Appendix

A. Proof of Theorem 1

This section provides the detailed proof of Theorem 1. The following error decomposition is proved in terms of the definitions of g_λ and $\hat{g}_\lambda$ in Section 3.

Proposition 1 *Denote $D(\lambda) = \inf_{g\in\mathcal{H}}\{\mathcal{E}(g) - \mathcal{E}(f_\rho) + \lambda\|g\|_{\mathcal{H}}^2\}$. For any $\mathbf{z} \in \mathcal{Z}^m$, the excess generalization error of $f_{\mathbf{z}}$ in Section 3 can be decomposed as below:*

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \leq E_1 + E_2 + D(\lambda),$$

where

$$E_1 = \mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}})) + \mathcal{E}_{\mathbf{z}}(\hat{g}_\lambda) - \mathcal{E}(\hat{g}_\lambda)$$

and

$$E_2 = \mathcal{E}(\hat{g}_\lambda) + \lambda m\|\hat{g}_\lambda\|_{\ell_2}^2 - \mathcal{E}(g_\lambda) - \lambda\|g_\lambda\|_{\mathcal{H}}^2.$$

Proof: It is easy to deduce that

$$\begin{aligned}
& \mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \\
\leq\ & \mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}})) + \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \\
\leq\ & \mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}})) + [\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda m\|f_{\mathbf{z}}\|_{\ell_2}^2 - (\mathcal{E}_{\mathbf{z}}(\hat{g}_\lambda) + \lambda m\|\hat{g}_\lambda\|_{\ell_2}^2)] \\
& + \mathcal{E}_{\mathbf{z}}(\hat{g}_\lambda) - \mathcal{E}(\hat{g}_\lambda) + \mathcal{E}(\hat{g}_\lambda) + \lambda m\|\hat{g}_\lambda\|_{\ell_2}^2 - \mathcal{E}(f_\rho) \\
\leq\ & E_1 + \mathcal{E}(\hat{g}_\lambda) + \lambda m\|\hat{g}_\lambda\|_{\ell_2}^2 - \mathcal{E}(f_\rho),
\end{aligned}$$

where the last inequality follows from the definitions of $f_{\mathbf{z}}, \hat{g}_\lambda$.

According to the definition g_λ, we can see that

$$\mathcal{E}(\hat{g}_\lambda) + \lambda m\|\hat{g}_\lambda\|_{\ell_2}^2 - \mathcal{E}(f_\rho) = E_2 + \mathcal{E}(g_\lambda) - \mathcal{E}(f_\rho) + \lambda\|g_\lambda\|_{\mathcal{H}}^2.$$

Combining the above two estimates, we get the desired result. □

The error term E_1 depends on the full samples and the subsampling set, which usually is called as the sample error in learning theory [1]. The error term E_2 is induced by the vary from the data dependent hypothesis space $\mathcal{H}_m$ to data independent hypothesis space $\mathcal{H}$, which is called the hypothesis error [2, 4]. The last term is the approximation error, which reflects the approximation ability of learning model to the regression function.

Before error estimates, we provide the upper bounds of $f_{\mathbf{z}}$ and f_λ.

Lemma 1 *For any $\mathbf{z} \in \mathcal{Z}^m$, there holds*

$$\|f_{\mathbf{z}}\|_{\ell_1} \leq \frac{1}{\sqrt{\lambda}} \ \ and \ \ \|f_\lambda\|_\infty \leq \frac{\sqrt{D(\lambda)}}{\lambda}.$$

Proof: From the definition of $f_{\mathbf{z}}$, we can see that $\lambda m\|f_{\mathbf{z}}\|_{\ell_2}^2 \leq \mathcal{E}_{\mathbf{z}}(0) \leq 1$, which means

$$\|f_{\mathbf{z}}\|_{\ell_1} \leq \sqrt{m}\|f_{\mathbf{z}}\|_{\ell_2} \leq \frac{1}{\sqrt{\lambda}}.$$

To estimate f_λ, we introduce an auxiliary function $\phi : \mathbb{R} \to \mathbb{R}$ defined as

$$\phi(\theta) = \int_{\mathcal{X}} (L_K(f_\lambda + \theta h) - f_\rho)^2 d\rho_{\mathcal{X}} + \lambda \int_{\mathcal{X}} (f_\lambda + \theta h)^2 d\rho_{\mathcal{X}},$$

where h is any fixed function in $L_{\rho_{\mathcal{X}}}^2$.

Note that $\phi(0) = D(\lambda) + \mathcal{E}(f_\rho)$. It means that $\frac{d\phi(\theta)}{d\theta}|_{\theta=0} = 0$. Then

$$\phi'(0) = 2\int_{\mathcal{X}} (L_K f_\lambda(x) - f_\rho(x))[\int_{\mathcal{X}} h(t)K(x,t)d\rho_{\mathcal{X}}(t)]d\rho_{\mathcal{X}}(x) + \lambda \int_{\mathcal{X}} f_\lambda(x)h(x)d\rho_{\mathcal{X}} = 0.$$

According to the arbitrariness of h, we get

$$\lambda f_\lambda(t) = -\int_{\mathcal{X}} (L_K f_\lambda(x) - f_\rho(x))K(x,t)d\rho_{\mathcal{X}}(x), \forall t \in \mathcal{X}.$$

Then, for any $t \in \mathcal{X}$,

$$\begin{aligned} |f_\lambda(t)| &= \lambda^{-1}|\int_{\mathcal{X}} (L_K f_\lambda(x) - f_\rho(x))K(x,t)d\rho_{\mathcal{X}}(x)| \\ &\leq \lambda^{-1}\sqrt{\int_{\mathcal{X}} (L_K f_\lambda(x) - f_\rho(x))^2 d\rho_{\mathcal{X}}(x)}\sqrt{\int_{\mathcal{X}} K^2(x,t)d\rho_{\mathcal{X}}(x)} \\ &\leq \frac{\sqrt{D(\lambda)}}{\lambda}. \end{aligned}$$

This completes the proof of Lemma 1. □

A.1. Hypothesis error estimate

The hypothesis error E_2 involves the random variables with values in Hilbert space. To bound this term, we introduce the following concentration inequality in [3].

Lemma 2 *Let $\mathcal{H}$ be a Hilbert space and ξ be independent random variable on $\mathcal{Z}$ with values in $\mathcal{H}$. Assume that $\|\xi\|_{\mathcal{H}} \leq \tilde{M} < \infty$ almost surely. Let $\{z_i\}_{i=1}^m$ be independent random samples from ρ. Then, for any $\delta \in (0,1)$,*

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

holds true with confidence $1-\delta$.

Proposition 2 *For any $0 < \delta < 1$, with confidence $1-\delta$, there holds*

$$\mathcal{E}(\hat{g}_\lambda) - \mathcal{E}(g_\lambda) \leq D(\lambda)\Big(1 + \frac{4\log(1/\delta)}{m\lambda} + \frac{8\log^2(1/\delta)}{m^2\lambda^2}\Big)$$

and

$$E_2 \leq D(\lambda)\Big(1 + \frac{6\log(2/\delta)}{m\lambda} + \frac{8\log^2(2/\delta)}{m^2\lambda^2} + \sqrt{\frac{2\log(2/\delta)}{m\lambda}}\Big).$$

Proof: From the definitions of g_λ and $\hat{g}_\lambda$, we have

$$m\lambda\|\hat{g}_\lambda\|_{\ell_2}^2 - \lambda\|g_\lambda\|_{\mathcal{H}}^2 = \lambda\Big(\frac{1}{m}\sum_{i=1}^m (f_\lambda(\bar{x}_i))^2 - \int_{\mathcal{X}} (f_\lambda(x))^2 d\rho_{\mathcal{X}}(x)\Big).$$

Let $\xi_1(x) = (f_\lambda(x))^2$. From Lemma 1, we get $|\xi_1(x)| \leq D(\lambda)/\lambda^2$ for any $x \in \mathcal{X}$ and

$$E\xi_1^2 \leq \frac{D(\lambda)}{\lambda^2}\int_{\mathcal{X}} (f_\lambda(x))^2 d\rho_{\mathcal{X}}(x) \leq \frac{(D(\lambda))^2}{\lambda^3}.$$

Applying Lemma 2 to random variable ξ_1, we get for any $\delta \in (0,1)$

$$m\lambda\|\hat{g}_\lambda\|_{\ell_2}^2 - \lambda\|g_\lambda\|_{\mathcal{H}}^2 \leq \frac{2D(\lambda)\log(1/\delta)}{m\lambda} + D(\lambda)\sqrt{\frac{2\log(1/\delta)}{m\lambda}} \tag{1}$$

with confidence $1-\delta$.

Now we turn to estimate $\mathcal{E}(\hat{g}_\lambda)-\mathcal{E}(g_\lambda)$. Observe that

$$\begin{aligned}\mathcal{E}(\hat{g}_\lambda)-\mathcal{E}(g_\lambda) &= \|\hat{g}_\lambda-f_\rho\|^2_{L^2_{\rho_\mathcal{X}}}-\|g_\lambda-f_\rho\|^2_{L^2_{\rho_\mathcal{X}}}\\ &\le \|\hat{g}_\lambda-g_\lambda\|_{L^2_{\rho_\mathcal{X}}}\cdot\|\hat{g}_\lambda+g_\lambda-2f_\rho\|_{L^2_{\rho_\mathcal{X}}}\\ &\le \|\hat{g}_\lambda-g_\lambda\|_{L^2_{\rho_\mathcal{X}}}\cdot(\|\hat{g}_\lambda-g_\lambda\|_{L^2_{\rho_\mathcal{X}}}+2\|g_\lambda-f_\rho\|_{L^2_{\rho_\mathcal{X}}})\\ &\le 2\|\hat{g}_\lambda-g_\lambda\|^2_{L^2_{\rho_\mathcal{X}}}+D(\lambda), \qquad (2)\end{aligned}$$

where the last inequality follows from

$$2\|g_\lambda-f_\rho\|_{L^2_{\rho_\mathcal{X}}}\cdot\|\hat{g}_\lambda-g_\lambda\|_{L^2_{\rho_\mathcal{X}}}\le\|g_\lambda-f_\rho\|^2_{L^2_{\rho_\mathcal{X}}}+\|\hat{g}_\lambda-g_\lambda\|^2_{L^2_{\rho_\mathcal{X}}}.$$

Denote $\xi_2(x)=f_\lambda(x)K(x,t)$. Then, $\|\xi_2\|_\infty\le\|f_\lambda\|_\infty\le\sqrt{D(\lambda)}/\lambda$ and $E\|\xi_2\|^2\le\|f_\lambda\|^2_{L^2_{\rho_\mathcal{X}}}\le D(\lambda)/\lambda$. Applying Lemma 2 to ξ_2, we get with confidence $1-\delta$,

$$\|\hat{g}_\lambda-g_\lambda\|_{L^2_{\rho_\mathcal{X}}}=\Big\|\frac{1}{m}\sum_{i=1}^m\xi_2(\bar{x}_i)-E\xi_2\Big\|_{L^2_{\rho_\mathcal{X}}}\le\frac{2\sqrt{D(\lambda)}}{m\lambda}\log(1/\delta)+\sqrt{\frac{2D(\lambda)\log(1/\delta)}{m\lambda}}.$$

Then, from (2), there holds

$$\mathcal{E}(\hat{g}_\lambda)-\mathcal{E}(g_\lambda)\le D(\lambda)+\frac{4D(\lambda)\log(1/\delta)}{m\lambda}+\frac{8D(\lambda)\log^2(1/\delta)}{m^2\lambda^2}. \qquad (3)$$

Combining the estimates (1) and (3), we get the desired upper bound of E_2. □

A.2. Sample error estimate

Given any $R>0$, define a class of functions as

$$\mathcal{B}_R=\Big\{f=\sum_{i=1}^m\alpha_iK(u_i,\cdot):\sum_{i=1}^m|\alpha_i|\le R,\{u_i\}_{i=1}^m\in\mathcal{X}^m\Big\}. \qquad (4)$$

Recently, some tight estimates have been established in [4, 5] to bound the empirical covering number of $\mathcal{B}_1$. Now recall some basic definitions for covering numbers.

Definition 1 *Let $(\mathcal{U},d)$ be a pseudo-metric space and denote a subset $S\subset\mathcal{U}$. For every $\epsilon>0$, the covering number $\mathcal{N}(S,\epsilon,d)$ of S with respect to ϵ,d is defined as the minimal number of balls of radius ϵ whose union covers S, that is,*

$$\mathcal{N}(S,\epsilon,d)=\min\Big\{l\in\mathbb{N}:S\subset\bigcup_{j=1}^l B(s_j,\epsilon)\textit{ for some }\{s_j\}_{j=1}^l\subset\mathcal{U}\Big\},$$

where $B(s_j,\epsilon)=\{s\in\mathcal{U}:d(s,s_j)\le\epsilon\}$ is a ball in $\mathcal{U}$.

The empirical covering number with ℓ_2 metric is defined as below.

Definition 2 *Let $\mathcal{F}$ be a set of functions on $\mathcal{X}$, $\mathbf{u}=(x_i)_{i=1}^k$ and $\mathcal{F}|_\mathbf{u}=\{(f(u_i))_{i=1}^k:f\in\mathcal{F}\}\subset\mathbb{R}^k$. Set $\mathcal{N}_{2,\mathbf{u}}(\mathcal{F},\epsilon)=\mathcal{N}(\mathcal{F}|_\mathbf{u},\epsilon,d_2)$. The ℓ_2 empirical covering number of $\mathcal{F}$ is defined by*

$$\mathcal{N}_2(\mathcal{F},\epsilon)=\sup_{k\in\mathbb{N}}\sup_{\mathbf{u}\in\mathcal{X}^k}\mathcal{N}_{2,\mathbf{u}}(\mathcal{F},\epsilon),\epsilon>0,$$

where ℓ_2 metric

$$d_2(\mathbf{a},\mathbf{b})=\Big(\frac{1}{k}\sum_{i=1}^k|a_i-b_i|^2\Big)^{\frac{1}{2}},\forall\mathbf{a}=(a_i)_{i=1}^k\in\mathbb{R}^k,\mathbf{b}=(b_i)_{i=1}^k\in\mathbb{R}^k.$$

The following concentration inequality, proved in [6], is used for our sample error analysis.

Lemma 3 *Assume that there are constants $B, c > 0$ and $\alpha \in [0,1]$ such that $\|f\|_\infty \leq B$ and $Ef^2 \leq c(Ef)^\alpha$ for every $f \in \mathcal{F}$. If for some $a > 0$ and $p \in (0,2)$,*

$$\log(\mathcal{N}_2(\mathcal{F}, \epsilon)) \leq a\epsilon^{-p}, \forall \epsilon > 0,$$

then there exists a constant c'_p depending only on p such that for any $t > 0$, with probability at least $1 - e^{-t}$, there holds

$$Ef - \frac{1}{n}\sum_{i=1}^n f(z_i) \leq \frac{1}{2}\eta^{1-\alpha}(Ef)^\alpha + c'_p\eta + 2\Big(\frac{ct}{n}\Big)^{\frac{1}{2-\alpha}} + \frac{18Bt}{n}, \forall f \in \mathcal{F},$$

where

$$\eta := \max\Big\{c^{\frac{2-p}{4-2\alpha+p\alpha}}\Big(\frac{a}{n}\Big)^{\frac{2}{4-2\alpha+p\alpha}}, B^{\frac{2-p}{2+p}}\Big(\frac{a}{n}\Big)^{\frac{2}{2+p}}\Big\}.$$

The sample error E_1 can be further decomposed as

$$E_1 = E_{11} + E_{12},$$

where

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

and

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

Now we provide the estimates for E_{11} and E_{12}.

Lemma 4 *Suppose that $\mathcal{X}$ is compact subset of $\mathbb{R}^d$ and $K \in C^s(\mathcal{X} \times \mathcal{X})$ for some $s > 0$. For any $\delta \in (0,1)$, with confidence $1 - \delta$, there holds*

$$E_{11} \leq \frac{1}{2}(\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho)) + c_1(\lambda^{\frac{p}{2+p}}n^{-\frac{2}{2+p}} + n^{-1}\log(1/\delta)),$$

where c_1 is a positive constant independent of n, λ, δ, and p is defined in Theorem 1.

Proof: From Lemma 1 and the definition of B_R in (4), we can see that $f_{\mathbf{z}} \in \mathcal{B}_R$ with $R = \lambda^{-\frac{1}{2}}$. Denote

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

It follows that for any $z \in \mathcal{Z}, g \in \mathcal{G}_R$

$$|g(z)| = |2y - \pi(f)(x) - f_\rho(x)| \cdot |\pi(f)(x) - f_\rho(x)| \leq 8 \tag{5}$$

and

$$Eg^2 = E|2y - \pi(f)(x) - f_\rho(x)|^2 \cdot |\pi(f)(x) - f_\rho(x)|^2 \leq 16Eg. \tag{6}$$

For any $f_1, f_2 \in \mathcal{B}_R$ and $z = (x,y) \in \mathcal{Z}$, denote $g_1(z) = (y - \pi(f_1)(x))^2 - (y - f_\rho(x))^2$ and $g_2(z) = (y - \pi(f_2)(x))^2 - (y - f_\rho(x))^2$. Then,

$$\begin{aligned} |g_1(z) - g_2(z)| &= |2y - \pi(f)(x) - f_\rho(x)| \cdot |\pi(f)(x) - f_\rho(x)| \\ &\leq 4|\pi(f_1)(x) - \pi(f_2)(x)| \\ &\leq 4|f_1(x) - f_2(x)|. \end{aligned}$$

It means that

$$\log \mathcal{N}_2(\mathcal{G}_R, \varepsilon) \leq \log \mathcal{N}_2(\mathcal{B}_R, \frac{\varepsilon}{4}) \leq \log \mathcal{N}_2(\mathcal{B}_1, \frac{\varepsilon}{4R}) \leq c_p(4R)^p\varepsilon^{-p}, \tag{7}$$

where the last inequality follows from Theorem 2 in [4] and Theorem 3 in [5].

The inequalities (5)-(7) tells us that Lemma 3 holds true for any $g \in \mathcal{G}_R$ with $a = c_p(4R)^p$, $c = 16$, $\alpha = 1$ and $B = 8$. Then, for any $g \in \mathcal{G}_R$, with confidence $1 - \delta$

$$Eg - \frac{1}{n}\sum_{i=1}^n g(z_i) \leq \frac{1}{2}(\mathcal{E}(\pi(f)) - \mathcal{E}(f_\rho)) + \tilde{c_1}((\lambda^{-\frac{p}{2+p}}n^{-\frac{2}{2+p}} + \log(1/\delta)n^{-1}),$$

where $\tilde{c_1}$ is a positive constant independent of n, δ, λ. This competes the proof. $\square$

Lemma 5 *Suppose that $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$ and $K \in C^s(\mathcal{X} \times \mathcal{X})$ for some $s > 0$. For any $\delta \in (0,1)$, with confidence $1-\delta$, there holds*

$$E_{12} \leq \frac{1}{2}(\mathcal{E}(\hat{g}_\lambda) - \mathcal{E}(g_\lambda)) + \frac{1}{2}D(\lambda) + c_2 D(\lambda)(\lambda^{-2}n^{-\frac{2}{2+p}} + \lambda^{-2}n^{-1}\log(1/\delta)),$$

where c_2 is a positive constant independent of λ, δ, n.

Proof: Denote

$$\mathcal{G} = \{\hat{g} : \hat{g} := \hat{g}_{\mathbf{v}}(\cdot) = \frac{1}{m}\sum_{i=1}^{m} f_\lambda(v_i)K(v_i,\cdot), \mathbf{v} = (v_i)_{i=1}^m \in \mathcal{X}^m\}$$

and

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

It is follows that $\hat{g}_\lambda \in \mathcal{G}$ by the definition of $\hat{g}_\lambda$ and using $v_i = \bar{x}_i, i \in \{1, ..., m\}$. For any $g \in \mathcal{G}$

$$\|g\|_\infty \leq \|f_\lambda\|_\infty \leq R := \frac{\sqrt{D(\lambda)}}{\lambda},$$

where the last inequality follows from Lemma 1. Moreover, for any $h \in \mathcal{H}$

$$\|h\|_\infty = \sup_{(x,y)} |2y - \hat{g}(x) - f_\rho(x)| \cdot |\hat{g}(x) - f_\rho(x)| \leq (3+R)^2 \tag{8}$$

and

$$Eh^2 \leq (3+R)^2 E(\hat{g}(x) - f_\rho(x))^2 \leq (3+R)^2 Eh.$$

For any $\hat{g}_1, \hat{g}_2 \in \mathcal{G}$, we have

$$|h_1(z) - h_2(z)| \leq 2(1+R)|\hat{g}_1(x) - \hat{g}_2(x)|.$$

This means

$$\log \mathcal{N}_2(\mathcal{H}, \varepsilon) \leq \log \mathcal{N}_2\Big(\mathcal{G}, \frac{\varepsilon}{2+2R}\Big) \leq \log \mathcal{N}_2\Big(\mathcal{B}_1, \frac{\varepsilon}{2R(1+R)}\Big) \leq c_s 2^p (R^2+R)^p \varepsilon^{-p}, \tag{9}$$

where the covering number bounds in [4, 5] are used for the last inequality.

The estimates (8)-(9) verifies the conditions of Lemma 3. Then, for any $0 < \delta < 1$, we get with confidence at least $1-\delta$,

$$\begin{aligned} E_{12} &\leq \frac{1}{2}\Big(\mathcal{E}(\hat{g}_\lambda) - \mathcal{E}(g_\lambda)\Big) + \tilde{c}\Big(R^2 n^{-\frac{2}{2+p}} + (R+3)^2 n^{-1}\log(1/\delta)\Big) \\ &\leq \frac{1}{2}\Big(\mathcal{E}(\hat{g}_\lambda) - \mathcal{E}(g_\lambda)\Big) + 2\tilde{c}\lambda^{-2}D(\lambda)\Big(n^{-\frac{2}{2+p}} + n^{-1}\log(1/\delta)\Big), \end{aligned}$$

where $\tilde{c}$ is a constant independence of n, δ, λ. This completes the proof. □

Combining Lemma 4 and Lemma 5, we obtain the estimate of sample error E_1.

Proposition 3 *Suppose that $\mathcal{X}$ is compact subset of $\mathbb{R}^d$ and $K \in C^s(\mathcal{X} \times \mathcal{X})$ for some $s > 0$. For any $\delta \in (0,1)$, with confidence $1-2\delta$, there holds*

$$\begin{aligned} E_1 &\leq \frac{1}{2}(\mathcal{E}(\pi(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) + \mathcal{E}(\hat{g}_\lambda) - \mathcal{E}(g_\lambda)) + c_1(\lambda^{-\frac{p}{2+p}} n^{-\frac{2}{2+p}} + n^{-1}\log(1/\delta)) \\ &\quad + c_2\lambda^{-2}D(\lambda)(n^{-\frac{2}{2+p}} + n^{-1}\log(1/\delta)). \end{aligned}$$

where c_1, c_2 is a positive constant independent of n, λ, δ, and p is defined in Theorem 1.

A.3. Proof of Theorem 1

The proof of Theorem 1 is provided as below.

Proof of Theorem 1. Combining Propositions 1-3, we have with confidence $1-4\delta$

$$\begin{aligned} \mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) &\leq \frac{1}{2}(\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(\pi(f_\rho))) + c_1(\lambda^{-\frac{p}{2+p}} n^{-\frac{2}{2+p}} + n^{-1}\log(1/\delta)) \\ &\quad + c_2\lambda^{-2}D(\lambda)(n^{-\frac{2}{2+p}} + n^{-1}\log(1/\delta)) \\ &\quad + D(\lambda)\Big(4 + \frac{16\log(2/\delta)}{m\lambda} + \frac{32\log^2(2/\delta)}{m^2\lambda^2} + \sqrt{\frac{\log(2/\delta)}{m\lambda}}\Big). \end{aligned}$$

By a direct computation and setting $\tilde{\delta} = 4\delta$, we obtain the desired result of Theorem 1. □

B. Proof of Theorem 2

Proof of Theorem 2. Under the approximation condition $D(\lambda) \leq c_\beta \lambda^\beta$, Theorem 1 yields with confidence $1-\delta$

$$\begin{aligned} & \mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \\ \leq \quad & \tilde{c}\log^2(8/\delta)\Big(\lambda^{-\frac{p}{2+p}} n^{-\frac{2}{2+p}} + \lambda^\beta + \lambda^{\beta-1}m^{-1} + \lambda^{\beta-2}m^{-2} + \lambda^{\beta-2}n^{-\frac{2}{2+p}}\Big). \end{aligned}$$

We have $m^{-2} \geq n^{-\frac{2}{2+p}}$ as $m \leq n^{\frac{1}{2+p}}$. Then, the above estimate implies that with confidence $1-\delta$

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \leq 2\tilde{c}\log^2(8/\delta)\Big(\lambda^{-\frac{p}{2+p}} m^{-2} + \lambda^\beta + \lambda^{\beta-1}m^{-1} + \lambda^{\beta-2}m^{-2}\Big).$$

Setting $\lambda = m^{-\theta}$ for some $\theta > 0$, we get with confidence $1-\delta$

$$\begin{aligned} \mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \quad & \leq \quad 2\tilde{c}\log^2(8/\delta)\Big(m^{-(2-\frac{p\theta}{2+p})} + m^{-\beta\theta} + m^{-(1-\theta+\beta\theta)} + m^{-(2-2\theta+\beta\theta)}\Big) \\ & \leq \quad 8\tilde{c}\log^2(8/\delta)m^{-\gamma}, \end{aligned}$$

where

$$\gamma = \min\Big\{2 - \frac{p\theta}{2+p}, 2+\beta\theta-2\theta, \beta\theta, 1+\beta\theta-\theta\Big\}.$$

This completes the proof of Theorem 2. □

C. Proof of Theorem 3

Proof of Theorem 3. Observe that $\sum_{i=1}^n p_i = 1$. Then

$$\begin{aligned} S(p_1, ..., p_n) \quad & = \quad \sum_{i=1}^n \frac{1-L_{ii}}{p_i}\|K_i\|_2^2 \cdot \sum_{i=1}^n p_i = \sum_{i=1}^n \Big(\frac{\sqrt{1-L_{ii}}}{\sqrt{p_i}}\|K_i\|_2\Big)^2 \cdot \sum_{i=1}^n (\sqrt{p_i})^2 \\ & \geq \quad \sum_{i=1}^n \Big(\sqrt{1-L_{ii}}\|K_i\|_2\Big)^2, \end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality.

Here the Cauchy-Schwarz inequality holds under equality when $\frac{\sqrt{1-L_{ii}}}{\sqrt{p_i}}\|K_i\|_2 = \sqrt{p_i}, i = 1, ..., n$. Thus, we can set $p_i = k^{-1}\|K_i\|_2\sqrt{1-L_{ii}}$ such that $\sum_{i=1}^n p_i = 1$. This yields the desired result in Theorem 3. □

References

[1] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc.*, 39: 1–49, 2002.

[2] Y. Feng, S. Lv, H. Huang, and J. Suykens. Kernelized elastic net reguularization: generalization bouunds and sparse recovery. *Neural Computat.*, 28: 1–38, 2016.

[3] I. Pinelis. Optimum bounds for the distribution of martingales in Banach spaces. *Ann. Probab.*, 22: 1679–1706, 1994.

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

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

[6] Q. Wu, Y. Ying, and D.X. Zhou. Multi-kernel regularized classfiers. *J. Complexity*, 23: 108–134, 2007.