
Fast Convergence on Perfect Classification for Functional Data

Supplementary Material

This supplementary material contains full proofs of the statements in the main text.

S1 Proof of Lemma 1

Proof of Lemma 1. f_0 minimizes $R(f)$, if $\mathrm{sign}(f_0(x)) = \mathrm{sign}(\Pr(Y=1|x) - \Pr(Y=-1|x))$ is satisfied. The Radon-Nikodym derivative for $\Pr(Y=1|X)\Pi(X) = \Pr(X|Y=1)\Pr(Y=1) = P_+(X)(1-w)$ in terms of Π implies $\Pr(Y=1|x) = (1-w)p_+(x)$. Similarly, we have $\Pr(Y=-1|x) = wp_-(x)$. Hence, f_0 has the desired property. □

S2 Note on Assumption 1

We firstly provide additional example on Assumption 1.

Example 1 (Monotone/Convex Path)**.** Assume $\mathcal{X}$ is a set of component-wise monotonic functions from $[0,1]^p$ to $[0,1]$ with $p \geq 2$. With $\gamma = 2(p-1)$, Assumption 1 follows from Theorem 1.1 in Gao and Wellner (2007). Alternatively, let $\mathcal{X}$ be a set of convex functions on $[0,1]^p$ that are uniformly bounded. From

Theorem 3.1 in Guntuboyina and Sen (2012) with setting $\gamma = p/2$, Assumption 1 holds.

Example 2 (Gaussian Process)**.** Let X be a Gaussian process on $[0,1]^p$ with a positive even p, and assume its covariance function $\mathrm{Cov}(t,t'), t,t' \in [-,1]^d$ is $\mathrm{Cov}(t,t')k_\alpha(\|t-t'\|_2)$ where k_α is Matérn kernel function ((4.14) in Williams and Rasmussen (2006)). In this case, with probability 1, a path of X is in a RKHS whose kernel is $k_{\alpha-p/2}$. Then, if $\mathcal{X}$ is a unit-ball of the RKHS in terms of an RKHS norm, we obtain that Assumption 1 holds with $\gamma = p/(\alpha - p/2)$. For details, see Corollary 4.15 in Kanagawa *et al.* (2018).

We also present the following result to show the validity of Example 4 on unbounded functions.

Proposition 1. *Let $\mathcal{F}$ be the set of functions with the form as in Example 4 with fixed $J \in \mathbb{N}$ and locations $t_1, ..., t_J \in [0,1]$. Then, there exists a constant C^* such that the following inequality holds for any $\varepsilon \in (0, \bar{\varepsilon})$ with existing $\bar{\varepsilon}$:*

$$\log \mathcal{N}(\varepsilon, \mathcal{W}^\alpha, \|\cdot\|_{L^2}) \leq V' \varepsilon^{-1/\alpha},$$

Proof of Proposition 1. Let $\mathcal{W}^\alpha$ be a unit-ball in the Sobolev space on $[0,1]$ with an order $\alpha \in \mathbb{N}$. By applying Theorem 4.3.26 in Giné and Nickl (2016), there exists an constant V' such that the following inequality

$$\log \mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_{L^2}) \leq C^* \varepsilon^{-1/\alpha},$$

for every $\varepsilon > 0$. Hence, we set $M_1 = M_1(\varepsilon) = \log \mathcal{N}(\varepsilon, \mathcal{W}^\alpha, \|\cdot\|_{L^2})$ and take a subset $\{g_m\}_{m=1}^{M_1} \subset \mathcal{W}^\alpha$ as centers of the ε-balls to cover $\mathcal{W}^\alpha$, that is, for any $g \in \mathcal{W}^\alpha$, there exists $g' \in \{g_m\}_{m=1}^{M_1}$ such that $\|g - g'\|_{L^2} \leq \varepsilon$.

We also consider a set of location parameters $a_j \in [0,1]$ and a covering number of a parameter space for the locations. Let $\mathcal{I} = [0,1]^J$ be the space for $A = (a_1, ..., a_J) \in \mathcal{I}$. We know that there exists a constant $C > 0$ such that

$$\mathcal{N}(\varepsilon, \mathcal{I}, \|\cdot\|) \leq \mathcal{N}(\varepsilon, [0,1], \|\cdot\|)^J \leq (C\varepsilon)^{-J}.$$

Then, let $\{A_m\}_{m=1}^{M_2}$ be subsets of size $M_2 = M_2(\varepsilon)$ such that there are the centers of the ε-balls to cover $\mathcal{I}$.

Fix a function f which has the form in Example 4 as

$$f(x; g, A) = g(x) + \sum_{j=1}^{J} \psi(x; a_j, t_j). \tag{S2.1}$$

Note that the locations $t_1, ..., t_J \in [0,1]$ are fixed. By the definition of the subsets, we can find g_m and $A_{m'}$ from the subsets such that $\|g - g'\|_{L^2} \leq \varepsilon$, and $\|(a_1, ..., a_J)^\top - A_{m'}\| \leq \varepsilon$ for each ε. Then, we define

$$\widehat{f}(x) := g_m(x) + \sum_{j=1}^{J} \psi(x; a_{m',j}, t_j),$$

where we write $A_{m'} = (a_{m',1}, ..., a_{m',J})^\top$. We can bound the following difference as

$$\|f - \widehat{f}\|_{L^2} \leq \|g - g_m\|_{L^2} + \left\| \sum_{j=1}^{J} \psi(\cdot; a_j, t_j) - \sum_{j=1}^{J} \psi(\cdot; a_{m',j}, t_j) \right\|_{L^2}$$

$$\leq \varepsilon + \sum_{j=1}^{J} \left\| \psi(\cdot; a_j, t_j) - \psi(\cdot; a_{m',j}, t_j) \right\|_{L^2}.$$

About the norm in the last term, we can bound it as

$$\begin{aligned}
&\|\psi(\cdot; a_j, t_j) - \psi(\cdot; a_{m',j}, t_j)\|_{L^2}^2 \\
&\leq \int_0^1 \left(\frac{a_j}{|x - t_j|^{1/3}} - \frac{a_{m',j}}{|x - t_j|^{1/3}} \right)^2 dx \\
&= (a_j - a_{m',j})^2 \int_0^1 \left(\frac{1}{|x - t_j|^{1/3}} \right)^2 dx \\
&= (a_j - a_{m',j})^2 3((1 - t_j)^{1/3} + t_j^{1/3}) \\
&\leq 6(a_j - a_{m',j})^2.
\end{aligned}$$

Combining the results and the Cauchy-Schwartz inequality, we obtain

$$\|f - \widehat{f}\|_{L^2} \leq \varepsilon + \sqrt{6} \sum_{j=1}^{J} |a_j - a_{m',j}| \leq \varepsilon + \sqrt{6}\sqrt{J} \|A - A_{m'}\| \leq (1 + \sqrt{6J})\varepsilon$$

Hence, we find that the product set of $\{g_m\}_{m=1}^{M_1}$ and $\{A_m\}_{m=1}^{M_2}$ can construct a $(1 + \sqrt{6J})\varepsilon$-covering set of a set of f with the form (S2.1). Then, we bound the covering number of $\mathcal{F}$ as

$$\begin{aligned}
\log \mathcal{N}((1 + \sqrt{6J})\varepsilon, \mathcal{F}, \|\cdot\|_{L^2}) &\leq \log \mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_{L^2}) + \log \mathcal{N}(\varepsilon, \mathcal{I}, \|\cdot\|) \\
&\leq V' \varepsilon^{-1/\alpha} + J \log(C\varepsilon^{-1}).
\end{aligned}$$

We update ε as $\varepsilon \leftarrow (1 + \sqrt{6J})\varepsilon$ and achieve C^* such that we can ignore the term with the order of $\log(1/\varepsilon)$, then obtain the statement. □

S3 Note on Assumption 2

Several distributions are known to satisfy Assumption 2. We develop the following simple example:

Example 3 (Uniformly distributed Fourier coefficients)**.** We consider a distribution Π of a function h on $[0,1]$ whose Fourier coefficients by a basis are uniformly distributed. Let $\{\varphi_j : [0,1] \to \mathbb{R}\}_{j=1,2,\ldots,\infty}$ be a trigonometric basis as an orthonormal basis (see Example 1.3 in Tsybakov (2008)). We set Π as a measure of h which has a form

$$h(\cdot) = \sum_{j=1}^{\infty} \theta_j \varphi_j(\cdot),$$

where θ_j is a random Fourier coefficient which independently follows a uniform distribution on $[-1/j, 1/j]$. Note that Parseval's equality yields $\|h\|_2^2 = \sum_{j=1}^{\infty} \theta_j^2 \leq \sum_{j=1}^{\infty} 1/j^2 = \pi^2/6$ almost surely, hence the support of Π is in the L^2 space. Furthermore, h belongs to the Sobolev space since the coefficients $\{\theta_j\}_{j=1}^{\infty}$ are in the Sovolev ellipsoid (for details, see Section 1.7.1 in Tsybakov (2008)), the support of Π satisfies Assumption 1.

We show that Π satisfies Assumption 2. Without loss of generality, we consider a ball $B(0,\delta)$ whose center is 0 with fixed $\delta > 0$. We define $C_{1.5} := \sum_{j=1}^{\infty} 1/j^{1.5} \approx 2.61238$. We study the measure as

$$\Pi(h \in B(0,\delta)) = \Pi(\|h\|^2 \leq \delta^2)$$

$$
\begin{aligned}
&= \Pi\left(\sum_{j=1}^{\infty} \theta_j^2 \leq \delta^2\right) \\
&= \Pi\left(\sum_{j=1}^{\infty} \theta_j^2 \leq \frac{\delta^2}{C_{1.5}} \sum_{j=1}^{\infty} j^{-1.5}\right) \\
&\geq \prod_{j=1}^{\infty} \Pi\left(\theta_j^2 \leq \frac{\delta^2}{C_{1.5} j^{1.5}}\right) \\
&= \prod_{j=1}^{J} \Pi\left(\theta_j^2 \leq \frac{\delta^2}{C_{1.5} j^{1.5}}\right),
\end{aligned}
$$

where $J = \max\{j \in \mathbb{N} \mid 1/j^2 \geq \delta^2/(C_{1.5} j^{1.5})\}$. The first inequality follows the independent property of θ_j, and the last equality follows that $\Pi(\theta_j^2 \leq \frac{\delta^2}{C_{1.5} j^{1.5}}) = 1$ for $j \geq J+1$. For $j \leq J$, $\Pi(\theta_j^2 \leq \frac{\delta^2}{C_{1.5} j^{1.5}})$ is positive since θ_j follows the uniform distribution, we obtain that $\Pi(h \in B(0, \delta)) > 0$. □

Another example is the truncated Gaussian as described below.

Example 4 (Small shifted ball probability with truncated Gaussian processes)**.** Let h be a Borel measurable centered Gaussian random element in a separable Hilbert space $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$. From Kuelbs *et al.* (1994), for any $x \in \mathbb{H}$, $\varepsilon > 0$, $0 \leq \alpha \leq 1$, it holds that

$$
\Pi(h : \|h - x\|_{\mathbb{H}} \leq \varepsilon) \geq \exp\left\{ - \inf_{x_0 : \in \mathbb{H} : \|x_0 - x\| \leq \alpha\varepsilon} \frac{\|x_0\|_{\mathbb{H}}^2}{2} + \log \Pi(\|h\|_{\mathbb{H}} < (1-\alpha)\varepsilon) \right\}. \tag{S3.2}
$$

To satisfy Assumption 1, we consider a probability measure of a truncated version of a Gaussian measure. Given a constant $c > 0$ as a truncation level, we define a ball $\mathbb{H}_c := \{h \in \mathbb{H} \mid \|h\|_{\mathbb{H}} \leq c\}$ such that $\bar{C} := \Pi(\mathbb{H}_c) > 0$. We, then,

consider a measure Π_c associated with the truncated Gaussian process, such that $H \in \sigma(\mathbb{H}_c)$ satisfies $\Pi_c(H) := \Pi(H \mid \mathbb{H}_c) = \Pi(H)/\bar{C}$. Using the inequality (S3.2), for any $x \in \mathbb{H}_c$, it holds that

$$\begin{aligned}
&\Pi_c(h : \|h - x\|_{\mathbb{H}} \leq \varepsilon) \\
&\geq \exp\left\{ - \inf_{x_0 : \in \mathbb{H} : \|x_0 - x\| \leq \alpha\varepsilon} \frac{\|x_0\|_{\mathbb{H}}^2}{2} + \log \Pi(\|h\|_{\mathbb{H}} < (1-\alpha)\varepsilon) \right\} \bar{C}^{-1},
\end{aligned}$$

for any $\alpha \in (0, 1)$. Hence, by setting $\mathbb{H}_c = L^2,\ \alpha = \frac{1}{2}$ and $\varepsilon = \frac{\delta}{2}$, we obtain

$$\begin{aligned}
\Pi_c\left(B\left(x; \frac{\delta}{2}\right)\right) &= \Pi\left(h : \|h - x\|_{L^2} \leq \frac{\delta}{2}\right) \bar{C}^{-1} \\
&\geq \exp\left\{ - \inf_{x_0 : \in L^2 : \|x_0 - x\| \leq \delta/4} \frac{\|x_0\|_{L^2}^2}{2} + \log \Pi\left(\|h\|_{L^2} < \frac{\delta}{4}\right) \right\} \bar{C}^{-1} \\
&\geq \Pi\left(h : \|h\|_{L^2} < \frac{\delta}{4}\right) \exp\left(-\frac{\|x\|_{L^2}^2}{2}\right) \bar{C}^{-1} \\
&\geq \Pi\left(h : \|h\|_{L^2} < \frac{\delta}{4}\right) \exp\left(-\frac{c^2}{2}\right) \bar{C}^{-1},
\end{aligned}$$

for any $x \in \mathbb{H}_c$. Since h is a centered Gaussian, a ball near 0 with positive radius has positive measure (Gao *et al.*, 2004). Then $\Pi(B(x; \delta/2)) > 0$ holds. □

S4 Proof of the Delaigle–Hall and hard-margin Condition

We start with the proof for connecting the Delaigle–Hall condition and the hard-margin condition, which is one of the key points of this study.

Proof of Proposition 2. We will develop an explicit classifier based on the Delaigle–Hall condition, then show that the classifier has a positive margin. Without loss

of generality, we set $\mu_- = 0$, hence $\mu_{-,j} = 0$ holds for all $j \in \mathbb{N}$. Hence, we have $\psi_M := \sum_{j=1}^{M} \theta_j^{-1} \mu_{+,j} \phi_j$, and $f_M^*(x) = (\langle x, \psi_M \rangle - \langle \mu_+, \psi_M \rangle)^2 - \langle x, \psi_M \rangle^2$ for $x \in \mathcal{X}$ and $M \in \mathbb{N}$. For X generated from P_-, $f_M^*(X)$ is written as

$$f_M^*(X) = \langle \mu_+, \psi_M \rangle^2 - 2\langle \mu_+, \psi_M \rangle \alpha_- Z_-,$$

where the random variable $Z_- = \langle X, \psi_M \rangle / \alpha_-$ and $\alpha_-^2 = \sum_{j=1}^{\infty} \theta_{-,j} \langle \psi_M, \phi_{-,j} \rangle^2$. Here, $E[Z_-] = 0$ and $E[Z_-^2] = 1$ hold. Similarly, for X generated from P_+, we obtain

$$f_M^*(X) = -\langle \mu_+, \psi_M \rangle^2 - 2\langle x - \mu_+, \psi_M \rangle \langle \mu_+, \psi_M \rangle = -\langle \mu_+, \psi_M \rangle^2 - 2\langle \mu_+, \psi_M \rangle \alpha_+ Z_+,$$

where $Z_+ = \langle X - \mu_+, \psi_M \rangle / \alpha_+$ and $\alpha_+^2 = \sum_{j=1}^{\infty} \theta_{+,j} \langle \psi_M, \phi_{+,j} \rangle^2$. Here, Z_+ satisfies $E[Z_+] = 0$ and $E[Z_+^2] = 1$.

Now, we evaluate the margin of the classifier f_M^* with the measure Π. For any $\delta > 0$, we bound it as

$$\begin{aligned}
&\Pi(\{x : |\, \|x - \mu_+\|^2 - \|x\|^2| \leq \delta\}) \\
&= \lim_{M \to \infty} \Pi(\{x : |f_M^*(x)| \leq \delta\}) \\
&= \lim_{M \to \infty} w P_-(|f_M^*(X)| \leq \delta) + (1 - w) P_+(|f_M^*(X)| \leq \delta) \\
&\leq \lim_{M \to \infty} w P_-(f_M^*(X) \leq \delta) + (1 - w) P_+(f_M^*(X) \geq -\delta) \\
&= \lim_{M \to \infty} w P_-(\langle \mu_+, \psi_M \rangle^2 - 2\langle \mu_+, \psi_M \rangle \alpha_- Z_- \leq \delta) \\
&\quad + (1 - w) P_+(-\langle \mu_+, \psi_M \rangle^2 - 2\langle \mu_+, \psi_M \rangle \alpha_+ Z_+ \geq -\delta)
\end{aligned}$$

$$= \lim_{M\to\infty} wP_-\left(Z_- \geq \frac{\langle \mu_+, \psi_M\rangle^2 - \delta}{2\alpha_-\langle \mu_+, \psi_M\rangle}\right) + (1-w)P_+\left(-Z_+ \geq \frac{\langle \mu_+, \psi_M\rangle^2 - \delta}{2\alpha_+\langle \mu_+, \psi_M\rangle}\right)$$

$$\leq \lim_{M\to\infty} \frac{\{4w\alpha_-^2 + 4(1-w)\alpha_+^2\}\langle \mu_+, \psi_M\rangle^2}{(\langle \mu_+, \psi_M\rangle^2 - \delta)^2} \quad (\because \text{Chebyshev's inequality})$$

$$= 0.$$

The last equality holds because of the following relation: for $\ell \in \{-, +\}$, we obtain

$$\begin{aligned}
\lim_{M\to\infty} \frac{\langle \mu_+, \psi_M\rangle^2}{\alpha_\ell^2} &= \lim_{M\to\infty} \frac{(\sum_{j=1}^M \theta_j^{-1}\mu_j^2)^2}{\sum_{j=1}^\infty \theta_{\ell,j}\langle \psi_M, \phi_{\ell,j}\rangle^2} \\
&= \lim_{M\to\infty} \frac{(\sum_{j=1}^M \theta_j^{-1}\mu_j^2)^2}{\sum_{j=1}^\infty \theta_{\ell,j}(\sum_{i=1}^M \theta_i^{-1}\mu_i\langle \phi_i, \phi_{\ell,j}\rangle)^2} \\
&= \infty,
\end{aligned}$$

by the Delaigle–Hall condition. □

S5 Proof of Convergence Analysis

S5.1 Additional Notation

For a function $f : \mathcal{X} \times \{-1, 1\} \to \mathbb{R}$, we employ the notation $(\ell \circ f)(x, y) = \ell(yf(x))$. Also, for $g = \ell \circ f$, its expectation and empirical mean with respect to P is written as $Pg = \mathbb{E}_{(X,Y)\sim P}[g(X, Y)]$ and $P_n f = \frac{1}{n}\sum_{i=1}^n g(X_i, Y_i)$ with the observed data $\{(X_i, Y_i) : i = 1, ..., n\}$.

We define an open ball $B(x; \delta') \subset \mathcal{X}$ of radius δ' centered at $x \in \mathcal{X}$ with

metric $\|\cdot\|$. We also define a set $\mathcal{H}(x,\delta') \subset \mathcal{H}$ which is a set of a map $h \in \mathcal{H}$ satisfying the following three conditions:

$$(i)\,\forall x' \in \mathcal{X}\ 0 \leq h(x') \leq 2\delta,\ (ii)\,h \geq \delta' \text{ on } B\left(x; \frac{\delta'}{2}\right), \text{ and}$$
$$(iii) \int_{B(x;\delta')^c} h d\Pi \leq \delta' \int_{\mathcal{X}} h d\Pi, \tag{S5.3}$$

where $B(x;\delta')^c := \mathcal{X} \backslash B(x;\delta')$. It is obvious to show $\mathcal{H}(x,\delta') \neq \emptyset$, since there exists a continuous f such that $0 \leq f \leq \frac{3}{2}\delta'$ on $B(x,\delta'/2)$ and $f = 0$ on $B(x,\delta')^c$ holds, and $\mathcal{H}$ is dense in $C(\mathcal{X})$.

We define $q(x,\delta') = \inf_{h \in \mathcal{H}(x,\delta')} \|h\|_{\mathcal{H}}$ and $\bar{q}(\delta')$ as its decreasing envelope such that $\bar{q}(\delta') \geq \sup_{x \in \mathcal{X}} q(x,\delta')$ holds. We also define $p(x,\delta') = (\delta')^2 \Pi(B(x;\delta'/2))$ and define its lower envelope function $\bar{p}$ as $p(x,\delta') \geq \bar{p}(\delta') > 0$ for all x such that $|\widetilde{f}^*(x)| \geq 1$ holds. This definition is related to the small shifted ball probability and it varies with the setting of Π and $\mathcal{X}$. Remark that the existence of a lower envelope is guaranteed by Assumption 2. Further, on the set $\{x : |\widetilde{f}^*(x)| \geq 1\}$, we consider a positive function $r : \mathbb{R}_+ \to \mathbb{R}_+$ such that $r(\delta') \geq \bar{p}(\delta')/\bar{q}(\delta') > 0$ holds.

S5.2 Full Proof

Proof of Theorem 1. This proof contains three steps: (i) a basis decomposition on the generalization error, (ii) bound a misclassification error with the bounded condition, and (iii) bound an unbounded probability. In the following, each step

is described in one subsection.

(i) Basic Decomposition: We start with a basic decomposition for the generalization error for the classification. To fit the situation with the Delaigle–Hall condition, we extend its formulation. In the following, $\Pr(\cdot)$ and $E[\cdot]$ denote a probability and an expectation with respect to the observed data from $P^{\otimes n}$.

Lemma 1. *Suppose the Delaigle–Hall condition holds. Then, the following equation holds:*

$$E[R(\widehat{f}_n) - R(\widetilde{f}^*)] \leq \int |\eta(x)| \Pr(\widehat{f}_n(x)\widetilde{f}^*(x) \leq 0)\, d\Pi(x).$$

Proof of Lemma 1. We transform the generalization error for any $f \in \mathcal{H}$ as

$$\begin{aligned}
R(f) - R(\widetilde{f}^*) &= \mathrm{E}_X[\mathrm{E}_Y[I_{\{Y \neq \mathrm{sign}(f(X))\}} - I_{\{Y \neq \mathrm{sign}(\widetilde{f}^*(X))\}} \,|\, X]] \\
&= \mathrm{E}_X[\{I_{\{1 \neq \mathrm{sign}(f(X))\}} - I_{\{1 \neq \mathrm{sign}(\widetilde{f}^*(X))\}}\} \cdot \Pr(Y = 1|X) \\
&\qquad + \{I_{\{-1 \neq \mathrm{sign}(f(X))\}} - I_{\{-1 \neq \mathrm{sign}(f^*(X))\}}\} \cdot \Pr(Y = -1|X)\,] \\
&\leq \mathrm{E}_X[I_{\{\mathrm{sign}(\widetilde{f}^*(X)) \neq \mathrm{sign}(f(X))\}}\, |\eta(X)|\,] \\
&= \int_{\{x \in \mathcal{X} : \mathrm{sign}(\widetilde{f}^*(x)) \neq \mathrm{sign}(f(x))\}} |\eta(x)|\, d\Pi(x).
\end{aligned}$$

We consider its expectation with $\widehat{f}_n$ and develop its upper bound as

$$\begin{aligned}
E[R(\widehat{f}_n) - R(\widetilde{f}^*)] &= E\left[\int_{\{x \in \mathcal{X} : \mathrm{sign}(\widetilde{f}^*(x)) \neq \mathrm{sign}(\widehat{f}_n(x))\}} |\eta(x)| d\Pi(x)\right] \\
&= E\left[\int_{\{x \in \mathcal{X} : \widehat{f}_n(x)\widetilde{f}^*(x) \leq 0\}} |\eta(x)| d\Pi(x)\right] \\
&= \int |\eta(x)|\, E[\, I_{\{x \in \mathcal{X} : \widehat{f}_n(x)\widetilde{f}^*(x) \leq 0\}}\,]\, d\Pi(x)
\end{aligned}$$

$$= \int |\eta(x)| \Pr(\widehat{f}_n(x)\widetilde{f}^*(x) \leq 0)\, d\Pi(x).$$

Then, we obtain the statement. □

Our next goal is to study the probability $\Pr(\widehat{f}_n(x)\widetilde{f}^*(x) \leq 0)$ in Lemma 1 for a given $x \in \mathcal{X}$. For any $x \in \mathcal{X}$ such that $\widetilde{f}^*(x) > 0$ holds, with the threshold U, we obtain

$$\begin{aligned}\Pr(\widehat{f}_n(x)\widetilde{f}^*(x) \leq 0) &= \Pr(\widehat{f}_n(x) \leq 0) \\ &\leq \underbrace{\Pr(\widehat{f}_n(x) \leq 0\,,\, \|\widehat{f}_n\|_{\mathcal{H}} \leq U)}_{=T_1} + \underbrace{\Pr(\|\widehat{f}_n\|_{\mathcal{H}} > U)}_{=T_2}.\end{aligned} \tag{S5.4}$$

If $\widetilde{f}^*(x) < 0$ holds, we obtain the similar bound. We will bound the terms T_1 and T_2, respectively.

(ii-1) Bound T_1 via hard-margin Condition: As preparation, we fix x such that $\widetilde{f}^*(x) \geq \delta = 1$ holds, which follows from $\operatorname{ess\,inf}_{x\in\mathcal{X}} |\widetilde{f}^*(x)| \geq \delta$ for any δ by Proposition 2. Also, we fix $\delta_0 > 0$ then pick $h \in \mathcal{H}(x, \delta_0)$ as (S5.3). We rewrite the empirical loss in (2.2) as

$$L_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i f(X_i)) + \lambda \|f\|_{\mathcal{H}}^2.$$

By Lemma 6, we obtain its functional derivative in terms of f at $\widehat{f}_n$ with direction h as $\nabla L_n(\widehat{f}_n) = \frac{1}{n}\sum_{i=1}^{n} \ell'(Y_i\widehat{f}_n(X_i))Y_i h(X_i) + 2\lambda\langle \widehat{f}_n, h\rangle_{\mathcal{H}}$. By the optimal condition of $\widehat{f}_n$, we have $\nabla L_n(\widehat{f}_n) = 0$.

We bound the term T_1 by combining a probability of the event with $\nabla L_n(\widehat{f}_n)$. Let $\mathcal{U}$ be an event $\{\widehat{f}_n(x) \leq 0, \|\widehat{f}_n\|_{\mathcal{H}} \leq U\}$. We simply obtain

$$\begin{aligned}
T_1 &= \Pr(\mathcal{U}, \nabla L_n(\widehat{f}_n) = 0) \\
&= \Pr(\nabla L_n(\widehat{f}_n) = 0 \mid \mathcal{U}) \Pr(\mathcal{U}) \\
&= \{1 - \Pr(\nabla L_n(\widehat{f}_n) \neq 0 \mid \mathcal{U})\} \Pr(\mathcal{U}) \\
&\leq \{1 - \Pr(\nabla L_n(\widehat{f}_n) < 0 \mid \mathcal{U})\} \Pr(\mathcal{U}) \\
&\leq 1 - P_L,
\end{aligned}$$

where we define $P_L = \Pr(\nabla L_n(\widehat{f}_n) < 0 \mid \mathcal{U})$. The first line follows the fact $\Pr(\nabla L_n(\widehat{f}_n) = 0) = 1$. To bound T_1, we will study P_L.

We consider an event $\mathcal{U}$, and study the derivative $\nabla L_n(\widehat{f}_n)$. We define $\nabla \widehat{L} = \frac{1}{n} \sum_{i=1}^n \ell'(Y_i \widehat{f}_n(X_i)) Y_i h(X_i)$ as a derivative of the loss function part from $\nabla L_n(\widehat{f}_n)$. By Lemma 7 associated with Lemma 8, we can bound tail probability of $\widehat{L}$ as

$$\begin{aligned}
\Pr\left(\nabla \widehat{L} < -\frac{1}{2}\delta_0 E[h(X)] \mid \mathcal{U}\right) &\geq 1 - 2\exp\left(-\frac{n\delta_0 E[h(X)]}{C_{L,U}}\right) \\
&\geq 1 - 2\exp\left(-\frac{np(x,\delta_0)}{C_{L,U}}\right),
\end{aligned}$$

which follows the relation $\delta_0 E[h(X)] \geq \delta_0^2 \Pi(B(x; \delta_0/2)) = p(x, \delta_0)$. By this result, we can also bound $\nabla L_n(\widehat{f}_n)$ as

$$\nabla L_n(\widehat{f}_n) = \nabla \widehat{L} + 2\lambda \langle \widehat{f}_n, h \rangle$$

$$\leq -\delta_0 E[h(X)]/2 + 2\lambda U \|h\|_{\mathcal{H}}$$

$$\leq -p(x,\delta_0)/2 + 2\lambda U q(x,\delta_0),$$

with probability at least $1-2\exp(-np(x,\delta_0)/C_{L,U})$. The first inequality follows the Cauchy-Schwartz inequality and $\|\widehat{f}_n\|_{\mathcal{H}} \leq U$. Since we set $\lambda < \frac{p(x,\delta_0)}{4Uq(x,\delta_0)} = \frac{r(x,\delta_0)}{4U}$, we obtain $\nabla L_n(\widehat{f}_n) < 0$ with the probability. Thus, we have

$$\begin{aligned} T_1 \leq 1 - P_L &\leq 1 - \left\{1 - 2\exp\left(-\frac{np(x,\delta_0)}{C_{L,U}}\right)\right\} \\ &\leq 2\exp\left(-\frac{np(x,\delta_0)}{C_{L,U}}\right) \leq 2\exp\left(-\frac{n\bar{p}(\delta_0)}{C_{L,U}}\right). \end{aligned} \tag{S5.5}$$

(ii-2) Bound T_2 via Metric Entropy of Functional Data Space: We bound T_2 in (S5.4) by using the *peeling* technique (for introduction, see Chapter 7 in Steinwart and Christmann (2008)).

As preparation, we derive an upper bound of $\|\widehat{f}_n\|_{\mathcal{H}}$. Since

$$\lambda\|\widehat{f}_n\|_{\mathcal{H}}^2 \leq P_n(\ell \circ \widehat{f}_n) + \lambda\|\widehat{f}_n\|_{\mathcal{H}}^2 \leq \ell(0) + \|0\|_{\mathcal{H}}^2 \leq 1,$$

where the second inequality is obtained by replacing $\widehat{f}_n$ by 0 and the optimality condition of $\widehat{f}_n$, and the last inequality follows the bounded condition on the loss function, we obtain $\bar{R} = \lambda^{-1/2}\ell(0)^{-1/2}$ as an upper bound of $\|\widehat{f}_n\|_{\mathcal{H}}$.

We decompose the term T_2. We remind the definition $f^\dagger = \mathrm{argmin}_{f\in\mathcal{H}} R(f)$, and consider a constant $R = \|f^\dagger\|_{\mathcal{H}}$ which is assumed to be no less than 1 with-

out loss of generality. We also define events $\mathcal{A}(R)$ and $\mathcal{E}(R)$ as

$$\mathcal{A}(R) = \left\{ \frac{R}{2} \leq \|\widehat{f}_n\|_{\mathcal{H}} \leq R \right\}, \text{ and } \mathcal{E}(R) = \left\{ \|\widehat{f}_n\|_{\mathcal{H}} \leq \frac{R}{2} \right\},$$

and a sequence $R_k = 2^k, k = 1, 2, ..., N$ where $N = \log_2 \bar{R} + 1$. For each $\lambda > \underline{\lambda}$ and sufficiently large n, we have

$$N = \log_2 \bar{R} + 1 = \frac{1}{2\log 2} \log \frac{\ell(0)}{\lambda} + 1 \leq \frac{1}{2\log 2} \log \frac{n}{C_{V,\gamma} \log\log n} + 1 \leq C'_{V,\gamma} \log n,$$

where $C'_{V,\gamma}$ is a constant depending on $C_{V,\gamma}$. We remark that $\cup_{k=1}^{N} \mathcal{A}(R_k) \supset \{U \leq \|\widehat{f}_n\|_{\mathcal{H}}\}$ since $\|\widehat{f}_n\|_{\mathcal{H}} \leq \bar{R}$ holds. Since $\mathcal{A}(R_k), k = 1, ..., N$ are disjoint up to null sets, we obtain

$$T_2 \leq \sum_{k=1}^{N} \Pr(U \leq \|\widehat{f}_n\|_{\mathcal{H}} | \mathcal{A}(R_k)) \Pr(\mathcal{A}(R_k)) \leq \sum_{k=1}^{N} \Pr(U \leq \|\widehat{f}_n\|_{\mathcal{H}} | \mathcal{A}(R_k)). \tag{S5.6}$$

Now, we will bound the probability $\Pr(U \leq \|\widehat{f}_n\|_{\mathcal{H}} | \mathcal{A}(R_k))$ in the following.

We investigate the event $\mathcal{E}(R)$ with conditional on $\mathcal{A}(R)$ and study the event $U \leq \|\widehat{f}_n\|_{\mathcal{H}}$. We set a constant $c_{V,\gamma} = 2\sqrt{R}(\sqrt{6} + \frac{1}{V3^\gamma}) \exp(V3^\gamma)$. An inequality $P_n(\ell \circ \widehat{f}_n) - \inf_{g \in \mathcal{H}: \|g\|_{\mathcal{H}} \leq R} P_n(\ell \circ g) \geq 0$ and a uniform bound defined by

$$\Delta(n, \gamma, t, R) = R c_{V,\gamma} (\log n)^{-1/\gamma} + \sqrt{2t/n}, \tag{S5.7}$$

and Lemma 2 implies

$$\lambda \|\widehat{f}_n\|_{\mathcal{H}}^2 \leq P_n(\ell \circ \widehat{f}_n) - \inf_{g \in \mathcal{H}: \|g\|_{\mathcal{H}} \leq R} P_n(\ell \circ g) + \lambda \|\widehat{f}_n\|_{\mathcal{H}}^2$$

$$
\begin{aligned}
&= \inf_{f\in\mathcal{H}:\|f\|_{\mathcal{H}}\leq R}\left\{P_n(\ell\circ f) - \inf_{g\in\mathcal{H}:\|g\|_{\mathcal{H}}\leq R} P_n(\ell\circ g) + \lambda\|f\|_{\mathcal{H}}^2\right\} \\
&\leq \inf_{f\in\mathcal{H}:\|f\|_{\mathcal{H}}\leq R}\left\{P(\ell\circ f) - \inf_{g\in\mathcal{H}:\|g\|_{\mathcal{H}}\leq R} P(\ell\circ g) + \lambda\|f\|_{\mathcal{H}}^2\right\} + 2\Delta(n,\gamma,t,R) \\
&\leq \lambda\|f^\dagger\|_{\mathcal{H}}^2 + 2\Delta(n,\gamma,t,R),
\end{aligned}
$$

with probability at least $1-\exp(-t)$ for any $t>0$. The last inequality holds by substituting $f^\dagger$. Combining an inequality $R/2 < \|\widehat{f}_n\|_{\mathcal{H}}$ with this result yields

$$
R^2/4 \leq \|\widehat{f}_n\|_{\mathcal{H}}^2 \leq \|f^\dagger\|_{\mathcal{H}}^2 + 2\Delta(n,\gamma,t,R)/\lambda.
$$

Solving this inequality with respect to R yields that

$$
\begin{aligned}
R &\leq 4c_{V,\gamma}\lambda^{-1}(\log n)^{-1/\gamma} + \sqrt{\left(4c_{V,\gamma}\lambda^{-1}(\log n)^{-1/\gamma}\right)^2 + 4\|f^\dagger\|_{\mathcal{H}}^2 + 8\lambda^{-1}\sqrt{2t/n}} \\
&\leq 8c_{V,\gamma}\lambda^{-1}(\log n)^{-1/\gamma} + 2\|f^\dagger\|_{\mathcal{H}} + 2\left(2t\right)^{1/4}\lambda^{-1/2}n^{-1/4} \\
&\leq C_{V,\gamma}(\|f^\dagger\|_{\mathcal{H}} \vee \lambda^{-1}(\log n)^{-1/\gamma} \vee t^{1/4}\lambda^{-1/2}n^{-1/4}),
\end{aligned}
$$

where $C_{V,\gamma}$ is a constant depending on $c_{V,\gamma}$. By setting $t = n\zeta^2$ and with sufficiently small $\zeta > 0$ which will be specified later, we obtain $R \leq C_{V,\gamma}\|f^\dagger\|_{\mathcal{H}} = U$ holds. Consequently, conditional on $\mathcal{A}(R)$, the event $\mathcal{E}(R)$ implies $R \leq U$ with probability at least $1-\exp(-n\zeta^2)$, which contradicts the setting of $R \geq U$. Hence, for any measurable event Ω, it holds that

$$
\Pr(\Omega \mid \mathcal{A}(R)) \leq \Pr(\mathcal{E}(R)^c \mid \mathcal{A}(R)) \leq 1-(1-\exp(-n\zeta^2)) = \exp(-n\zeta^2).
$$

We put this inequality with setting $\Omega = \{U \leq \|\widehat{f}_n\|_{\mathcal{H}}\}$ into (S5.6). Then we

obtain

$$T_2 \leq \sum_{k=1}^{N} \Pr(\mathcal{E}(R_k)^c \mid \mathcal{A}(R_k)) \leq N \exp(-n\zeta^2) \leq e^{-n\zeta} C'_{V,\gamma} \log n$$
$$\leq e^{-n\zeta} C'_{V,\gamma} \exp(n\varepsilon / C_{V,\gamma}) \leq C'_{V,\gamma} \exp\{-n\zeta(1 - C_{V,\gamma}^{-1})\} \leq \exp(-n\zeta/2), \tag{S5.8}$$

The last third inequality follows the setting of ζ as following $\zeta \geq C'_{V,\gamma} \geq C'_{V,\gamma} \frac{\log\log n}{n}$.

(iii) Combining Results: We utilize the derived bounds for T_1 in (S5.5) and for T_2 in (S5.8) into (S5.4), then obtain the following inequality:

$$\Pr(\widehat{f}_n(x)\widetilde{f}^*(x) \leq 0) \leq 2\exp\left(-\frac{n\bar{p}(\delta_0)}{C_{L,U}}\right) + \exp\left(-\frac{n\zeta}{2}\right) \leq \exp(-\beta n),$$

by selecting β depending on $\bar{p}(\delta_0), C_{L,U}$ and ζ. Finally, we obtain $E[R(\widehat{f}_n) - \inf_{f\in\mathcal{H}} R(f)] \leq E[R(\widehat{f}_n) - R(\widetilde{f}^*)]$ by the fact $R(\widetilde{f}^*) \leq \inf_{f\in\mathcal{H}} R(f)$, and the above results yield the statement. □

S6 Entropy Analysis for Functional Data

We provide several technical results with empirical process techniques.

Lemma 2. *Recall the definition of* $\Delta(n, \gamma, t, R)$ *in* (S5.7). *For any* $f' \in \mathcal{H}$, *we obtain*

$$P_n(\ell \circ f') - \inf_{f:\|f\|_{\mathcal{H}} \leq R} P_n(\ell \circ f) \leq P(\ell \circ f') - \inf_{f:\|f\|_{\mathcal{H}} \leq R} P(\ell \circ f) + 2\Delta(n, \gamma, t, R)$$

with probability at least $1 - e^t$ *with* $t > 0$.

Proof of Lemma 2. Simply, we obtain

$$
\begin{aligned}
&P_n(\ell \circ f') - \inf_{f:\|f\|_{\mathcal{H}} \leq R} P_n(\ell \circ f) \\
&= P_n(\ell \circ f') - P(\ell \circ f) + P(\ell \circ f') \\
&\quad - \inf_{f:\|f\|_{\mathcal{H}} \leq R} P(\ell \circ f) + \inf_{f:\|f\|_{\mathcal{H}} \leq R} P(\ell \circ f) - \inf_{f:\|f\|_{\mathcal{H}} \leq R} P_n(\ell \circ f) \\
&\leq \{P_n(\ell \circ f) - P(\ell \circ f)\} \\
&\quad + \left\{ P(\ell \circ f) - \inf_{f:\|f\|_{\mathcal{H}} \leq R} P(\ell \circ f) \right\} + \{P(\ell \circ f^\dagger) - P_n(\ell \circ f^\dagger)\}.
\end{aligned}
$$

By Lemma 3, $\Delta(n, \gamma, t, R)$ bounds the last two terms with probability at least $1 - t$. □

To complete Lemma 2, we provide the following lemma. This result is a well-known result with the Rademacher complexity, but we provide it for the sake of completeness. We introduce a ball in $\mathcal{H}$ with radius R as $\mathcal{H}_R = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq R\}$.

Lemma 3. *Define* $c_{V,\gamma} = 2\sqrt{R}(\sqrt{6} + \frac{1}{V3^\gamma}) \exp(V3^\gamma)$. *For any* $t > 0$ *and any* $n \in \mathbb{N}$, *we obtain*

$$
\Pr\left(\sup_{f \in \mathcal{H}_R} |P_n(\ell \circ f) - P(\ell \circ f)| \leq R c_{V,\gamma} (\log n)^{-1/\gamma} + \sqrt{2t/n} \right) \geq 1 - \exp(-t).
$$

Proof of Lemma 3. We firstly bound the term $\sup_{f \in \mathcal{H}_R} |P_n(\ell \circ f) - P(\ell \circ f)|$ by its expectation and others. Since a variation of the term is at most $2/n$ when one

pair of $\{(X_i, Y_i)\}_{i=1}^n$ changes, the McDiarmid's inequality (Theorem 3.3.14 in Giné and Nickl (2016)) implies that with probability at least $1 - e^{-t}$,

$$\sup_{f \in \mathcal{H}_R} |P_n(\ell \circ f) - P(\ell \circ f)| \leq E\left[\sup_{f \in \mathcal{H}_R} |P_n(\ell \circ f) - P(\ell \circ f)|\right] + \sqrt{\frac{2t}{n}}. \tag{S6.9}$$

Secondly, to bound the expectation term, we define the conditional Rademacher complexity of a class of functions $\mathcal{G}$ as follows:

$$\mathcal{R}_n(\mathcal{G}) = \frac{1}{n} E_\sigma \left[\sup_{f \in \mathcal{G}} \sum_{i=1}^n \sigma_i f(x_i)\right],$$

where $\sigma_1, \ldots, \sigma_n$ are independent random variables which is 1 with probability $1/2$ and -1 otherwise. We introduce $(X_1', Y_1'), \cdots, (X_n', Y_n')$ as independent pairs of random variables with the same distribution as (X, Y). We apply the independent pairs and bound the expectation term in (S6.9) as

$$\begin{aligned}
&E\left[\sup_{f \in \mathcal{H}_R} |P_n(\ell \circ f) - P(\ell \circ f)|\right] \\
&= E\left[\sup_{f \in \mathcal{H}_R} E\left[\left|\frac{1}{n}\sum_{i=1}^n \ell(Y_i f(X_i)) - \frac{1}{n}\sum_{i=1}^n \ell(Y_i' f(X_i'))\right| \mid \{(X_i, Y_i)\}_{i=1}^n\right]\right] \\
&\leq E\left[\sup_{f \in \mathcal{H}_R} \left|\frac{1}{n}\sum_{i=1}^n \ell(Y_i f(X_i)) - \frac{1}{n}\sum_{i=1}^n \ell(Y_i' f(X_i'))\right|\right] \\
&= E_\sigma\left[\sup_{f \in \mathcal{H}_R} \frac{1}{n}\sum_{i=1}^n |\sigma_i\{\ell(Y_i f(X_i)) - \ell(Y_i' f(X_i'))\}|\right] \\
&\leq E_\sigma\left[\sup_{f \in \mathcal{H}_R} \frac{1}{n}\sum_{i=1}^n \sigma_i \ell(Y_i f(X_i))\right] + E_\sigma\left[\sup_{f \in \mathcal{H}_R} \frac{1}{n}\sum_{i=1}^n \sigma_i \ell(Y_i' f(X_i'))\right] \\
&= 2\mathcal{R}_n(\ell \circ \mathcal{H}_R),
\end{aligned}$$

where $\ell \circ \mathcal{H}_R = \{\ell \circ f : f \in \mathcal{H}_R\}$. The first inequality following the Jensen's inequality, and the third equality follows the distribution equivalence by the random variable σ. We further apply the Ledoux-Talagrand contraction inequality (Theorem 3.2.1 in Giné and Nickl (2016)) with the 1-Lipschitz continuity of ℓ yields $\mathcal{R}_n(\ell \circ \mathcal{H}_R) \leq \mathcal{R}_n(\mathcal{H}_R)$. Combining the result with (S6.9), we obtain

$$\sup_{f \in \mathcal{H}_R} |P_n(\ell \circ f) - P(\ell \circ f)| \leq 2\mathcal{R}_n(\ell \circ \mathcal{H}_R) + \sqrt{2t/n} \leq 2\mathcal{R}_n(\mathcal{H}_R) + \sqrt{2t/n}. \tag{S6.10}$$

with probability at least $1 - e^{-t}$.

Finally, we apply Lemma 5 and bound $\mathcal{R}_n(\mathcal{H}_R)$. Then, we obtain the statement. □

To complete the empirical process result, we develop the following covering number result, which is a key term to study the convergence of the classifier with functional data.

Lemma 4. *There exists a constant $\bar{c} > 0$ such that for any $\varepsilon \in (0, \bar{c})$ such that the following holds:*

$$\log \mathcal{N}(\varepsilon, \mathcal{H}_R, \|\cdot\|_n) \leq \frac{6R^2}{\varepsilon} + 4R\left[\exp\left\{V\left(\frac{3R}{\varepsilon}\right)^{\gamma}\right\} - 1\right].$$

Proof of Lemma 4. As preparation, we consider an ε-covering set $\mathcal{X}$ of functions $\{x_1, \ldots x_m\}$ for $\mathcal{X}$ with $m = m(\varepsilon)$, that is, for any $x \in \mathcal{X}$, there exists x_j from the set such that $\|x - x_j\| \leq \varepsilon$ holds. By Assumption 1, we have $m \leq \exp(c\varepsilon^{-\gamma})$.

We consider grids in $\mathcal{H}_R$. For each $f \in \mathcal{H}_R$, we define a vector

$$A_f = (\lfloor f(x_1)/\varepsilon \rfloor\, ,\, \lfloor f(x_2)/\varepsilon \rfloor, ..., \lfloor f(x_n)/\varepsilon \rfloor)^\top \in \mathbb{R}^m.$$

For any pair $f, g \in \mathcal{H}_R$ such that $\max_{i=1,...,m} |f(x_i) - g(x_i)| < \varepsilon$ holds, we obtain $A_f = A_g$ since $\lfloor f(x_i)/\varepsilon \rfloor = \lfloor g(x_i)/\varepsilon \rfloor$ holds. We also mention the following difference: for any $x \in \mathcal{X}$ and $f \in \mathcal{H}_R$, we obtain $|f(x) - f(x_i)| \leq \|f\|_{\mathcal{H}} \|x - x_i\| \leq R\|x - x_i\| \leq R\varepsilon$, by the property of (2.1) and that of the covering set.

With these results, we bound the following distance with the pair f, g and any $x \in \mathcal{X}$:

$$\begin{aligned} |f(x) - g(x)| &= |f(x) - f(x_i) + f(x_i) - g(x_i) + g(x_i) - g(x)| \\ &\leq |f(x) - f(x_i)| + |f(x_i) - g(x_i)| + |g(x_i) - g(x)| \\ &\leq (2R + 1)\varepsilon, \end{aligned}$$

hence we have $\|f - g\|_{L^\infty} \leq (2R + 1)\varepsilon$.

From the above discussion, the covering number $\mathcal{N}((2R+1)\varepsilon, \mathcal{H}_R, \|\cdot\|_{L^\infty})$ is bounded by the number of different A_f when f ranges over $\mathcal{H}_R$. Since $|f(x)| \leq \|f\|_{\mathcal{H}_R} \leq R$ for any $x \in \mathcal{X}$ by (2.1), the number of possible values of each element of A_f is bounded by $(2R/\varepsilon + 1)$. Assume the covering set $x_1, \ldots, x_m$ is ordered such that $i < j$ implies $\|x_i - x_j\| < 2\varepsilon$, then we obtain

$|f(x_j) - f(x_i)| \leq R\|x_j - x_i\| < 2R\varepsilon$. Therefore, we obtain

$$-2R\varepsilon + f(x_i) < f(x_j) < 2R\varepsilon + f(x_i).$$

It implies that for given $f(x_i)$ the number of possible values of $\lfloor f(x_j)/\varepsilon \rfloor$ is at most $4R + 1$. Hence, we can bound the covering number as

$$\begin{aligned}\mathcal{N}((2R+1)\varepsilon, \mathcal{H}_R, \|\cdot\|_{L^\infty}) &= |\{A_f : f \in \mathcal{H}_R\}| \\ &\leq (2R/\varepsilon + 1)(4R+1)^{m-1} \\ &\leq (2R/\varepsilon + 1)(4R+1)^{\exp(c\varepsilon^{-\gamma})-1}\end{aligned}$$

As a result, we obtain the following bound in the norm $\|\cdot\|_{L^\infty}$:

$$\begin{aligned}&\log N(\varepsilon, \mathcal{H}_R, \|\cdot\|_{L^\infty}) \\ &\leq \log\left\{\frac{2R(2R+1)}{\varepsilon} + 1\right\} + \left[\exp\left\{V\left(\frac{2R+1}{\varepsilon}\right)^\gamma\right\} - 1\right] \log(4R+1).\end{aligned}$$

Since the empirical norm $\|\cdot\|_n$ possesses the Riesz property (e.g. page 83 in Van Der Vaart and Wellner (1996)), we obtain

$$\begin{aligned}&\log N(\varepsilon, \mathcal{H}_R, \|\cdot\|_n) \\ &\leq (2R(2R+1)/\varepsilon + 1) + \left[\exp\left\{V\left(\frac{2R+1}{\varepsilon}\right)^\gamma\right\} - 1\right] \log(4R+1) \\ &\leq 6\frac{R^2}{\varepsilon} + 4R\left[\exp\left\{V\left(\frac{3R}{\varepsilon}\right)^\gamma\right\} - 1\right],\end{aligned}$$

by the setting $R \geq 1$. Then, we obtain the statement. □

The following result is to bound the Rademacher complexity by the covering

number. Although technique follows a standard discussion by the Dudley's integral, the covering number for functional data analysis has a specific role from functional data.

Lemma 5. *Suppose Assumption 1 holds. Then, we obtain*

$$\mathcal{R}_n(\mathcal{H}_R) \leq Rc_{V,\gamma}(\log n)^{-1/\gamma}.$$

Proof of Lemma 5. Now, we bound $\mathcal{R}_n(\mathcal{H}_R)$ using the following inequality learned from Srebro and Sridharan (2010):

$$R_n(\mathcal{H}_R) \leq \inf_{\alpha \geq 0} \left\{ 4\alpha + 12 \int_{\alpha}^{\sup_{f\in\mathcal{H}_R}\sqrt{P_n f^2}} \sqrt{\frac{\log N(\varepsilon, \mathcal{H}_R, \|\cdot\|_n)}{n}}\, d\varepsilon \right\}$$

As $\sup_{x\in\mathcal{X}} |f(x)| \leq \|f\|_{\mathcal{H}_R} \leq R$ for all $f \in \mathcal{H}_R$, we obtain $\sup_{f\in\mathcal{H}_R}\sqrt{P_n f^2} = R$. Lemma 4 yields

$$\begin{aligned}
&\int_{\alpha}^{R} \sqrt{\frac{\log \mathcal{N}(\varepsilon, \mathcal{H}_R, \|\cdot\|_n)}{n}}\, d\varepsilon \\
&\leq \frac{1}{\sqrt{n}} \int_{\alpha}^{R} \sqrt{\frac{6R^2}{\varepsilon} + 4R\left[\exp\left\{V\left(\frac{3R}{\varepsilon}\right)^{\gamma}\right\} - 1\right]} d\varepsilon \\
&\leq \frac{1}{\sqrt{n}} \int_{1}^{\frac{1}{\varepsilon_0}} \sqrt{6R\tau + 4R\left\{\exp\left(V3^{\gamma}\tau^{\gamma}\right) - 1\right\}} \frac{R}{\tau^2} d\tau \\
&\leq \frac{R}{\sqrt{n}} \int_{1}^{\frac{1}{\varepsilon_0}} \sqrt{6R}\tau^{-3/2} + 2\sqrt{R}\tau^{-2}\{\exp\left(V3^{\gamma}\tau^{\gamma}\right) - 1\} d\tau \\
&\leq \frac{R}{\sqrt{n}} \int_{1}^{\frac{1}{\varepsilon_0}} \sqrt{6R}\tau^{-3/2} + 2\sqrt{R}\tau^{\gamma-1}\exp\left(V3^{\gamma}\tau^{\gamma}\right) d\tau \\
&\leq Rn^{-1/2}\{2\sqrt{6R} + 2\sqrt{R}\exp(V3^{\gamma}\varepsilon_0^{-\gamma})/(V3^{\gamma})\} \\
&\leq Rn^{-1/2}\{2\sqrt{R}(\sqrt{6} + \frac{1}{V3^{\gamma}})\exp(V3^{\gamma})\}\exp(\varepsilon_0^{-\gamma}).
\end{aligned}$$

Here, we substitute $\tau = R/\varepsilon$ and define $\varepsilon_0 = \alpha/R$. Then, we have

$$\begin{aligned}\mathcal{R}_n(\mathcal{H}_R) \leq\ & R \inf_{0\leq\varepsilon_0\leq 1}[4\varepsilon_0 + n^{-1/2}\{2\sqrt{R}(\sqrt{6} + \frac{1}{V3^\gamma})\exp(V3^\gamma)\}\exp(\varepsilon_0^{-\gamma})] \\ &\leq R[4(\log n^{1/4})^{-1/\gamma} + n^{-1/4}\{2\sqrt{R}(\sqrt{6} + \frac{1}{V3^\gamma})\exp(V3^\gamma)\}] \\ &\leq R\{2\sqrt{R}(\sqrt{6} + \frac{1}{V3^\gamma})\exp(V3^\gamma)\}(\log n)^{-1/\gamma}.\end{aligned}$$

In the second inequality, we substitute ε_0 to $(\log n^{1/4})^{-1/\gamma}$. Then, we obtain the statement. □

S7 Technical Lemma

We provide several technical results for the proof of the main theorem.

Lemma 6. *We obtain the following equality:*

$$\nabla L_n(\widehat{f}_n) = \frac{1}{n}\sum_{j=1}^{n} \ell'(Y_j\widehat{f}_n(X_j))Y_j h(X_j) + 2\lambda\langle\widehat{f}_n, h\rangle_{\mathcal{H}}.$$

Proof of Lemma 6. We study the optimization problem in (2.2) by considering its functional derivative in the Fréchet sense. For a coefficient $\alpha > 0$, we rewrite the target function in (2.2) as

$$L_n(\alpha) = P_n\{\ell \circ (\widehat{f}_n + \alpha h)\} + \lambda\|\widehat{f}_n + \alpha h\|_{\mathcal{H}}^2 = \frac{1}{n}\sum_{j=1}^{n} \ell\{Y_j(\widehat{f}_n + \alpha h)(X_j)\} + \lambda\|\widehat{f}_n + \alpha h\|_{\mathcal{H}}^2.$$

Since $\widehat{f}_n$ is the minimizer of the problem in (2.2), a derivative of $L_n(\alpha)$ is 0 with $f = \widehat{f}_n$ and $\alpha = 0$. By the differentiability of ℓ, we obtain the following

derivative

$$\frac{dL_n}{d\alpha}(0) = \frac{1}{n}\sum_{j=1}^{n} \ell'\{Y_j \widehat{f}_n(X_j)\} Y_j h(X_j) + 2\lambda \langle \widehat{f}_n, h \rangle_{\mathcal{H}},$$

then we obtain the statement. □

Lemma 7. *Suppose* $\widehat{f}_n(x) \leq 0$ *and* $\|\widehat{f}_n\|_{\mathcal{H}} < U$ *hold. Then, for* $x \in \mathcal{X}$ *such that* $f_0(x) = \delta > 0$ *holds, we set* $S = B(x; \delta_0)$ *obtain*

$$\frac{1}{n}\sum_{j=1}^{n} \ell'\{Y_j \widehat{f}_n(X_j)\} Y_j h(X_j) \leq \frac{1}{n}\sum_{j=1}^{n} \xi_j,$$

where $\xi_j := 2U\delta_0 h(X) I_S(X_j) + \ell'(0) Y h(X) I_S(X_j) + |\ell'(-U)| h(X) I_{S^c}(X_j)$.

Proof of Lemma 7. We prepare some inequalities. It should be noted that

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n} \ell'\{Y_j \widehat{f}_n(X_j)\} Y_j h(X_j) \\
&= \frac{1}{n}\sum_{j:Y_j=+1}^{n} \ell'\{\widehat{f}_n(X_j)\} h(X_j) - \frac{1}{n}\sum_{j:Y_j=-1}^{n} \ell'\{-\widehat{f}_n(X_j)\} h(X_j).
\end{aligned}$$

Also, we recall that ℓ' is negative and increasing, h is nonnegative. For any $x' \in S$, the RKHS property (2.1) provides $\widehat{f}_n(x') \leq \|\widehat{f}_n\|_{\mathcal{H}} \delta_0 \leq U\delta_0$, then we have $\ell'\{\widehat{f}_n(x')\} \leq \ell'(U\delta_0),$ and $\ell'\{-\widehat{f}_n(x')\} \geq \ell'(-U\delta_0)$. Also, $|\widehat{f}_n(x')| \leq \|\widehat{f}_n\|_{\mathcal{H}} \leq U$ holds for all $x' \in S$ suggests that $|\ell'\{\widehat{f}_n(y)\}| \leq |\ell'(-U)|,$ and $|\ell'\{-\widehat{f}_n(y)\}| \leq |\ell'(-U)|$.

Now, we are ready to bound the target value. For $j = 1, ..., n$, we define $Z_j := h(X_j) I_S(X_j)$ and $Z_j^c := h(X_j) I_{S^c}(X_j)$ for brevity. We bound the value

as

$$
\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n}\ell'\{Y_j\widehat{f}_n(X_j)\}Y_j h(X_j) \\
&\leq \frac{\ell'(U\delta_0)}{n}\sum_{j:X_j\in S, Y_j=+1} h(X_j) - \frac{\ell'(-U\delta_0)}{n}\sum_{j:X_j\in S, Y_j=-1} h(X_j) + \frac{|\ell'(-U)|}{n}\sum_{j:X_j\in S^c} h(X_j) \\
&= \frac{\ell'(U\delta_0)}{n}\sum_{j:X_j\in S}\frac{1+Y_j}{2}h(X_j) - \frac{\ell'(-U\delta_0)}{n}\sum_{j:X_j\in S}\frac{1-Y_j}{2}h(X_j) + \frac{|\ell'(-U)|}{n}\sum_{j:X_j\in S^c} h(X_j) \\
&= \frac{\ell'(U\delta_0)-\ell'(-U\delta_0)}{2n}\sum_{j=1}^{n} Z_j + \frac{\ell'(U\delta_0)+\ell'(-U\delta_0)}{2n}\sum_{j=1}^{n} Y_j Z_j + \frac{|\ell'(-U)|}{n}\sum_{j=1}^{n} Z_j^c.
\end{aligned}
$$

About the coefficient terms, we obtain

$$
\begin{aligned}
\left|\frac{\ell'(U\delta_0)+\ell'(-U\delta_0)}{2} - \ell'(0)\right| &\leq \frac{|\ell'(U\delta_0)-\ell'(0)|}{2} + \frac{|\ell'(-U\delta_0)-\ell'(0)|}{2} \\
&\leq \ell''(0)U\delta_0 \leq U\delta_0,
\end{aligned}
$$

and similarly

$$
\frac{|\ell'(U\delta_0)-\ell'(-U\delta_0)|}{2} \leq \frac{|\ell'(U\delta_0)-\ell'(0)|}{2} + \frac{|\ell'(-U\delta_0)-\ell'(0)|}{2} \leq U\delta_0.
$$

Using the inequalities, we further bound the target value as

$$
\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n}\ell'\{Y_j\widehat{f}_n(X_j)\}Y_j h(X_j) \\
&\leq \frac{U\delta_0}{n}\sum_{j=1}^{n} Z_j + \frac{U\delta_0+\ell'(0)}{n}\sum_{j=1}^{n} Y_j Z_j + \frac{|\ell'(-U)|}{n}\sum_{j=1}^{n} Z_j^c \\
&\leq 2\frac{U\delta_0}{n}\sum_{j=1}^{n} Z_j + \frac{\ell'(0)}{n}\sum_{j=1}^{n} Y_j Z_j + \frac{|\ell'(-U)|}{n}\sum_{j=1}^{n} Z_j^c.
\end{aligned}
$$

Then, we obtain the statement by the definition of ξ_j. □

Lemma 8. *Consider the same setting with Lemma 7. Then, we get the following inequality:*

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\xi_j \geq -\frac{1}{2}\delta_0 E[h(X)]\right) \leq 2\exp\left\{-\frac{n\delta_0 E[h(X)]}{C_{U,L}}\right\}.$$

Proof of Lemma 8. We firstly bound an expectation of ξ_j,

$$\begin{aligned}E[\xi_j] &= 2U\delta_0 E[h(X)I_S(X)] + \ell'(0)E[Yh(X)I_S(X)] + |\ell'(-U)|E[h(X)I_{S^c}(X)] \\ &\leq 2U\delta_0 E[h(X)] + \ell'(0)E[Yh(X)I_S(X)] + |\ell'(-U)|\delta E[h(X)],\end{aligned}$$

by the conditions of $\mathcal{H}(x,\delta)$ presented in (S5.3). We define $\bar{f}(x) = \widetilde{f}^*(x) \vee 1$. For the term $E[Yh(X)I_S(X)]$, we approximate it as

$$\begin{aligned}E[Yh(X)I_S(X)] &\geq E[\bar{f}(X)h(X)I_S(X)] \\ &\geq (1-L\delta_0)E[h(X)I_S(X)] \\ &\geq (1-L\delta_0)(1-\delta_0)E[h(X)].\end{aligned}$$

The first equality holds since $\widetilde{f}^*$ is a perfect classifier. The second equality follows the hard-margin condition on $\widetilde{f}^*$ and $\inf_{x'\in S}\widetilde{f}^*(x') \geq \delta - L\delta_0$ by the Lipschitz constant L of $\widetilde{f}^*$. For the last inequality, we apply the condition *(iii)* for $\mathcal{H}(x,\delta_0)$ in (S5.3) and obtain

$$\delta_0\int_{\mathcal{X}} h d\Pi \geq \int_{S^c} h d\Pi = E[h] - \int_S h d\Pi,$$

then we have $E[h(X)I_S(X)] \geq (1-\delta_0)E[h(X)]$. Since $\ell'(0) < 0$ holds, we

substitute the bound for $E[Yh(X)I_S(X)]$ and obtain

$$E[\xi_j] \leq \{2U\delta_0 + \ell'(0)(1 - L\delta_0)(1 - \delta_0) + |\ell'(-U)|\delta_0\}E[h(X)] \leq -\delta_0 E[h(X)].$$

The last inequality follows by selecting a sufficiently small $\delta_0 > 0$ as $\delta_0 \leq 1/(L + 4U + 12)$.

We finally bound a tail probability of $n^{-1}\sum_{j=1}^n \xi_j$ by the Bernstein inequality (Theorem 3.1.7 in Giné and Nickl (2016)). Using an elementary inequality $(a+b+c)^2 \leq 3a^2 + 3b^2 + 3c^2$, we have $E(\xi^2) \leq C\delta_0 E(h(X))$. In addition, it is clear that $|\xi| \leq C_{U,L}\delta_0$. Then, by the Bernstein's inequality, we obtain

$$\begin{aligned}\Pr\left(\frac{1}{n}\sum_{j=1}^n \xi_j \geq -\frac{1}{2}\delta_0 E[h(X)]\right) &\leq 2\exp\left\{-\frac{n^2\delta_0^2 E[h(X)]^2/8}{\sum_{i=1}^n E[\xi_i^2] - nC_{U,L}\delta_0^2 E[h(X)]/6}\right\} \\ &\leq 2\exp\left\{-\frac{n\delta_0 E[h(X)]}{C_{U,L}}\right\}.\end{aligned}$$

Then, we obtain the statement. □

S8 Additional Experiment: Different Eigenfunctions

We implement an additional experiment to validate the main result when the covariance functions between different labels do not have the same eigenfunctions. The setting in this section does not strictly satisfy our assumptions, therefore it is outside the scope of our theory. However, to investigate the potential applicability of our theory, we perform this experiment with different hyper-parameter choices.

S8.1 Experimental Setting

As in Section 4, we generate functional data from two groups with labels $\{-1, 1\}$. For each group, we generate n functions on $\mathcal{T} = [0, 1]$ with two orthogonal bases:

$$\begin{cases} \phi_0(t) = 1, \quad \phi_j(t) = \sqrt{2}\sin(\pi j t), \quad \forall j \geq 1, \\ \psi_0(t) = 1, \quad \psi_j(t) = \sqrt{2}\cos(\pi j t), \quad \forall j \geq 1. \end{cases}$$

n is set from 1 to 3000. For a label $+1$, we generate functional data $X_{i+}(t) = \sum_{j=0}^{50}(\theta_j^{1/2} Z_{j+} + \mu_{j+})\phi_j(t)$ with random variables Z_{j+} and coefficients θ_j, μ_{j+} for $j = 0, 1, ..., 50$ and $i = 1, ..., n$. For a label -1, we generate $X_{i-}(t) = \sum_{j=0}^{50}(\theta_j^{1/2} Z_{j-} + \mu_{j-})\psi_j(t)$ with random variables Z_{j-} and coefficients μ_{j-}.

That is, in *Scenario 1*, we set $\theta_j = j^{-2}$, $\mu_{j-} = 0$, and change $\mu_{j+} = j^{-\gamma}$ and draw Z_{j+}, Z_{j-} from standard normal Gaussian. Then we determined the DH condition based on whether the gamma was greater or less than $3/2$. In *Scenario 2*, we set $\theta_j = j^{-2}$, $\mu_{j-} = 0$ and adjust $\mu_{j+} = \mathbf{1}\{j = 0\}\mu$, and let Z_{j+}, Z_{j-} be sample from uniform distribution on $[-1/2, 1/2]$. Although it is analytically challenging to specify when the HM condition is violated because of the different basis functions, we present the results of our experiments for various μ. In Scenario 2, because of the difficulty of rigorously checking the DH condition in the setting, we examined a broader range of $\mu \in \{1.5, 1.7, 1.9, 2.1\}$.

Other settings are the same as Section 4. For each n, we newly generate

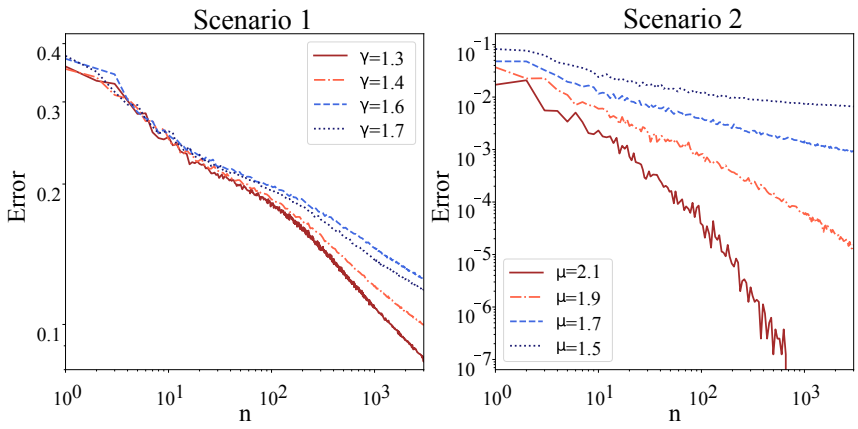

Figure 1: Error (logarithm of misclassification error rate) by the RKHS against $\log n$. Left: Scenario 1 for the Delaigle–Hall condition with $\gamma \in \{1.3\,(\text{solid}), 1.4\,(\text{dashes}), 1.6\,(\text{dots}), 1.7\,(\text{dotdash})\}$. Right: Scenario 2 for the hard-margin condition with $\mu \in \{2.1\,(\text{solid}), 1.9\,(\text{dashes}), 1.7\,(\text{dot}), 1.5\,(\text{dotdash})\}$.

1000 test data and calculate misclassification rates with the test data. Each simulation experiment is repeated 200 times, and the average value is reported. We investigate the classification error of the RKHS classifier with the Gaussian kernel and the logit loss. The tuning parameters are chosen by cross-validation.

S8.2 Result

The results are shown in Figure 1. In Scenario 1, we see a difference in convergence speed for each value of γ, although the difference is not so clear as in Figure 1 in the main text. The reason is that the conditions we are actually checking are different from those we should impose, so there must be a differ-

ence in scale. However, as n increases to some extent, e.g. $n \geq 100$, we can observe an exponential-like fast convergence. In Scenario 2, the results confirm exponential convergence for large μ, and as μ gets smaller, the rate falls off as in polynomial convergence.

Bibliography

Gao, F., Hannig, J., Lee, T.-Y. and Torcaso, F. (2004) Exact l 2 small balls of gaussian processes, *Journal of Theoretical Probability*, **17**, 503–520.

Gao, F. and Wellner, J. A. (2007) Entropy estimate for high-dimensional monotonic functions, *Journal of Multivariate Analysis*, **98**, 1751–1764.

Giné, E. and Nickl, R. (2016) *Mathematical foundations of infinite-dimensional statistical models*, Cambridge University Press.

Guntuboyina, A. and Sen, B. (2012) Covering numbers for convex functions, *IEEE Transactions on Information Theory*, **59**, 1957–1965.

Kanagawa, M., Hennig, P., Sejdinovic, D. and Sriperumbudur, B. K. (2018) Gaussian processes and kernel methods: A review on connections and equivalences, *arXiv preprint arXiv:1807.02582*.

Kuelbs, J., Li, W. V. and Linde, W. (1994) The gaussian measure of shifted balls, *Probability Theory and Related Fields*, **98**, 143–162.

Srebro, N. and Sridharan, K. (2010) Note on refined dudley integral covering number bound, *Unpublished results. http://ttic. uchicago. edu/karthik/dudley. pdf*.

Steinwart, I. and Christmann, A. (2008) *Support vector machines*, Springer Science & Business Media.

Tsybakov, A. B. (2008) Introduction to nonparametric estimation.

Van Der Vaart, A. W. and Wellner, J. A. (1996) *Weak convergence and empirical processes*, Springer.

Williams, C. K. and Rasmussen, C. E. (2006) *Gaussian processes for machine learning*, vol. 2, MIT press Cambridge, MA.