Journal of Machine Learning Research 23 (2022) 1-24 Submitted 8/21; Revised 3/22; Published 5/22

Statistical Rates of Convergence for Functional Partially Linear Support Vector Machines for Classification

Yingying Zhang
Academy of Statistics and Interdisciplinary Sciences, KLATASDS-MOE
East China Normal University, Shanghai, China

Yan-Yong Zhao
School of Statistics and Data Science
Nanjing Audit University, Nanjing, China

Heng Lian HENGLIAN@CITYU.EDU.HK
City University of Hong Kong Shenzhen Research Institute
Shenzhen, China
and
Department of Mathematics
City University of Hong Kong
Kowloon Tong, Hong Kong, China

Editor: Corinna Cortes

Abstract

In this paper, we consider the learning rate of support vector machines with both a functional predictor and a high-dimensional multivariate vectorial predictor. Similar to the literature on learning in reproducing kernel Hilbert spaces, a source condition and a capacity condition are used to characterize the convergence rate of the estimator. It is highly non-trivial to establish the possibly faster rate of the linear part. Using a key basic inequality comparing losses at two carefully constructed points, we establish the learning rate of the linear part which is the same as if the functional part is known. The proof relies on empirical processes and the Rademacher complexity bound in the semi-nonparametric setting as analytic tools, Young's inequality for operators, as well as a novel "approximate convexity" assumption.

Keywords: Convergence rate; Prediction risk; Rademacher complexity; Support vector classification.

1. Introduction

Binary classification based on support vector machines (SVM) is by now a popular and mature statistical tool for pattern recognition. It was originally posed as a margin-maximization procedure and later found to be consistent with the standard loss+penalty paradigm often appearing in traditional statistical procedures. In particular, based on the penalized hinge loss formulation, various penalties have been used to deal with different problem structures (Wu and Zhou, 2006; Luo et al., 2015), with the ridge penalty and the squared RKHS (reproducing kernel Hilbert space) penalty corresponding to the standard linear and kernel SVM, respectively.

In this paper, we consider the setting that the features for prediction consist of a functional part (an observed curve) and a high-dimensional multivariate part. For example, in a disease presence prediction study, a high-dimensional vectorial predictor may come from genetic sequence information (such as SNPs) while other measurements such as brain imaging data may be regarded as the (two or three-dimensional) functional predictor. Kong et al. (2016) considered a regression problem where daily concentration measurements of PM2.5 is used as a functional predictor and the scalar covariates are obtained from the U.S. census for each city including factors such as land area per individual and water area per individual, among many others. Ma et al. (2019) established the association between the mini-mental state examination scores (response variable) and the brain volume imaging data of 20 regions of interest (ROI) (functional predictor), while accounting for the scalar covariates, including the selected 1071 SNPs and some prognostic-related covariates.

Classification with functional data has been widely studied in the statistical and machine learning literature (Delaigle and Hall, 2012; Preda et al., 2007; Chamroukhi et al., 2013), while here we are particularly interested in the learning rate when using an SVM as the prediction tool, which has not been studied before.

More specifically, given an i.i.d. sample of observations $(X_i, \mathbf{z}_i, y_i)$, $i = 1, \ldots, n$, with $X_i \in L^2(\mathcal{T})$, $\mathbf{z}_i \in \mathbb{R}^p$, $y_i \in \{-1, 1\}$, and $\mathcal{T}$ is a compact set of a certain Euclidean space (without much of loss of generality, we will always assume $\mathcal{T} = [0, 1]$), we consider the following optimization problem

$$(\widehat{f}, \widehat{\boldsymbol{\beta}}) = \arg\min_{f, \boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} h(y_i(\langle X_i, f\rangle + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta})) + \lambda_1 \|f\|^2 + \lambda_2 \|\boldsymbol{\beta}\|_1.$$

In the above, $h(x) = (1-x)_+$ is the hinge loss, $\langle X_i, f\rangle = \int_{\mathcal{T}} X_i(t) f(t) dt$ is the functional part and $\boldsymbol{\beta} \in \mathbb{R}^p$ is the coefficient in the linear part. A quadratic penalty on $\|f\|^2 = \int_{\mathcal{T}} f^2(t) dt$ is used for regularization which is important since $L^2(\mathcal{T})$ is an infinite-dimensional space, and optimization without penalty will make the problem ill-posed (Ramsay and Dalzell, 1991). Since we mainly consider the case that p is large (possibly larger than n), a lasso penalty exploiting the potential sparseness property of the unknown high-dimensional coefficient vector is suitable and popularly used in high-dimensional learning.

Our theoretical results are in some sense similar to the recent work Xia et al. (2021), which studied non-functional partially linear SVM classification that considered a nonparametric function $f(x_i)$ in a reproducing kernel Hilbert space framework, in place of our $\langle X_i, f\rangle$. It is known that functional regression and nonparametric kernel regression has strong similarities (see Lin and Rosasco (2017) which proposed a uniform treatment of both). However, we note the following key differences and our distinctive contributions.

- We can establish a non-trivial learning rate even without any smoothness assumption on f (except it is in $L^2(\mathcal{T})$). We will show later that we can also adapt a RKHS framework for f in our model. Smoothness assumption for our f as the coefficient for the functional predictor is much more natural than the corresponding assumption in Xia et al. (2021) for the nonparametric function estimation problem. With details presented in Section 5, we show that in the setting of Xia et al. (2021), their assumption that f is in a RKHS is usually very stringent, even when they consider it only as

an approximation to the true function. While in the functional setting, we naturally can construct nontrivial examples that f can be smooth.

- For the nonparametric part, the rates in Xia et al. (2021) are characterized by a single parameter indicating the capacity of the function space. We instead use an additional source condition that characterizes the smoothness of f. Such a source condition is often used in previous theoretical studies of kernel learning problems, but restricted to the least-squares loss (Gu, 2013). As far as we know, our work is the first study that uses a source condition in non-least-squares loss which is incorporated into our theoretical investigations using Young's inequality for operators (Suzuki and Sugiyama, 2013).

- Most importantly, Xia et al. (2021) only established an overall rate for the sum of the nonparametric and the linear part (they stated the rate for the linear part but it is the same as the overall rate). Indeed, they considered a rate for the linear part that does not depend on the capacity parameter of the RKHS as an open problem in their discussion section at the end of the paper. It's non-trival to remove the effect of the functional part. To achieve this goal, we propose a novel "approximate convexity" assumption and based on this assumption, the key innovative step is to use the basic inequality comparing losses at two carefully constructed points. We are not aware of similar constructions in the literature concerning functional data classification.

The rest of the article is organized as follows. In Section 2, we establish the overall rate of estimation based on empirical processes and Rademacher complexity. Section 3 establishes the rate for the linear part, which is the same as if the functional part is known. Section 4 provides clarifications and sufficient conditions for a key local strong convexity assumption used. In Section 5, we conclude the paper and discuss a RKHS framework for the functional coefficient f and argue that the same proof also works for this setting. We also carefully discuss the problem considered in Xia et al. (2021) in relation to ours, in particular stating the rates we can obtain using their setting, which extends the results in Xia et al. (2021).

2. Functional estimation in the L^2 space with a high-dimensional linear part

Without loss of generality we assume $E[X] = 0$ and define the covariance operator $\Gamma = E[X \otimes X]$, where for $f, g \in L^2([0,1])$, $f \otimes g$ denotes the operator that maps $h \in L^2([0,1])$ to $\langle g, h\rangle f \in L^2([0,1])$. We assume $E\|X\|^2 < \infty$ which guarantees that Γ is a compact trace operator. By Mercer's theorem, we have the spectral decomposition

$$\Gamma = \sum_{j=1}^{\infty} s_j e_j \otimes e_j,$$

where s_j is the sequence of eigenvalues decreasing to zero, and $\{e_j\}$ is an orthonormal basis of $L^2([0,1])$.

For any $(f, \boldsymbol{\beta})$, with $(f_0, \boldsymbol{\beta}_0)$ denoting the target "true" parameters (defined in assumption (A1) later), the error for which we will establish the learning rate is given by

$E[(\langle X^*, f - f_0\rangle + \mathbf{z}^{*T}(\boldsymbol{\beta} - \boldsymbol{\beta}_0))^2]$ where the expectation is taken over $(X^*, \mathbf{z}^*)$ which is an independent copy of $(X_i, \mathbf{z}_i)$. For the functional term, the error will depend on two quantities. One is the smoothness of f_0 and the other is the capacity of the hypothesis space which is the space of the linear functions of X of the form $\langle X, f\rangle, f \in L^2([0,1])$. For the former, we assume the so-called "source condition" that for some constant $C > 0$ (throughout the paper C will denote a generic positive constant),

$$\|\Gamma^{-r} f_0\| \leq C \text{ for some } r \in [0, 1/2]. \tag{1}$$

For the latter, we define $\mathcal{R}(u) = \sqrt{\frac{1}{n}\sum_j \min\{s_j, u^2\}}$, which turns out to be an upper bound of the local Rademacher complexity of the space of linear functions mentioned above (Lemma 1). We note that for $r = 0$, we do not assume any further smoothness beyond $f_0 \in L^2([0,1])$. Writing $f_0 = \sum_j f_{0j} e_j$ with f_{0j} the Fourier coefficients in the basis $\{e_j\}$, the source condition is equivalent to $\sum_j f_{0j}^2 / s_j^{2r} < \infty$ and it is readily seen that larger r requires faster convergence of the Fourier coefficients to zero, and thus can be regarded as a stronger smoothness condition for f_0. Increasing r beyond $r = 1/2$ does not make the rate faster which is a well-known saturation effect in the regularized estimation (Lin et al., 2017; Lin and Rosasco, 2017; Lin and Cevher, 2020). Since we assume $f_0 \in L^2$, we only study the attainable case here. The non-attainable case corresponds to the setting that $f_0 \notin L^2$, which seems significantly harder and we leave it for future investigation.

Formally, we take up the following technical assumptions.

(A1) $(X_i, \mathbf{z}_i, y_i)$ is an i.i.d. sample and $(f_0, \boldsymbol{\beta}_0) = \arg\min_{f,\boldsymbol{\beta}} E[h(y_i(\langle X_i, f\rangle + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}))]$.

(A2) Both $\|X\|$ and components of $\mathbf{z}$ are sub-Gaussian.

(A3) f_0 satisfies the source condition (1) with some $r \in [0, 1/2]$. $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})^{\mathrm{T}}$ is a s-sparse vector with support $S = \{j : \beta_{0j} \neq 0\}$.

(A4) There exist constants $C_1, C_2 > 0$ and $\xi \geq 2$, such that for any $f, \boldsymbol{\beta}$, $E[h(y(\langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}))] - E[h(y(\langle X, f_0\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0))] \geq C_1 E[(\langle X, f - f_0\rangle + \mathbf{z}^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0))^2]^{\frac{\xi}{2}}$, and $E[(\langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta})^2] \asymp \|\Gamma^{1/2} f\|^2 + \|\boldsymbol{\beta}\|^2$, if $E[(\langle X, f - f_0\rangle + \mathbf{z}^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0))^2] \leq C_2$.

Existence and uniqueness of minimizer of $E[h(y_i(\langle X_i, f\rangle + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}))]$ provide a target for estimation. For linear models, existence and uniqueness was investigated in Koo et al. (2008). Following that work, later we also give a concrete example when $(X, \mathbf{z})$ is Gaussian for which the closed-form expression of $(f_0, \boldsymbol{\beta}_0)$ is available. Sub-Gaussian assumption as (A2) is commonly-used in high dimensional analysis (Loh and Wainwright, 2015) for theoretical convenience. (A2) is satisfied if $(X, \mathbf{z})$ is jointly Gaussian, but also allows other short-tailed distributions. The first part of (A3) imposes smoothness condition on f_0 while the second part imposes sparseness of the high-dimensional vector. As mentioned in the introduction, we believe smoothness of f_0 in the functional setting is much more reasonable (as illustrated by the concrete Gaussian example) than the corresponding similar assumption for the standard non-functional nonlinear SVM, with the detailed arguments presented in Section 5. (A4) was also used in Xia et al. (2021) where further references are provided with sufficient conditions for it to hold. We believe that $\xi = 2$ is the more typical case, assuming the risk difference to be locally strongly convex.

Although we do not explicitly impose assumptions on $\mathcal{R}(u)$ which is related to the capacity of the hypothesis space, the learning rate is dependent on $\mathcal{R}(u)$ through the positive scalar u_n that satisfies $\mathcal{R}(u_n) = u_n^{2+2r}$ (there is a unique such positive value (Bartlett et al., 2005)).

Theorem 1 *Under assumptions (A1)-(A4), setting $\lambda_1 = Cu_n^2$, and $\lambda_2 = C\sqrt{\log p/n}$ for C sufficiently large, we have*

$$\begin{aligned}\|\Gamma^{1/2}(\widehat{f} - f_0)\|^2 + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 &\lesssim \left(E[h(y(\langle X, \widehat{f}\rangle + \mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}))] - E[h(y(\langle X, f_0\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0))]\right)^{\frac{2}{\xi}} \\ &= O_p\left(\left(u_n^{2+4r} + \frac{s\log p}{n}\right)^{\frac{1}{\xi-1}}\right).\end{aligned}$$

As a special case, if $s_j \asymp j^{-\alpha}$ for some $\alpha > 1$, we have

$$\|\Gamma^{1/2}(\widehat{f} - f_0)\| + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p\left(\left(n^{-\frac{\alpha(1+2r)}{2(\alpha(1+2r)+1)}} + \sqrt{\frac{s\log p}{n}}\right)^{\frac{1}{\xi-1}}.\right)$$

The rest of this section is devoted to the proof of Theorem 1. We first state and prove the following bound for a semiparametric version of Rademacher complexity.

Lemma 1 *For any $u > 0, v > 0$,*

$$E\left[\sup_{\substack{f\in L^2([0,1])\\ \boldsymbol{\beta}\in\mathbb{R}^p}} \frac{(1/n)\sum_i \sigma_i(\langle X_i, f\rangle + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta})}{\|\Gamma^{1/2}f\|/u + \|f\| + \|\boldsymbol{\beta}\|_1/v}\right] \le C\left(\mathcal{R}(u) + \sqrt{\frac{\log p}{n}}v\right),$$

where $\sigma_i \in \{-1, 1\}, i = 1, \ldots, n$ are i.i.d. Rademacher variables and

$$\mathcal{R}(u) = \left(\frac{1}{n}\sum_{j=1}^{\infty}\min\{s_j, u^2\}\right)^{1/2}.$$

We also have

$$E\left[\sup_{\boldsymbol{\beta}, f\in L^2([0,1])}\left|(P - P_n)\frac{h(y(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta} + \langle X, f\rangle)) - h(y(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0 + \langle X, f_0\rangle))}{\|\Gamma^{1/2}(f - f_0)\|/u + \|f - f_0\| + \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1/v}\right|\right] \le C\left(\mathcal{R}(u) + \sqrt{\frac{\log p}{n}}v\right),$$

where P_n is the empirical measure on the observed data and P is the corresponding population measure.

Proof of Lemma 1. First, we have

$$E\left[\sup_{\substack{\|\Gamma^{1/2}f\|\le u\\ \|f\|\le 1}} (1/n)\sum_i \sigma_i\langle X_i, f\rangle\right] \le C\mathcal{R}(u). \tag{2}$$

The proof for the above is basically the same as Theorem 41 of Mendelson (2002) adapted to the functional context, which we also present here for completeness. Indeed, we have the the Karhunen-Loéve expansion $X_i = \sum_j \eta_{ij} e_j$ with $E\eta_{ij}^2 = s_j$, and we can also expand $f = \sum_j f_j e_j$ with $f_j = \langle f, e_j \rangle$ being the Fourier coefficients of f. Thus we have

$$\begin{aligned} & E\left[\sup_{f:\|\Gamma^{1/2}f\|\le u, \|f\|\le 1} \frac{1}{n}\sum_{i=1}^n \sigma_i \langle X_i, f\rangle\right] \\ = \; & E\left[\sup_{f:\sum_j s_j f_j^2 \le u^2, \sum_j f_j^2 \le 1} \frac{1}{n}\sum_{i=1}^n \sigma_i \langle \sum_j \eta_{ij} e_j, \sum_j f_j e_j\rangle\right] \\ = \; & E\left[\sup_{f:\sum_j s_j f_j^2 \le u^2, \sum_j f_j^2 \le 1} \frac{1}{n}\sum_{j=1}^\infty \sum_{i=1}^n \sigma_i \eta_{ij} f_j\right]. \end{aligned}$$

We bound the first moment of the above by the squared root of the second moment. Since $\{f : \sum_j s_j f_j^2 \le u^2, \sum_j f_j^2 \le 1\} \subseteq \{f : \sum_j \max\{1, \frac{s_j}{u^2}\} f_j^2 \le 2\}$, denoting $\nu_j = \max\{1, \frac{s_j}{u^2}\}$, we have that, using the Cauchy-Schwarz inequality,

$$\begin{aligned} & E\left[\left(\sup_{f:\sum_j s_j f_j^2 \le u^2, \sum_j f_j^2 \le 1} \frac{1}{n}\sum_{j=1}^\infty \sum_{i=1}^n \sigma_i \eta_{ij} f_j\right)^2\right] \\ \le \; & E\left[\left(\sup_{f:\sum_j \nu_j f_j^2 \le 2} \frac{1}{n}\sum_{j=1}^\infty \sum_{i=1}^n \frac{\sigma_i \eta_{ij}}{\sqrt{\nu_j}} \sqrt{\nu_j} f_j\right)^2\right] \\ \le \; & \frac{2}{n^2} E\left[\sum_{j=1}^\infty \left(\sum_{i=1}^n \frac{\sigma_i \eta_{ij}}{\sqrt{\nu_j}}\right)^2\right] \\ = \; & \frac{2}{n}\sum_j \frac{s_j}{\nu_j} = \frac{2}{n}\sum_j \min\{s_j, u^2\}. \end{aligned}$$

This finishes the proof of (2).

For the linear part, we have

$$\begin{aligned} & E\left[\sup_{\|\boldsymbol{\beta}\|_1 \le v} (1/n)\sum_i \sigma_i \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\beta}\right] \\ \le \; & E\left[\|(1/n)\sum_i \sigma_i \mathbf{z}_i\|_\infty\right] \sup_{\|\boldsymbol{\beta}\|_1 \le v} \|\boldsymbol{\beta}\|_1 \\ \le \; & C\sqrt{\frac{\log p}{n}} v, \end{aligned} \tag{3}$$

using the sub-Gaussianity of the components of $\mathbf{z}$.

Then, the standardized version $f' := \frac{f}{\|\Gamma^{1/2}f\|/u+\|f\|}$ satisfies $\|\Gamma^{1/2}f'\| \le u$ and $\|f'\| \le 1$, and thus (2) leads to

$$E\left[\sup_f \frac{(1/n)\sum_i \sigma_i\langle X_i, f\rangle}{\|\Gamma^{1/2}f\|/u+\|f\|}\right] \le E\left[\sup_{\substack{\|\Gamma^{1/2}f'\|\le u\\ \|f'\|\le 1}} (1/n)\sum_i \sigma_i\langle X_i, f'\rangle\right] \le C\mathcal{R}(u). \tag{4}$$

Similarly (3) leads to

$$E\left[\sup_{\boldsymbol{\beta}} \frac{(1/n)\sum_i \sigma_i \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_1/v}\right] \le C\sqrt{\frac{\log p}{n}}v. \tag{5}$$

The first part of the lemma is proved by combining (4) and (5).

To obtain the second part of the lemma, we use the symmetrization argument (Pollard, 2012) and the contraction inequality for the Rademacher complexity (Theorem 2.2 of Koltchinskii (2011)) to get

$$\begin{aligned}
& E\left[\sup_{\boldsymbol{\beta},f}\left|(P-P_n)\frac{h(y(\langle X_i,f\rangle+\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}))-h(y(\langle X_i,f_0\rangle+\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0))}{u^{-1}\|\Gamma^{1/2}(f-f_0)\|+\|f-f_0\|+v^{-1}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1}\right|\right] \\
\le\ & CE\left[\sup_{\boldsymbol{\beta},f}\left|\frac{(1/n)\sum_i \sigma_i\{h(y(\langle X_i,f\rangle+\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}))-h(y(\langle X_i,f_0\rangle+\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0))\}}{u^{-1}\|\Gamma^{1/2}(f-f_0)\|+\|f-f_0\|+v^{-1}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1}\right|\right] \\
\le\ & CE\left[\sup_{\boldsymbol{\beta},f}\left|\frac{(1/n)\sum_i \sigma_i(\langle X_i,f-f_0\rangle+\mathbf{z}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0))}{u^{-1}\|\Gamma^{1/2}(f-f_0)\|+\|f-f_0\|+v^{-1}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1}\right|\right] \\
\le\ & C\left(\mathcal{R}(u)+\sqrt{\frac{\log p}{n}}v\right).
\end{aligned}$$

This finishes the proof. □

The next lemma uses concentration inequality to obtain the high probability bound corresponding to the expectation bound of Lemma 1. For technical reasons, an additional term $\sqrt{s}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|/v$ in the denominator is necessary.

Lemma 2 *With probability at least* $1-\exp\{-\min\{n\frac{H(u,v)}{(1+v)\log(p\vee n)}, n\frac{H^2(u,v)}{(u+v/\sqrt{s})^2}\}\}$,

$$\sup_{\boldsymbol{\beta},f\in\mathcal{H}}\left|(P-P_n)\frac{h(y(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}+\langle X,f\rangle))-h(y(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0+\langle X,f_0\rangle))}{\|\Gamma^{1/2}(f-f_0)\|/u+\|f-f_0\|+\sqrt{s}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|/v+\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1/v}\right| \le CH(u,v),$$

where $H(u,v)=\mathcal{R}(u)+\sqrt{\frac{\log p}{n}}v$.

Proof of Lemma 2. Since the hinge loss is Lipschitz, using that

$$\begin{aligned}
& \left|\frac{h(y(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}+\langle X,f\rangle))-h(y(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0+\langle X,f_0\rangle))}{\|\Gamma^{1/2}(f-f_0)\|/u+\|f-f_0\|+\sqrt{s}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|/v+\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1/v}\right| \\
\le\ & C\left|\frac{\langle X,f-f_0\rangle+\mathbf{z}^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}{\|\Gamma^{1/2}(f-f_0)\|/u+\|f-f_0\|+\sqrt{s}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|/v+\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1/v}\right| \\
\le\ & C(\|X\|+v\|\mathbf{z}\|_\infty),
\end{aligned}$$

and

$$\begin{aligned} & Var\left(\frac{h(y(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}+\langle X,f\rangle))-h(y(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0+\langle X,f_0\rangle))}{\|\Gamma^{1/2}(f-f_0)\|/u+\|f-f_0\|+\sqrt{s}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|/v+\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1/v}\right) \\ \leq & CVar\left(\frac{\langle X,f-f_0\rangle+\mathbf{z}^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}{\|\Gamma^{1/2}(f-f_0)\|/u+\|f-f_0\|+\sqrt{s}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|/v+\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1/v}\right) \\ \leq & C(u^2+v^2/s), \end{aligned}$$

by the Adamczak bound (pages 24–25 of Koltchinskii (2011)), we have

$$\begin{aligned} & \sup_{\boldsymbol{\beta},f}\left|(P-P_n)\frac{h(y(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}+\langle X,f\rangle))-h(y(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0+\langle X,f_0\rangle))}{\|\Gamma^{1/2}(f-f_0)\|/u+\|f-f_0\|+\sqrt{s}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|/v+\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1/v}\right| \\ \leq & CE\left[\sup_{\boldsymbol{\beta},f}\left|(P-P_n)\frac{h(y(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}+\langle X,f\rangle))-h(y(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0+\langle X,f_0\rangle))}{\|\Gamma^{1/2}(f-f_0)\|/u+\|f-f_0\|+\sqrt{s}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|/v+\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1/v}\right|\right] \\ & +C(u+v/\sqrt{s})\sqrt{t/n}+C(\|\max_i\|X_i\|\|_{\psi_1}+v\|\max_i\|\mathbf{z}_i\|_\infty\|_{\psi_1})(t/n), \end{aligned}$$

with probability at least $1-e^{-t}$, where $\|.\|_{\psi_1}$ is the Orlicz norm associated with the function $\psi_1(x)=e^x-1$. By Lemma 2.2.2 of Vaart and Wellner (1996), we have $\|\max_i\|\mathbf{z}_i\|_\infty\|_{\psi_1}\leq C\log(n\vee p)$ and $\|\max_i\|X_i\|\|_{\psi_1}\leq C\log n$. By setting $t=\min\{n\frac{H(u,v)}{(1+v)\log(p\vee n)},n\frac{H^2(u,v)}{(u+v/\sqrt{s})^2}\}$, we complete the proof. □

Our last lemma is the key lemma that takes into account the source condition.

Lemma 3 *Assume that* (1) *holds for some* $r\in[0,1/2]$. *For any* $f\in L^2([0,1])$ *and* $\lambda>0$, $|\lambda\langle f_0,f-f_0\rangle|\leq C(\lambda^{\frac{1}{2}+r}\|\Gamma^{1/2}(f-f_0)\|+\lambda^{1+r}\|f-f_0\|)$.

Proof of Lemma 3.

$$\begin{aligned} |\lambda\langle f_0,f-f_0\rangle| &= \lambda|\langle\Gamma^{-r}f_0,\Gamma^r(f-f_0)\rangle| \\ &\leq C\lambda^{\frac{1}{2}+r}\|\lambda^{\frac{1}{2}-r}\Gamma^r(f-f_0)\| \\ &\leq C\lambda^{\frac{1}{2}+r}\sqrt{\langle f-f_0,((1-2r)\lambda+2r\Gamma)(f-f_0)\rangle} \\ &\leq C(\lambda^{\frac{1}{2}+r}\|\Gamma^{1/2}(f-f_0)\|+\lambda^{1+r}\|f-f_0\|), \end{aligned}$$

where the second line uses assumption (1), and the third line uses Young's inequality for positive operators $\lambda^{1-2r}\Gamma^{2r}\leq(1-2r)\lambda+2r\Gamma$ for $r\in[0,1/2]$. □

Proof of Theorem 1. We have

$$\begin{aligned} & \frac{1}{n}\sum_i h(y(\mathbf{z}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}+\langle X_i,\widehat{f}\rangle))+\lambda_1\|\widehat{f}\|^2+\lambda_2\|\widehat{\boldsymbol{\beta}}\|_1 \\ \leq & \frac{1}{n}\sum_i h(y(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}_0+\langle X_i,f_0\rangle))+\lambda_1\|f_0\|^2+\lambda_2\|\boldsymbol{\beta}_0\|_1. \end{aligned}$$

Using Lemma 2 and 3, with probability at least $1-\exp\{-\min\{n\frac{H(u,v)}{(1+v)\log(p\vee n)}, n\frac{H^2(u,v)}{(u+v/\sqrt{s})^2}\}\}$,

$$\begin{aligned}
& E[h(y(\mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}+\langle X,\widehat{f}\rangle))]-E[h(y(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0+\langle X,f_0\rangle))]+\lambda_2\|\widehat{\boldsymbol{\beta}}\|_1 \\
\leq\ & \lambda_1\|f_0\|^2-\lambda_1\|\widehat{f}\|^2+\lambda_2\|\boldsymbol{\beta}_0\|_1+C\frac{H(u,v)}{u}\|\Gamma^{1/2}(\widehat{f}-f_0)\|+CH(u,v)\|\widehat{f}-f_0\| \\
& +C\frac{H(u,v)}{v/\sqrt{s}}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|+C\frac{H(u,v)}{v}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1 \\
=\ & -2\lambda_1\langle f_0,\widehat{f}-f_0\rangle-\lambda_1\|\widehat{f}-f_0\|^2+\lambda_2\|\boldsymbol{\beta}_0\|_1+C\frac{H(u,v)}{u}\|\Gamma^{1/2}(\widehat{f}-f_0)\|+CH(u,v)\|\widehat{f}-f_0\| \\
& +C\frac{H(u,v)}{v/\sqrt{s}}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|+C\frac{H(u,v)}{v}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1 \\
\leq\ & C\lambda_1^{1/2+r}\|\Gamma^{1/2}(\widehat{f}-f_0)\|+C\lambda_1^{1+r}\|\widehat{f}-f_0\|-\lambda_1\|\widehat{f}-f_0\|^2+\lambda_2\|\boldsymbol{\beta}_0\|_1 \\
& +C\frac{H(u,v)}{u}\|\Gamma^{1/2}(\widehat{f}-f_0)\|+CH(u,v)\|\widehat{f}-f_0\|+C\frac{H(u,v)}{v/\sqrt{s}}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|+C\frac{H(u,v)}{v}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1 \\
\leq\ & C(\lambda_1^{1/2+r}+\frac{H(u,v)}{u})\|\Gamma^{1/2}(\widehat{f}-f_0)\|+C(\lambda_1^{1+2r}+\frac{H^2(u,v)}{\lambda_1})-\frac{\lambda_1}{2}\|\widehat{f}-f_0\|^2+\lambda_2\|\boldsymbol{\beta}_0\|_1 \\
& +C\frac{H(u,v)}{v/\sqrt{s}}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|+C\frac{H(u,v)}{v}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1, \qquad (6)
\end{aligned}$$

where the last inequality uses $uv\leq u^q/q+v^p/p$ with $1/q+1/p=1$. Setting $u=u_n$ (u_n as defined just before the statement of Theorem 1) and $v=u_n^{2+2r}\sqrt{n/\log p}$, with $\lambda_1=Cu_n^2$, $\lambda_2=C\sqrt{\log p/n}$, it is easy to see that

$$C\frac{H(u_n,v)}{v}=C\left(\frac{\mathcal{R}(u_n)+\sqrt{\log p/n}v}{v}\right) \ = \ C\left(\frac{u_n^{2+2r}}{u_n^{2+2r}\sqrt{n/\log p}}+\sqrt{\log p/n}\right)\leq\lambda_2/2.$$

Thus (6) implies

$$\begin{aligned}
& E[h(y(\mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}+\langle X,\widehat{f}\rangle))]-E[h(y(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0+\langle X,f_0\rangle))]+\lambda_2\|\widehat{\boldsymbol{\beta}}\|_1+\frac{\lambda_1}{2}\|\widehat{f}-f_0\|^2 \\
\leq\ & C(\lambda_1^{1/2+r}+\frac{H(u_n,v)}{u_n})\|\Gamma^{1/2}(\widehat{f}-f_0)\|+C(\lambda_1^{1+2r}+\frac{H^2(u_n,v)}{\lambda_1})+\lambda_2\|\boldsymbol{\beta}_0\|_1 \\
& +C\lambda_2\sqrt{s}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|+\frac{\lambda_2}{2}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1.
\end{aligned}$$

Using $\|\widehat{\boldsymbol{\beta}}\|_1=\|\widehat{\boldsymbol{\beta}}_S\|_1+\|\widehat{\boldsymbol{\beta}}_{S^c}\|_1$, $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1=\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|_1+\|\widehat{\boldsymbol{\beta}}_{S^c}-\boldsymbol{\beta}_{0S^c}\|_1$ and $\|\boldsymbol{\beta}_{0S}\|_1-\|\widehat{\boldsymbol{\beta}}_S\|_1\leq\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|_1$, we then have

$$\begin{aligned}
& E[h(y(\mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}+\langle X,\widehat{f}\rangle))]-E[h(y(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0+\langle X,f_0\rangle))]+\lambda_2\|\widehat{\boldsymbol{\beta}}_{S^c}\|_1+\frac{\lambda_1}{2}\|\widehat{f}-f_0\|^2 \\
\leq\ & C(\lambda_1^{1/2+r}+\frac{H(u_n,v)}{u_n})\|\Gamma^{1/2}(\widehat{f}-f_0)\|+C(\lambda_1^{1+2r}+\frac{H^2(u_n,v)}{\lambda_1})+\lambda_2\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|_1 \\
& +C\lambda_2\sqrt{s}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|+\frac{\lambda_2}{2}\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|_1+\frac{\lambda_2}{2}\|\widehat{\boldsymbol{\beta}}_{S^c}-\boldsymbol{\beta}_{0S^c}\|_1,
\end{aligned}$$

leading to

$$\begin{aligned} & \gamma(\widehat{f},\widehat{\boldsymbol{\beta}}) + \frac{\lambda_2}{2}\|\widehat{\boldsymbol{\beta}}_{S^c}\|_1 + \frac{\lambda_1}{2}\|\widehat{f} - f_0\|^2 \\ \leq \;& C(\lambda_1^{1/2+r} + \frac{H(u_n,v)}{u_n})\|\Gamma^{1/2}(\widehat{f} - f_0)\| \\ & + C(\lambda_1^{1+2r} + \frac{H^2(u_n,v)}{\lambda_1}) + C\lambda_2\sqrt{s}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|, \end{aligned} \tag{7}$$

using $\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|_1 \leq \sqrt{s}\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\| \leq \sqrt{s}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|$, where $\gamma(\widehat{f},\widehat{\boldsymbol{\beta}}) = E[h(y(\mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \langle X, \widehat{f}\rangle))] - E[h(y(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0 + \langle X, f_0\rangle))]$ is the risk difference. By some re-arrangement, we have

$$\gamma(\widehat{f},\widehat{\boldsymbol{\beta}}) \leq C(\lambda_1^{1/2+r} + \frac{H(u_n,v)}{u_n} + \lambda_2\sqrt{s})(\|\Gamma^{1/2}(\widehat{f} - f_0)\| + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|) + C(\lambda_1^{1/2+r} + \frac{H(u_n,v)}{u_n})^2.$$

By assumption (A4) and using Young's inequality that $ab \leq a^{\frac{\xi}{\xi-1}}/\frac{\xi}{\xi-1} + b^{\xi}/\xi$, we then have

$$\gamma(\widehat{f},\widehat{\boldsymbol{\beta}}) \leq C(\lambda_1^{\frac{1}{2}+r} + \frac{H(u_n,v)}{u_n} + \lambda_2\sqrt{s})^{\frac{\xi}{\xi-1}} + c\gamma(\widehat{f},\widehat{\boldsymbol{\beta}}),$$

where $\xi \geq 2$ is defined in assumption (A4) and c is a constant that can be made smaller than 1 when applying Young's inequality. Thus

$$\gamma(\widehat{f},\widehat{\boldsymbol{\beta}}) \leq C(\lambda_1^{\frac{1}{2}+r} + \frac{H(u_n,v)}{u_n} + \lambda_2\sqrt{s})^{\frac{\xi}{\xi-1}} = C(u_n^{1+2r} + \sqrt{s\mathrm{log}p/n})^{\frac{\xi}{\xi-1}}.$$

This also implies

$$\|\Gamma^{1/2}(\widehat{f} - f_0)\| + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \leq C(u_n^{1+2r} + \sqrt{s\mathrm{log}p/n})^{\frac{1}{\xi-1}}.$$

Finally, note that when $s_j \asymp j^{-\alpha}$, we can easily calculate $\mathcal{R}(u) \asymp u^{1-1/\alpha}/\sqrt{n}$ and then $u_n \asymp n^{-\frac{\alpha}{2(\alpha(1+2r)+1)}}$. The bound above then becomes $(u_n^{1+2r} + \sqrt{\frac{s\mathrm{log}p}{n}})^{\frac{1}{\xi-1}} \asymp (n^{-\frac{\alpha(1+2r)}{2(\alpha(1+2r)+1)}} + \sqrt{\frac{s\mathrm{log}p}{n}})^{\frac{1}{\xi-1}}$. □

3. Rate for the linear part

The main purpose of this section is to show that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(\sqrt{s\mathrm{log}p/n})$. As discussed in the introduction, it is a non-trivial task to remove the effect of the functional part. We need the following additional assumptions.

(B1) Let $\mathbf{g}_0 = (g_{01}, \ldots, g_{0p})^{\mathrm{T}} \in (L^2([0,1]))^p$ with $g_{0j} = \arg\min_g E[\delta(1 - y(\langle X, f_0\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0))(z_j - \langle X, g\rangle)^2]$ where $\delta(.)$ is the Dirac delta function. We assume $\max_j \|g_{0j}\| \leq C$.

(B2) There exists some constant $C > 0$ such that, for any $f, \boldsymbol{\beta}$ with $\|\Gamma^{1/2}(f - f_0)\| + \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq \delta_n := Cn^{-1/4}$, we have $E[h(y(\langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0 + (\mathbf{z} - \langle X, \mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)))] - E[h(y(\langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0))] \geq C\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 - \delta_n^2\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|$, where $\langle X, \mathbf{g}_0\rangle = (\langle X, g_{01}\rangle, \ldots, \langle X, g_{0p}\rangle)^{\mathrm{T}}$.

Assumptions similar to (B1) often appear in theoretical studies of semiparametric models to make it possible to transform linear part to make it orthogonal to the nonparametric part (Li, 2000; Wang et al., 2009). Such orthogonalization is also implicit in (B2) when we subtract the effect of the functional part from $\mathbf{z}$. Some technical issues are hidden in the assumption (B2). Note that the delta function that appears in the definition of g_{0j} is important, which is related to the second derivative of the loss function. The critical role of the delta function is also highlighted in Section 4 where we discuss why this naturally appears to make proper orthogonalization. Seemingly Strangely, the way $\mathbf{g}_0$ is defined plays no role in our proof of Theorem 2. However, this issue is clarified in Section 4 that shows that (B2) can be expected to be satisfied only for such $\mathbf{g}_0$ as defined in (B1). As demonstrated in Section 4, the term $\delta_n^2\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|$ seems necessary to avoid the pathological case that $\|\Gamma^{1/2}(\widehat{f}-f_0)\|$ is much larger than $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|$ which would destroy the local convexity at $(f_0,\boldsymbol{\beta}_0)$. Such kind of "approximate convexity" also appears in works such as Loh and Wainwright (2015), where it arises for entirely different reasons.

We could assume $E[h(y(\langle X,f\rangle+\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0+(\mathbf{z}-\langle X,\mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)))]-E[h(y(\langle X,f\rangle+\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0))]\geq C\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^{\xi}-\delta_n^2\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|$ for some $\xi\geq 2$ as in (A4) with small changes in proof and convergence rate below, but $\xi=2$ seems to be most typical (a locally quadratic expected loss) and our concrete example to be presented in Section 4 shows the assumption (B2) is naturally satisfied. Thus in this section we only consider $\xi=2$.

Theorem 2 *Under the same assumptions as in Theorem 1 with $\xi=2$, and $u_n^2=O(\sqrt{\log p/n})$, and (B1) and (B2) hold, then $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|=O_p\left(\sqrt{\frac{s\log p}{n}}\right)$ and $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1=O_p\left(s\sqrt{\frac{\log p}{n}}\right)$.*

Proof of Theorem 2. First, we assume $u_n^{2+4r}\leq s\log p/n$. In this case, Theorem 1 immediately implies $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|=O_p(\sqrt{s\log p/n})$. Furthermore, (7) implies $\lambda_2\|\widehat{\boldsymbol{\beta}}_{S^c}\|_1\leq Cu_n^{2+4r}+C\lambda_2\sqrt{s}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|\leq Cs\log p/n$. Combining this with $\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|_1\leq\sqrt{s}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|$, we get $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1=O_p(s\sqrt{\log p/n})$.

Thus in the rest of the proof, we assume $u_n^{2+4r}>s\log p/n$ (the linear part has a faster rate). In this case, (7) implies $\|\widehat{f}-f_0\|$ is bounded. To ease notation, let $\mathcal{L}(f,\boldsymbol{\beta})=\frac{1}{n}\sum_i h(y_i(\langle X_i,f\rangle+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}))+\lambda_1\|f\|^2+\lambda_2\|\boldsymbol{\beta}\|_1$. The key innovative step for proving the rate of the linear part is to use the *basic inequality* (as Geer et al. (2000) calls such kind of comparison inequalities for other models)

$$\mathcal{L}(\widehat{f},\widehat{\boldsymbol{\beta}})\leq\mathcal{L}(\widehat{f}+\mathbf{g}_0^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0),\boldsymbol{\beta}_0),\tag{8}$$

which is true since $(\widehat{f},\widehat{\boldsymbol{\beta}})$ minimizes $\mathcal{L}$ and $g_{0j}\in L^2([0,1])$. The above can be written as

$$\begin{aligned}&P_nh(y(\langle X,\widehat{f}+\mathbf{g}_0^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\rangle+\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0+(\mathbf{z}-\langle X,\mathbf{g}_0\rangle)^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)))\\&-P_nh(y(\langle X,\widehat{f}+\mathbf{g}_0^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\rangle+\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0))+\lambda_2\|\widehat{\boldsymbol{\beta}}\|_1\\\leq\;&\lambda_1\|\widehat{f}+\mathbf{g}_0^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\|^2-\lambda_1\|\widehat{f}\|^2+\lambda_2\|\boldsymbol{\beta}_0\|_1.\end{aligned}$$

The next goal is to replace the empirical measure P_n above by the population measure P. Let $g(X,y,\mathbf{z};f,\boldsymbol{\beta})=h(y(\langle X,\widehat{f}+\mathbf{g}_0^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\rangle+\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0+(\mathbf{z}-\langle X,\mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)))-$

$h(y(\langle X, \widehat{f}+\mathbf{g}_0^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\rangle+\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0))$. We have, similar to the proof in Lemma 1, for any $u>0$,

$$\begin{aligned}
& E\Big[\sup_{\substack{\|\Gamma^{1/2}(f-f_0)\|\le u,\\ \|f-f_0\|\le C,\,\boldsymbol{\beta}}}(P_n-P)\frac{g(X,y,\mathbf{z};f,\boldsymbol{\beta})}{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1}\Big] \\
\le\ & CE\Big[\sup_{\substack{\|\Gamma^{1/2}(f-f_0)\|\le u,\\ \|f-f_0\|\le C,\,\boldsymbol{\beta}}}\frac{1}{n}\sum_i\sigma_i\frac{g(X_i,y_i,\mathbf{z}_i;f,\boldsymbol{\beta})}{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1}\Big] \\
\le\ & CE\Big[\sup_{\boldsymbol{\beta}}\frac{1}{n}\sum_i\sigma_i\frac{(\mathbf{z}_i-\langle X_i,\mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0))}{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1}\Big] \\
\le\ & C\sqrt{\frac{\log p}{n}}.
\end{aligned}$$

Using the concentration inequality (the Adamczak bound again), by $\|g\|^2\le C(\|X\|^2+\|\mathbf{z}\|^2)\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2$ and $|g|\le\|\mathbf{z}-\langle X,\mathbf{g}_0\rangle\|_\infty\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1$, we get that with probability $1-e^{-t}$,

$$\begin{aligned}
& \sup_{\substack{\|\Gamma^{1/2}(f-f_0)\|\le u,\\ \|f-f_0\|\le C,\,\boldsymbol{\beta}}}(P_n-P)\frac{g(X,y,\mathbf{z};f,\boldsymbol{\beta})}{\sqrt{s}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|+\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1} \\
\le\ & CE\Big[\sup_{\substack{\|\Gamma^{1/2}(f-f_0)\|\le u,\\ \|f-f_0\|\le C,\,\boldsymbol{\beta}}}(P_n-P)\frac{g(X,y,\mathbf{z};f,\boldsymbol{\beta})}{\sqrt{s}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|+\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1}\Big] \\
& +C\sqrt{\frac{t}{sn}}+C\|\max_i\|\mathbf{z}_i-\langle X_i,\mathbf{g}_0\rangle\|_\infty\|_{\psi_1}\frac{t}{n}.
\end{aligned}$$

Setting $t=\min\{s\log p,\frac{\sqrt{n\log p}}{\log(p\vee n)}\}$, we get

$$\sup_{\substack{\|f-f_0\|\le u,\\ \|f-f_0\|\le C,\,\boldsymbol{\beta}}}(P_n-P)\frac{g(X,y,\mathbf{z};f,\boldsymbol{\beta})}{\sqrt{s}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|+\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1}\le C\sqrt{\frac{\log p}{n}}.$$

Our assumption that $\|X\|$ is sub-Gaussian in particular means Γ is a trace operator with $\sum_{j=1}^\infty s_j<\infty$. It is then easy to verify that $\mathcal{R}(n^{-1/4})=O(n^{-1/2})$ and by the definition of u_n, our estimator always has convergence rate $O_p(n^{-1/4})$. Then, by assumption (B2),

$$\begin{aligned}
& C\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2-\delta_n^2\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|+\lambda_2\|\widehat{\boldsymbol{\beta}}\|_1 \\
\le\ & \lambda_1\|\widehat{f}+\mathbf{g}_0^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\|^2-\lambda_1\|\widehat{f}\|^2 \\
& +\lambda_2\|\boldsymbol{\beta}_0\|_1+C\sqrt{\frac{\log p}{n}}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1+C\sqrt{\frac{s\log p}{n}}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|. \qquad (9)
\end{aligned}$$

Noting that $\|\widehat{f}-f_0\|$ is bounded and f_0 is fixed implies $\|\widehat{f}\|$ is bounded, $\lambda_1\|\widehat{f}+\mathbf{g}_0^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\|^2-\lambda_1\|\widehat{f}\|^2=\lambda_1\|\mathbf{g}_0^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\|^2+2\lambda_1\langle\widehat{f},\mathbf{g}_0^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\rangle\le C\lambda_1\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1\le C\sqrt{\frac{\log p}{n}}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1$.

Note that $\delta_n \le Cn^{-1/4}$. From (9) and the Cauchy-Schwartz inequality $\sqrt{\frac{s\log p}{n}}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \le C\frac{s\log p}{n} + \frac{1}{4C}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2$, we have

$$C\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2 + \lambda_2\|\widehat{\boldsymbol{\beta}}\|_1 \le \lambda_2\|\boldsymbol{\beta}_0\|_1 + C\sqrt{\frac{\log p}{n}}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1 + C\frac{s\log p}{n}.$$

Using $\|\boldsymbol{\beta}\|_1 = \|\boldsymbol{\beta}_S\|_1 + \|\boldsymbol{\beta}_{S^c}\|_1$, the above displayed implies

$$\begin{aligned} & C\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2 + \lambda_2\|\widehat{\boldsymbol{\beta}}_S\|_1 + \lambda_2\|\widehat{\boldsymbol{\beta}}_{S^c}\|_1 \\ \le\ & \lambda_2\|\boldsymbol{\beta}_{0S}\|_1 + C\sqrt{\frac{\log p}{n}}\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|_1 + C\sqrt{\frac{\log p}{n}}\|\widehat{\boldsymbol{\beta}}_{S^c}\|_1 + C\frac{s\log p}{n}. \end{aligned}$$

Moving the second term on the left-hand side to the right side and moving the third term on the right-hand side to the left side, using that $\lambda_2 = C\sqrt{\log p/n}$ for a sufficiently large C, we get

$$C\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2 + C\sqrt{\frac{\log p}{n}}\|\widehat{\boldsymbol{\beta}}_{S^c}\|_1 \le C\sqrt{\frac{\log p}{n}}\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|_1 + C\frac{s\log p}{n}, \tag{10}$$

which in turn yields

$$\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1 \le C\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|_1 + s\sqrt{\frac{\log p}{n}} \le C\sqrt{s}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\| + s\sqrt{\frac{\log p}{n}}. \tag{11}$$

Using (11) in (10), we immediately get $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2 \le Cs\log p/n$, and using (11) again we get $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1 \le C(s\sqrt{\log p/n})$.

4. Discussions on assumption (B2)

First, we argue informally why we use $E[h(y(\langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0 + (\mathbf{z}-\langle X, \mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)))] - E[h(y(\langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0))] \ge C\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2$ (ignoring the term $\delta_n^2\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|$ for now). In particular, we argue why assumption (B2) cannot be changed to $E[h(y(\langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0 + \mathbf{z}^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)))] - E[h(y(\langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0))] \ge C\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2$ which looks simpler. In the following, we show that the latter cannot be expected to be satisfied unless X and $\mathbf{z}$ are independent.

We first develop intuition by considering a convex smooth loss $\ell(y, \langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta})$ instead of the hinge loss. The parallel derivations for the hinge loss follows similarly. Using Taylor's expansion, and that $(f_0, \boldsymbol{\beta}_0)$ minimizes $\ell(y, \langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta})$, we have

$$\begin{aligned} & E[\ell(y, \langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta})] - E[\ell(y, \langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)] \\ \approx\ & E[\ell'(y, \langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{z}^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)] + 0.5E[\ell''(y, \langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)\{\mathbf{z}^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\}^2] \\ \approx\ & E[\{\ell'(y, \langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0) - \ell'(y, \langle X, f_0\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)\}\mathbf{z}^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)] \\ & +0.5E[\ell''(y, \langle X, f_0\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)\{\mathbf{z}^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\}^2] \\ \approx\ & E[\{\ell''(y, \langle X, f_0\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)\langle X, f-f_0\rangle\mathbf{z}^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)] \\ & +0.5E[\ell''(y, \langle X, f_0\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)\{\mathbf{z}^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\}^2], \end{aligned} \tag{12}$$

where ℓ' and ℓ'' denote the first and the second derivative with respect to the second argument, respectively, and in the second step we use that $E[\ell'(y, \langle X, f_0\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{z}^{\mathrm{T}}] = 0$

which is a first order condition for the optimality of $(f_0, \boldsymbol{\beta}_0)$. The second term in (12) is naturally expected to be quadratic since $\ell'' \geq 0$ due to convexity, but in general the first term is hard to control and thus $E[\ell(y, \langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta})] - E[\ell(y, \langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)] \geq C\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2$ in general will not hold when $f \neq f_0$. On the other hand, if we define $\mathbf{g}_0 = \arg\min_{\mathbf{g}=(g_1,\ldots,g_p)^{\mathrm{T}}} E[\ell''(y, \langle X, f_0\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)(\mathbf{z} - \langle X, \mathbf{g}\rangle)^2]$, we have by the same arguments

$$\begin{aligned} & E[\ell(y, \langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0 + (\mathbf{z} - \langle X, \mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0))] - E[\ell(y, \langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)] \\ \approx\ & E[\{\ell''(y, \langle X, f_0\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)\langle X, f - f_0\rangle(\mathbf{z} - \langle X, \mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)] \\ & + 0.5E[\ell''(y, \langle X, f_0\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)\{(\mathbf{z} - \langle X, \mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\}^2], \end{aligned}$$

and the first term on the right-hand side above is exactly zero due to that the definition of $\mathbf{g}_0$ as a (weighted) projection making $E[\{\ell''(y, \langle X, f_0\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)X(\mathbf{z} - \langle X, \mathbf{g}_0\rangle)^{\mathrm{T}}] = 0$. Thus we naturally can expect

$$E[\ell(y, \langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0 + (\mathbf{z} - \langle X, \mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0))] - E[\ell(y, \langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)] \geq C\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2.$$

For the hinge loss, although the loss is not differentiable, the expected loss is and the second derivative is formally given by an expression involving the delta function, and thus assumption (B2) is reasonable.

In the second part of this section, we provide a concrete example when (B2) is satisfied, by making the above discussions rigorous under the Gaussian assumption for $(X, \mathbf{z})$. We use the following specific setup.

Let $W = (X, \mathbf{z}) \in L^2([0,1]) \times \mathbb{R}^p$ considered to be in a product Hilbert space. Assume W given $y = 1$ has a Gaussian distribution in the sense that $\langle W, F\rangle = \langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}$ is Gaussian for all $F = (f, \boldsymbol{\beta})$, and similarly for $y = -1$. The mean and the covariance operator of W (conditional on $y = 1$) is denoted by $\mu = (\mu_1, \mu_2)$ and $\Sigma = \{\Sigma_{jj'}\}_{j,j'=1}^2$ based on the partition $W = (X, \mathbf{z})$.

As shown in Koo et al. (2008), both the first the second derivatives of the expected hinge loss are well-defined. We have

$$\begin{aligned} & E[h(y(\langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0 + (\mathbf{z} - \langle X, \mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)))] - E[h(y(\langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0))] \\ =\ & -E[I\{y(\langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0) \leq 1\}y(\mathbf{z} - \langle X, \mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\ & + 0.5E[\delta(1 - y\langle X, f^*\rangle - y\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}^*)\{(\mathbf{z} - \langle X, \mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\}^2] \\ =\ & -E[(I\{y(\langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0) \leq 1\} - I\{y(\langle X, f_0\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0) \leq 1\})y(\mathbf{z} - \langle X, \mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)] \\ & + 0.5E[\delta(1 - y\langle X, f^*\rangle - y\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}^*)\{(\mathbf{z} - \langle X, \mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\}^2] \\ =\ & E[\delta(1 - y(\langle X, f^*\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0))\langle X, f - f_0\rangle \cdot (\mathbf{z} - \langle X, \mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)] \\ & + 0.5E[\delta(1 - y\langle X, f^*\rangle - y\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}^*)\{(\mathbf{z} - \langle X, \mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\}^2] \\ =\ & E[\delta(1 - y(\langle X, f_0\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0))\langle X, f - f_0\rangle \cdot (\mathbf{z} - \langle X, \mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)] \\ & + 0.5E[\delta(1 - \langle X, f_0\rangle - \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)\{(\mathbf{z} - \langle X, \mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\}^2] \\ & + 0.5E[\{\delta(1 - y\langle X, f^*\rangle - y\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}^*) - \delta(1 - y\langle X, f_0\rangle - y\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)\}\{(\mathbf{z} - \langle X, \mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\}^2] \\ & + E[\{\delta(1 - y(\langle X, f^*\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)) - \delta(1 - y(\langle X, f_0\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0))\} \\ & \qquad \cdot\langle X, f - f_0\rangle \cdot (\mathbf{z} - \langle X, \mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)], \end{aligned} \tag{13}$$

with $\boldsymbol{\beta}^*$ lying between $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}$, f^* between f_0 and f (different appearances of f^* may mean different values in the above), where the second equality is due to the first-order optimality conditions $E[I\{y(\langle X, f_0\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0) \le 1\}y\mathbf{z}] = 0$ and $E[I\{y(\langle X, f_0\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0) \le 1\}yX] = 0$, the first and the third equality use Taylor's expansion, and the last equality is obtained simply by adding and subtracting the same terms.

Importantly, the first term of (13) is zero by the first-order optimality condition of the weighted quadratic problem that defines $\mathbf{g}_0$ in (B1). The second term can naturally be assumed to be bounded below by $C\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2$. The third term and the fourth term can be bounded in similar ways as follows.

Let $F^* = (f^*, \boldsymbol{\beta}^*)$, $F_0 = (f_0, \boldsymbol{\beta}_0)$. Then we write

$$\begin{aligned} & E[\{\delta(1 - y\langle X, f^*\rangle - y\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}^*) - \delta(1 - y\langle X, f_0\rangle - y\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)\}\{(\mathbf{z} - \langle X, \mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\}^2] \\ = \; & E[\{\delta(1 - y\langle W, F^*\rangle) - \delta(1 - y\langle W, F_0\rangle)\}g(W, \boldsymbol{\beta})], \end{aligned}$$

where $g(W, \boldsymbol{\beta}) = \{(\mathbf{z} - \langle X, \mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\}^2$. By the calculations in the Appendix B of Lian and Fan (2018), we have in the Gaussian setting

$$\begin{aligned} & E[\{\delta(1 - y\langle W, F^*\rangle)\}g(W, \boldsymbol{\beta})] \\ = \; & p_0 E[\{\delta(1 + \langle W, F^*\rangle)\}g(W, \boldsymbol{\beta})|y = -1] + p_1 E[\{\delta(1 - \langle W, F^*\rangle)\}g(W, \boldsymbol{\beta})|y = 1] \\ = \; & p_0 q_*(-1) E[g(W, \boldsymbol{\beta})|y = -1, \langle W, F^*\rangle = -1] + p_1 q_*(1) E[g(W, \boldsymbol{\beta})|y = 1, \langle W, F^*\rangle = 1], \end{aligned}$$

and similarly

$$\begin{aligned} & E[\{\delta(1 - y\langle W, F_0\rangle)\}g(W, \boldsymbol{\beta})] \\ = \; & p_0 q_0(-1) E[g(W, \boldsymbol{\beta})|y = -1, \langle W, F_0\rangle = -1] + p_1 q_0(1) E[g(W, \boldsymbol{\beta})|y = 1, \langle W, F_0\rangle = 1], \end{aligned}$$

where $p_0 = P(y = -1)$, $p_1 = P(y = 1)$, $q_*(1)$ ($q_*(-1)$) is the (Gaussian) density of $\langle W, F^*\rangle$ at 1 given $y = 1$ (the density of $\langle W, F^*\rangle$ at -1 given $y = -1$), and $q_0(1)$ ($q_0(-1)$) is the density of $\langle W, F_0\rangle$ at 1 given $y = 1$ (the density of $\langle W, F_0\rangle$ at -1 given $y = -1$). Without loss of generality we only consider expectations conditional on $y = 1$ below since the case for $y = -1$ is the same. In the following, the distributions are always conditional on $y = 1$.

Due to the Gaussianity assumption, these conditional expectations have closed-form expressions. More specifically, the joint distribution of $(\langle W, F^*\rangle, \mathbf{z} - \langle X, \mathbf{g}_0\rangle)$ is

$$\begin{pmatrix} \langle W, F^*\rangle \\ \mathbf{z} - \langle X, \mathbf{g}_0\rangle \end{pmatrix} \sim \left(\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}, \begin{pmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{pmatrix} \right),$$

where $\eta_1 = \langle \mu_1, f^*\rangle + \mu_2^{\mathrm{T}}\boldsymbol{\beta}^*$, $\eta_2 = \mu_2 - \langle \mu_1, \mathbf{g}_0\rangle$, $\Theta_{11} = \langle f^*, \Sigma_{11} f^*\rangle + \boldsymbol{\beta}^{*\mathrm{T}}\Sigma_{22}\boldsymbol{\beta}^* + 2\boldsymbol{\beta}^{*\mathrm{T}}\Sigma_{21} f^*$, $\Theta_{21} = -\mathbf{g}_0^{\mathrm{T}}\Sigma_{11} f^* - \mathbf{g}_0^{\mathrm{T}}\Sigma_{12}\boldsymbol{\beta}^* + \Sigma_{21} f^* + \Sigma_{22}\boldsymbol{\beta}^*$, $\Theta_{12} = \Theta_{21}^{\mathrm{T}}$, $\Theta_{22} = \Sigma_{22} + \mathbf{g}_0^{\mathrm{T}}\Sigma_{11}\mathbf{g}_0 - \mathbf{g}_0^{\mathrm{T}}\Sigma_{12} - \Sigma_{21}\mathbf{g}_0$ (here we regard $\mathbf{g}_0$ naturally as an operator from $\mathbb{R}^p$ to $(L^2([0,1]))^p$). And we have the conditional distribution

$$(\mathbf{z} - \langle X, \mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)|\langle W, F^*\rangle = 1 \sim N(a_1, b_1),$$

where $a_1 = (\eta_2 - \Theta_{21}\Theta_{11}^{-1}(1 - \eta_1))^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$, $b_1 = (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^{\mathrm{T}}(\Theta_{22} - \Theta_{21}\Theta_{11}^{-1}\Theta_{12})(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ and $E[g(W, \boldsymbol{\beta})|y = 1, \langle W, F^*\rangle = 1] = a_1^2 + b_1$. Similarly $E[g(W, \boldsymbol{\beta})|y = 1, \langle W, F_0\rangle = 1] = a_2^2 + b_2$ where (a_2, b_2) differs from (a_1, b_1) by changing f^* to f_0 and changing $\boldsymbol{\beta}^*$ to $\boldsymbol{\beta}_0$ in the various

expressions above. Furthermore, $q_*(1) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{(1-\eta_1)^2}{2\Theta_{11}^2}\}$, for example. From these expressions, and that $\|\Sigma_{11}(f-f_0)\| + \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \le \delta_n$, we can easily get if $\mu_1 \in \Sigma_{11}^{1/2} L^2([0,1])$, $\Sigma_{11}^{-1/2}\Sigma_{12}$ is a bounded operator, $f_0\Sigma_{11}f_0 > 0$ then

$$|E[\{\delta(1-y\langle W, F^*\rangle)\}g(W,\boldsymbol{\beta})] - E[\{\delta(1-y\langle W, F_0\rangle)\}g(W,\boldsymbol{\beta})]| \le C\delta_n\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2.$$

Similarly the fourth term of (13) is $O(\delta_n^2\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|)$, which gives the additional term in the lower bound leading to "approximate convexity".

Summarizing the above, we have proved the following proposition.

Proposition 1 *Assume the following assumptions:*

(C1) Conditional on $y = 1$ (similary for $y = -1$), $(X, \mathbf{z})$ is jointly Gaussian with mean (μ_1, μ_2) and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$,

(C2) $\langle f_0, \Sigma_{11} f_0\rangle > 0$, $\Sigma_{11}^{-1/2}\mu_1 \in L^2([0,1])$, $\Sigma_{11}^{-1/2}\Sigma_{12}$ is a bounded operator, and $E[\delta(1 - y(\langle X, f_0\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)(\mathbf{z} - \langle X, \mathbf{g}_0\rangle)^{\otimes 2}]$ is also positive-definite.

Then we have that there exists some constant $c > 0$ such that for any $f, \boldsymbol{\beta}$ with $\|\Gamma^{1/2}(f - f_0)\| + \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le \delta_n = Cn^{-1/4}$,

$$\begin{aligned} &E[h(y(\langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0 + (\mathbf{z} - \langle X, \mathbf{g}_0\rangle)^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)))] - E[h(y(\langle X, f\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0))] \\ &\ge c\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 - \delta_n^2\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|. \end{aligned}$$

From the above calculations, we can see that the reason for the appearance of the additional term $\delta_n^2\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|$ can roughly be explained as follows. When we use Taylor's expansion to approximate the risk difference, the main term is the quadratic term $C\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2$, while the higher order term involves terms such as $\|\Gamma^{1/2}(f-f_0)\|^2\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|$. However, if $\|\Gamma^{1/2}(f-f_0)\|^2 > \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|$, this term is not actually "higher-order". To deal with this, we relax the condition to be that the risk difference is bounded below by $C\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2 - \delta_n^2\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|$ which then reasonably holds based on Taylor's expansion. We would automatically have the quadratic lower bound $C\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2$ if the loss were the least squares loss instead of the hinge loss.

5. Discussions on the RKHS setting

In this paper, we have established the learning rate of functional partially linear SVM, including the overall rate and the possibly faster rate for the high-dimensional linear part.

We can also put the functional coefficient f in a RKHS, with few modifications in the proof. We assume f_0 is in an RKHS $\mathcal{H} \subseteq L^2([0,1])$, characterized by a bivariate positive definite kernel function $K(.,.)$. K also denotes the operator $Kf = \int_{[0,1]} K(t,.)f(t)dt$. We assume $\iint_{[0,1]^2} K^2(s,t)\,ds\,dt < \infty$, which guarantees that the operator K is compact. Under the additional assumption that $\sup_{s,t} K(s,t) < \infty$, it is also a trace operator. The spectral theory for $T := K^{1/2}\Gamma K^{1/2}$ yields

$$T = \sum_{j=1}^{\infty} s_j e_j \otimes e_j,$$

with eigenvalues $s_1 \geq s_2 \geq \cdots > 0$ and eigenfunctions $e_1, e_2, \ldots$. Our estimator is now

$$(\widehat{f}, \widehat{\boldsymbol{\beta}}) = \underset{f \in \mathcal{H}, \boldsymbol{\beta}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} h(y_i(\langle X_i, f \rangle + \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\beta})) + \lambda_1 \|f\|_{\mathcal{H}}^2 + \lambda_2 \|\boldsymbol{\beta}\|_1,$$

where $\|.\|_{\mathcal{H}}$ is the RKHS norm (the RKHS inner product is denoted by $\langle ., . \rangle_{\mathcal{H}}$).

We can show under straightforward modifications of the assumptions, most notably that the source condition becomes $\|T^{-r} f_0\|_{\mathcal{H}} \leq C$ for some $r \in [0, 1/2]$, and also $\|g_{0j}\|_{\mathcal{H}} \leq C$, we have $\|\Gamma^{1/2}(f - f_0)\| + \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq C(u_n^{1+2r} + \sqrt{s \log p / n})$ (with $\xi = 2$ in assumption (A4)) and $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq C\sqrt{s \log p / n}$.

To see that we only need minor modifications in the proof to deal with this RKHS setting, note that since $\mathcal{H} = K^{1/2}(L^2([0,1]))$ and the mapping $K^{1/2}$ is isometric from $L^2([0,1])$ to $\mathcal{H}$ (Wahba, 1990), we have $\langle X, f \rangle = \langle K^{1/2} X, K^{-1/2} f \rangle$ and $\|f\|_{\mathcal{H}}^2 = \|K^{-1/2} f\|^2$. Then the proofs in Section 2 and Section 3 can be followed line by line by regarding $K^{1/2} X$ as the functional predictor, and the RKHS norm or inner product involving f can all be converted to L^2 norm and L^2 inner product involving $K^{-1/2} f$. Note that the covariance operator of $K^{1/2} X$, given by $E[(K^{1/2} X) \otimes (K^{1/2} X)]$, is exactly T. Thus T plays the role of Γ in Section 2. The details are contained in the appendix.

We now discuss the setup of Xia et al. (2021), which considered the non-functional setting by replacing $\langle X, f \rangle$ in our model by $f(x)$ (for simplicity we again assume the domain of f is $[0,1]$ but this is non-essential). They define the estimation target in an RKHS $\mathcal{H}$ to be

$$(f^*, \boldsymbol{\beta}^*) = \underset{f \in \mathcal{H}, \boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \; E[h(y(f(x) + \mathbf{z}^{\mathrm{T}} \boldsymbol{\beta}))]. \tag{14}$$

Note the minimization is over the RKHS $\mathcal{H}$, and f^* is considered to be a certain approximation to the "true" function f_0 which may be outside $\mathcal{H}$. However, the well-posedness of (14) is a very strong assumption even in the nonparametric case with $\boldsymbol{\beta}_0 = 0$. The reason is that it is known that under mild measurability assumptions the minimizer of $\arg\min_{f \in L^2([0,1])} E[h(yf(x))]$ is $f = \mathrm{sign}\{P(y = 1|x) - 1/2\}$ except on the set $P(y = 1|x) = 1/2$ (Bartlett et al., 2006). Thus, due to the discontinuity of the sign function, in general, one would expect f is not smooth and only in $L^2([0,1])$. Furthermore, since $\mathcal{H}$ is typically dense in $L^2([0,1])$, this means $\inf_{f \in \mathcal{H}} E[h(yf(x))]$ is not achieved for a $f \in \mathcal{H}$, and in typical situations one has a sequence $f_n \in \mathcal{H}$ such that $E[h(yf_n(x))] \to \arg\min_{f \in L^2([0,1])} E[h(yf(x))]$ with $\|f_n\|_{\mathcal{H}} \to \infty$ (if the sequence were bounded, by the weak compactness of the unit ball in a Hilbert space, there would be a converging subsequence in $\mathcal{H}$ converging to a minimizer in $\mathcal{H}$ with bounded norm, which would lead to a contradiction). One could deal with this problem by considering the minimization over $\{f : \|f\| \leq C\}$. However, unless $C \to \infty$, the approximation error of this estimator does not converge to zero. If $\inf_{f \in \mathcal{H}} E[h(yf(x))]$ is achievable, then our Theorem 1 indeed produces the same learning rate as Xia et al. (2021) (note that $f(x) = \langle f, K(x, .) \rangle_{\mathcal{H}}$ and thus the proof is basically the same as in the functional case with the functional predictor replaced by $K(x, .)$), although their result is limited to $r = 0$ and only for the overall rate combining the nonparametric and the linear parts. Alternatively, Wu and Zhou (2006); Steinwart and Scovel (2007) used a regularized loss to make the minimization problem in $\mathcal{H}$ well-defined.

In our functional setting, for the problem $\arg\min_{f \in L^2([0,1])} E[h(y\langle X, f \rangle)]$, the minimizer is achievable and f can also be a smooth function. To see a concrete example, again consider

the Gaussian setting for X conditional on y. Following the same calculations as in Koo et al. (2008) (in particular their equation (8)), the minimizer is $f_0 \propto \Gamma^{-1}(E[X|y=1) - E[X|y=-1])$ (when this expression is well-defined), and thus one can certainly construct concrete example such that $f_0 \in \Gamma^r(L^2([0,1]))$ for any $r \geq 0$, as long as $E[X|y]$ is assumed to be sufficiently smooth.

6. Simulations

We illustrate the performances of the functional partial linear SVM for classification via simulations. The simulated data are generated from the following model. First, y_i are generated from the binary distribution $P(y_i = 1) = P(y_i = -1) = 0.5$. Given $y_i = 1$, $\mathbf{z}_i$ is generated from a multivariate normal distribution with mean $\mu = (0.1, 0.2, 0.3, 0.4, 0.5, 0, \ldots, 0)^T$ and covariance matrix $\Sigma = (\sigma_{jj'})$ with $\sigma_{jj} = 1$ for all j, and $\sigma_{jj'} = -0.2$ if $j \leq 5, j' \leq 5, j \neq j'$ and $\sigma_{jj'} = 0$ otherwise. The functional predictor $X_i(t)$ is generated by $X_i(t) = \sum_{k=1}^{50} \xi_{ik}\phi_k(t)$ where $\phi_{2l-1} = 2^{-1/2}\cos\{(2l-1)\pi t\}$ and $\phi_{2l} = 2^{-1/2}\sin\{(2l-1)\pi t\}(l = 1, \ldots, 25; t \in T = [0,1])$ are Fourier basis functions.

Case 1: ξ_{ik} are independent and identically distributed as $N(b_k, 16k^{-2})$ for different i where $b_1 = 0.1, b_2 = 0.2, b_3 = 0.3, b_4 = 0.4, b_k = 0.8/(k-2)^4$.

Case 2: This case is the same as Case 1 expcet that we allow moderate correlation between $X_i(t)$ and $\mathbf{z}_i$ by giving a correlation structure specified by $\text{corr}(\xi_{ik}, z_{il}) = r^{|k-l|+1}$ for $k = 1, \ldots, 4,\ l = 1, \ldots, 5$, with $r = 0.2$.

Case 3: ξ_{ik} are independent and identically distributed as $N(b_k, 4k^{-1})$ for different i where $b_1 = 0.1, b_2 = 0.2, b_3 = 0.3, b_4 = 0.4, b_k = 0.8/(k-2)^4$.

Given $y_i = -1$, $\mathbf{z}_i$ is generated from a multivariate normal distribution with mean $-\mu$ and covariance matrix Σ. The functional predictor $X_i(t)$ is generated in the same way as given $y_i = 1$ but with ξ_{ik} from normal with negative mean in all three cases. The true functional coefficient has the form $f(t) = \sum_{k=1}^{50} f_k\phi_k(t)$, a linear combination of the eigenbasis. Therefore, we numerically minimize

$$\min_{f_k, \boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n} h(y_i(\sum_{k=1}^{50}\xi_{ik}f_k + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta})) + \lambda_1\sum_{k=1}^{50} f_k^2 + \lambda_2\|\boldsymbol{\beta}\|_1.$$

By the equations (28) and (29) in Peng et al. (2016), the true parameters can be numerically calculated. The minimization problem is solved by the semismooth newton coordinate descent algorithm proposed by Yi and Huang (2017). We consider sample sizes $n = 200, 400, 600, 800, 1000, 1500$ with $p = 1000$ and use an independent tuning data set of size $10n$ to choose the tuning parameters by minimizing the prediction error. For each simulation scenario, we use 200 repetitions. The estimation accuracy is measured by three criteria:

- Prediction error on an independent test dataset with $10n$ sample size;
- L_2 estimation error for the linear part $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2/\|\boldsymbol{\beta}_0\|_2$;
- L_2 estimation error for the functional part $\|\Gamma^{1/2}(\widehat{f}(t) - f_0(t))\|/\|\Gamma^{1/2}f_0(t)\|$ where $\|\Gamma^{1/2}(\widehat{f}(t) - f_0(t))\| = \{\sum_{k=1}^{50} s_k(\widehat{f}_k - f_{0k})^2\}^{1/2}$, $\|\Gamma^{1/2}f_0(t)\| = \{\sum_{k=1}^{50} s_k f_{0k}^2\}^{1/2}$ and s_k is the variance of ξ_{ik}.

We normalize the L_2 estimation error by the norm of the true parameters to make it more comparable for different cases. The prediction error is showed in Figure 1 for Case 1 and Case 2. Since the figure for Case 3 is very similar with Case 1 and we omit it. The oracle prediction error shown in Figure 1 is the predictor error obtained by the true $(f_0, \boldsymbol{\beta}_0)$. The results show that the prediction error becomes closer to the oracle one as sample size increases.

Next we plot the L_2 error for the linear part and functional part vesus $\sqrt{\log p/n}$. In Figure 2, we see that the linear error is indeed proportional to $\sqrt{\log p/n}$ in all cases. Comparing Case 1 with Case 2, the errors for the latter are larger, which is possibly due to the correlation between the linear part and the functional part making the estimation more difficult. Comparing Case 1 with Case 3, we see the errors for the functional part for the latter is larger, due to that the eigenvalues of the covariance operator Γ decay slower for Case 3, which is consistent with our theoretical results.

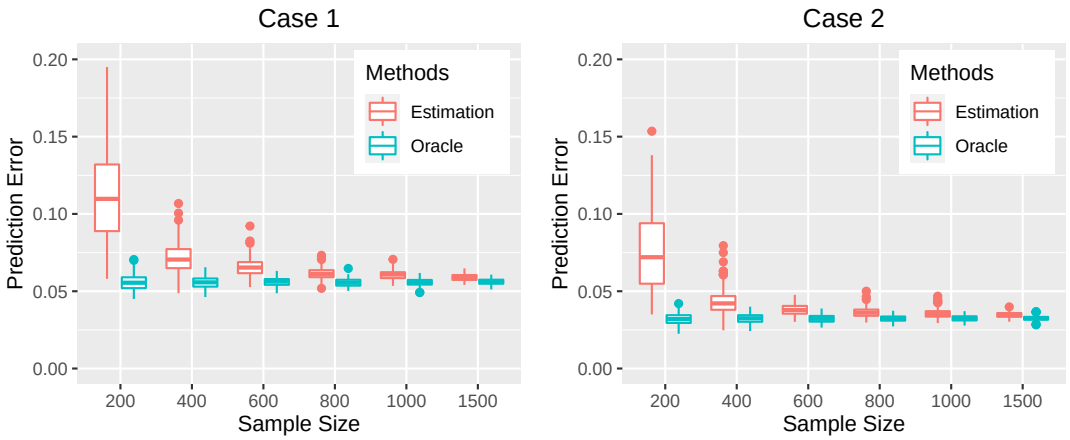

Figure 1: Boxplot of prediction errors for Case 1 and Case 2.

Acknowledgements

The authors sincerely thank the editor, the associate editor and three anonymous reviewers for their insightful comments that improved the manuscript. Yan-Yong Zhao's research is supported by National Natural Science Foundation of China under Grants No. 12071220, 11701286, National Statistical Research Project of China under Grants No. 2020LZ35, and Social Science Foundation of Jiangsu Province under Grants No. 20EYC008. Heng Lian's research is supported in part by the NSFC Project 11871411 at the Shenzhen Research Institute, City University of Hong Kong; and in part by the Hong Kong Research Grants Council (RGC) General Research Fund under Grant 11301718, Grant 11300519, Grant 11300721, and Grant 11311822.

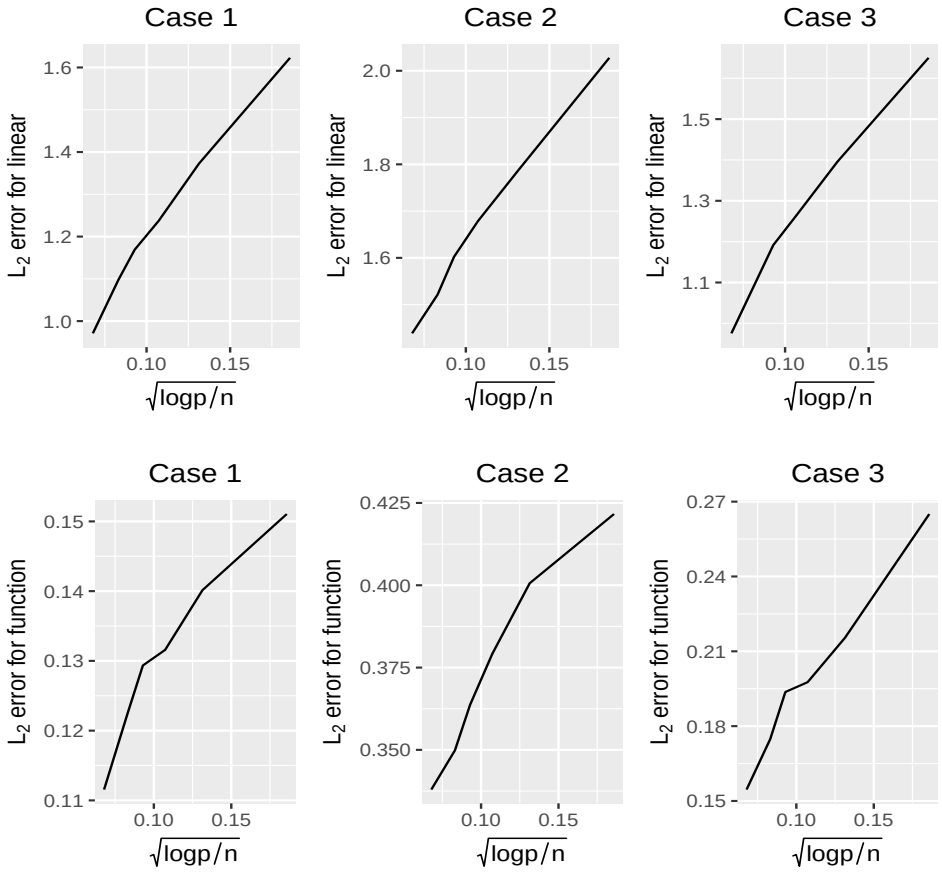

Figure 2: L_2 error for the linear part and the functional part in Cases 1-3 with x-axis being $\sqrt{\log p/n}$. The first row is the errors for the linear part, and the second row is the errors for the functional part.

Appendix: Details on the RKHS setting

We assume that f_0 is in a RKHS $\mathcal{H} \subseteq L^2([0,1])$, characterized by a bivariate positive definite kernel function $K(.,.)$. Our estimator is now

$$(\widehat{f}, \widehat{\boldsymbol{\beta}}) = \arg \min_{f \in \mathcal{H}, \boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} h(y_i(\langle X_i, f \rangle + \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\beta})) + \lambda_1 \|f\|_{\mathcal{H}}^2 + \lambda_2 \|\boldsymbol{\beta}\|_1,$$

where $\|.\|_{\mathcal{H}}$ is the RKHS norm (the RKHS inner product is denoted by $\langle ., . \rangle_{\mathcal{H}}$).

Let K also denote the integral operator $Kf = \int_{[0,1]} K(t,.) f(t) dt$. We assume that $\iint_{[0,1]^2} K^2(s,t)\, ds\, dt < \infty$, which guarantees that the operator K is compact. Under the additional assumption that $\sup_s K(s,s) < \infty$, it is also a trace operator. By the Riesz representation theorem, the operator K satisfies $\langle f, Kg \rangle_{\mathcal{H}} = \langle f, g \rangle$.

Note that since the square-root operator $K^{1/2}$ is isometric from $L^2([0,1])$ to $\mathcal{H}$ (Wahba, 1990), we have $\mathcal{H} = K^{1/2}(L^2([0,1]))$. To facilitate theoretical analysis, we reformulate the problem. Since we have $\langle X, f \rangle = \langle K^{1/2} X, K^{-1/2} f \rangle$ and $\|f\|_{\mathcal{H}}^2 = \|K^{-1/2} f\|^2$, our estimator is equivalent to

$$(\widehat{f}, \widehat{\boldsymbol{\beta}}) = \arg \min_{f \in \mathcal{H}, \boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} h(y_i(\langle K^{1/2} X_i, K^{-1/2} f \rangle + \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\beta})) + \lambda_1 \|K^{-1/2} f\|^2 + \lambda_2 \|\boldsymbol{\beta}\|_1.$$

Let $g = K^{-1/2} f$, and the above is equivalent to finding an optimization in $L^2([0,1])$ such that

$$(\widehat{g}, \widehat{\boldsymbol{\beta}}) = \arg \min_{g \in L^2([0,1]), \boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} h(y_i(\langle K^{1/2} X_i, g \rangle + \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\beta})) + \lambda_1 \|g\|^2 + \lambda_2 \|\boldsymbol{\beta}\|_1.$$

and the desired estimator in the RKHS is $\widehat{f} = K^{1/2} \widehat{g}$. Regarding $K^{1/2} X$ as the new functional predictor, we can imitate the proofs in Section 2 and 3. Define a new operator

$$T = E[(K^{1/2} X) \otimes (K^{1/2} X)] = K^{1/2} \Gamma K^{1/2},$$

which plays the role of Γ in Section 2 and 3. By the Mercer's Theorem, T has a spectral expansion given by

$$T = \sum_{j=1}^{\infty} s_j e_j \otimes e_j,$$

where $s_1 \geq s_2 \geq \cdots > 0$ are the eigenvalues with $s_j \to 0$ and $\{e_j\}$ are the orthonormalized eigenfunctions in $L^2([0,1])$. We shall see that the statistical convergence rate depends on the decay rate of eigenvalues of T. Now the source condition becomes

$$\|T^{-r} f_0\|_{\mathcal{H}} \leq C, \text{for some } r \in [0, 1/2]. \tag{15}$$

Then the proofs in Section 2 can be followed line by line by regarding g as the target and we have

$$\|T^{1/2}(g - g_0)\| + \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq C(u_n^{1+2r} + \sqrt{s \log p / n}). \tag{16}$$

Furthermore,

$$\begin{aligned}\|T^{1/2}(g-g_0)\| &= \langle T^{1/2}K^{-1/2}(f-f_0), T^{1/2}K^{-1/2}(f-f_0)\rangle \\ &= \langle TK^{-1/2}(f-f_0), K^{-1/2}(f-f_0)\rangle \\ &= \langle K^{1/2}\Gamma(f-f_0), K^{-1/2}(f-f_0)\rangle \\ &= \langle \Gamma(f-f_0), (f-f_0)\rangle = \|\Gamma^{1/2}(f-f_0)\|.\end{aligned}$$

Thus (16) is equivalent to

$$\|\Gamma^{1/2}(f-f_0)\| + \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \leq C(u_n^{1+2r} + \sqrt{s\log p/n}).$$

At last, for the convergence rate of linear part, we should modify assumption (B1) to

(B1′) Let $\mathbf{g}_0 = (g_{01}, \ldots, g_{0p})^{\mathrm{T}}$ with $g_{0j} = \arg\min_g E[\delta(1-y(\langle X, f_0\rangle + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0))(z_j - \langle X, g\rangle)^2]$ where $\delta(.)$ is the Dirac delta function. We assume that $g_{0j} \in \mathcal{H}, \forall j$ and $\max_j \|g_{0j}\|_{\mathcal{H}} \leq C$.

Following the proofs line by line in Section 3, we can obtain $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \leq C\sqrt{s\log p/n}$ for $\xi = 2$ in assumption (A4). Formally, we state the result as follows.

Theorem 3 *Under the same assumptions as for Theorem 1, except that we now assume $f_0 \in \mathcal{H}$, (A4) is assumed to hold with $f \in \mathcal{H}$, and the source condition is replaced by* (15)*, we have*

$$\|\Gamma^{1/2}(f-f_0)\| + \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| = O_p(u_n^{1+2r} + \sqrt{s\log p/n}).$$

Furthermore, under the additional assumptions (B1′) and (B2) (with f in (B2) assumed to be in $\mathcal{H}$), and $\xi = 2$, we have $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\| = O_p\left(\sqrt{\frac{s\log p}{n}}\right)$ and $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1 = O_p\left(s\sqrt{\frac{\log p}{n}}\right)$.

References

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Faicel Chamroukhi, Hervé Glotin, and Allou Samé. Model-based functional mixture discriminant analysis with hidden process regression for curve classification. *Neurocomputing*, 112:153–163, 2013.

Aurore Delaigle and Peter Hall. Achieving near-perfect classification for functional data. *Journal of the Royal Statistical Society Series B*, 74(2):267–286, 2012.

Sara A. Geer, Sara van de Geer, and D. Williams. *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge, 2000.

Chong Gu. *Smoothing spline ANOVA models*. Springer, New York, 2013.

Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*. Springer, New York, 2011.

Dehan Kong, Kaijie Xue, Fang Yao, and Hao H. Zhang. Partially functional linear regression in high dimensions. *Biometrika*, 103(1):147–159, 2016.

Ja-Yong Koo, Yoonkyung Lee, Yuwon Kim, and Changyi Park. A Bahadur representation of the linear support vector machine. *Journal of Machine Learning Research*, 9(44): 1343–1368, 2008.

Qi Li. Efficient estimation of additive partially linear models. *International Economic Review*, 41(4):1073–1092, 2000.

Heng Lian and Zengyan Fan. Divide-and-conquer for debiased l1-norm support vector machine in ultra-high dimensions. *Journal of Machine Learning Research*, 18(182):1–26, 2018.

Junhong Lin and Volkan Cevher. Optimal convergence for distributed learning with stochastic gradient methods and spectral algorithms. *Journal of Machine Learning Research*, 21 (20):1–44, 2020.

Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(97):1–47, 2017.

Shao-Bo Lin, Xin Guo, and Ding-Xuan Zhou. Distributed learning with regularized least squares. *Journal of Machine Learning Research*, 18(92):1–31, 2017.

Po-Ling Loh and Martin J. Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16(19):559–616, 2015.

Luo Luo, Yubo Xie, Zhihua Zhang, and WuJun Li. Support matrix machines. In *the 32nd International Conference on Machine Learning*, 2015.

Haiqiang Ma, Ting Li, Hongtu Zhu, and Zhongyi Zhu. Quantile regression for functional partially linear model in ultra-high dimensions. *Computational Statistics and Data Analysis*, 129:135–147, 2019.

Shahar Mendelson. Geometric parameters of kernel machines. In *International Conference on Computational Learning Theory*, 2002.

Bo Peng, Lan Wang, and Yichao Wu. An error bound for l1-norm support vector machine coefficients in ultra-high dimension. *Journal of Machine Learning Research*, 17(233):1–26, 2016.

David Pollard. *Convergence of stochastic processes*. Springer, New York, 2012.

Cristian Preda, Gilbert Saporta, and Caroline Lévéder. PLS classification of functional data. *Computational Statistics*, 22(2):223–235, 2007.

James O. Ramsay and C.J. Dalzell. Some tools for functional data analysis. *Journal of the Royal Statistical Society Series B*, 53(3):539–572, 1991.

Ingo Steinwart and Clint Scovel. Fast rates for support vector machines using Gaussian kernels. *The Annals of Statistics*, 35(2):575–607, 2007.

Taiji Suzuki and Masashi Sugiyama. Fast learning rate of multiple kernel learning: Trade-off between sparsity and smoothness. *The Annals of Statistics*, 41:1381–1405, 2013. ISSN 0090-5364. doi: 10.1214/13-AOS1095.

Aad W. Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer, New York, 1996.

Grace Wahba. *Spline models for observational data*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1990.

Huixia Judy Wang, Zhongyi Zhu, and Jianhui Zhou. Quantile regression in partially linear varying coefficient models. *The Annals of Statistics*, 37(6B):3841–3866, 2009.

Qiang Wu and Ding-Xuan Zhou. Analysis of support vector machine classification. *Journal of Computational Analysis and Applications*, 8(2):99–119, 2006.

Yifan Xia, Yongchao Hou, and Shaogao Lv. Learning rates for partially linear support vector machine in high dimensions. *Analysis and Applications*, 19(1):167–182, 2021.

Congrui Yi and Jian Huang. Semismooth Newton coordinate descent algorithm for elastic-net penalized Huber loss regression and quantile regression. *Journal of Computational and Graphical Statistics*, 26(3):547–557, 2017.