The Annals of Statistics
2000, Vol. 28, No. 3, 734–755

TENSOR PRODUCT SPACE ANOVA MODELS

BY YI LIN

University of Wisconsin, Madison

To deal with the curse of dimensionality in high-dimensional nonparametric problems, we consider using tensor product space ANOVA models, which extend the popular additive models and are able to capture interactions of any order. The multivariate function is given an ANOVA decomposition, that is, it is expressed as a constant plus the sum of functions of one variable (main effects), plus the sum of functions of two variables (two-factor interactions) and so on. We assume the interactions to be in tensor product spaces. We show in both regression and white noise settings, the optimal rate of convergence for the TPS-ANOVA model is within a *log* factor of the one-dimensional optimal rate, and that the penalized likelihood estimator in TPS-ANOVA achieves this rate of convergence. The quick optimal rate of the TPS-ANOVA model makes it very preferable in high-dimensional function estimation. Many properties of the tensor product space of Sobolev–Hilbert spaces are also given.

1. Introduction. Much progress has been made in the nonparametric estimation of univariate functions. When it comes to multivariate function estimation, however, extra difficulties are encountered. One major difficulty is the curse of dimensionality caused by the fact that even a reasonably large number of data points can be sparse in a high-dimensional space. It requires far more data to get a decent estimate in high-dimensional problems. This is reflected in the optimal rate of convergence: it is generally much slower for high-dimensional problems than for one- (low-) dimensional problems. Furthermore, a general multivariate function is hard to visualize and does not usually give a good idea of the effect of each covariate. Hence it poses problems for interpretation.

Several different models have been proposed to bypass these difficulties. The additive models, proposed by Stone (1985) and developed by Hastie and Tibshirani (1990), are one popular choice. Additive models assume the high-dimensional function to be a sum of one-dimensional functions. By doing so, additive models effectively reduce the "working dimension" of the problem to one. Stone (1985) showed that the optimal convergence rate for additive models is the same as that for univariate function estimation problems. Thus, in a sense, the additive models overcome the curse of dimensionality. Additive models are also easy to interpret since they give a direct description of the effect of each covariate. The fitting of additive models is manageable. Hastie and Tibshirani (1990) gave a detailed discussion of the application of

Received September 1998; revised January 2000.

AMS 1991 *subject classifications*. Primary 62G07; secondary 62J20.

Key words and phrases. Functional ANOVA, tensor product space, white noise model, rate of convergence, optimal rate of convergence, penalized likelihood estimation, interaction, curse of dimensionality, smoothing splines.

the backfitting algorithm in fitting additive models. There are also other feasible methods to fit additive models. Because of these reasons, additive models are successful in a variety of problems.

In order to increase the flexibility of additive models to accommodate situations where interactions among the covariates may be present, it is desirable to extend the additive methodology by considering functional ANOVA models, the analogues of parametric ANOVA models. The functional ANOVA models assume that the high-dimensional function to be estimated is a sum of one-dimensional functions (main effects), two-dimensional functions (two-way interactions), and so on. That is, we decompose the d-dimensional function f as

$$f(x_1, x_2, \ldots, x_d) = \text{constant} + \sum_{i=1}^{d} f_i(x_i) + \sum_{i<j} f_{ij}(x_i, x_j) + \cdots,$$

where the components satisfy side conditions that guarantee uniqueness, and the series is truncated in some manner.

Different strategies have been adopted to model the interactions in functional ANOVA models. We will consider the tensor product space strategy. After determining the function space of each main effect, this strategy models an interaction as lying in the tensor product space of the function spaces of the interacting main effects. That is, if we assume $f_1(x_1)$ to be in a Hilbert space E_1 of functions of x_1 and $f_2(x_2)$ to be in a Hilbert space E_2 of functions of x_2, then we can model f_{12} as in $E_1 \otimes E_2$, the tensor product space of E_1 and E_2. Higher order interactions are modeled similarly.

For a Hilbert space E_1 of functions of x_1 and a Hilbert space E_2 of functions of x_2, $E_1 \otimes E_2$ is defined as the completion of the class of functions of the form

$$\sum_{i=1}^{k} f_i(x_1) g_i(x_2), \qquad f_i \in E_1,\ g_i \in E_2,\ k \text{ is any positive integer}$$

under the norm induced by the norms in E_1 and E_2. The norm in $E_1 \otimes E_2$ satisfies

$$\langle f_1(x_1)g_1(x_2), f_2(x_1)g_2(x_2)\rangle_{E_1 \otimes E_2} = \langle f_1(x_1), f_2(x_1)\rangle_{E_1} \langle g_1(x_2), g_2(x_2)\rangle_{E_2}.$$

For an introduction to the basics of tensor product space of general Hilbert spaces, see Kadison and Ringrose (1993), or Section 1 of Appendix A in Lin (1998).

The TPS-ANOVA strategy for modeling interactions can be motivated by the ideas used in parametric problems. In multiple linear regression, we often use the product of two variables, $x_i x_j$, to model the interaction between the two variables, and use the product of three variables to model three-way interaction and so on. In the semiparametric case, the usual model without interaction assumes

$$f(x_1, x_2, \ldots, x_d) = f_1(x_1) + a_2 x_2 + \cdots + a_d x_d$$

and we can use $x_i g(x_1)$ to model the interaction between x_i and x_1, where we assume g and f_1 lie in the same function space. The varying coefficient

model considered in Hastie and Tibshirani (1993) is an example of this type of structure. Notice in both cases we are actually assuming that the function space for interaction is the tensor product space of the function spaces for main effects.

The tensor product space ANOVA modeling is capable of dealing with interactions of all orders in a flexible way, thus vastly extending the additive methodology. Wahba and her colleagues have successfully applied the TPS-ANOVA models in many practical situations. They considered fitting the models with the penalized likelihood method, also known as the smoothing spline method. In fact, when it was originally proposed in the nonparametric settings, the TPS-ANOVA model was motivated by the use of the smoothing spline method and was called by Wahba and her colleagues the smoothing spline ANOVA model. Reproducing kernel Hilbert space (RKHS) plays an essential role in the smoothing spline methods. "The idea behind smoothing spline ANOVA model is to construct an RKHS of functions so that the components of the ANOVA decomposition represent an orthogonal decomposition of f in the RKHS. (Then RKHS methods can be used to find the smoothing spline estimator)" [quoted from Wahba, Wang, Gu, Klein and Klein (1995)]. Tensor product structure achieves this goal naturally, and the fact that the reproducing kernel of a tensor product space of RKHSs is simply the product of the reproducing kernels of the component spaces makes the computation of the estimator via the smoothing spline method straightforward.

Wahba, Wang, Gu, Klein and Klein (1995) gave an extensive discussion about many aspects of the fitting of the TPS-ANOVA model via penalized likelihood method. They proposed methods for model selection and for making confidence statements, developed practical algorithms, and provided public software. They pointed out that

> ... it is tantalizing to conjecture the circumstances under which Stone's convergence rates could be obtained in the smoothing (spline) context ....

Chen (1991) proved in the regression setting that, for the smoothing spline estimator, when the data form an equidistant grid, the expected squared error averaged over the data points, which is an approximation to the integrated squared error, goes to zero at the rate $O(n^{-2m/(2m+1)})$. It needs to be pointed out that there is a small error in that paper; the rate actually differs from $O(n^{-2m/(2m+1)})$ by a *log* factor. Since this result was proved for a very special design, it was not clear whether a similar result is valid under general conditions. Our results include a corrected version of Chen (1991) as a special case.

In this paper, we study the optimal rate of the TPS-ANOVA model and the rate of convergence of the penalized likelihood estimator in fitting the TPS-ANOVA model under general conditions. We concentrate on regression and white noise settings. The rate of convergence of the penalized likelihood estimator in fitting TPS-ANOVA in other settings (such as generalized regression, density estimation and hazard regression) will be established in a separate paper.

We show that the minimax mean integrated squared error for the TPS-ANOVA model goes to 0 at a rate that is within a *log* factor of the one-dimensional optimal rate. This quick optimal rate for TPS-ANOVA makes the TPS-ANOVA model very preferable in high-dimensional function estimation. We also show that the penalized likelihood estimator in TPS-ANOVA achieves this rate. This means that the penalized likelihood method has very good statistical properties.

Now let us explain intuitively how it can be that the TPS-ANOVA model includes higher order interactions and still has an optimal rate that is close to the one-dimensional optimal rate. For this, let us introduce some concepts first.

For a nonnegative integer m, the Sobolev–Hilbert space of univariate functions with order m and domain [0, 1], denoted by $H^m([0,1])$, is defined by

$$H^m([0,1]) = \{f | f^{(\nu)} \text{ abs. cont.}, \nu = 0, 1, \ldots, m-1; f^{(m)} \in L_2\}.$$

In the nonparametric estimation, it is typical to impose the mth order smoothness condition on a univariate function by assuming it is in $H^m([0,1])$. In TPS-ANOVA, when we assume that the main effects are in $H^m([0,1])$, the kth order interactions lie in $\otimes^k H^m([0,1])$, the tensor product space of k $H^m([0,1])$ spaces. Since any function in $\otimes^k H^m([0,1])$ has one derivative of order km (order m in each direction), we can see that TPS-ANOVA puts higher order smoothness conditions on interactions than on main effects, and the order of the smoothness condition imposed on an interaction increases with the order of the interaction. The resulting models enjoy an optimal rate that is very close to the optimal rate of one-dimensional problems. This reveals an intuitively appealing aspect of the tensor product strategy in nonparametric function estimation: starting from an additive model, when we make the model more complex by throwing in higher order interaction terms, we assume appropriately stronger smoothness conditions on the new terms to keep the model manageable, yet do not change the smoothness conditions on existing terms. The resulting model retains a favorable optimal rate.

The main body of the paper is in the next three sections. In Section 2, the tensor product space of Sobolev–Hilbert spaces and the function space for the TPS-ANOVA model are studied. This lays the groundwork for further study of TPS and TPS-ANOVA models. (The TPS model can be seen as a special case of the TPS-ANOVA model, the saturated TPS-ANOVA model.) Sections 3 and 4 are based on Section 2 and are the more statistical parts.

Many nonparametric function estimation problems have white noise counterparts. Brown and Low (1996) showed the asymptotic equivalence between a regression problem and its corresponding white noise problem. Nussbaum (1996) showed the asymptotic equivalence between a density estimation problem and its corresponding white noise problem. These results suggest the importance of considering white noise problems: the treatments in the white noise problem often illustrate the treatments in other types of nonparametric problems. Section 3 establishes the exact rate of the minimax mean integrated squared error for the TPS-ANOVA white noise model, and shows that the penalized likelihood estimator achieves this rate.

In the white noise TPS-ANOVA model, we observe a surface $z_{\mathbf{x}}$ represented as [$\mathbf{x}$ stands for the d-dimensional vector $(x_1, \ldots, x_d)$]

$$dz_{\mathbf{x}} = f_0(\mathbf{x})d\mathbf{x} + n^{-1/2}\omega(\mathbf{x})dB_{\mathbf{x}} \tag{1}$$

where $B_{\mathbf{x}}$ is the d-dimensional Brownian motion, $\omega(\mathbf{x})$ is defined on $[0,1]^d$, and $0 < c_1 < \omega(\mathbf{x}) < c_2 < \infty$ on $[0,1]^d$. Here f_0 is the unknown function of interest. We assume f_0 has a TPS-ANOVA structure, the main effects are in $H^m(0,1)$, and the highest order of interaction is r.

Note. A list of the notation that may not be defined explicitly in the paper is at the end of the paper.

The main result of Section 3 is the theorem.

THEOREM 1.1. *For the TPS-ANOVA white noise model, the rate of the minimax mean integrated squared error is* $[n(\log n)^{1-r}]^{-2m/(2m+1)}$ *and the penalized likelihood estimator achieves this rate.*

Notice this optimal rate is very close to the optimal rate for one-dimensional model. Due to the close relationship of regression models and white noise models [see Brown and Low (1996)], Theorem 1.1 proves that the rate of minimax mean integrated squared error for the TPS-ANOVA regression model is also $[n(\log n)^{1-r}]^{-2m/(2m+1)}$ when the errors are Gaussian and $m > d/2$.

Since the optimal rate is a property of the model itself and does not depend on the method used to fit the model, and the formulation of the TPS-ANOVA models actually does not depend on the penalized likelihood method, we may also fit the model with nonparametric schemes other than the penalized likelihood method, and it is reasonable to hypothesize that some other nonparametric methods can also perform well in this model.

Section 4 establishes the rate of convergence of the penalized likelihood estimator in the TPS-ANOVA regression model. In this model we observe

$$y_i = f_0(x_{1i}, x_{2i}, \ldots, x_{di}) + \varepsilon_i, \qquad i = 1, \ldots, n \tag{2}$$

$(x_{1i}, x_{2i}, \ldots, x_{di})$, $i = 1, \ldots, n$ iid with density $p(\mathbf{x})$. The ε_i's are independent of the $\mathbf{x}_i$'s and independent of each other. $E\,\varepsilon_i = 0$, and $\operatorname{var}\varepsilon_i = \sigma^2$. The regression function f_0 has a TPS-ANOVA structure, the main effects are in $H^m(0,1)$, and the highest order of interaction is r. We further assume that $\mathbf{x}$ takes values only in the unit cube $[0,1]^d$ and that the density, $p(\mathbf{x})$, is bounded away from zero and infinity in the unit cube, that is, $0 < C_1 \le p(\mathbf{x}) \le C_2 < \infty$.

The main result in Section 4 can be stated as Theorem 1.2.

THEOREM 1.2. *Suppose* $m > 1$, *then in the TPS-ANOVA regression model, the (uniform) rate of convergence of the penalized likelihood estimator and its component functions for estimating the regression function* f_0 *and its component functions is* $[n(\log n)^{1-r}]^{-2m/(2m+1)}$ *when* λ, *the smoothing parameter for the roughness penalty, is appropriately chosen.*

Similar results on the rate of convergence of related estimates of derivatives of f_0 are also obtained.

The proof of Theorem 1.2 follows an approach commonly used in the study of the rate of convergence of the penalized likelihood estimators. This approach was first utilized in Silverman (1982). Cox and O'Sullivan (1990) provided a general framework for this approach. Their examples include (one-dimensional) density estimation, hazard estimation, and nonparametric logistic regression. O'Sullivan (1993) applied this framework to the proportional hazards model. Gu and Qiu (1993, 1994) and Gu (1996) provided a simpler analysis in the same line under strong assumptions. They considered the rate of convergence of the penalized likelihood estimators in density estimation, (generalized) regression and hazard estimation. If we combine Gu and Qiu (1994) and the argument in Example 3 of Gu (1996), we get that in the two-dimensional tensor product space regression models, the rate of convergence of the penalized likelihood estimator is $O_p(n^{\varepsilon-4/5})$, $\forall\ \epsilon > 0$ (in their analysis $m = 2$). One drawback of the analysis in Gu and Qiu (1993, 1994) and Gu (1996) is that it makes an assumption on eigenfunctions that is almost impossible to check. Theorem 1.2 represents a sharper result on the rate of convergence of the penalized likelihood estimator in high-dimensional tensor product space ANOVA regression models under general conditions.

2. The parameter space.

2.1. *Tensor product space of Sobolev–Hilbert spaces.* For any nonnegative integer m, the Sobolev–Hilbert space of univariate functions with order m and domain $[0, 1]$, denoted by $H^m([0, 1])$, is defined by

$$H^m([0, 1]) = \{f | f^{(\nu)} \text{ abs. cont.}, \nu = 0, 1, \ldots, m-1; f^{(m)} \in L_2\}$$

with the Sobolev norm

$$\int_0^1 [f(u)]^2\,du + \cdots + \int_0^1 [f^{(m)}(u)]^2\,du.$$

We can see $H^0([0, 1])$ is just the L_2 space on $[0, 1]$.

REMARK. The above definitions are only for integer m. The extensions to fractional m is possible through interpolation. See Oden and Reddy (1976).

Let $\otimes^d H^m$ denote the completed tensor product space of $H^m([0, 1])$ with itself d times. For simplicity, the norm on $\otimes^d H^m$ induced by the Sobolev norm on $H^m([0, 1])$ will be called the Sobolev norm on $\otimes^d H^m$. In later writing, we will denote the Sobolev norms by $\|\cdot\|$ with subscripts. The subscript shows which space the Sobolev norm is on. The corresponding inner product will be denoted $\langle\cdot, \cdot\rangle$ with subscripts. The following two lemmas extend the properties of the univariate Sobolev spaces to the tensor product space of univariate Sobolev spaces. For a proof of these two lemmas, see Lin (1998).

LEMMA 2.1. *For any $s > 1/2$, and any $f \in \otimes^d H^s$, there exists a constant C not depending on f, such that* $\sup|f(x_1, x_2, \ldots, x_d)| \le C\|f\|_{\otimes^d H^s}$.

LEMMA 2.2. *For any $s > 1/2$, there exists a constant C depending only on s, such that for any $f, g \in \otimes^d H^s$,*

$$\|fg\|_{\otimes^d H^s} \le C\|f\|_{\otimes^d H^s}\|g\|_{\otimes^d H^s}.$$

2.2. *ANOVA decomposition.* With the Sobolev norm, $H^m([0,1])$ can be decomposed as the direct sum of two orthogonal Hilbert subspaces,

$$H^m([0,1]) = \{1\} \oplus H_0^m([0,1]),$$

where $\{1\}$ is the space of scalars. $H_0^m([0,1])$ is the subspace (orthogonal to $\{1\}$) satisfying $\int_0^1 f(x)\,dx = 0$. Thus we have

$$\bigotimes^d H^m = \bigotimes^d \left[\{1\} \oplus H_0^m([0,1])\right].$$

Identify the tensor product of $\{1\}$ with any Hilbert space with that Hilbert space itself, then $\otimes^d H^m$ is the direct sum of all the subspaces of the form $H_0^m(x_{j_1}) \otimes H_0^m(x_{j_2}) \otimes \cdots \otimes H_0^m(x_{j_k})$ and $\{1\}$, where $\{j_1, j_2, \ldots, j_k\}$ is a subset of $\{1, 2, \ldots, d\}$, and the subspaces in the decomposition are all orthogonal to each other.

We can now study the TPS-ANOVA structure with this decomposition. In the TPS-ANOVA models we represent a d-dimensional function as

$$f(x_1, x_2, \ldots, x_d) = \text{constant} + \sum_{i=1}^{d} f_i(x_i) + \sum_{i<j} f_{ij}(x_i, x_j) + \cdots. \tag{3}$$

Let r be the highest order of interaction in the model. We will impose the smoothness condition on the main effects by assuming $f_i(x_i) \in H_0^m(x_i)$ for $i = 1, 2, \ldots, d$. Due to the TPS structure, this assumption effectively determines the function space for the function f. Denote this function space by F. Then F is the direct sum of some set of the orthogonal subspaces in the decomposition of $\otimes^d H^m$. Let F_0 be the direct sum of the corresponding set of subspaces in the corresponding decomposition of $\otimes^d H^0$. The restriction of the Sobolev norm on $\otimes^d H^m$ ($\otimes^d H^0$) to F (F_0) will be called the Sobolev norm on F (F_0), and will be denoted by $\|\cdot\|_F$ ($\|\cdot\|_{F_0}$).

2.3. *Eigensystem.* Consider the spaces of univariate functions $H_0^m([0,1]) \subset H_0^0([0,1])$. In $H_0^m([0,1])$, the quadratic form $\langle f, f\rangle_{H^0}$ is completely continuous with respect to $\langle f, f\rangle_{H^m}$. Apply the theory in Weinberger [(1974), Section 3.3], and denote the eigenvalues and eigenvectors of the Rayleigh quotient of the two quadratic forms by $\{\mu_i\}$ and $\{\mu_i^{1/2}\varphi_i\}$, $i = 2, 3, \ldots$. Then $1 \ge \mu_2 \ge \mu_3 \ge \cdots$ and $\langle \varphi_i, \varphi_j\rangle_{H^m} = \mu_i^{-1}\delta_{ij}$, $\langle \varphi_i, \varphi_j\rangle_{H^0} = \delta_{ij}$.

Since $\langle f, f\rangle_{H^0}$ is positive definite, by Section 3.3 of Weinberger (1974), $\{\varphi_i\}$ actually form an orthogonal basis in $H_0^m([0,1])$. Therefore, $\varphi_1 = 1, \varphi_2, \ldots$ form an orthogonal basis in $H^m([0,1]) = \{1\} \oplus H_0^m([0,1])$, and $1 = \mu_1 \ge \mu_2 \ge \cdots$ are the eigenvalues of the Rayleigh quotient $\|\cdot\|_{H^0}^2/\|\cdot\|_{H^m}^2$ in $H^m([0,1])$. Since $H^m([0,1])$ is dense in $H^0([0,1])$, and since $\varphi_1, \varphi_2, \ldots$ form an orthonormal

system in $H^0([0,1])$, we know that $\varphi_1, \varphi_2, \ldots$ also form an orthonormal basis in $H^0([0,1])$.

From Silverman [(1982), Section 5] or Cox (1988), $\mu_i \sim i^{-2m}$.

Denote $\varphi_{i_1}(x_1)\varphi_{i_2}(x_2)\cdots\varphi_{i_d}(x_d)$ by $\varphi_{i_1 i_2 \cdots i_d}$ and $\mu_{i_1}\mu_{i_2}\cdots\mu_{i_d}$ by $\mu_{i_1 i_2 \cdots i_d}$. Since $\{\varphi_i\}$ is an orthogonal basis in H^m, and an orthonormal basis in H^0, $\{\varphi_{i_1 i_2 \cdots i_d}\}$ form an orthogonal basis in $\otimes^d H^m$, and an orthonormal basis in $\otimes^d H^0$. It is easy to see that $\{\mu_{i_1 i_2 \cdots i_d}\}$, where i_j goes from 1 to ∞, $j = 1, 2, \ldots, d$, are the eigenvalues of the Rayleigh quotient $\langle f, f\rangle_{\otimes^d H^0}/\langle f, f\rangle_{\otimes^d H^m}$ in $\otimes^d H^m$.

By the decomposition of F and F_0, a subset of $\{\varphi_{i_1 i_2 \cdots i_d}\}$ forms an orthogonal basis in F and an orthonormal basis in F_0. Order the corresponding subset of $\{\mu_{i_1 i_2 \cdots i_d}\}$ from large to small, write them as $\{\nu_i\}$. Then $\{\nu_i\}$ is the eigenvalues of the Rayleigh quotient $\langle f, f\rangle_{F_0}/\langle f, f\rangle_F$ in F.

2.4. *A norm related to the penalized likelihood estimator.* The penalized likelihood method augments the negative log-likelihood with a roughness penalty, and minimizes the penalized negative log-likelihood. The roughness penalty is a quadratic functional on F.

DEFINITION 2.1. A standard roughness penalty is a quadratic functional $J(\cdot)$ such that $\int f^2 + J(f)$ is a norm equivalent to the Sobolev norm.

Most of the commonly used penalties are standard. Later we will restrict our attention to standard penalties.

Note. The most commonly used roughness penalty is introduced in Wahba (1990). It can be shown that this roughness penalty is standard. See Lin (1998).

Let $p(\mathbf{x})$ be a function supported on the unit cube $[0,1]^d$ and bounded away from zero and infinity in the unit cube, that is, $0 < c_1 \le p(\mathbf{x}) \le c_2$. Given any standard penalty $J(\cdot)$, we can define a new norm on F:

$$\|f\|^2 = \int f^2(\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x} + J(f) \qquad \forall\, f \in F.$$

The corresponding inner product is denoted by $\langle\cdot,\cdot\rangle$.

Note. The notation for the norms are a little confusing. While $\|\cdot\|$ without any subscript is the new norm on F defined above, $\|\cdot\|$ with subscripts are the Sobolev norms, with the subscripts indicating the space which the Sobolev norms are on. However, since $\|\cdot\|$ and the Sobolev norm $\|\cdot\|_F$ are equivalent norms on F, and are largely interchangeable in later developments, the risk of a confusion is not serious. From now on, we will be mainly using the new norm $\|\cdot\|$ on F, and will state clearly each time the Sobolev norm is used.

Similarly, we can define a new norm on F_0,

$$\|f\|_0^2 = \int f^2(\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x} \qquad \forall\, f \in F_0.$$

The corresponding inner product is denoted by $\langle\cdot,\cdot\rangle_0$. This new norm $\|\cdot\|_0$ is equivalent to the Sobolev norm $\|\cdot\|_{F_0}$ on F_0.

Denote the eigenvalues and eigenvectors of the Rayleigh quotient $\langle f, f\rangle_0/\langle f, f\rangle$ in F by $\{(1+\rho_k)^{-1}\}$ and $\{(1+\rho_k)^{-1/2}\phi_k\}$, $k = 1, 2, \ldots$. Then $\langle \phi_i, \phi_j\rangle = (1+\rho_i)\delta_{ij}$, $\langle \phi_i, \phi_j\rangle_0 = \delta_{ij}$. Since $\|\cdot\| \sim \|\cdot\|_F$, and $\|\cdot\|_0 \sim \|\cdot\|_{F_0}$, by the mapping principle [or Theorem 3.8.1 in Weinberger (1974)], we have $(1+\rho_i)^{-1} \sim \nu_i$.

The following quantity plays an important role in determining the rate of convergence of penalized likelihood type estimators and will appear frequently in later derivation:

$$N_b(\lambda) = \sum_{i=1}^{\infty}(1+\rho_i)^b(1+\lambda\rho_i)^{-2}.$$

LEMMA 2.3. *For any $b \in [0, 2-1/2m)$, we have $N_b(\lambda) = O\big[\lambda^{-(b+1/2m)}\times \big(\log(1/\lambda)\big)^{r-1}\big]$ as λ goes to zero. Here r is the highest order of interaction in the TPS-ANOVA model.*

The proof of the lemma is deferred to the Appendix.

3. The TPS-ANOVA white noise model. We now consider the white noise TPS-ANOVA model defined at (1). To obtain uniform results, we consider only functions satisfying $\|f_0\| < B$.

Let θ_{fi} be the coefficients when we expand $f \in F$ in terms of the basis $\{\phi_i\}$ of F defined in Section 2.4 with $p(\mathbf{x}) = \omega^{-2}(\mathbf{x})$, the TPS-ANOVA white noise model is equivalent to the following Gaussian shift model:

$$y_i = \theta_{f_0 i} + \varepsilon_i, \qquad i = 1, 2, \ldots,$$

where $\varepsilon_i \sim N(0, 1/n)$ are independent noises, $\theta_{f_0 i}$'s are parameters of interest, satisfying $\sum(1+\rho_i)\theta_{f_0 i}^2 = \|f_0\|^2 < B^2$.

3.1. *Penalized likelihood estimator.* For any $f \in F$, we have $J(f) = \|f\|^2 - \|f\|_0^2 = \sum \rho_i \theta_{fi}^2$. Hence the penalized likelihood estimator for $\theta_{f_0 i}$ is the minimizer of the following:

$$\sum(y_i - \theta_i)^2 + \lambda \sum \rho_i \theta_i^2.$$

Solving the minimization problem, we get the penalized likelihood estimator $\hat\theta_i = (1+\lambda\rho_i)^{-1} y_i$. Therefore,

$$\sum(E\hat\theta_i - \theta_{f_0 i})^2 = \sum \lambda^2\rho_i^2(1+\lambda\rho_i)^{-2}\theta_{f_0 i}^2 \le \tfrac{1}{4}\lambda \sum \rho_i \theta_{f_0 i}^2 \le \tfrac{1}{4}\lambda B^2,$$

$$\sum \operatorname{var}\hat\theta_i = 1/n \sum(1+\lambda\rho_i)^{-2} = N_0(\lambda)/n \le Cn^{-1}\lambda^{-1/2m}\left(\log\frac{1}{\lambda}\right)^{r-1}.$$

Hence we have

$$E\int(\hat f - f_0)^2\, d\mathbf{x} \sim E\sum(\hat\theta_i - \theta_{f_0 i})^2 \le C\left(\lambda + n^{-1}\lambda^{-1/2m}\left(\log\frac{1}{\lambda}\right)^{r-1}\right).$$

The constant C here is uniform for all $\|f_0\| < B$. If $\lambda \sim [n(\log n)^{1-r}]^{-2m/(2m+1)}$, then $E\int(\hat f - f_0)^2\, d\mathbf{x} \le C[n(\log n)^{1-r}]^{-2m/(2m+1)}$. This shows that the penalized likelihood estimator achieves a uniform rate of convergence $[n(\log n)^{1-r}]^{-2m/(2m+1)}$.

3.2. *The minimax rate.* We now show that $[n(\log n)^{1-r}]^{-2m/(2m+1)}$ is the rate of minimax mean integrated squared error for the TPS-ANOVA white noise model. Since the penalized likelihood estimator can achieve this rate, all we need to show is that the minimax rate of the TPS-ANOVA white noise model is at least $[n(\log n)^{1-r}]^{-2m/(2m+1)}$. For this we only need to show the minimax rate for the r-dimensional full TPS white noise model is at least $[n(\log n)^{1-r}]^{-2m/(2m+1)}$. In the following we consider the r-dimensional full TPS white noise model.

Again consider the Gaussian shift model. In this situation, $\{\nu_i, i = 1, 2, \ldots\}$ is the same set as $\{\mu_{i_1} \cdots \mu_{i_r}, i_j = 1, 2, \ldots; j = 1, \ldots, r\}$. Since $(1 + \rho_i)^{-1} \sim \nu_i$, and $\mu_i \sim i^{-2m}$, without loss of generality, we can assume the set $\{1 + \rho_i, i = 1, 2, \ldots\}$ is the same as the set $\{i_1^{2m} \cdots i_r^{2m}, i_j = 1, 2, \ldots; j = 1, \ldots r\}$, and $B = 1$.

By Lemma 6, Theorem 7 and the proof of Theorem 7 in Donoho, Liu and MacGibbon (1990), the difficulty (the minimax risk) of the r-dimensional full TPS Gaussian shift model, is larger than 80% of the difficulty, for linear estimates, of the hardest rectangle subproblem, and the latter difficulty, which we will denote by R^*, is

$$\max\left[\sum n^{-2}\left(\pi_i^2 + \frac{1}{n}\right)^{-2}\pi_i^2 + \frac{1}{n}\pi_i^4\left(\pi_i^2 + \frac{1}{n}\right)^{-2}\right]$$

or

$$\max\left[\sum \frac{1}{n}\pi_i^2\left(\pi_i^2 + \frac{1}{n}\right)^{-1}\right].$$

The maximization is taken over π under the constraint $\sum(1 + \rho_i)\pi_i^2 = 1$.

For notational simplicity, we will write π_i^2 as γ_i. We use the Lagrange multiplier method to find nonnegative $\bar{\gamma}_i$ that maximize $[\sum(1/n)\gamma_i(\gamma_i + (1/n))^{-1}]$ under the constraint

$$\sum(1 + \rho_i)\gamma_i = 1. \tag{4}$$

Let

$$A = \sum \gamma_i\left(\gamma_i + \frac{1}{n}\right)^{-1} - a(1 + \rho_i)\gamma_i,$$

where a is a scalar. Then

$$\frac{\partial A}{\partial \gamma_i} = n^{-1}\left(\gamma_i + \frac{1}{n}\right)^{-2} - a(1 + \rho_i).$$

Maximizing A under the constraint $\gamma_i \geq 0$, $\forall\, i$, we get

$$\bar{\gamma}_i = \left[b(1 + \rho_i)^{-1/2} - \frac{1}{n}\right]_+,$$

where $b = (na)^{-1/2}$ is a scalar.

By (4), we have

$$\sum(1+\rho_i)\left[b(1+\rho_i)^{-1/2}-\frac{1}{n}\right]_+=1.$$

That is,

$$\sum_{i_1,\,i_2,\ldots,i_r} i_1^{2m}i_2^{2m}\cdots i_r^{2m}\left[bi_1^{-m}i_2^{-m}\cdots i_r^{-m}-\frac{1}{n}\right]_+=1.$$

From now on we will use multiindex notation. For example, the expression $i_1^k i_2^k\cdots i_r^k$ will be written as $\underline{i}^k$. We have

$$\sum_{\underline{i}^m\le(nb)}\underline{i}^{2m}\left[b\underline{i}^{-m}-\frac{1}{n}\right]=1,$$

$$\sum_{\underline{i}^1\le(nb)^{1/m}} b\underline{i}^m-\frac{1}{n}\underline{i}^{2m}=1.$$

From this, we can see that $nb\to\infty$. Using the integral approximation, we have

$$\int_{\mathbf{x}^1\le(nb)^{1/m},\,x_i\ge1,\,i=1,\ldots,r} b\mathbf{x}^m-\frac{1}{n}\mathbf{x}^{2m}d\mathbf{x}\sim1$$

Changing the variable in the integral, $z_i=\Pi_{j\le i}x_j$, $i=1,2,\ldots,r$, we get

$$\int_1^{(nb)^{1/m}}\left[\int_1^{z_r}\cdots\int_1^{z_2}\left(bz_r^m-\frac{1}{n}z_r^{2m}\right)z_1^{-1}\cdots z_{r-1}^{-1}\,dz_1\cdots dz_{r-1}\right]dz_r\sim1,$$

that is,

$$\int_1^{(nb)^{1/m}}\left[\left(bz_r^m-\frac{1}{n}z_r^{2m}\right)\left(\log z_r\right)^{r-1}\right]dz_r\sim1.$$

Integrating by parts, since $nb\to\infty$, the left-hand side can be shown to be of the order $n^{(m+1)/m}b^{(2m+1)/m}[\log(nb)]^{r-1}$. Then from $n^{(m+1)/m}b^{(2m+1)/m}\times[\log(nb)]^{r-1}\sim1$, we get

$$b\sim n^{-(m+1)/(2m+1)}[\log n]^{-m(r-1)/(2m+1)}. \tag{5}$$

Now we have

$$\begin{aligned}R^*&=\sum\frac{1}{n}\overline{\gamma}_i\left(\overline{\gamma}_i+\frac{1}{n}\right)^{-1}\\&=\sum_{\underline{i}^1\le(nb)^{1/m}}\frac{1}{n}\left[b\underline{i}^{-m}-\frac{1}{n}\right]\Big/(b\underline{i}^{-m})\\&=\sum_{\underline{i}^1\le(nb)^{1/m}}\frac{1}{n}\left[1+\frac{1}{nb}\underline{i}^m\right]\\&\sim\int_{\mathbf{x}^1\le(nb)^{1/m},\,x_i\ge1,\,i=1,\ldots,r}\frac{1}{n}\left[1+\frac{1}{nb}\mathbf{x}^m\right]d\mathbf{x}.\end{aligned}$$

Again changing variables in the integral, and then integrating by parts, we get $R^* \sim n^{-(m-1)/m} b^{1/m} (\log(nb))^{r-1}$. By (5), we get $R^* \sim [n(\log n)^{1-r}]^{-2m/(2m+1)}$.

To summarize, from the results above, we get Theorem 1.1 as stated in Section 1.

4. The penalized likelihood estimator in the TPS-ANOVA regression model. We now consider the TPS-ANOVA regression model as defined in (2).

To obtain uniform results, we consider the regression functions satisfying $\|f_0\| < B$ for some positive constant B.

The penalized likelihood estimator for the regression function is the minimizer of $l_{n,\lambda}(f) = l_n(f) + \lambda J(f)$ in F, where

$$l_n(f) = \frac{1}{n} \sum_{i=1}^{n} (f(x_{1i}, \ldots, x_{di}) - y_i)^2$$

and $J(\cdot)$ is a standard roughness penalty. The smoothing parameter λ depends on n, that is, $\lambda = \lambda_{(n)}$.

For large n, with probability 1, $l_{n,\lambda}(f)$ is a positive definite quadratic form of f; it has a unique minimizer in F, which we denote by $\hat{f}$. We have

$$Dl_{n,\lambda}(\hat{f}) = 0 \tag{6}$$

here D is the notation for the Fréchet derivative. For a definition of the Fréchet derivative, see Huber (1981), for example.

Let $l_\infty(f) = \int (f(\mathbf{x}) - f_0(\mathbf{x}))^2 p(\mathbf{x})\, d\mathbf{x} + \sigma^2$, $l_{\infty,\lambda}(f) = l_\infty(f) + \lambda J(f)$. Since $l_{\infty,\lambda}(f)$ is a positive definite quadratic form in f, it has a unique minimizer in F, which we denote by $\bar{f}$. We have

$$Dl_{\infty,\lambda}(\bar{f}) = 0.$$

Notice, $\bar{f} - f_0$ is the deterministic part of the estimation error, and $\hat{f} - \bar{f}$ is the stochastic part. We will study them separately.

4.1. *Intermediate spaces.* In order to study the estimation of the derivatives of the unknown function, and for some technical reasons, we need to introduce the intermediate spaces between F and F_0. The proof of the following proposition is in the Appendix.

PROPOSITION 4.1. *The natural injection from F with norm $\|\cdot\|$ to F_0 with norm $\|\cdot\|_0$ is continuous and dense.*

This proposition makes sure that the concept of intermediate spaces applies for F and F_0. By Section 4.8 of Oden and Reddy (1976), we can define the intermediate spaces between F_0 and F as follows.

For $b \in [0, 1]$, and $\theta \in F_0$, let

$$\|\theta\|_b = \left\{ \sum_{i=1}^{\infty} (1 + \rho_i)^b \langle \theta, \phi_i \rangle_0^2 \right\}^{1/2}.$$

Let Θ_b be the normed linear space $\{\theta \in F_0 : \|\theta\|_b < \infty\}$, with $\|\cdot\|_b$ norm, then Θ_b, $0 \le b \le 1$, is the intermediate spaces between F_0 and F. Θ_1 is the same as F, and Θ_0 is the same as F_0. Θ_b, $0 \le b \le 1$, is a Hilbert space with inner product

$$\langle \theta, \xi \rangle_b = \sum_{i=1}^{\infty} (1+\rho_i)^b \langle \theta, \phi_i \rangle_0 \langle \xi, \phi_i \rangle_0.$$

If $b \le a$, then $\Theta_a \subset \Theta_b$, and Θ_a has a stronger norm.

PROPOSITION 4.2. *For any $b \in [0,1]$, we have $\Theta_b \subset \otimes^d H^{mb}([0,1])$. Also, for any $f \in F$, and its component functions defined by* (3)*, we have*

$$\|f\|_b^2 \sim c^2 + \sum_i \|f_i\|^2_{H^{mb}([0,1])} + \sum_{i<j} \|f_{ij}\|^2_{\otimes^2 H^{mb}([0,1])} + \cdots .$$

Here $H^{mb}([0,1])$ is the (fractional) Sobolev space of order mb, $\|\cdot\|_{H^{mb}([0,1])}$ and $\|\cdot\|_{\otimes^2 H^{mb}([0,1])}$ are Sobolev norms. The proof of this proposition is given in the Appendix.

4.2. *The deterministic error.* Let $a_{f1}, a_{f2}, \ldots$ be the coefficients when we expand $f \in F$ in terms of ϕ_i, $i = 1, 2, \ldots$; then

$$\begin{aligned} f_0 &= \sum_{i=1}^{\infty} a_{f_0 i} \phi_i, \\ f &= \sum_{i=1}^{\infty} a_{fi} \phi_i, \\ l_{\infty,\lambda}(f) &= \int (f(\mathbf{x}) - f_0(\mathbf{x}))^2 p(\mathbf{x})\, d\mathbf{x} + \sigma^2 + \lambda J(f) \\ &= \sigma^2 + \sum_{i=1}^{\infty} (a_{fi} - a_{f_0 i})^2 + \lambda \sum_{i=1}^{\infty} \rho_i a_{fi}^2, \end{aligned}$$

so the minimizer $\bar{f}$ has $a_{\bar{f}i} = a_{f_0 i}/(1+\lambda\rho_i)$. Hence,

$$\begin{aligned} \|\bar{f} - f_0\|_b^2 &= \sum_{i=1}^{\infty} [a_{f_0 i} - a_{f_0 i}(1+\lambda\rho_i)^{-1}]^2 (1+\rho_i)^b \\ &\le \lambda^{1-b} \sum_{i=1}^{\infty} (1+\rho_i) a_{f_0 i}^2 = \lambda^{1-b} \|f_0\|^2 \le \lambda^{1-b} B^2. \end{aligned}$$

Therefore we have the following lemma.

LEMMA 4.1. $\|\bar{f} - f_0\|_b \le \lambda^{(1/2)(1-b)} \|f_0\| < \lambda^{(1/2)(1-b)} B \qquad \forall\ 0 \le b \le 1.$

4.3. *The stochastic error.* Direct calculation yields the following derivatives. Lemma 2.1 shows that evaluation is a continuous linear functional in Θ_b for $b > 1/2m$. Hence these derivatives are bounded linear functionals:

$$Dl_\infty(f)g = 2\int (f - f_0)(\mathbf{x})g(\mathbf{x})p(\mathbf{x})\,d\mathbf{x},$$

$$Dl_n(f)g = -\frac{2}{n}\sum_{i=1}^{n}[(y_i - f(x_{1i}, \ldots, x_{di}))g(x_{1i}, \ldots, x_{di})],$$

$$D^2 l_\infty(f)gh = 2\int g(\mathbf{x})h(\mathbf{x})p(\mathbf{x})\,d\mathbf{x} = 2\langle g, h\rangle_0,$$

$$D^2 l_n(f)gh = \frac{2}{n}\sum_{i=1}^{n}[g(x_{1i}, \ldots, x_{di})h(x_{1i}, \ldots, x_{di})].$$

By Oden and Reddy (1976), there exists a bounded linear operator U from F_0 into F, such that, $\langle f, Ug\rangle = \langle f, g\rangle_0$, $\forall\ f \in F$, $g \in F_0$. The restriction of U to F is self-adjoint and positive definite.

Since $J(f) = \langle f, f\rangle - \langle f, f\rangle_0$, we have

$$D^2 l_{\infty,\lambda}(f)gh = 2\langle (U + \lambda(I - U))g, h\rangle.$$

Since F is a Hilbert space with inner product $\langle\cdot,\cdot\rangle$, the conjugate space of F can be identified with F. We will not distinguish the two in later writing. Let $G_\lambda = U + \lambda(I - U)$. The equation above implies, for any $g \in F$,

$$G_\lambda g = (1/2)D^2 l_{\infty,\lambda}(\bar f)g.$$

By Theorem 3.1.1 of Weinberger (1974), we have $U\phi_i = (1+\rho_i)^{-1}\phi_i$. Hence $G_\lambda\phi_i = (1+\rho_i)^{-1}(1+\lambda\rho_i)\phi_i$. By the Lax–Milgram theorem [see Aubin (1979)], G_λ as an operator from $F \to F$ has a bounded inverse on F. We have $G_\lambda^{-1}\phi_i = (1 + \rho_i)(1 + \lambda\rho_i)^{-1}\phi_i$, and for any $\theta \in F$,

$$\|G_\lambda^{-1}\theta\|_b^2 = \sum_{i=1}^{\infty}(1 + \rho_i)^b(1 + \lambda\rho_i)^{-2}\langle\theta, \phi_i\rangle^2.$$

For notational purpose, let $\tilde f = \bar f - \frac{1}{2}G_\lambda^{-1}Dl_{n,\lambda}(\bar f)$. This is an approximation of $\hat f$: Since $l_{n,\lambda}(f)$ is a quadratic form of f, a Taylor expansion of $Dl_{n,\lambda}(\hat f)$ around $\bar f$ shows that (6) is equivalent to

$$Dl_{n,\lambda}(\bar f) + D^2 l_{n,\lambda}(\bar f)(\hat f - \bar f) = 0, \tag{7}$$

and by definition $\tilde f$ satisfies a similar equation,

$$Dl_{n,\lambda}(\bar f) + D^2 l_{\infty,\lambda}(\bar f)(\tilde f - \bar f) = 0. \tag{8}$$

We will study $\tilde f - \bar f$ first, then $\hat f - \tilde f$.

LEMMA 4.2. $\|\tilde f - \bar f\|_b^2 = O_p[n^{-1}\lambda^{-(b+1/2m)}(\log(1/\lambda))^{r-1}] \qquad \forall\ b \in [0, 1].$

PROOF. Notice $Dl_{n,\lambda}(\bar{f}) = Dl_{n,\lambda}(\bar{f}) - Dl_{\infty,\lambda}(\bar{f}) = Dl_n(\bar{f}) - Dl_\infty(\bar{f})$. Direct calculation gives, for any nonrandom $g \in F$,

$$
\begin{aligned}
&E\left(\tfrac{1}{2}Dl_{n,\lambda}(\bar{f})g\right)^2 \\
&\quad = E\left(\tfrac{1}{2}Dl_n(\bar{f})g - \tfrac{1}{2}Dl_\infty(\bar{f})g\right)^2 \\
&\quad = \frac{1}{n}\mathrm{var}[(Y - \bar{f}(X_1,\ldots,X_d))g(X_1,\ldots,X_d)] \\
&\quad \le \frac{1}{n}\{E[(f_0(X_1,\ldots,X_d) - \bar{f}(X_1,\ldots,X_d))^2 g^2(X_1,\ldots,X_d)] \\
&\qquad\qquad + \sigma^2 E g^2(X_1,\ldots,X_d)\} \\
&\quad \le \frac{1}{n}\{E[C^2\|f_0 - \bar{f}\|^2 g^2(X_1,\ldots,X_d)] + \sigma^2 E g^2(X_1,\ldots,X_d)\} \\
&\qquad\qquad \text{(by Lemma 2.1)} \\
&\quad \le \frac{1}{n}\{E[C^2 B^2 g^2(X_1,\ldots,X_d)] + \sigma^2 E g^2(X_1,\ldots,X_d)\} \\
&\qquad\qquad \text{(by Lemma 4.1)} \\
&\quad < Cn^{-1}\|g\|_0^2.
\end{aligned}
$$

Thus we have

$$
\begin{aligned}
E\|\tilde{f} - \bar{f}\|_b^2 &= E\left\|\frac{1}{2}G_\lambda^{-1}Dl_{n,\lambda}(\bar{f})\right\|_b^2 \\
&= E\left[\sum_{i=1}^{\infty}(1+\lambda\rho_i)^{-2}\left(\frac{1}{2}Dl_{n,\lambda}(\bar{f})\phi_i\right)^2(1+\rho_i)^b\right] \\
&\le \sum_{i=1}^{\infty}(1+\lambda\rho_i)^{-2}(1+\rho_i)^b E\left(\frac{1}{2}Dl_{n,\lambda}(\bar{f})\phi_i\right)^2 \\
&\le \frac{C}{n}\sum_{i=1}^{\infty}(1+\lambda\rho_i)^{-2}(1+\rho_i)^b \\
&= Cn^{-1}N_b(\lambda) \\
&\le Cn^{-1}\lambda^{-(b+1/2m)}\left(\log\frac{1}{\lambda}\right)^{r-1},
\end{aligned}
$$

which implies $\|\tilde{f} - \bar{f}\|_b^2 = O_p[n^{-1}\lambda^{-(b+1/2m)}(\log(1/\lambda))^{r-1}]$. □

PROPOSITION 4.3.

$$\hat{f} - \tilde{f} = G_\lambda^{-1}[1/2D^2 l_\infty(\bar{f})(\hat{f} - \bar{f}) - 1/2D^2 l_n(\bar{f})(\hat{f} - \bar{f})].$$

PROOF. By (7) and (8), we have $1/2D^2 l_{\infty,\lambda}(\bar{f})(\tilde{f}-\bar{f}) = 1/2D^2 l_{n,\lambda}(\bar{f})(\hat{f}-\bar{f})$, and therefore

$$\begin{aligned} G_\lambda(\hat{f}-\tilde{f}) &= 1/2D^2 l_{\infty,\lambda}(\bar{f})(\hat{f}-\tilde{f}) \\ &= 1/2D^2 l_{\infty,\lambda}(\bar{f})(\hat{f}-\bar{f}) - 1/2D^2 l_{n,\lambda}(\bar{f})(\hat{f}-\bar{f}) \\ &= 1/2D^2 l_{\infty}(\bar{f})(\hat{f}-\bar{f}) - 1/2D^2 l_{n}(\bar{f})(\hat{f}-\bar{f}) \end{aligned}$$

and the conclusion follows. □

LEMMA 4.3. *If* $n^{-1}\lambda^{-(2b+1/2m)}(\log(1/\lambda))^{r-1} \to 0$ *and* $1 \geq b > 1/2m$, *then for* $a \in [0, b]$,

$$\|\hat{f}-\tilde{f}\|_a^2 = o_p\left[n^{-1}\lambda^{-(a+1/2m)}\left(\log\frac{1}{\lambda}\right)^{r-1}\right]. \tag{9}$$

PROOF. First notice that (9) follows from the following: for any $a \in [0, b]$,

$$\|\hat{f}-\tilde{f}\|_a^2 = O_p\left[n^{-1}\lambda^{-(a+b+1/2m)}\left(\log\frac{1}{\lambda}\right)^{r-1}\right]\|\hat{f}-\bar{f}\|_b^2. \tag{10}$$

This is because once (10) is established, we have, by plugging $a = b$ in (10),

$$\|\hat{f}-\tilde{f}\|_b^2 = O_p\left[n^{-1}\lambda^{-(2b+1/2m)}\left(\log\frac{1}{\lambda}\right)^{r-1}\right]\|\hat{f}-\bar{f}\|_b = o_p(1)\|\hat{f}-\bar{f}\|_b^2.$$

By the triangle inequality, we have

$$\begin{aligned} &\|\tilde{f}-\bar{f}\|_b \geq \|\hat{f}-\bar{f}\|_b - \|\hat{f}-\tilde{f}\|_b = (1-o_p(1))\|\hat{f}-\bar{f}\|_b, \\ &\|\hat{f}-\bar{f}\|_b^2 = O_p(\|\tilde{f}-\bar{f}\|_b^2) = O_p\left[n^{-1}\lambda^{-(b+1/2m)}\left(\log\tfrac{1}{\lambda}\right)^{r-1}\right]. \end{aligned} \tag{11}$$

The second equality follows from Lemma 4.2. Combining (10) and (11), we get (9).

Now we set out to prove (10). By Proposition 4.3,

$$\begin{aligned} \|\hat{f}-\tilde{f}\|_a^2 &= \|G_\lambda^{-1}[1/2D^2 l_\infty(\bar{f})(\hat{f}-\bar{f}) - 1/2D^2 l_n(\bar{f})(\hat{f}-\bar{f})]\|_a^2 \\ &= \sum_{i=1}^{\infty}(1+\rho_i)^a(1+\lambda\rho_i)^{-2} \\ &\qquad \times\left[\frac{1}{2}D^2 l_n(\bar{f})(\hat{f}-\bar{f})\phi_i - \frac{1}{2}D^2 l_\infty(\bar{f})(\hat{f}-\bar{f})\phi_i\right]^2 \qquad (12) \\ &= \sum_{i=1}^{\infty}(1+\rho_i)^a(1+\lambda\rho_i)^{-2} \\ &\qquad \times\left[\frac{1}{n}\sum_{j=1}^{n}(\hat{f}-\bar{f})(\mathbf{x}_j)\phi_i(\mathbf{x}_j) - \int(\hat{f}-\bar{f})(\mathbf{x})\phi_i(\mathbf{x})p(\mathbf{x})\right]^2. \end{aligned}$$

Let $g = (\hat{f}-\bar{f})\phi_i$. By Lemma 2.2,

$$\|g\|_{\otimes^d H^{mb}([0,1])} \leq C\|\hat{f}-\bar{f}\|_b\|\phi_i\|_b.$$

Now recall $\varphi_{i_1\cdots i_d}$ and $\mu_{i_1\cdots i_d}$ from Section 2.3. Let $Q_{gi_1\cdots i_d}$ be the coefficients in the expansion of g in terms of $\varphi_{i_1\cdots i_d}$. We will use the multiindex notation in the following (write $i_1\cdots i_d$ as $\underline{i}$):

$$g = \sum_{\underline{i}} Q_{g\underline{i}}\varphi_{\underline{i}}.$$

Then we have

$$\begin{aligned}
&\left[\frac{1}{n}\sum_{j=1}^{n}(\hat f-\bar f)(\mathbf{x}_j)\phi_i(\mathbf{x}_j) - \int(\hat f-\bar f)(\mathbf{x})\phi_i(\mathbf{x})p(\mathbf{x})\right]^2 \\
&\quad= \left[\frac{1}{n}\sum_{j=1}^{n} g(\mathbf{x}_j) - \int g(\mathbf{x})p(\mathbf{x})\right]^2 \\
&\quad= \left[\sum_{\underline{i}}\left(Q_{g\underline{i}}\frac{1}{n}\sum_{j=1}^{n}\varphi_{\underline{i}}(\mathbf{x}_j)\right) - \sum_{\underline{i}}\left(Q_{g\underline{i}}\int\varphi_{\underline{i}}(\mathbf{x})p(\mathbf{x})\right)\right]^2 \\
&\quad= \left[\sum_{\underline{i}} Q_{g\underline{i}}\left[\frac{1}{n}\sum_{j=1}^{n}\varphi_{\underline{i}}(\mathbf{x}_j) - \int\varphi_{\underline{i}}(\mathbf{x})p(\mathbf{x})\right]\right]^2 \\
&\quad\le \left[\sum_{\underline{i}} Q_{g\underline{i}}^2\mu_{\underline{i}}^{-b}\right]\left[\sum_{\underline{i}}\mu_{\underline{i}}^{b}\left[\frac{1}{n}\sum_{j=1}^{n}\varphi_{\underline{i}}(\mathbf{x}_j) - \int\varphi_{\underline{i}}(\mathbf{x})p(\mathbf{x})\right]^2\right] \\
&\quad= \|g\|^2_{\otimes^d H^{mb}([0,1])}\left[\sum_{\underline{i}}\mu_{\underline{i}}^{b}\left[\frac{1}{n}\sum_{j=1}^{n}\varphi_{\underline{i}}(\mathbf{x}_j) - \int\varphi_{\underline{i}}(\mathbf{x})p(\mathbf{x})\right]^2\right] \\
&\quad\le C\|\hat f-\bar f\|_b^2\,\|\phi_i\|_b^2\left[\sum_{\underline{i}}\mu_{\underline{i}}^{b}\left[\frac{1}{n}\sum_{j=1}^{n}\varphi_{\underline{i}}(\mathbf{x}_j) - \int\varphi_{\underline{i}}(\mathbf{x})p(\mathbf{x})\right]^2\right] \\
&\quad= C(1+\rho_i)^b\|\hat f-\bar f\|_b^2\left[\sum_{\underline{i}}\mu_{\underline{i}}^{b}\left[\frac{1}{n}\sum_{j=1}^{n}\varphi_{\underline{i}}(\mathbf{x}_j) - \int\varphi_{\underline{i}}(\mathbf{x})p(\mathbf{x})\right]^2\right]
\end{aligned} \tag{13}$$

We also have

$$\begin{aligned}
&E\left[\sum_{\underline{i}}\mu_{\underline{i}}^{b}\left[\frac{1}{n}\sum_{j=1}^{n}\varphi_{\underline{i}}(\mathbf{x}_j) - \int\varphi_{\underline{i}}(\mathbf{x})p(\mathbf{x})\right]^2\right] \\
&\quad\le \sum_{\underline{i}}\mu_{\underline{i}}^{b}E\left[\frac{1}{n}\sum_{j=1}^{n}\varphi_{\underline{i}}(\mathbf{x}_j) - \int\varphi_{\underline{i}}(\mathbf{x})p(\mathbf{x})\right]^2 \\
&\quad\le \sum_{\underline{i}}\mu_{\underline{i}}^{b}\left[\frac{1}{n}\int\varphi_{\underline{i}}^2 p\right] \\
&\quad\le \frac{C}{n}\sum_{\underline{i}}\mu_{\underline{i}}^{b}\int\varphi_{\underline{i}}^2 \\
&\quad= \frac{C}{n}\sum_{\underline{i}}\mu_{\underline{i}}^{b} = \frac{C}{n}\sum_{\underline{i}}\mu_{i_1}^{b}\cdots\mu_{i_d}^{b} = \frac{C}{n}\left(\sum_{i_1}\mu_{i_1}^{b}\right)^d \\
&\quad= \frac{C}{n}.
\end{aligned} \tag{14}$$

The last step follows from $b > 1/(2m)$, and $\mu_{i_1} \sim i_1^{-2m}$.

Combining (12), (13), (14), we get

$$\|\hat{f} - \tilde{f}\|_a^2 = O_p\big[n^{-1}\|\hat{f} - \bar{f}\|_b^2 N_{a+b}(\lambda)\big],$$

and (10) follows. Hence the lemma is proved. □

It is easily seen by examining the above proof that the conclusion of the lemma holds uniformly for all f_0 satisfying $\|f_0\| < B$.

4.4. *Rate of convergence of the penalized likelihood estimator in regression.* Combining Lemmas 4.1, 4.2 and 4.3, we get the theorem.

THEOREM 4.1. *If* $1 \geq b > 1/2m$ *and* $n^{-1}\lambda^{-(2b+1/2m)}(\log(1/\lambda))^{r-1} \to 0$, *then for* $a \in [0, b]$,

$$\|\hat{f} - f_0\|_a^2 = O(\lambda^{1-a}) + O_p\left[n^{-1}\lambda^{-(a+1/2m)}\left(\log\frac{1}{\lambda}\right)^{r-1}\right]$$

uniformly over any f_0 *satisfying* $\|f_0\| < B$.

Many results can be derived from Theorem 4.1. Setting $b = 1/2m + \varepsilon/2$ and $a = 0$ in Theorem 4.1, it is easy to get one corollary.

COROLLARY 4.1. *If* $n^{-1}\lambda^{-(3/2m+\varepsilon)} \to 0$ *for some* $\varepsilon > 0$. *Then*

$$\int(\hat{f} - f_0)^2 p = O(\lambda) + O_p\left[n^{-1}\lambda^{-1/2m}\left(\log\frac{1}{\lambda}\right)^{r-1}\right].$$

PROOF OF THEOREM 1.2. When $m > 1$, let $\lambda \sim [n(\log n)^{1-r}]^{-2m/(2m+1)}$, then the condition of the above corollary is satisfied. Hence by Corollary 4.1 we have $\|\hat{f} - f_0\|_0^2 = O_p([n(\log n)^{1-r}]^{-2m/(2m+1)})$. Recalling the ANOVA decomposition defined by (3) in Section 2.2, and by Proposition 4.2, we know that

$$\int_0^1(\hat{f}_i - f_{0i})^2\,dx_i = \|\hat{f}_i - f_{0i}\|_{H^0([0,1])}^2 = O_p([n(\log n)^{1-r}]^{-2m/(2m+1)}),$$

$$\int_0^1\int_0^1(\hat{f}_{ij} - f_{0ij})^2\,dx_i\,dx_j = \|\hat{f}_{ij} - f_{0ij}\|_{\otimes^2 H^0([0,1])}^2 = O_p([n(\log n)^{1-r}]^{-2m/(2m+1)})$$

and so on. These mean the integrated squared errors of estimating f_0 and its component functions by $\hat{f}$ and its component functions go to 0 at a rate of

$$O_p([n(\log n)^{1-r}]^{-2m/(2m+1)}),$$

and the rate is uniform for all f_0 satisfying $\|f_0\| < B$. Summing up, we get Theorem 1.2 as stated in Section 1. □

Setting $b = a = k/m$ in Theorem 4.1, we get the corollary.

COROLLARY 4.2. *For positive integer $k < m$, if*

$$n^{-1}\lambda^{-(4k+1)/2m}(\log(1/\lambda))^{r-1} \to 0,$$

then

$$\|\hat{f} - f_0\|_{k/m}^2 = O(\lambda^{1-k/m}) + O_p[n^{-1}\lambda^{-((2k+1)/2m)}(\log(1/\lambda))^{r-1}].$$

When $k < m/2$, let $\lambda \sim [n(\log n)^{1-r}]^{-2m/(2m+1)}$, the condition of the above corollary is satisfied. Hence by Corollary 4.2 we have

$$\|\hat{f} - f_0\|_{k/m}^2 = O_p([n(\log n)^{1-r}]^{-2(m-k)/(2m+1)})$$

By Proposition 4.2, we know that

$$\|\hat{f}_i - f_{0i}\|_{H^k([0,1])}^2 = O_p([n(\log n)^{1-r}]^{-2(m-k)/(2m+1)}),$$
$$\|\hat{f}_{ij} - f_{0ij}\|_{\otimes^2 H^k([0,1])}^2 = O_p([n(\log n)^{1-r}]^{-2(m-k)/(2m+1)})$$

and so on. But we also know that $\int_0^1[(d^k/dx_i^k)(\hat{f}_i(x_i) - f_{0i}(x_i))]^2\,dx_i \le \|\hat{f}_i - f_{0i}\|_{H^k([0,1])}^2$, so we have the theorem.

THEOREM 4.2. *For positive integer $k < m/2$, let $\lambda \sim [n(\log n)^{1-r}]^{-2m/(2m+1)}$, then*

$$\int_0^1\left[\frac{d^k}{dx_i^k}(\hat{f}_i(x_i) - f_{0i}(x_i))\right]^2 dx_i = O_p([n(\log n)^{1-r}]^{-2(m-k)/(2m+1)}).$$

Obviously we can obtain similar results for derivatives of the interaction terms.

APPENDIX

PROOF OF LEMMA 2.3. Since $(1+\rho_i)^{-1} \sim \nu_i$, $\{\nu_i\}$ is a particular subset of $\{\mu_{i_1 i_2 \cdots i_d}\}$, with at most r of $i_1, i_2, \ldots, i_d$ not equal to 1, and $\mu_i \sim i^{-2m}$, we have

$$\begin{aligned} N_b(\lambda) &\sim \sum_{i_1=1}^{\infty}\cdots\sum_{i_r=1}^{\infty}(i_1^{2m}\cdots i_r^{2m})^b(1+\lambda i_1^{2m}\cdots i_r^{2m})^{-2} \\ &\sim \int_1^{\infty}\cdots\int_1^{\infty}(x_1^{2m}\cdots x_r^{2m})^b(1+\lambda x_1^{2m}\cdots x_r^{2m})^{-2}\,dx_1\cdots dx_r \\ &\sim \int_1^{\infty}\cdots\int_1^{\infty}(1+\lambda y_1^{\beta}\cdots y_r^{\beta})^{-2}\,dy_1\cdots dy_r, \end{aligned}$$

where $\beta = 2m/(2mb+1)$. Set $z_i = \Pi_{j\le i} y_j$, $i = 1, 2, \ldots, r$, we get

$$\begin{aligned}
&\int_1^\infty \cdots \int_1^\infty (1+\lambda y_1^\beta \cdots y_r^\beta)^{-2}\, dy_1 \cdots dy_r \\
&\quad = \int_1^\infty \left[\int_1^{z_r} \left[\int_1^{z_{r-1}} \cdots \int_1^{z_2} (1+\lambda z_r^\beta)^{-2} z_1^{-1} \cdots z_{r-1}^{-1}\, dz_1 \cdots dz_{r-2} \right] dz_{r-1} \right] dz_r \\
&\quad = \int_1^\infty (1+\lambda z_r^\beta)^{-2} \left[\int_1^{z_r} \cdots \int_1^{z_2} z_1^{-1} \cdots z_{r-1}^{-1}\, dz_1 \cdots dz_{r-1} \right] dz_r \\
&\quad = \int_1^\infty (1+\lambda z_r^\beta)^{-2} [(r-1)!]^{-1} (\log z_r)^{r-1}\, dz_r \\
&\quad \sim \lambda^{-1/\beta} \int_{\lambda^{1/\beta}}^\infty (1+x^\beta)^{-2} \left[\log x + \frac{1}{\beta} \log \frac{1}{\lambda} \right]^{r-1} dx \\
&\quad = \lambda^{-1/\beta} \left[O\left(\log \frac{1}{\lambda} \right)^{r-1} + O\left(\log \frac{1}{\lambda} \right)^{r-2} + \cdots + O\left(\log \frac{1}{\lambda} \right) + O(1) \right] \\
&\quad = \lambda^{-1/\beta} O\left(\log \frac{1}{\lambda} \right)^{r-1} \\
&\quad = O\left[\lambda^{-(b+1/2m)} \left(\log \frac{1}{\lambda} \right)^{r-1} \right]. \qquad \square
\end{aligned}$$

PROOF OF PROPOSITION 4.1. Since $\|\cdot\| \sim \|\cdot\|_F$ and $\|\cdot\|_0 \sim \|\cdot\|_{F_0}$, we only need to show the injection from F with $\|\cdot\|_F$ to F_0 with $\|\cdot\|_{F_0}$ is continuous and dense. That the injection is continuous is obvious, because F has a stronger norm. With the aid of the decomposition of F and F_0, we see that the injection is dense follows easily from the following two statements:

1. $\otimes^r H_0^m$ is dense in $\otimes^r H_0^0$.
2. If A_1 is a dense subset in Hilbert space B_1, A_2 is a dense subset in Hilbert space B_2, and B_1 and B_2 are orthogonal to each other. Then $A_1 + A_2$ is dense in $B_1 + B_2$.

With the fact that $H_0^m([0,1])$ is dense in $H_0^0([0,1])$, the proofs of the two statements are straightforward. $\square$

PROOF OF PROPOSITION 4.2. We can introduce the intermediate space Θ_{F_b} with norm $\|\cdot\|_{F_b}$ between F_0 with $\|\cdot\|_{F_0}$ and F with $\|\cdot\|_F$. Since $\|\cdot\| \sim \|\cdot\|_F$ and $\|\cdot\|_0 \sim \|\cdot\|_{F_0}$, we have $\Theta_b = \Theta_{F_b}$ and $\|\cdot\|_b \sim \|\cdot\|_{F_b}$. Now the proposition follows from Proposition 2.3.1 in Lin (1998) and the discussion preceding it. $\square$

Notation and conventions. By a constant C, we denote a generic finite positive constant. It does not depend on the sample size n in any way, and it does not depend on the unknown function being estimated. The same is true for constants C_1 and C_2. Even successive appearances of such constants may not denote the same number.

A d-dimensional vector $(x_1, x_2, \ldots, x_d)$ is also written as $\mathbf{x}$, and $d\mathbf{x} = dx_1 dx_2 \cdots dx_d$ when we write expressions like $\int f(\mathbf{x})\, d\mathbf{x}$.

$a_i \sim b_i$ means the ratio of a_i and b_i is bounded away from zero and infinity, that is, the ratio is between two positive constants not depending on i.

$\|\cdot\|_1 \sim \|\cdot\|_2$ means that the two norms on the same Hilbert space are equivalent, that is, the ratio of $\|f\|_1$ and $\|f\|_2$ is between two positive constants not dependent on f.

For two Hilbert spaces A_1 and A_2, $A_1 = A_2$ means that the two spaces are the same as sets and have equivalent norms.

A sequence $\{f_n\}$ of estimators for f_0 is said to achieve a (uniform) rate of convergence $\{r_n\}$ for some positive sequence $\{r_n\}$ if

$$\lim_{a\to\infty} \limsup_n \sup_{\|f_0\|\le B} P_{f_0}\left(\int (f_n - f_0)^2 > a r_n \right) = 0.$$

See Stone (1980, 1982).

Acknowledgments. This paper is part of the author's Ph.D. dissertation. The author expresses his sincere appreciation for the guidance and the encouragement of his thesis advisor, Professor Larry Brown. The author also thanks Professor Charles Epstein for his help.

REFERENCES

AUBIN, J. P. (1979). *Applied Functional Analysis*. Wiley, New York.

BROWN, L. D. and LOW, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398.

CHEN, Z. (1991). Interaction spline models and their convergence rates. *Ann. Statist.* **19** 1855–1868.

COX, D. D. (1988). Approximation of method of regularization estimators. *Ann. Statist.* **16** 694–712.

COX, D. D. and O'SULLIVAN, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *Ann. Statist.* **18** 1676–1695.

DONOHO, D. L., LIU, R. C. and MACGIBBON, B. (1990). Minimax risk over hyperrectangles, and implications. *Ann. Statist.* **18** 1416–1437.

GU, C. and QIU, C. (1993). Smoothing spline density estimation: theory. *Ann. Statist.* **21** 217–234.

GU, C. and QIU, C. (1994). Penalized likelihood regression: a simple asymptotic analysis. *Statist. Sinica* **4** 297–304.

GU, C. (1996). Penalized likelihood hazard estimation: a general procedure. *Statist. Sinica* **6** 861–876.

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

HASTIE, T. J. and TIBSHIRANI, R. J. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55** 757–796.

HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York.

KADISON, R. V. and RINGROSE, J. R. (1997). *Fundamentals of the Theory of Operator Algebras*. Amer. Math. Soc., Providence, RI.

LIN, Y. (1998). Tensor product space ANOVA models in multivariate function estimation. Ph.D. dissertation, Dept. Statistics, Univ. Pennsylvania.

NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24** 2399–2430.

ODEN, J. T. and REDDY, J. N. (1976). *An Introduction to the Mathematical Theory of Finite Elements*. Wiley, New York.

O'SULLIVAN, F. (1993). Nonparametric estimation in the Cox model. *Ann. Statist.* **21** 124–145.

SILVERMAN, B. W. (1982). On the estimation of a probability density function by the maximum penalised likelihood method. *Ann. Statist.* **10** 795–810.

STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705.

WAHBA, G. (1990). *Spline models for Observational Data*. SIAM, Philadelphia.

WAHBA, G., WANG, Y., GU, C., KLEIN, R. and KLEIN, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy. *Ann. Statist.* **23** 1865–1895.

WEINBERGER, H. F. (1974). *Variational Methods for Eigenvalue Approximation*. SIAM, Philadelphia.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN, MADISON
1210 WEST DAYTON STREET
MADISON, WISCONSIN 53706-1685
E-MAIL: ylin@stat.wisc.edu