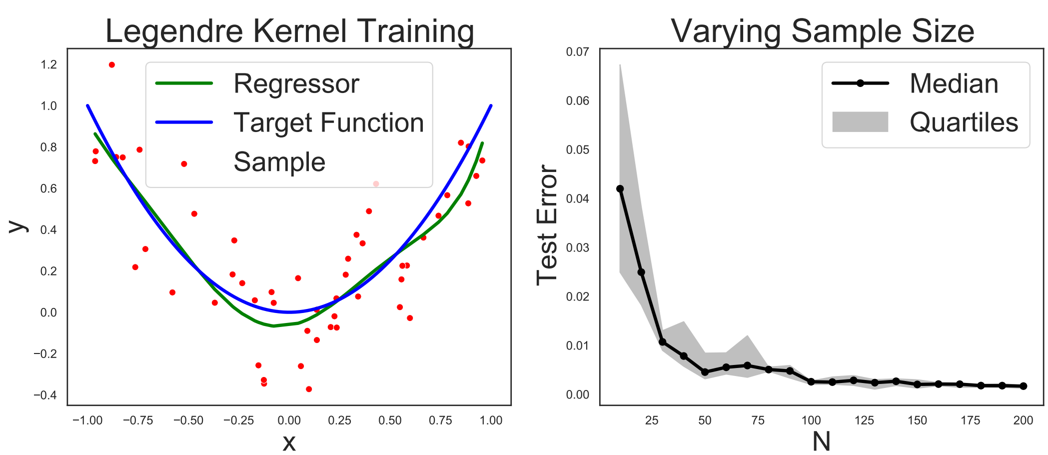

Figure 7: (left): LK training; (right): the decay of test error as N varies. Same as Figure 1.

Upper bounds In Figure 8, the expression of Bach's and our upper bounds are directly computed:

$$\text{Bach's upper bound} = 4\lambda \|\tilde{f}\|_{\mathcal{H}}^2 + \frac{8\sigma^2 R^2}{\lambda N}(1 + 2\log N)$$
$$\text{Our upper bound without residue} = \lambda \|\tilde{f}\|_{\mathcal{H}}^2 \left(1 + 2\sqrt{\frac{\log N}{N}}\right),$$

where the constants $\|\tilde{f}\|_{\mathcal{H}}^2$ and R^2 can be computed directed from the choice of kernel and target function.

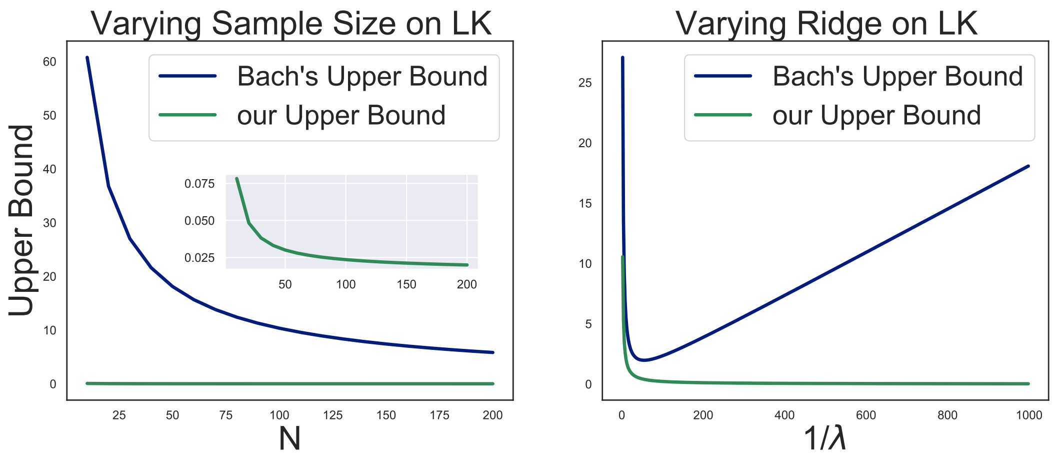

Figure 8: Bound improvement on LK. Same as Figure 2.

Lower Bound Last but not least, we need to show our lower bound is valid. To see this clearly, we need to write the bound in exact sums instead of in HKRS norm square $\|\tilde{f}\|_{\mathcal{H}}^2$: namely, we compute I

$$\frac{\lambda^2 \lambda_M}{(\lambda_M + \lambda)^2}\|\tilde{f}\|_{\mathcal{H}}^2 \le I = \lambda^2 \sum_{k=1}^{M} \frac{\tilde{\gamma}_k^2}{(\lambda_k + \lambda)^2} \le \lambda \|\tilde{f}\|_{\mathcal{H}}^2, \tag{40}$$

instead of using the inequality (40) in Lemma C.19; and

$$M \frac{\lambda_M^2}{(\lambda_M + \lambda)^2} \le \sum_{k=1}^{M} \frac{\lambda_k^2}{(\lambda_k + \lambda)^2} \le M, \tag{41}$$

instead of using the inequality (40) in Theorem C.21. Then we can compute our bounds as:

$$\text{Our upper bound without residue} = \lambda^2 I \left(1 + 2\sqrt{\frac{\log N}{N}}\right) + \frac{\sigma^2}{N} \sum_{k=1}^{M} \frac{\lambda_k^2}{(\lambda_k + \lambda)^2} \left(1 + \sqrt{\frac{\log N}{N}}\right),$$
$$\text{Our lower bound without residue} = \lambda^2 I \left(1 - 2\sqrt{\frac{\log N}{N}}\right) + \frac{\sigma^2}{N} \sum_{k=1}^{M} \frac{\lambda_k^2}{(\lambda_k + \lambda)^2} \left(1 - \sqrt{\frac{\log N}{N}}\right),$$

and we drop the residue terms $C_1 \frac{\log N}{N}$ and $C_2 \frac{\sigma^2}{N} M \frac{\log N}{N}$ by the same reason as before. From Figure 3, we can see that our bounds precisely describe the decay of the test error. Our bounds are not 'bounding' the test errors in smaller instances due to the absence of the residue terms, which increases the interval of confidence of our approximation. But for larger instances, say $N > 100$, all upper and lower bounds, and the averaged test error converge to the same limit.

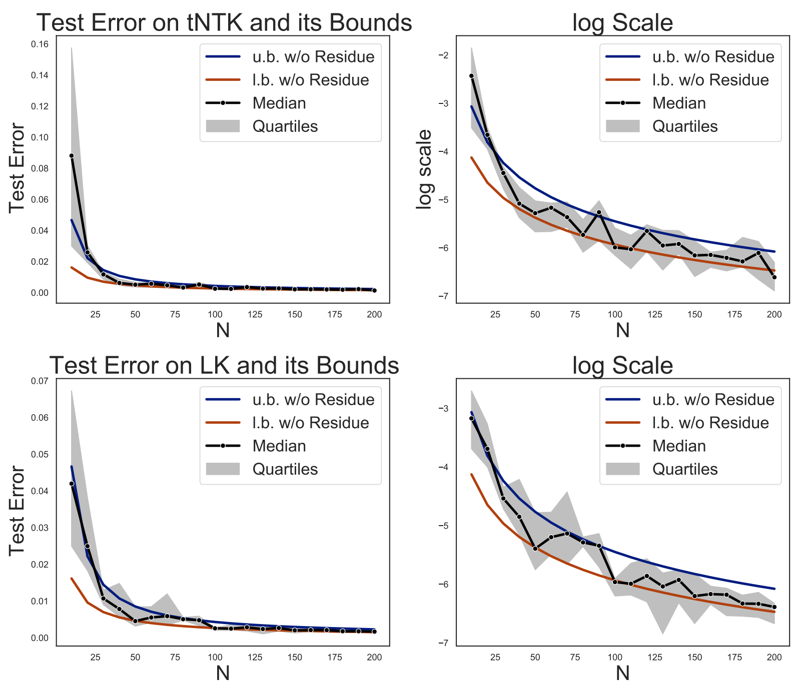

Figure 9: Our bounds comparing to the averaged test error with varying N, over 10 iterations. Same as Figure 3.