Supplementary Material of "Sparse Shrunk Additive Models"

Guodong Liu * 1 **Hong Chen** * 2 **Heng Huang** 1

A. Proofs of Theorems 1 and 2

A.0.Error decomposition

The proofs of Theorems 1 and 2 involve a integration of techniques for error analysis with integral operator approximation (Smale & Zhou, 2007; Sun & Wu, 2011; Shi, 2013; Nie & Wang, 2015) and the empirical process theory for analyzing kernel methods (Pinelis, 1994; Wu et al., 2007; Christmann & Zhou, 2016). The proof of Theorem 3 follows the analysis technique for sparse characterization (Shi et al., 2011).

The key to bound $\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho)$ is a novel error decomposition, where some intermediate functions are constructed as the stepping stone functions. Then, we bound the decomposed terms respectively in terms of operator approximation and concentration equalities for empirical processes.

From Proposition 1 in (Shi, 2013), we know that $L^T_{K^{(j)}} = UL^{\frac{1}{2}}_{\tilde{K}^{(j)}}$ and $L_{K^{(j)}} = L^{\frac{1}{2}}_{\tilde{K}^{(j)}}U^T$ for each $j \in \{1, 2, ..., d\}$, where U is a partial isometry on $L^2_{\rho_{\mathcal{X}^{(j)}}}$ with U^TU being the orthogonal prediction onto the RKHS $\mathcal{H}_{\tilde{K}^{(j)}}$.

For any $j \in \{1, 2, ..., d\}$, define the intermediate function $f^{(j)}_\lambda$ by

$$\begin{aligned} f^{(j)}_\lambda &= \arg\min_{f \in L^2_{\rho_{\mathcal{X}^{(j)}}}} \Big\{ \|L_{K^{(j)}} f^{(j)} - f^{(j)}_\rho\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}} \\ &\quad + \lambda \|U^T f^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}} \Big\}. \end{aligned} \tag{1}$$

Denote $f_\lambda = \sum_{j=1}^d f^{(j)}_\lambda$ and $g_\lambda = \sum_{j=1}^d g^{(j)}_\lambda$ with $g^{(j)}_\lambda = L_{K^{(j)}} f^{(j)}_\lambda$.

*Equal contribution [1]Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, United States [2]Department of Mathematics and Statistics, College of Science, Huazhong Agricultural University, Wuhan, China. Correspondence to: Hong Chen <chenh@mail.hzau.edu.cn>, Guodong Liu <guodong.liu.e@pitt.edu>.

Proceedings of the 37th International Conference on Machine Learning, Online, PMLR 119, 2020.

Define the empirical version of g_λ as

$$\hat{g}_\lambda(x) = \frac{1}{m}\sum_{i=1}^m \sum_{j=1}^d f^{(j)}_\lambda(x^{(j)}_i) K^{(j)}(x^{(j)}_i, x^{(j)}), x \in \mathcal{X}. \tag{2}$$

Now we give the following error decomposition.

Proposition 1. *For $f_{\mathbf{z}}, \hat{g}_\lambda$, there holds*

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \le E_1 + E_2 + E_3,$$

where

$$\begin{aligned} E_1 &= \mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}})) + \mathcal{E}_{\mathbf{z}}(\hat{g}_\lambda) - \mathcal{E}(\hat{g}_\lambda), \\ E_2 &= \mathcal{E}(\hat{g}_\lambda) - \mathcal{E}(g_\lambda) + \lambda\|\hat{g}_\lambda\|_{\ell_1}, \end{aligned}$$

and

$$E_3 = \mathcal{E}(g_\lambda) - \mathcal{E}(f_\rho).$$

Proof. According the definition of $f_{\mathbf{z}}$, we have

$$\begin{aligned} & \mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \\ \le\ & \mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}})) + \mathcal{E}_{\mathbf{z}}(\hat{g}_\lambda) - \mathcal{E}(f_\rho) + \lambda\|\hat{g}_\lambda\|_{\ell_1} \\ & + \Big\{\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\|f_{\mathbf{z}}\|_{\ell_1} - (\mathcal{E}_{\mathbf{z}}(\hat{g}_\lambda) + \lambda\|\hat{g}_\lambda\|_{\ell_1})\Big\} \\ \le\ & \mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}})) + \mathcal{E}_{\mathbf{z}}(\hat{g}_\lambda) - \mathcal{E}(f_\rho) + \lambda\|\hat{g}_\lambda\|_{\ell_1}. \end{aligned} \tag{3}$$

Note that

$$\begin{aligned} \mathcal{E}_{\mathbf{z}}(\hat{g}_\lambda) - \mathcal{E}(f_\rho) &= (\mathcal{E}(\hat{g}_\lambda) - \mathcal{E}(g_\lambda)) + \mathcal{E}(g_\lambda) - \mathcal{E}(f_\rho) \\ &\quad + \mathcal{E}_{\mathbf{z}}(\hat{g}_\lambda) - \mathcal{E}(\hat{g}_\lambda). \end{aligned} \tag{4}$$

Combining both (3) and (4), we get the desired decomposition.

The error term E_1 measures the divergence between the empirical risk and the corresponding expected risk, which usually is called sample error in learning theory. In terms of recent theoretical progress for learning with data dependent hypothesis spaces (Shi et al., 2011; Shi, 2013; Feng et al., 2016), we can bound sample error E_1 via concentration inequality associated with empirical covering numbers (Wu et al., 2007; Christmann & Zhou, 2016). The error term E_2 reflects the drift risk for learning with hypothesis spaces $\mathcal{H}_{\mathbf{z}}$ and $\mathcal{H}$, and hence is called as the hypothesis error.

By relating $\mathcal{E}(\hat{g}_\lambda)-\mathcal{E}(g_\lambda)$ with $\sum_{j=1}^d \|\hat{g}_\lambda^{(j)}-g_\lambda^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}}$, we can estimate this hypothesis error through the inequality in Hilbert space (Pinelis, 1994; Smale & Zhou, 2007). The error term E_2 is called the approximation error, which describes the approximation ability of regularized scheme. Following the approximation analysis with integral operator in (Smale & Zhou, 2007; Shi, 2013; Nie & Wang, 2015), we derive the upper bound of E_2 based on the properties of $L_{\tilde{K}^{(j)}}, 1\le j\le d$.

A.1. Estimate of Approximation Error E_3

In this paper, we use the analysis techniques in (Smale & Zhou, 2007; Shi, 2013) to bound the approximation error E_3.

The following lemma is used in our analysis, which is proved in Proposition 2 in (Shi, 2013).

Lemma 1. *From the definition of $f_\lambda^{(j)}$ and $g_\lambda^{(j)} = L_{K^{(j)}} f_\lambda^{(j)}$, $j\in\{1,2,...,d\}$, there are*

$$f_\lambda^{(j)} = U(\lambda I + L_{\tilde{K}^{(j)}})^{-1} L_{\tilde{K}^{(j)}}^{\frac{1}{2}} f_\rho^{(j)}$$

and

$$\|f_\lambda^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}} = \|U^T f_\lambda^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}}.$$

Lemma 2. *Under Assumption 1, there holds*

$$\begin{aligned}&\|L_{\tilde{K}^{(j)}} f_\lambda^{(j)} - f_\rho^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}} + \lambda\|f_\lambda^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}}\\ \le\ & \lambda^{\min\{1,2r\}}\|g_\rho^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}}(2+\|L_{\tilde{K}^{(j)}}^{r-\frac{1}{2}}\|^2+\|L_{\tilde{K}^{(j)}}^{r-1}\|^2).\end{aligned}$$

Proof. Recall that $\{\lambda_i^{(j)},\psi_i^{(j)}\}_{i\ge1}$ are the normalized eigenpairs of the integral operator $L_{\tilde{K}^{(j)}}$ and $\{\psi\}_{i\ge1}$ form an orthogonal basis of $L^2_{\rho_{\mathcal{X}^{(j)}}}$. Let $g_\rho^{(j)} = L_{\tilde{K}^{(j)}}^{-r} f_\rho^{(j)} = \sum_{t=1}^\infty a_t\psi_t^{(j)}$. Then $\|g_\rho^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}} = \sum_{t=1}^\infty (a_t^{(j)})^2<\infty$.

If Assumption 1 holds for some $r\in(0,\frac{1}{2})$, then from Lemma 1 we have

$$\begin{aligned}&\|f_\lambda^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}} = \|U^T f_\lambda^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}}\\ =\ & \|U^T U(\lambda I + L_{\tilde{K}^{(j)}})^{-1} L_{\tilde{K}^{(j)}}^{\frac{1}{2}} f_\rho^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}}\\ =\ & \|(\lambda I + L_{\tilde{K}^{(j)}})^{-1} L_{\tilde{K}^{(j)}}^{\frac{1}{2}} f_\rho^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}}\\ =\ & \|(\lambda I + L_{\tilde{K}^{(j)}})^{-1} L_{\tilde{K}^{(j)}}^{\frac{1}{2}+r} L_{\tilde{K}^{(j)}}^{-r} f_\rho^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}}\end{aligned}$$

Moreover,

$$\begin{aligned}&\lambda\|f_\lambda^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}}\\ =\ & \lambda\|(\lambda I + L_{\tilde{K}^{(j)}})^{-1} L_{\tilde{K}^{(j)}}^{\frac{1}{2}+r} L_{\tilde{K}^{(j)}}^{-r} f_\rho^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}}\\ =\ & \lambda\Big\|\sum_{t\ge1}\frac{(\lambda_t^{(j)})^{\frac{1}{2}+r}}{\lambda_t^{(j)}+\lambda}a_t^{(j)}\psi_t^{(j)}\Big\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}}\\ =\ & \lambda\sum_{t\ge1}\frac{\lambda_t^{(j)}}{\lambda_t^{(j)}+\lambda}\cdot\frac{\lambda_t^{2r}}{\lambda_t+\lambda}(a_t^{(j)})^2\\ \le\ & \lambda^{2r}\sum_{t\ge1}\frac{\lambda_t^{(j)}}{\lambda_t^{(j)}+\lambda}(a_t^{(j)})^2\\ \le\ & \lambda^{2r}\|g_\rho^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}},\end{aligned}\tag{5}$$

where the first inequality follows from Lemma 1 in (Nie & Wang, 2015) and the second inequality is obtained based on the definition of $g_\rho^{(j)}$.

If Assumption 1 is true for some $r\ge\frac{1}{2}$,

$$\begin{aligned}&\lambda\|f_\lambda^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}}\\ =\ & \lambda\|(\lambda I + L_{\tilde{K}^{(j)}})^{-1} L_{\tilde{K}^{(j)}} L_{\tilde{K}^{(j)}}^{r-\frac{1}{2}} L_{\tilde{K}^{(j)}}^{-r} f_\rho^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}}\\ \le\ & \lambda\|L_{\tilde{K}^{(j)}}^{r-\frac{1}{2}}\|^2\cdot\|L_{\tilde{K}^{(j)}}^{-r} f_\rho^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}}\\ \le\ & \lambda\|L_{\tilde{K}^{(j)}}^{r-\frac{1}{2}}\|^2\cdot\|g_\rho^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}}.\end{aligned}\tag{6}$$

Now turn to bound $\|g_\lambda^{(j)}-f_\rho^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}}$. From Lemma 1, we can deduce that

$$\begin{aligned}g_\lambda^{(j)} &= L_{K^{(j)}} f_\lambda^{(j)} = L_{\tilde{K}^{(j)}}(\lambda I + L_{\tilde{K}^{(j)}})^{-1} f_\rho^{(j)}\\ &= (\lambda I + L_{\tilde{K}^{(j)}})^{-1} L_{\tilde{K}^{(j)}} f_\rho^{(j)}\end{aligned}$$

and

$$\|g_\lambda^{(j)}-f_\rho^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}} = \lambda^2\|(\lambda I + L_{\tilde{K}^{(j)}})^{-1} f_\rho^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}}.$$

For $r\in(0,1)$, we have

$$\begin{aligned}&\lambda\|g_\lambda^{(j)}-f_\rho^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}}\\ =\ & \lambda^2\|(\lambda I + L_{\tilde{K}^{(j)}})^{-1} L_{\tilde{K}^{(j)}}^{r} L_{\tilde{K}^{(j)}}^{-r} f_\rho^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}}\\ \le\ & \lambda^2\sum_{t\ge1}(a_t^{(j)})^2\Big(\frac{(\lambda_t^{(j)})^r}{\lambda_t^{(j)}+\lambda}\Big)^2\\ \le\ & \lambda^{2r}\|g_\rho^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}}.\end{aligned}\tag{7}$$

For $r \geq 1$, we get

$$
\begin{aligned}
& \lambda \|g_\lambda^{(j)} - f_\rho^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}} \\
= \;& \lambda^2 \|(\lambda I + L_{\tilde{K}^{(j)}})^{-1} L_{\tilde{K}^{(j)}} L^{r-1}_{\tilde{K}^{(j)}} L^{-r}_{\tilde{K}^{(j)}} f_\rho^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}} \\
\leq \;& \lambda^2 \|L^{r-1}_{\tilde{K}^{(j)}}\|^2 \|L^{-r}_{\tilde{K}^{(j)}} f_\rho^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}} \\
\leq \;& \lambda^2 \|L^{r-1}_{\tilde{K}^{(j)}}\|^2 \|g_\rho^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}}. \qquad (8)
\end{aligned}
$$

Combining (5)-(8), we get the desired result. $\square$

Lemma 3. *For $j \in \{1, 2, ..., d\}$ and $g_\lambda^{(j)} = L_{K^{(j)}} f_\lambda^{(j)}$ with $f_\lambda^{(j)}$ defined in Section 4, there hold*

$$
\|f_\lambda^{(j)}\|_\infty \leq \sqrt{2 + \|L^{r-\frac{1}{2}}_{\tilde{K}^{(j)}}\|^2 + \|L^{r-1}_{\tilde{K}^{(j)}}\|^2} \cdot \lambda^{\min\{-\frac{1}{2}, r-1\}} \|g_\rho^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}}
$$

and

$$
\|f_\lambda^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}} \leq \sqrt{2 + \|L^{r-\frac{1}{2}}_{\tilde{K}^{(j)}}\|^2 + \|L^{r-1}_{\tilde{K}^{(j)}}\|^2} \cdot \lambda^{\min\{0, r-\frac{1}{2}\}} \|g_\rho^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}}.
$$

Proof. Note that

$$
\begin{aligned}
f_\lambda^{(j)} &= U(\lambda I + L_{\tilde{K}^{(j)}})^{-1} L^{\frac{1}{2}}_{\tilde{K}^{(j)}} f_\rho^{(j)} \\
&= L^T_{K^{(j)}} (\lambda I + L_{\tilde{K}^{(j)}})^{-1} f_\rho^{(j)}
\end{aligned}
$$

and

$$
\|g_\lambda^{(j)} - f_\rho^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}} = \lambda \|(\lambda I + L_{\tilde{K}^{(j)}})^{-1} f_\rho^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}}.
$$

Therefore,

$$
\begin{aligned}
\|f_\lambda^{(j)}\|_\infty &\leq \|(\lambda I + L_{\tilde{K}^{(j)}})^{-1} f_\rho^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}} \\
&= \lambda^{-1} \|L_{K^{(j)}} f_\lambda^{(j)} - f_\rho^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}} \\
&\leq \sqrt{2 + \|L^{r-\frac{1}{2}}_{\tilde{K}^{(j)}}\|^2 + \|L^{r-1}_{\tilde{K}^{(j)}}\|^2} \cdot \lambda^{\min\{-\frac{1}{2}, r-1\}} \|g_\rho^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}}.
\end{aligned}
$$

The second statement follows directly form the result of Lemma 2. $\square$

Proposition 2. For $g_\lambda = \sum_{j=1}^d g_\lambda^{(j)} = \sum_{j=1}^d L_{\tilde{K}^{(j)}} f_\lambda^{(j)}$, there holds

$$
E_3 \leq \lambda^{\min\{1, 2r\}} \Big(2 + \|L^{r-\frac{1}{2}}_{\tilde{K}^{(j)}}\|^2 + \|L^{r-1}_{\tilde{K}^{(j)}}\|^2\Big) \cdot \Big(\sum_{j=1}^d \|g_\rho^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}}\Big)^2.
$$

Proof. Based on Cauchy-Schwarz inequality, we can observe that

$$
\begin{aligned}
\sqrt{E_3} &= \Big(\int_{\mathcal{Z}} (g_\lambda(x) - f_\rho(x))^2 d\rho(x, y)\Big)^{\frac{1}{2}} \\
&= \Big(\int_{\mathcal{Z}} \Big(\sum_{j=1}^d (g_\lambda^{(j)}(x^{(j)}) - f_\rho^{(j)}(x^{(j)})\Big)^2 d\rho(x, y)\Big)^{\frac{1}{2}} \\
&\leq \sum_{j=1}^d \|L_{\tilde{K}^{(j)}} f_\lambda^{(j)} - f_\rho^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}} \qquad (9)
\end{aligned}
$$

Lemma 2 tells us that $\forall j \in \{1, 2, ..., d\}$

$$
\begin{aligned}
& \|L_{\tilde{K}^{(j)}} f_\lambda^{(j)} - f_\rho^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}} \\
\leq \;& \sqrt{2 + \|L^{r-\frac{1}{2}}_{\tilde{K}^{(j)}}\|^2 + \|L^{r-1}_{\tilde{K}^{(j)}}\|^2} \lambda^{\min\{\frac{1}{2}, r\}} \|g_\rho^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}}.
\end{aligned}
$$

Combining this estimate with (9), we get the desired upper bound on E_3. $\square$

A.2. Estimate of Hypothesis Error E_2

The hypothesis error reflects the divergence between $\hat{g}_\lambda$ and g_λ on the expected risk and regularization. The following inequality from (Pinelis, 1994; Smale & Zhou, 2007) is used to bound the divergence.

Lemma 4. *Let $\mathcal{H}$ be a Hilbert space. For an independent random variable ξ on $\mathcal{Z}$ with values in $\mathcal{H}$, assume that $\|\xi\|_{\mathcal{H}} \leq M < \infty$ almost surely. For any given independent identical distributed samples $\{z_i\}_{i=1}^m \subset \mathcal{Z}$ and any $\delta \in (0, 1)$, there holds*

$$
\begin{aligned}
& \Big\|\frac{1}{m}\sum_{i=1}^m \xi(z_i) - E\xi\Big\|_{\mathcal{H}} \\
\leq \;& \frac{2M\log(2/\delta)}{m} + \sqrt{\frac{2E\|\xi\|^2_{\mathcal{H}}\log(2/\delta)}{m}}
\end{aligned}
$$

with confidence at least $1 - \delta/2$.

Proposition 3. For any $0 < \delta < 1$, with confidence $1 - \delta$, we have

$$
\begin{aligned}
& \mathcal{E}(\hat{g}_\lambda) - \mathcal{E}(g_\lambda) \\
\leq \;& 16\sqrt{c}\lambda^{\min\{0, r-\frac{1}{2}\}} \Big(\frac{\log(2/\delta)}{m} + \sqrt{\frac{\log(2/\delta)}{m}}\Big) \\
& \cdot \Big(\sum_{j=1}^d \|g_\rho^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}} + (\sum_{j=1}^d \|g_\rho^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}})^2\Big)
\end{aligned}
$$

and

$$
E_2 \leq c_2 \Big(\lambda^{\min\{0, r-\frac{1}{2}\}} \sqrt{\frac{\log(2/\delta)}{m}} + \lambda^{\min\{\frac{1}{2}, r\}}\Big),
$$

where $c = 2 + \|L^{r-\frac{1}{2}}_{\tilde{K}^{(j)}}\|^2 + \|L^{r-1}_{\tilde{K}^{(j)}}\|^2$ and c_2 is a positive constant independent of m, δ.

Proof. From Cauchy-Schwarz inequality, we can see that

$$\begin{aligned}
&\mathcal{E}(\hat{g}_\lambda) - \mathcal{E}(g_\lambda) \\
\leq\ & \Big(\int_{\mathcal{Z}} (2y - \hat{g}_\lambda(x) - g_\lambda(x))^2 d\rho(x,y)\Big)^{\frac{1}{2}} \\
&\cdot\Big(\int_{\mathcal{Z}} (\hat{g}_\lambda(x) - g_\lambda(x))^2 d\rho(x,y)\Big)^{\frac{1}{2}} \\
\leq\ & \Big(8 + 2\int_{\mathcal{Z}} (\hat{g}_\lambda(x) - g_\lambda(x))^2 d\rho(x,y)\Big)^{\frac{1}{2}} \\
&\cdot\Big(\int_{\mathcal{Z}} (\hat{g}_\lambda(x) - g_\lambda(x))^2 d\rho(x,y)\Big)^{\frac{1}{2}} \\
\leq\ & \Big(\sqrt{8} + \sqrt{2}\sum_{j=1}^{d} \|\hat{g}_\lambda^{(j)} - g_\lambda^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}}\Big) \\
&\cdot \sum_{j=1}^{d} \|\hat{g}_\lambda^{(j)} - g_\lambda^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}}. \qquad (10)
\end{aligned}$$

Denote $\xi^{(j)} = f_\lambda^{(j)}(x^{(j)})K(x^{(j)}, u)$ for any $j \in \{1, 2, ..., d\}$. Then, from Lemma 3, we can deduce that

$$\|\xi^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}} \leq \|f_\lambda^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}} \leq \sqrt{c}\lambda^{\min\{0, r-\frac{1}{2}\}}\|g_\rho^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}}$$

and

$$E\|\xi^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}} \leq \|f_\lambda^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}} \leq c\lambda^{\min\{0, 2r-1\}}\|g_\rho^{(j)}\|^2_{L^2_{\rho_{\mathcal{X}^{(j)}}}}.$$

Moreover, for any $j \in \{1, ..., d\}$ and $u \in \mathcal{X}^{(j)}$,

$$\begin{aligned}
&\|\hat{g}_\lambda^{(j)} - g_\lambda^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}} \\
=\ & \Big\|\frac{1}{m}\sum_{i=1}^{m} \xi_i^{(j)} - E\xi^{(j)}\Big\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}} \\
\leq\ & \frac{2\sqrt{c}\lambda^{\min\{0, r-\frac{1}{2}\}}\|g^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}} \log(2/\delta)}{m} \\
&+\lambda^{\min\{0, r-\frac{1}{2}\}}\|g_\rho^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}}\sqrt{\frac{2c\log(2/\delta)}{m}}, \qquad (11)
\end{aligned}$$

where the last inequality is derived from Lemma 4. Then, we obtain the first statement by combining the estimates (10) and (11).

Now consider the upper bound of $\lambda\|\hat{g}_\lambda\|_{\ell_1}$. From the definition of $\hat{g}_\lambda$, we have

$$\begin{aligned}
\lambda\|\hat{g}_\lambda\|_{\ell_1} &\leq \lambda\sum_{j=1}^{d}\|f_\lambda^{(j)}\|_\infty \leq \sum_{j=1}^{d}\|L_{K^{(j)}}f_\lambda^{(j)} - f_\rho^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}} \\
&\leq \sqrt{c}\lambda^{\min\{0, r-\frac{1}{2}\}}\sum_{j=1}^{d}\|g^{(j)}\|_{L^2_{\rho_{\mathcal{X}^{(j)}}}}.
\end{aligned}$$

Combining this estimate with the first statement, we derive the desired upper bound of E_2. □

A.3. Estimate of Sample Error E_1

In this paper, the sample error is estimated by the analysis technique associated with the empirical covering numbers. The empirical covering numbers with ℓ_2-metric is denoted by $\mathcal{N}_2(\mathcal{F}, \varepsilon)$ and its detail definition can be founded in (Van der Vaart & Wellner, 1996; Shi et al., 2011).

Definition 1. *For a function set $\mathcal{F}$ and $\mathbf{u} = (u_i)_{i=1}^k \in \mathcal{X}$, the metric $d_{2,\mathbf{u}}$ is defined by*

$$d_{2,\mathbf{u}}(f, g) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(f(u_i) - g(u_i))^2}, \forall f, g \in \mathcal{F}.$$

For every $\varepsilon > 0$, the empirical covering number is defined as $\mathcal{N}_2(\mathcal{F}, \varepsilon) = \sup_{k\in\mathbb{N}}\sup_{\mathbf{u}\in\mathcal{X}^k}\mathcal{N}_{2,\mathbf{u}}(\mathcal{F}, \varepsilon)$, where

$$\begin{aligned}
\mathcal{N}_{2,\mathbf{u}}(\mathcal{F}, \varepsilon) \;=\; &\inf\Big\{l \in \mathbb{N} : \exists\{f_i\}_{i=1}^l \textit{ such that} \\
&\mathcal{F} \subset \cup_{i=1}^l\{f \in \mathcal{F} : d_{2,\mathbf{u}}(f, f_i) \leq \varepsilon\}\Big\}.
\end{aligned}$$

The following concentration inequality is established in (Wu et al., 2007).

Lemma 5. *Let $\mathcal{F}$ be a measurable function set on $\mathcal{Z}$. Assume that, for any $f \in \mathcal{F}$, $\|f\|_\infty \leq B$ and $E(f^2) \leq cEf$ for some positive constants B, c. If for some $a > 0$ and $s \in (0, 2)$, $\log\mathcal{N}_2(\mathcal{F}, \varepsilon) \leq a\varepsilon^{-p}$ for any $\varepsilon > 0$, then there exists a constant c_p' such that for any $\delta \in (0, 1)$,*

$$\begin{aligned}
&\Big|Ef - \frac{1}{m}\sum_{i=1}^{m}f(z_i)\Big| \\
\leq\ & \frac{1}{2}Ef + c_p'\max\{c^{\frac{2-p}{2+p}}, B^{\frac{2-p}{2+p}}\}(\frac{a}{m})^{\frac{2}{2+p}} \\
&+\frac{(2c + 18B)\log(1/\delta)}{m}
\end{aligned}$$

with confidence at least $1 - 2\delta$.

For any $R > 0$, denote

$$\begin{aligned}
\mathcal{B}_R^{(j)} \;=\; &\Big\{f^{(j)} = \sum_{i=1}^{m}\alpha_i^{(j)}K^{(j)}(u_i^{(j)}, \cdot) \in \mathcal{H}^{(j)} : \\
&\|f^{(j)}\|_{\ell_1} \leq R\Big\}
\end{aligned}$$

and

$$\mathcal{B}_R = \Big\{f = \sum_{j=1}^{d}f^{(j)} : \|f\|_{\ell_1} \leq R\Big\},$$

where

$$\|f\|_{\ell_1} = \inf\Big\{\sum_{j=1}^{d}\|f^{(j)}\|_{\ell_1} : f = \sum_{j=1}^{d}f^{(j)}, f^{(j)} \in \mathcal{H}^{(j)}\Big\}.$$

Now we state the estimate on the empirical covering numbers of $\mathcal{B}_1$. Similar analysis can be found in (Christmann & Zhou, 2016) for $\mathcal{B}_1$ in reproducing kernel Hilbert spaces.

Lemma 6. *For any $j \in \{1, 2, ..., d\}$, assume that $K^{(j)} \in C^s$ for some $s > 0$. Then,*

$$\log \mathcal{N}_2(\mathcal{B}_1, \varepsilon) \le d^{1+p} c_p \varepsilon^{-p},$$

where p is defined in Section 3 and c_p is a constant independent of ε.

Proof. For every $j \in \{1, 2, ..., d\}$ and $\mathbf{x}^{(j)} \in (\mathcal{X}^{(j)})^S$, there exists a set $\{f_i^{(j)}\}_{i=1}^{N_j}$ with $N_i = \mathcal{N}_2(\mathcal{B}_1^{(j)}, \varepsilon)$ such that

$$\begin{aligned} \forall f^{(j)} \in \mathcal{B}_1^{(j)}, \quad \exists \quad & i_j \in \{1, 2, ..., N_j\}, s.t., \\ & d_{2,\mathbf{x}^{(j)}}(f^{(j)} - f_{i_j}^{(j)}) \le \varepsilon. \end{aligned}$$

For $f = \sum_{j=1}^d f^{(j)} \in \mathcal{B}_1$, we know $f^{(j)} \in \mathcal{B}_1^{(j)}$. For every $\mathbf{x} = (x_\ell)_{\ell=1}^S \in \mathcal{X}^S$, we have $\mathbf{x}^{(j)} = (x_\ell^{(j)})_{\ell=1}^S \in (\mathcal{X}^{(j)})^S, j \in \{1, 2, ..., d\}$. Let $\tilde{f} = \sum_{j=1}^d f_{i_j}^{(j)}$. Then

$$\begin{aligned} & d_{2,\mathbf{x}}(f, \tilde{f}) \\ = & \Big\{\frac{1}{S}\sum_{\ell=1}^S (f(x_\ell) - \tilde{f}(x_\ell))^2\Big\}^{\frac{1}{2}} \\ = & \Big\{\frac{1}{S}\sum_{\ell=1}^S (\sum_{j=1}^d f^{(j)}(x_\ell^{(j)}) - \sum_{j=1}^d \tilde{f}_{i_j}^{(j)}(x_\ell^{(j)}))^2\Big\}^{\frac{1}{2}} \\ \le & \sum_{j=1}^d \Big\{\frac{1}{S}\sum_{l=1}^S (f^{(j)}(x_\ell^{(j)}) - \tilde{f}_{i_j}^{(j)}(x_\ell^{(j)}))^2\Big\}^{\frac{1}{2}} \\ \le & \sum_{j=1}^d d_{2,\mathbf{x}^{(j)}}(f^{(j)} - f_{i_j}^{(j)}) \\ \le & d\varepsilon. \end{aligned}$$

Therefore,

$$\log \mathcal{N}_2(\mathcal{B}_1, d\varepsilon) \le \sum_{j=1}^d \log \mathcal{N}_2(\mathcal{B}_1^{(j)}, d\varepsilon).$$

According to Theorem 2 in (Shi et al., 2011) (also see Lemmas 2 and 3 in (Shi, 2013)) and considering $\|f^{(j)}\|_{\ell_1} \le \sqrt{m\|f^{(j)}\|_{\ell_2}^2}$, we further get

$$\log \mathcal{N}_2(\mathcal{B}_1, d\varepsilon) \le d c_p \varepsilon^{-p}.$$

Setting $\tilde{\varepsilon} = d\varepsilon$, we get the desired result. □

Proposition 4. Under Assumptions 1 and 2, for any $\delta \in (0, 1)$, there holds

$$\begin{aligned} E_1 \quad \le \quad & \frac{1}{2}(\mathcal{E}(\pi(f_\mathbf{z})) - \mathcal{E}(f_\rho)) + \frac{1}{2}(\mathcal{E}(\hat{g}_\lambda) - \mathcal{E}(g_\lambda)) \\ & + E_3 + C_1 \log(2/\delta)\big(\lambda^{-\frac{2p}{2+p}} m^{-\frac{2}{2+p}} \\ & + \lambda^{\min\{-1, 2r-2\}} m^{-\frac{2}{2+p}} + m^{-1}\big) \end{aligned}$$

with confidence $1 - \delta$, where C_1 is a positive constant independent m, λ, δ, and p is defined in Section 3.

Proof. The sample error E_1 can be decomposed as

$$E_{11} = \mathcal{E}(\pi(f_\mathbf{z})) - \mathcal{E}(f_\rho) - (\mathcal{E}_\mathbf{z}(\pi(f_\mathbf{z})) - \mathcal{E}_\mathbf{z}(f_\rho))$$

and

$$E_{12} = \mathcal{E}_\mathbf{z}(\hat{g}_\lambda) - \mathcal{E}_\mathbf{z}(f_\rho) - (\mathcal{E}(\hat{g}_\lambda) - \mathcal{E}(f_\rho)).$$

In the sequel, we will bound E_{11} and E_{12} respectively.

Denote

$$\mathcal{G}_R = \{g(z) = (y - \pi(f)(x))^2 - (y - f_\rho(x))^2 : f \in \mathcal{B}_R\}.$$

For any $g \in \mathcal{G}_R$, we can deduce that $|g(z)| \le 8$ and $Eg^2 \le 16Eg$. Let $g_1, g_2 \in \mathcal{G}_R$ associated with f_1, f_2 respectively. It can be seen that

$$\begin{aligned} |g_1(z) - g_2(z)| \quad & \le \quad 4|\pi(f_1)(x) - \pi(f_2)(x)| \\ & \le \quad 4|f_1(x) - f_2(x)|. \end{aligned}$$

This means

$$\begin{aligned} \log \mathcal{N}_2(\mathcal{G}_R, \varepsilon) \quad & \le \quad \log \mathcal{N}_2(\mathcal{B}_R, \frac{\varepsilon}{4}) \le \log \mathcal{N}_2(\mathcal{B}_1, \frac{\varepsilon}{4R}) \\ & \le \quad c_p d^{1+p}(4R)^p \varepsilon^{-p}, \end{aligned}$$

where the last inequality follows form Lemma 6.

Applying Lemma 5 to $\mathcal{G}_R$, we have with confidence $1 - \frac{\delta}{2}$

$$\begin{aligned} & Eg - \frac{1}{m}\sum_{i=1}^m g(z_i) \le \frac{1}{2}(\mathcal{E}(\pi(f)) - \mathcal{E}(f_\rho)) \\ & + \tilde{c}_1(R^{\frac{2p}{2+p}} m^{-\frac{2}{2+p}} + m^{-1}\log(2/\delta)), \forall g \in \mathcal{G}_R, \end{aligned}$$

where $\tilde{c}_1$ is a constant independent of m, δ.

From the definition of $f_\mathbf{z}$ in Section 2, we know $f_\mathbf{z} \in \mathcal{B}_R$ with $R = \lambda^{-1}$. Then

$$\begin{aligned} E_{11} \quad \le \quad & \frac{1}{2}(\mathcal{E}(\pi(f_\mathbf{z})) - \mathcal{E}(f_\rho)) \\ & + \tilde{c}_1(\lambda^{-\frac{2p}{2+p}} m^{-\frac{2}{2+p}} + m^{-1}\log(1/\delta)) \quad (12) \end{aligned}$$

with confidence $1 - \frac{\delta}{2}$.

Now we turn to bound E_{12}. Denote

$$\hat{\mathcal{G}} = \Big\{\hat{g} = \sum_{j=1}^d \hat{g}_\lambda^{(j)} : \hat{g}_\lambda^{(j)} = \frac{1}{m}\sum_{i=1}^m f_\lambda^{(j)}(v_i^{(j)})K(v_i^{(j)}, \cdot)\Big\}$$

and

$$\hat{\mathcal{H}} = \Big\{h : h(z) = (y - \hat{g}(x))^2 - (y - f_\rho(x))^2, \hat{g} \in \hat{\mathcal{G}}\Big\}.$$

We can verify that

$$\begin{aligned} \|h\|_\infty &= \sup |2y - \hat{g}(x) - f_\rho(x)| \cdot |\hat{g}(x) - f_\rho(x)| \\ &\leq (3 + \sum_{j=1}^d \|f_\lambda^{(j)}\|_\infty)^2 \end{aligned}$$

and

$$Eh^2 \leq (3 + \sum_{j=1}^d \|f_\lambda^{(j)}\|_\infty)^2 Eh.$$

For any given $\hat{g}_1, \hat{g}_2 \in \hat{\mathcal{H}}$, the corresponding $h_1, h_2 \in \hat{\mathcal{H}}$ satisfy

$$|h_1(z) - h_2(z)| \leq 2(1 + \sum_{j=1}^d \|f_\lambda^{(j)}\|_\infty)|\hat{g}_1(x) - \hat{g}_2(x)|.$$

Then, from Lemma 6 and $\hat{g} \in \mathcal{B}_R$ with $R = \sum_{j=1}^d \|f_\lambda^{(j)}\|_\infty$, we have

$$\begin{aligned} & \log \mathcal{N}_2(\hat{\mathcal{H}}, \varepsilon) \\ \leq & \log \mathcal{N}_2\Big(\hat{\mathcal{G}}, \frac{\varepsilon}{2(1 + \sum_{j=1}^d \|f_\lambda^{(j)}\|_\infty)}\Big) \\ \leq & \log \mathcal{N}_2\Big(\mathcal{B}_1, \frac{\varepsilon}{2\sum_{j=1}^d \|f_\lambda^{(j)}\|_\infty(1 + \sum_{j=1}^d \|f_\lambda^{(j)}\|_\infty)}\Big) \\ \leq & c_p d^{1+p} 2^p (\sum_{j=1}^d \|f_\lambda^{(j)}\|_\infty + \sum_{j=1}^d \|f_\lambda^{(j)}\|_\infty^2)^p \varepsilon^{-p}. \end{aligned}$$

Applying Lemma 5 to $\hat{\mathcal{H}}$, with confidence $1 - \frac{\delta}{2}$ we have

$$\begin{aligned} E_{12} &= \sum_{i=1}^m h(z_i) - Eh \leq \frac{1}{2}(\mathcal{E}(\hat{g}_\lambda) - \mathcal{E}(f_\rho)) \\ &\quad + \tilde{c}_2 \|f_\lambda\|_\infty^2 (m^{-\frac{2}{2+p}} + m^{-1}\log(2/\delta)) \\ &\leq \frac{1}{2}(\mathcal{E}(\hat{g}_\lambda) - \mathcal{E}(g_\lambda) + E_3) \\ &\quad + d\tilde{c}_2' \lambda^{\min\{-1, 2r-2\}}(m^{-\frac{2}{2+p}} + m^{-1}\log(2/\delta)), \end{aligned}$$

where the last inequality follows from Lemma 3 and $\tilde{c}_2, \tilde{c}_2'$ are some positive constants.

Combining this with the estimates of E_{11} in (12), we get the upper bound on E_1. □

A.4. Proof of Theorem 1

Proof. Combining Propositions 1-4, we have with confidence $1 - 4\delta$

$$\begin{aligned} & \mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \\ \leq & C\log(2/\delta)(\lambda^{\min\{1, 2r\}} + \lambda^{\min\{0, r - \frac{1}{2}\}} m^{-\frac{1}{2}} \\ & + \lambda^{\min\{-1, 2r-2\}} m^{-\frac{2}{2+p}} + \lambda^{-\frac{2p}{2+p}} m^{-\frac{2}{2+p}}). \end{aligned}$$

When $r \in (0, \frac{1}{2})$, by setting $\lambda = m^{-\theta_1}$ with $0 < \theta_1 < \min\{\frac{1}{p}, \frac{1}{(2+p)(1-r)}\}$, we get with confidence $1 - 4\delta$

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \leq 4C\log(2/\delta)m^{-\gamma_1},$$

where $\gamma_1 = \min\big\{2r\theta_1, \frac{1}{2} + (r - \frac{1}{2})\theta_1, \frac{2}{2+p} - (2 - 2r)\theta_1, \frac{2}{2+p} - \frac{2p\theta_1}{2+p}\big\}$.

When $r \geq \frac{1}{2}$, taking $\lambda = m^{-\theta_2}$ with some $0 < \theta_2 < \min\{\frac{1}{p}, \frac{2}{2+p}\}$, we have with confidence $1 - 4\delta$

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \leq 4C\log(2/\delta)m^{-\gamma_2},$$

where

$$\gamma_2 = \min\Big\{\theta_2, \frac{1}{2}, \frac{2}{2+p} - \theta_2, \frac{2}{2+p} - \frac{2p\theta_2}{2+p}\Big\},$$

This completes the proof. □

A.5. Proof of Theorem 2

Theorem 2 is dependent on much stronger conditions on f_ρ than Theorem 1. The proof can be obtained directly by the estimate of E_{11} in Proposition 4.

Proof. Since $f_\rho^{(j)} \in \mathcal{H}^{(j)}$ for each $j \in \{1, 2, ..., d\}$, we know that $f_\rho \in \mathcal{H}$. Then,

$$\begin{aligned} & \mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \\ \leq & \mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}})) \\ & + \{\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\|f_{\mathbf{z}}\|_{\ell_1} - (\mathcal{E}_{\mathbf{z}}(f_\rho) + \lambda\|f_\rho\|_{\ell_1})\} \\ \leq & \mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) - (\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(f_\rho)) + \lambda\|f_\rho\|_{\ell_1} \\ = & E_{11} + \lambda\|f_\rho\|_{\ell_1}. \end{aligned}$$

From the estimate of E_{11} in (12), with confidence $1 - \delta$ we have

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \leq \bar{c}\log(1/\delta)(\lambda^{-\frac{2p}{2+p}} m^{-\frac{2}{2+p}} + \lambda),$$

where $\bar{c}$ is a positive constant independent of m, λ.

Taking λ such that $\lambda^{-\frac{2p}{2+p}} m^{-\frac{2}{2+p}} = \lambda$, we get the desired result. □

B. Proof of Theorem 3

Proof. Denote $\alpha = (\alpha_t^{(j)})_{t,j} \in \mathbb{R}^{md}$, where $t \in \{1, 2, ..., m\}$ and $j \in \{1, 2, ..., d\}$. Define

$$G(\alpha) = \frac{1}{m}\sum_{i=1}^{m}\big(y_i - \sum_{j=1}^{d}\sum_{t=1}^{m}\alpha_t^{(j)}K^{(j)}(x_t^{(j)}, x_i^{(j)})\big)^2 + \lambda\sum_{j=1}^{d}\sum_{t=1}^{m}|\alpha_t^{(j)}|.$$

Recall that $f_{\mathbf{z}} = \sum_{j=1}^{d}\sum_{t=1}^{m}\hat{\alpha}_t^{(j)}K^{(j)}(x_t^{(j)}, \cdot)$ and $\hat{\alpha} = (\hat{\alpha}_t^{(j)})_{t,j}$ is the maximizer of $G(\alpha)$.

Let $I_+ = \{(t,j) : \hat{\alpha}_t^{(j)} > 0\}$, $I_- = \{(t,j) : \hat{\alpha}_t^{(j)} < 0\}$, and $I_0 = \{(t,j) : \hat{\alpha}_t^{(j)} = 0\}$.

For $(t,j) \in I_+$, we get

$$\begin{aligned} &\frac{\partial G(\alpha)}{\partial \alpha_t^{(j)}}\Big|_{\alpha=\hat{\alpha}} \\ = &-\frac{2}{m}\sum_{i=1}^{m}(y_i - f_{\mathbf{z}}(x_i))K^{(j)}(x_t^{(j)}, x_i^{(j)}) + \lambda \\ = &\ 0 \end{aligned}$$

This means

$$\frac{1}{m}\sum_{i=1}^{m}(y_i - f_{\mathbf{z}}(x_i))K^{(j)}(x_t^{(j)}, x_i^{(j)}) = \frac{\lambda}{2}.$$

Similarly, for $(t,j) \in I_-$, there exists

$$\frac{1}{m}\sum_{i=1}^{m}(y_i - f_{\mathbf{z}}(x_i))K^{(j)}(x_t^{(j)}, x_i^{(j)}) = -\frac{\lambda}{2}.$$

For $(t,j) \in I_0$, there holds

$$\begin{aligned} -\frac{2}{m}\sum_{i=1}^{m}(y_i - f_{\mathbf{z}}(x_i))K^{(j)}(x_t^{(j)}, x_i^{(j)}) - \lambda &\leq \frac{\partial G(\alpha)}{\partial \alpha_t^{(j)}}\Big|_{\alpha=\hat{\alpha}} \\ &\leq -\frac{2}{m}\sum_{i=1}^{m}(y_i - f_{\mathbf{z}}(x_i))K^{(j)}(x_t^{(j)}, x_i^{(j)}) + \lambda. \end{aligned}$$

This means, for any $(t,j) \in I_0$,

$$\Big|\frac{1}{m}\sum_{i=1}^{m}(y_i - f_{\mathbf{z}}(x_i))K^{(j)}(x_t^{(j)}, x_i^{(j)})\Big| < \frac{\lambda}{2}.$$

This completes the proof. $\square$

References

Christmann, A. and Zhou, D. X. Learning rates for the risk of kernel-based quantile regression estimators in additive models. *Analysis and Applications*, 14(3):449–477, 2016.

Feng, Y., Lv, S., Hang, H., and Suykens, J. A. Kernelized elastic net regularization: Generalization bounds, and sparse recovery. *Neural Comput.*, 28(3):525–562, 2016.

Nie, W. and Wang, C. Constructive analysis for coefficient regularization regression algorithms. *Journal of Mathematical Analysis and Applications*, 431(2):1153–1171, 2015.

Pinelis, I. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pp. 1679–1706, 1994.

Shi, L. Learning theory estimates for coefficient-based regularized regression. *Appl. Comput. Harmon. Anal.*, 34 (2):252–265, 2013.

Shi, L., Feng, Y., and Zhou, D. X. Concentration estimates for learning with ℓ_1-regularizer and data dependent hypothesis spaces. *Appl. Comput. Harmon. Anal.*, 31(2): 286–302, 2011.

Smale, S. and Zhou, D. X. Learning theory estimates via integral operators and their approximations. *Constr. Approx.*, 26(2):153–172, 2007.

Sun, H. and Wu, Q. Least square regression with indefinite kernels and coefficient regularization. *Appl. Comput. Harmon. Anal.*, 30(1):96–109, 2011.

Van der Vaart, A. and Wellner, J. *Weak Convergence and Empirical Processes*. Springer, 1996.

Wu, Q., Ying, Y., and Zhou, D.-X. Multi-kernel regularized classifiers. *J. Complexity*, 23(1):108–134, 2007.