A. Proofs of Theorems 1 and 2

A.0. Error decomposition

The proofs of Theorems 1 and 2 involve an 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 $E(\pi(f_\lambda)) - E(f_\rho)$ is a novel error decomposition, where some intermediate functions are constructed as the stepping stones. 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 $L^T_{K(i)} = U L^2_{K(i)} U^T$ for each $j \in \{1, 2, ..., d\}$, where $U$ is a partial isometry on $L^2_{\nu_K(i)}$ with $U^T U$ being the orthogonal projection onto the RKHS $\mathcal{H}_K(i)$.

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

$$ f^{(j)}_\lambda = \arg \min_{f \in L^2_{\nu_K(i)}} \left\{ \| L_{K(i)} f^{(j)} - f^{(j)}_\rho \|^2, \right. $$

$$\left. + \lambda \| U^T f^{(j)} \|^2 \right\}. $$

(1)

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

Define the empirical version of $g_\lambda$ as

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

(2)

Now we give the following error decomposition.

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

$$ E(\pi(f_\lambda)) - E(f_\rho) \leq E_1 + E_2 + E_3, $$

where

$$ E_1 = E(\pi(f_\lambda)) - E_\pi(\hat{g}_\lambda) + E_\pi(\hat{g}_\lambda) - E(\hat{g}_\lambda), $$

$$ E_2 = E(\hat{g}_\lambda) - E(g_\lambda) + \lambda \| \hat{g}_\lambda \|_{\ell_1}, $$

and

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

**Proof.** According the definition of $f_\lambda$, we have

$$ E(\pi(f_\lambda)) - E(f_\rho) \leq E(\pi(f_\lambda)) - E_\pi(\hat{g}_\lambda) + E_\pi(\hat{g}_\lambda) - E(f_\rho) + \lambda \| \hat{g}_\lambda \|_{\ell_1} $$

$$ + \left\{ E_\pi(f_\lambda) + \lambda \| f_\lambda \|_{\ell_1} - (E_\pi(\hat{g}_\lambda) + \lambda \| \hat{g}_\lambda \|_{\ell_1}) \right\}. $$

$$ \leq E(\pi(f_\lambda)) - E_\pi(\pi(f_\lambda)) + E_\pi(\hat{g}_\lambda) - E(f_\rho) + \lambda \| \hat{g}_\lambda \|_{\ell_1} $$

Note that

$$ E_\pi(\hat{g}_\lambda) - E(f_\rho) = (E(\hat{g}_\lambda) - E(g_\lambda)) + E(g_\lambda) - E(f_\rho) $$

$$ + \epsilon(\hat{g}_\lambda) - E(\hat{g}_\lambda). $$

(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 $\mathcal{H}_\pi$ and $\mathcal{H}_i$, and hence is called as the hypothesis error.
By relating $\mathcal{E}(\hat{g}_\lambda) - \mathcal{E}(g_\lambda)$ with $\sum_{j=1}^d \|\hat{g}_\lambda^{(j)} - g_\lambda^{(j)}\|_2^2$, 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 $L_{K^{(j)}}, 1 \leq j \leq d$.

### 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

$$f_\lambda^{(j)} = U (\lambda I + L_{K^{(j)}})^{-1} L_{K^{(j)}}^{1/2} f^{(j)}$$

and

$$\|f_\lambda^{(j)}\|_{L_2^{(j)}}^2 = \|U^T f_\lambda^{(j)}\|_{L_2^{(j)}}^2.$$

**Lemma 2.** Under Assumption 1, there holds

$$\|L_{K^{(j)}} f_\lambda^{(j)} - f_{\rho}^{(j)}\|_{L_2^{(j)}}^2 + \lambda \|f_\lambda^{(j)}\|_{L_2^{(j)}}^2 \leq \lambda_{\min}^{(1,2r)} \|g_{\rho}^{(j)}\|_{L_2^{(j)}}^2 \left(2 + \|L_{K^{(j)}}^{-1} f_{\rho}\|_2^2 + \|L_{K^{(j)}}^{-1}\|_2^2\right).$$

**Proof.** Recall that $\left\{\lambda_{(j)}^{(1)}, \psi^{(1)}_{(j)}\right\}_{j \geq 1}$ are the normalized eigenpairs of the integral operator $\tilde{L}_{K^{(j)}}$ and $\{\psi^{(j)}\}_{j \geq 1}$ form an orthogonal basis of $L_2^{(j)}$. Let $g_{\rho}^{(j)} = L_{K^{(j)}}^{-1} f_{\rho}^{(j)} = \sum_{t=1}^\infty a_t \psi^{(j)}_t$. Then $\|g_{\rho}^{(j)}\|_{L_2^{(j)}}^2 = \sum_{t=1}^\infty (a_t)^2 < \infty$.

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

$$\|f_\lambda^{(j)}\|_{L_2^{(j)}}^2 = \|U^T f_\lambda^{(j)}\|_{L_2^{(j)}}^2 = \|U^T (\lambda I + L_{K^{(j)}})^{-1} L_{K^{(j)}}^{1/2} f^{(j)}\|_{L_2^{(j)}}^2$$

Moreover,

$$\lambda \|f_\lambda^{(j)}\|_{L_2^{(j)}}^2 = \lambda \|L_{K^{(j)}}\|_2^{-1} \lambda (\lambda I + L_{K^{(j)}}) L_{K^{(j)}}^{1/2} f_{\rho}^{(j)}\|_{L_2^{(j)}}^2$$

and

$$\lambda \|L_{K^{(j)}}^{-1} L_{K^{(j)}}^{-1} \lambda_{\min}^{(1,2r)} \|g_{\rho}^{(j)}\|_{L_2^{(j)}}^2 \leq \lambda^2 \sum_{t=1}^\infty (a_t)^2 \left(\frac{\lambda_{(j)}^{(1)}}{\lambda^{(j)} + \lambda}\right)^2 \|g_{\rho}^{(j)}\|_{L_2^{(j)}}^2.$$
For $r \geq 1$, we get
\[
\lambda \| g_{\lambda}^{(j)} - f_{\rho}^{(j)} \|_{L_{\infty}^2(x^{(j)})}^2 = \lambda^2 \|(A + L_{K^{(j)}})^{-1} f_{\rho}^{(j)} \|_{L_{\infty}^2(x^{(j)})}^2 \leq \lambda^2 \| L_{K^{(j)}}^{-1} \|_2^2 \| g_{\rho}^{(j)} \|_{L_{\infty}^2(x^{(j)})}^2 \leq \lambda^2 \| L_{K^{(j)}}^{-1} \|_2^2 \| g_{\rho}^{(j)} \|_{L_{\infty}^2(x^{(j)})}^2 \cdot \lambda \|(A + L_{K^{(j)}})^{-1} f_{\rho}^{(j)} \|_{L_{\infty}^2(x^{(j)})}^2.
\]
Combining (5)-(8), we get the desired result.

**Lemma 3.** For $j \in \{1, 2, \ldots, d\}$ and $g_{\lambda}^{(j)} = L_{K^{(j)}} f_{\lambda}^{(j)}$ with $f_{\lambda}^{(j)}$ defined in Section 4, there hold
\[
\| f_{\lambda}^{(j)} \|_{L_{\infty}^2(x^{(j)})} \leq \sqrt{2 + \| L_{K^{(j)}}^{-1} \|_2^2 + \| L_{K^{(j)}}^{-1} \|_2^2} \cdot \lambda \min_{0, r-1} \| g_{\rho}^{(j)} \|_{L_{\infty}^2(x^{(j)})}^2
\]
and
\[
\| f_{\lambda}^{(j)} \|_{L_{\infty}^2(x^{(j)})} \leq \sqrt{2 + \| L_{K^{(j)}}^{-1} \|_2^2 + \| L_{K^{(j)}}^{-1} \|_2^2} \cdot \lambda \min_{0, r-1} \| g_{\rho}^{(j)} \|_{L_{\infty}^2(x^{(j)})}^2.
\]

**Proof.** Note that
\[
f_{\lambda}^{(j)} = U(A + L_{K^{(j)}})^{-1} L_{K^{(j)}}^{-1} f_{\rho}^{(j)} = L_{K^{(j)}}^T (A + L_{K^{(j)}})^{-1} f_{\rho}^{(j)}
\]
and
\[
\| g_{\lambda}^{(j)} - f_{\rho}^{(j)} \|_{L_{\infty}^2(x^{(j)})} = \lambda \|(A + L_{K^{(j)}})^{-1} f_{\rho}^{(j)} \|_{L_{\infty}^2(x^{(j)})}.
\]
Therefore,
\[
\| f_{\lambda}^{(j)} \|_{L_{\infty}^2(x^{(j)})} \leq \|(A + L_{K^{(j)}})^{-1} f_{\rho}^{(j)} \|_{L_{\infty}^2(x^{(j)})} \leq \sqrt{2 + \| L_{K^{(j)}}^{-1} \|_2^2 + \| L_{K^{(j)}}^{-1} \|_2^2} \cdot \lambda \min_{0, r-1} \| g_{\rho}^{(j)} \|_{L_{\infty}^2(x^{(j)})}.
\]

The second statement follows directly from the result of Lemma 2.

**Proposition 2.** For $g_{\lambda} = \sum_{j=1}^d g_{\lambda}^{(j)} = \sum_{j=1}^d L_{K^{(j)}} f_{\lambda}^{(j)}$, there holds
\[
E_3 \leq \lambda \min_{1,2r} \left(2 + \| L_{K^{(j)}}^{-1} \|_2^2 + \| L_{K^{(j)}}^{-1} \|_2^2\right) \cdot \left(\sum_{j=1}^d \| g_{\rho}^{(j)} \|_{L_{\infty}^2(x^{(j)})}^2\right)^2.
\]

**A.2. Estimate of Hypothesis Error $E_2$**

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

**Lemma 4.** Let $H$ be a Hilbert space. For an independent random variable $\xi$ on $\mathcal{Z}$ with values in $H$, assume that $\|\xi\|_H \leq M < \infty$ almost surely. For any given independent distributed samples $\{z_i\}_{i=1}^m \subset \mathcal{Z}$ and any $\delta \in (0, 1)$, there holds
\[
\left\| \frac{1}{m} \sum_{i=1}^m \xi(z_i) - E\xi \right\|_H \leq \frac{2M \log(2/\delta)}{m} + \sqrt{\frac{2E\xi\|H\|^2 \log(2/\delta)}{m}}
\]
with confidence at least $1 - \delta/2$.

**Proposition 3.** For any $0 < \delta < 1$, with confidence $1 - \delta$, we have
\[
E(\hat{g}_{\lambda}) - E(g_{\lambda}) \leq \frac{16}{c} \lambda \min_{0, r-1} \left(\frac{\log(2/\delta)}{m} + \sqrt{\frac{\log(2/\delta)}{m}}\right) \cdot \left(\sum_{j=1}^d \| g_{\rho}^{(j)} \|_{L_{\infty}^2(x^{(j)})} + \sum_{j=1}^d \| g_{\rho}^{(j)} \|_{L_{\infty}^2(x^{(j)})}^2\right)^2
\]
and
\[
E_2 \leq c_2 \lambda \min_{0, r-1} \left(\frac{\log(2/\delta)}{m} + \lambda \min_{0, r-1}\right),
\]
where $c = 2 + \| L_{K^{(j)}}^{-1} \|_2^2 + \| L_{K^{(j)}}^{-1} \|_2^2$ and $c_2$ is a positive constant independent of $m, \delta$. 
Proof. From Cauchy-Schwarz inequality, we can see that

\[ E(\hat{g}_\lambda) - E(g_\lambda) \leq \left( \int_Z (2y - \hat{g}_\lambda(x) - g_\lambda(x))^2 \rho(x, y) \right)^{\frac{1}{2}} \]
\[ \cdot \left( \int_Z (\hat{g}_\lambda(x) - g_\lambda(x))^2 \rho(x, y) \right)^{\frac{1}{2}} \]
\[ \leq \left( 8 + 2 \int_Z (\hat{g}_\lambda(x) - g_\lambda(x))^2 \rho(x, y) \right)^{\frac{1}{2}} \]
\[ \cdot \left( \int_Z (\hat{g}_\lambda(x) - g_\lambda(x))^2 \rho(x, y) \right)^{\frac{1}{2}} \]
\[ \leq \left( \sqrt{8} + \sqrt{2} \sum_{j=1}^d \|g_{\lambda}^{(j)} - g_{\lambda}^{(j)}\|_{L^2_{\alpha^{(j)}}} \right) \]
\[ \cdot \sum_{j=1}^d \|\hat{g}_{\lambda}^{(j)} - g_{\lambda}^{(j)}\|_{L^2_{\alpha^{(j)}}}. \quad (10) \]

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

\[ \|\xi^{(j)}\|_{L^2_{X^{(j)}}} \leq \|f^{(j)}_\lambda\|_{L^2_{X^{(j)}}} \leq \sqrt{c_\lambda}\min\{0, r - \frac{1}{2}\} \|g_{\rho}^{(j)}\|_{L^2_{\rho^{(j)}}} \]

and

\[ E[\|\xi^{(j)}\|_{L^2_{X^{(j)}}}^2] \leq \|f^{(j)}_\lambda\|_{L^2_{X^{(j)}}}^2 \leq c_\lambda\min\{0, 2r - 1\} \|g_{\rho}^{(j)}\|_{L^2_{\rho^{(j)}}}^2 \]

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

\[ \|\hat{g}_{\lambda}^{(j)} - g_{\lambda}^{(j)}\|_{L^2_{\alpha^{(j)}}} \]
\[ = \left\| \frac{1}{m} \sum_{i=1}^m \xi_i^{(j)} - E\xi^{(j)} \right\|_{L^2_{X^{(j)}}} \]
\[ \leq 2\sqrt{c_\lambda}\min\{0, r - \frac{1}{2}\} \|g_{\rho}^{(j)}\|_{L^2_{\rho^{(j)}}} \log(2/\delta) \]
\[ \leq m \lambda\min\{0, r - \frac{1}{2}\} \|g_{\rho}^{(j)}\|_{L^2_{\rho^{(j)}}} \sqrt{\frac{2c\log(2/\delta)}{m}}, \quad (11) \]

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

\[ \lambda\|\hat{g}_{\lambda}\|_{\ell_1} \leq \lambda \sum_{j=1}^d \|f^{(j)}_\lambda\|_{\ell_1} \leq \sum_{j=1}^d \|L_{K^{(j)}} f^{(j)}_\lambda - f^{(j)}_\lambda\|_{L^2_{\alpha^{(j)}}} \]
\[ \leq \sqrt{c_\lambda}\min\{0, r - \frac{1}{2}\} \sum_{j=1}^d \|g^{(j)}\|_{L^2_{\alpha^{(j)}}}. \]

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 \( \ell_2 \)-metric is denoted by \( N_2(F, \varepsilon) \) and its detail definition can be founded in (Van der Vaart & Wellner, 1996; Shi et al., 2011).

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

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

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

\[ \mathcal{N}_{2,u}(F, \varepsilon) = \inf \left\{ l \in \mathbb{N} : \exists \{f_i\}_{i=1}^l \text{ such that } F \subset \bigcup_{i=1}^l \{f \in F : d_{2,u}(f, f_i) \leq \varepsilon\} \right\}. \]

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

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

\[ |Ef - \frac{1}{m} \sum_{i=1}^m f(z_i)| \]
\[ \leq \frac{1}{2} Ef + c' \max\{e^{\frac{2s}{s+2}}, B\} \left( \frac{a}{m} \right)^{\frac{s}{s+2}} \]
\[ + (2c + 18B) \log(1/\delta) \]

with confidence at least \( 1 - 2\delta \).

For any \( R > 0 \), denote

\[ B_{R}^{(j)} = \left\{ f^{(j)} = \sum_{i=1}^m \alpha_i^{(j)} K^{(j)}(u_i^{(j)}, \cdot) \in \mathcal{H}^{(j)} : \|f^{(j)}\|_{\ell_1} \leq R \right\}, \]

and

\[ B_{R} = \left\{ f = \sum_{j=1}^d f^{(j)} : \|f\|_{\ell_1} \leq R \right\}, \]

where

\[ \|f\|_{\ell_1} = \inf \left\{ \sum_{j=1}^d \|f^{(j)}\|_{\ell_1} : f = \sum_{j=1}^d f^{(j)}, f^{(j)} \in \mathcal{H}^{(j)} \right\}. \]
Now we state the estimate on the empirical covering numbers of $B_1$. Similar analysis can be found in (Christmann & Zhou, 2016) for $B_1$ in reproducing kernel Hilbert spaces.

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

$$\log \mathcal{N}_2(B_1, \varepsilon) \leq d^{1+p}c_p\varepsilon^{-p},$$

where $p$ is defined in Section 3 and $c_p$ is a constant independent of $\varepsilon$.

**Proof.** For every $j \in \{1, 2, \ldots, d\}$ and $x^{(j)} \in (\mathcal{X}^{(j)})^S$, there exists a set $\{f^{(j)}_i\}_{i=1}^{N_j}$ with $N_j = \mathcal{N}_2(B_1, \varepsilon)$ such that

$$\forall f^{(j)} \in B_1^{(j)}, \quad \exists i_j \in \{1, 2, \ldots, N_j\}, \text{s.t.},
\sum_{j=1}^d d_{2,\mathcal{X}^{(j)}}(f^{(j)} - f^{(j)}_{i_j}) \leq \varepsilon.$$

For $f = \sum_{j=1}^d f^{(j)} \in B_1$, we know $f^{(j)} \in B_1^{(j)}$. For every $x = (x^{(j)})_{j=1}^d \in \mathcal{X}^S$, we have $x^{(j)} = (x^{(j)}_i)_{i=1}^S \in (\mathcal{X}^{(j)})^S$, $j \in \{1, 2, \ldots, d\}$. Let $\hat{f} = \sum_{j=1}^d f^{(j)}_{i_j}$. Then

$$d_{2,\mathcal{X}}(f, \hat{f}) = \left\{ \frac{1}{S} \sum_{\ell=1}^S (f(x^{(j)}_\ell) - f(x^{(j)}_\ell))^2 \right\}^{\frac{1}{2}}$$

$$= \left\{ \frac{1}{S} \sum_{\ell=1}^S \left( \sum_{i=1}^d f^{(j)}(x^{(j)}_\ell) - \sum_{j=1}^d f^{(j)}(x^{(j)}_\ell) \right) \right\}^{\frac{1}{2}}$$

$$\leq \sum_{j=1}^d \left\{ \frac{1}{S} \sum_{i=1}^S \left( f^{(j)}(x^{(j)}_\ell) - f^{(j)}_{i_j}(x^{(j)}_\ell) \right) \right\}^{\frac{1}{2}}$$

$$\leq \sum_{j=1}^d d_{2,\mathcal{X}^{(j)}}(f^{(j)} - f^{(j)}_{i_j})$$

$$\leq d\varepsilon.$$

Therefore,

$$\log \mathcal{N}_2(B_1, \varepsilon) \leq \sum_{j=1}^d \log \mathcal{N}_2(B_1^{(j)}, d\varepsilon).$$

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

$$\log \mathcal{N}_2(B_1, \varepsilon) \leq dc_p\varepsilon^{-p}.$$

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

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

$$E_1 \leq \frac{1}{2}(\mathcal{E}(\pi(f_\delta)) - \mathcal{E}(f_\rho)) + \frac{1}{2}(\mathcal{E}(\delta_\lambda) - \mathcal{E}(\lambda_\mu))$$

$$+ E_3 + C_1 \log(2/\delta)(\lambda^{-\frac{2p}{2p+3}}m^{-\frac{2}{2p+3}})$$

$$+ \lambda^{\min\{-1, 2r-2\}}m^{-\frac{2}{2p+3} + m^{-1}}$$

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

**Proof.** The sample error $E_1$ can be decomposed as

$$E_1 = \mathcal{E}(\pi(f_\delta)) - \mathcal{E}(f_\rho) - (\mathcal{E}(\delta_\lambda) - \mathcal{E}(\lambda_\mu))$$

and

$$E_{12} = \mathcal{E}(\delta_\lambda) - \mathcal{E}(f_\rho) - (\mathcal{E}(\delta_\lambda) - \mathcal{E}(f_\rho)).$$

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

Denote

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

For any $g \in \mathcal{G}_R$, we can deduce that $|g(z)| \leq 8$ and $E_2 \leq 16E_2$. Let $g_1, g_2 \in \mathcal{G}_R$ associated with $f_1, f_2$ respectively. It can be seen that

$$|g_1(z) - g_2(z)| \leq 4|\pi(f_1)(x) - \pi(f_2)(x)|$$

$$\leq 4|f_1(x) - f_2(x)|.$$

This means

$$\log \mathcal{N}_2(\mathcal{G}_R, \varepsilon) \leq \log \mathcal{N}_2(B_R, \varepsilon) \leq \log \mathcal{N}_2(B_1, \frac{\varepsilon}{4R})$$

$$\leq c_p d^{1+p}(4R)^p \varepsilon^{-p},$$

where the last inequality follows from Lemma 6.

Applying Lemma 5 to $\mathcal{G}_R$, we have with confidence $1 - \frac{4}{2}$

$$E_1 = \frac{1}{m} \sum_{i=1}^m g(z_i) \leq \frac{1}{2} \left( \mathcal{E}(\pi(f)) - \mathcal{E}(f_\rho) \right)$$

$$+ \mathcal{E}(\delta_\lambda)R^{-\frac{2p}{2p+3}}m^{-\frac{2}{2p+3} + m^{-1}} \log(2/\delta), \forall g \in \mathcal{G}_R,$$

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

From the definition of $f_\delta$ in Section 2, we know $f_\delta \in B_R$ with $R = \lambda^{-1}$. Then

$$E_{11} \leq \frac{1}{2}(\mathcal{E}(\pi(f_\delta)) - \mathcal{E}(f_\rho))$$

$$+ \tilde{c}_1 \left( \lambda^{-\frac{2p}{2p+3}}m^{-\frac{2}{2p+3} + m^{-1}} \log(1/\delta) \right)$$

with confidence $1 - \frac{4}{2}$. □
Now we turn to bound $E_{12}$. Denote
\[
\hat{G} = \left\{ \hat{g} = \sum_{j=1}^{d} \hat{g}_\lambda^{(j)} : \hat{g}_\lambda^{(j)} = \frac{1}{m} \sum_{i=1}^{m} f_i^{(j)}(\phi_i^{(j)}) K(\phi_i^{(j)}, \cdot) \right\}
\]
and
\[
\hat{H} = \left\{ h : h(z) = (y - \hat{g}(x))^2 - (y - f_\rho(x))^2, \hat{g} \in \hat{G} \right\}.
\]
We can verify that
\[
\|h\|_\infty = \sup |2y - \hat{g}(x) - f_\rho(x)| |\hat{g}(x) - f_\rho(x)|
\leq (3 + \sum_{j=1}^{d} \|f_j^{(j)}\|_\infty)^2
\]
and
\[
E h^2 \leq (3 + \sum_{j=1}^{d} \|f_j^{(j)}\|_\infty)^2 E h.
\]

For any given $\hat{g}_1, \hat{g}_2 \in \hat{H}$, the corresponding $h_1, h_2 \in \hat{H}$ satisfy
\[
|h_1(z) - h_2(z)| \leq 2(1 + \sum_{j=1}^{d} \|f_j^{(j)}\|_\infty) |\hat{g}_1(x) - \hat{g}_2(x)|.
\]
Then, from Lemma 6 and $\hat{g} \in B_R$ with $R = \sum_{j=1}^{d} \|f_j^{(j)}\|_\infty$, we have
\[
\log N_2(\hat{H}, \varepsilon) \leq \log N_2 \left( \hat{g}, \frac{\varepsilon}{2(1 + \sum_{j=1}^{d} \|f_j^{(j)}\|_\infty)} \right)
\leq \log N_2 \left( B_1, \varepsilon \right)
\leq C_\rho d^{1+2p} \left( \sum_{j=1}^{d} \|f_j^{(j)}\|_\infty + \sum_{j=1}^{d} \|f_j^{(j)}\|_\infty^2 \right) \varepsilon \beta^{-p}.
\]

Applying Lemma 5 to $\hat{H}$, with confidence $1 - \delta$ we have
\[
E_{12} = \sum_{i=1}^{m} h(z_i) - E h \leq \frac{1}{2} (\varepsilon \hat{g}_\lambda - \varepsilon f_\rho)
+ \varepsilon_2\|f_\lambda\|_\infty^2 (m^{-\frac{2}{2+p}} + m^{-1} \log(2/\delta))
\leq \frac{1}{2} (\varepsilon \hat{g}_\lambda - \varepsilon g_\lambda) + \varepsilon_3
+ d \varepsilon_2 \min \{1, \frac{2}{1+2p} \} \left( m^{-\frac{2}{2+p}} + \log(2/\delta) \right),
\]
where the last inequality follows from Lemma 3 and $\varepsilon_2, \varepsilon_3$ 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 $1 - 4\delta$
\[
\mathcal{E}(\pi(f_\rho)) - \mathcal{E}(f_\rho)
\leq C \log(2/\delta) (\lambda^{\min\{1, 2r\}} + \lambda^{\min\{0, r - \frac{1}{2}\}} m^{-\frac{1}{2}}
+ \lambda^{\min\{-1, 2r - 2\}} m^{-\frac{2}{2+p}} + \lambda^{\frac{2p}{2+p}} m^{-\frac{2}{2+p}}).
\]

When $\rho \in (0, \frac{1}{2})$, by setting $\lambda = m^{-\theta_1}$ with $0 < \theta_1 < \min \{ \frac{1}{p}, \frac{2}{2+p} \}$, we get with confidence $1 - 4\delta$
\[
\mathcal{E}(\pi(f_\rho)) - \mathcal{E}(f_\rho) \leq 4 C \log(2/\delta) m^{-\gamma_1},
\]
where $\gamma_1 = \min \{ 2r \theta_1, \frac{1}{2} + (r - \frac{1}{2}) \theta_1, \frac{2}{2+p} - (2 - 2r) \theta_1, \frac{2}{2+p} - \frac{2p \theta_1}{2+p} \}$.

When $\rho \geq \frac{1}{2}$, taking $\lambda = m^{-\theta_2}$ with some $0 < \theta_2 < \min \{ \frac{1}{p}, \frac{2}{2+p} \}$, we have with confidence $1 - 4\delta$
\[
\mathcal{E}(\pi(f_\rho)) - \mathcal{E}(f_\rho) \leq 4 C \log(2/\delta) m^{-\gamma_2},
\]
where
\[
\gamma_2 = \min \{ \theta_2, \frac{1}{2} + \frac{2}{2+p} - \theta_2, \frac{2}{2+p} - \frac{2p \theta_2}{2+p} \}.
\]
This completes the proof.

### A.5. Proof of Theorem 2

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

#### Proof.

Since $f_\lambda^{(j)} \in H^{(j)}$ for each $j \in \{1, 2, ..., d\}$, we know that $f_\rho \in \mathcal{H}$. Then,
\[
\mathcal{E}(\pi(f_\rho)) - \mathcal{E}(f_\rho)
\leq \mathcal{E}(\pi(f_\lambda) - \mathcal{E}(f_\rho))
+ \{ \mathcal{E}_\lambda(f_\rho) + \lambda \|f_\rho\|_{\ell_1} - (\mathcal{E}_\lambda(f_\rho) + \lambda \|f_\rho\|_{\ell_1}) \}
\leq \mathcal{E}(\pi(f_\rho)) - \mathcal{E}(f_\rho) - (\mathcal{E}_\lambda(f_\rho)) + \lambda \|f_\rho\|_{\ell_1}
= E_{11} + \lambda \|f_\rho\|_{\ell_1}.
\]

From the estimate of $E_{11}$ in (12), with confidence $1 - \delta$ we have
\[
\mathcal{E}(\pi(f_\rho)) - \mathcal{E}(f_\rho) \leq \tilde{c} \log(1/\delta) (\lambda^{\frac{2p}{2+p}} m^{-\frac{2}{2+p}} + \lambda),
\]
where $\tilde{c}$ is a positive constant independent of $m, \lambda$.

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

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

Recall that \( f_\alpha = \sum_{j=1}^{d} \sum_{t=1}^{m} \alpha_t^{(j)} K^{(j)}(x_t^{(j)}, \cdot) \) 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

\[
\frac{\partial G(\alpha)}{\partial \alpha_t^{(j)}} \bigg|_{\alpha = \hat{\alpha}} = -\frac{2}{m} \sum_{i=1}^{m} (y_i - f_\hat{\alpha}(x_i)) K^{(j)}(x_t^{(j)}, x_i^{(j)}) + \lambda = 0.
\]

This means

\[
\frac{1}{m} \sum_{i=1}^{m} (y_i - f_\hat{\alpha}(x_i)) K^{(j)}(x_t^{(j)}, x_i^{(j)}) = \frac{\lambda}{2}.
\]

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

\[
\frac{1}{m} \sum_{i=1}^{m} (y_i - f_\hat{\alpha}(x_i)) K^{(j)}(x_t^{(j)}, x_i^{(j)}) = -\frac{\lambda}{2}.
\]

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

\[
-\frac{2}{m} \sum_{i=1}^{m} (y_i - f_\hat{\alpha}(x_i)) K^{(j)}(x_t^{(j)}, x_i^{(j)}) - \lambda \leq \frac{\partial G(\alpha)}{\partial \alpha_t^{(j)}} \bigg|_{\alpha = \hat{\alpha}} \leq -\frac{2}{m} \sum_{i=1}^{m} (y_i - f_\hat{\alpha}(x_i)) K^{(j)}(x_t^{(j)}, x_i^{(j)}) + \lambda.
\]

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

\[
\left| \frac{1}{m} \sum_{i=1}^{m} (y_i - f_\hat{\alpha}(x_i)) K^{(j)}(x_t^{(j)}, x_i^{(j)}) \right| < \frac{\lambda}{2}.
\]

This completes the proof. \(\square\)