October 2019

Joint Asymptotics for Smoothing Spline Semiparametric Nonlinear Models

Jiahui Yu
University of Massachusetts Amherst

JOINT ASYMPTOTICS FOR SMOOTHING SPLINE SEMIPARAMETRIC NONLINEAR MODELS

A Dissertation Presented

by

JIAHUI YU

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2019

Mathematics and Statistics

JOINT ASYMPTOTICS FOR SMOOTHING SPLINE SEMIPARAMETRIC NONLINEAR MODELS

A Dissertation Presented

by

JIAHUI YU

Approved as to style and content by:

Anna Liu, Chair

John Staudenmayer, Member

Markos Katsoulakis, Member

Jing Qian, Member

Nathaniel Whitaker, Department Chair
Mathematics and Statistics

DEDICATION

To my husband, Dom.

ACKNOWLEDGMENTS

Throughout my PhD study and the writing of this dissertation, I have received a great deal of support and guidance from many people.

I would first like to express my most sincere appreciation and gratitude to my advisor, Professor Anna Liu, who not only provided me valuable guidance for my thesis work, but also served as a role model as a female researcher in the field of mathematics and statistics. This dissertation would not have been possible without her continuous support and expert advice. Her encouragement and belief in me at times when I doubted myself enabled me to overcome many obstacles and grow as a researcher. I am also forever grateful to be introduced to the research area of nonparametric and semiparametric statistics.

I would also like to thank Professor Yuedong Wang and the rest of my thesis committee: Professor John Staudenmayer, Professor Markos Katsoulakis, and Professor Jing Qian for reading this dissertation and providing many insightful comments and discussions. I also thank them for their excellent suggestions on future topics as well as research areas related to my interests.

I want to dedicate special thanks to Professor Richard Ellis, although no longer with us, for both his inspirational support during the most difficult time of my graduate study and his great instruction in mathematical analysis and probability theory, which greatly shaped the mathematical foundation for this thesis study and my future research. My sincere thanks extend to the Graduate Program Director, Professor Tom Weston, and the Department

Chair, Professor Nathaniel Whitaker for their caring advice and assistance through this entire process.

Last but not the least, I want to thank my family and friends. Words cannot express how grateful I am to my parents and my husband for their unconditional love and their endless support and encouragement for me to pursue my dreams. I thank my friends Kostantinos Gourgoulias, Yue Tang, Haitian Yue, Zijing Zhang, and many others for their great company and many insightful discussions. Their friendship has made me truly cherish my time at graduate school.

ABSTRACT

JOINT ASYMPTOTICS FOR SMOOTHING SPLINE SEMIPARAMETRIC NONLINEAR MODELS

SEPTEMBER 2019

JIAHUI YU

B.S., UNIVERSITY OF CALIFORNIA LOS ANGELES

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Anna Liu

We study the joint asymptotics of general smoothing spline semiparametric models in the settings of density estimation and regression. We provide a systematic framework which incorporates many existing models as special cases, and further allows for nonlinear relationships between the finite-dimensional Euclidean parameter and the infinite-dimensional functional parameter.

For both density estimation and regression, we establish the local existence and uniqueness of the penalized likelihood estimators for our proposed models. In the density estimation setting, we prove joint consistency and obtain the rates of convergence of the joint estimator in an appropriate norm. The convergence rate of the parametric component in the standard Euclidean norm and the convergence for the overall density function in the symmetric

Kullback-Leibler (SKL) metric are also established. Finally, for our regression model, we obtain the joint consistency and rates of convergence in parallel to those for the density estimation model. In addition, we investigate a doubly penalized likelihood estimator in terms of joint consistency, parameter estimation consistency, and model selection consistency.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

I.1 Semiparametric statistical models

A key theme throughout almost every aspect of statistical analysis is the trade-off between bias and variance, or similarly, between underfitting and overfitting data. Approaches to statistical modeling that exemplify this tension are the parametric and nonparametric models. Parametric models reduce noise by imposing strong assumptions on the model, which greatly simplifies the estimation but can introduce bias. A standard example is the linear regression model. Nonparametric models, on the other hand, have the flexibility to adapt to the data, thereby reducing bias at the expense of more challenging estimation and the potential to overfit data. A well-known example in the context of density estimation is the rescaled histogram.

Semiparametric models exist to combine the advantages of parametric and nonparametric models. Domain knowledge can be incorporated into the model through assumptions reflected in the parametric component, while the nonparametric component allows the flexibility to capture unexpected features of the data.

Throughout the introduction, we denote a generic semiparametric model by $f(x) = \eta(x; \theta, h)$, where $\theta \in \mathbb{R}^p$ is a finite-dimensional Euclidean parameter and an infinite-dimensional functional parameter $h \in \mathcal{H}$ for some appropriate function space $\mathcal{H}$. We emphasize that the functional form of η is assumed to be known, but we do not specify this form. Rather, we iden-

tify reasonable assumptions on η that are sufficient to make the asymptotic analysis work. Many existing models therefore fit into our framework, including parametric and nonparametric models, but also *partially linear models* $\eta(x;\theta,h) = \theta^T x + h(x)$ and nonlinear models (as long as η is not "too nonlinear" as specified in the assumptions - see Chapter II for details).

In practice, the parametric component θ is often of primary interest and the nonparametric component h is treated as a nuisance parameter. In general, however, the estimation of semiparametric models is of comparable difficulty as the estimation of nonparametric models. It is not surprising therefore that many estimation methods for nonparametric models are adapted for semiparametric models. One such example is the estimation method we use throughout this dissertation, smoothing splines, discussed below.

I.2 Smoothing spline estimation

Smoothing spline estimation is a procedure for estimating nonparametric or semiparametric models by framing the estimation as an optimization problem: we wish to minimize some measure of the lack-of-fit of our model to the given data plus some measure of the variability of our model. This gives another perspective on the bias-variance trade-off.

The canonical example of smoothing spline estimation is the cubic smoothing spline, a nonparametric regression model. Suppose we wish to estimate a function f given stochastic data (x_i, y_i), $i = 1, \ldots, n$, where y_i is a noisy observation of $f(x_i)$. Let $Y_i = f(x_i) + \epsilon_i$, where $x_i \in [0,1]$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \sim N(0, \sigma^2 I)$, and let $\mathcal{H}$ denote the Sobolev space $H^2([0,1])$. Consider minimizing the *penalized least squares function*

$$\ell_{n,\lambda}(h \mid \text{data}) = \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - h\left(x_i\right)\right)^2 + \lambda \int_0^1 \left(h''(x)\right)^2 \, dx,$$

where $h \in \mathcal{H}$ and $\lambda > 0$ is called the *smoothing parameter* or *regularization parameter*. If $\hat{h}_\lambda = \arg\min_{h \in \mathcal{H}} \ell_{n,\lambda}(h \mid \text{data})$, the first term on the right ensures that $\hat{h}_\lambda$ approximates f at the points x_i, while the second term ensures that $\hat{h}_\lambda$ is reasonably smooth. As $\lambda \to 0$, $\hat{h}_\lambda$ approaches the function of minimum curvature which interpolates the given data. As $\lambda \to \infty$, the roughness penalty dominates, and $\hat{h}_\lambda$ approaches the best least squares approximation to f in the space $\{h \in \mathcal{H} : h'' = 0\}$, that is, $\hat{h}_\lambda$ approaches the linear regression function. For a fixed $\lambda > 0$, $\hat{h}_\lambda$ is called the *cubic smoothing spline*, and it can be shown that it coincides with the natural cubic spline with knots at the points x_i.

We generalize the least squares criterion, the roughness penalty, and the function space $\mathcal{H}$ in the cubic smoothing spline example, and apply it to the semiparametric setting. Thus, to estimate a semiparametric model depending on $\theta \in \mathbb{R}^p$ and $h \in \mathcal{H}$, where $\mathcal{H}$ now denotes a reproducing kernel Hilbert space, we minimize a *penalized likelihood function*

$$\ell_{n,\lambda}(\theta, h \mid \text{data}) = \ell_n(\theta, h \mid \text{data}) + \frac{\lambda}{2} J(h),$$

where the *loss functional* $\ell_n(\theta, h \mid \text{data})$ assesses how well the model fits the data and the *penalty function* J is a quadratic functional with a null space of finite dimension (so that as $\lambda \to \infty$, the method reduces to parametric estimation). In particular examples, the loss functional is often taken to be the negative log likelihood (hence the name) or the mean square error, while the penalty functional often represents a roughness penalty on h. The *joint smoothing spline estimator*, also called the *joint penalized likelihood estimator*, is the pair

$$(\hat{\theta}, \hat{h}) = \arg\min_{(\theta,h)} \ell_{n,\lambda}(\theta, h \mid \text{data}).$$

Much of this dissertation concerns the existence, consistency, and rate of convergence of the above joint estimator to the unobservable *true parameters* of the density and regression models under consideration. Smoothing spline estimation applies to density and conditional density estimation, hazard function estimation, regression, and general nonlinear inverse problems (Cox and O'Sullivan, 1990). However, we focus primarily on density estimation and regression.

I.3 Parameter selection

Note that the penalty functional in the semiparametric framework is independent of the Euclidean parameter $\theta = (\theta^1, \ldots, \theta^p)$ in $\mathbb{R}^p$. While the parametric component of a model is often parsimonious, many recent applications consider high-dimensional parameters where p is large, especially in the context of machine learning. If the dimension of the parameter space is large and the underlying true parameter θ_0 is sparse, one may be interested in a parameter reduction process to identify the significant parameters θ^j.

For a semiparametric model with partially linear structure, for example $\eta(x; \theta, h) = h(x^T\theta)$, identifying the significant parameters coincides with the well-known *variable selection* problem. In our general model $\eta(x; \theta, h)$, the number of parameters p need not correspond to the number of variables, that is, the dimension of the domain $\mathcal{X}$. We therefore refer to sparsity-encouraged regularization methods as *parameter selection* in our framework.

There are various classical parameter selection procedures, often based on computationally expensive combinatorial search to identify the "best" subset of parameters to include in the model. However, such procedures are inefficient, especially in high-dimensional settings where efficiency is most imporant, and Breiman (1996) shows that such procedures tend to suffer

from model selection instability. Both issues are alleviated by a regularization procedure, which fits nicely into the smoothing spline estimation framework.

In the context of regression and in addition to penalized likelihood estimation, we also consider the following *doubly penalized* likelihood function

$$L_n(\theta, h \mid \text{data}) = \ell_n(\theta, h \mid \text{data}) + \frac{\lambda}{2} J(h) + \sum_{j=1}^{p} q_{\lambda_j}\left(|\theta^j|\right),$$

where $\theta = (\theta^1, \ldots, \theta^p)$ and q_{λ_j} are specified *sparsity-encouraging penalty functionals* with *regularization parameters* $\lambda_j \geq 0$ for $j = 1, \ldots, p$. The idea is that for certain choices of penalty functionals, minimizing the corresponding doubly penalized likelihood function produces joint parameter estimates with some components of the Euclidean parameter θ exactly equal to 0.

Some simple examples of such penalty functionals are given by $q_{\lambda_j} = \lambda_j |\theta^j|^q$, which are called *bridge* penalties. The particular case $q = 1$ is called the *LASSO* penalty and $q = 2$ is referred to as the *ridge* penalty. However, only values of q satisfying $0 < q \leq 1$ are useful for parameter selection (Fan and Li, 2001). We note that for $0 < q < 1$, the l^q norm is not convex, and thus the resulting optimization problem is more challenging. More sophisticated penalties include the adaptive LASSO, elastic net, and SCAD penalties, discussed in greater detail in the literature review section.

I.4 Density estimation models

Having discussed semiparametric models in general, as well as their estimation techniques, we now turn to one of two classes of models that form the core of this dissertation.

Suppose we wish to estimate an unobservable probability density function f given observed data $x_1, \ldots, x_n$, assumed to be a random sample from the

corresonding probability distribution. Density estimation plays a fundamental role in many areas of statistics and machine learning. It is often used in exploratory data analysis because an estimate of the density function can provide valuable indication of features such as skewness and multimodality. Moreover, the density function can also be used to calculate probabilities of events, for likelihood functions, or for nonparametric or semiparametric discriminant analysis.

A special consideration for estimation is that a density function f must satisfy two constraints, $f \geq 0$ and $\int f \; dx = 1$. Such constraints naturally present mathematical and computational difficulties. However, Leonard (1978) and Silverman (1982) provide two different techniques to formulate density estimation as an unconstrained optimization problem. Their analysis is restricted to the nonparametric setting, but is equally applicable to semiparametric models. Most relevent to our work is the approach of Leonard (1978), who introduces the logistic density transform $f = e^h / \int e^h \; dx$ and proposes to estimate h by minimizing

$$\ell_{n,\lambda}(h) = -\frac{1}{n} \sum_{i=1}^{n} h(x_i) + \log \int e^h \; dx + \frac{\lambda}{2} J(h).$$

This formulation is unconstrained, but does not satisfy the important property of *identifiability* because estimators differing by a constant yield the same density function, $e^h / \int e^h \; dx = e^{h+c} / \int e^{h+c} \; dx$ for any constant c. In fact, since $J(h)$ is often taken to be a roughness penalty involving the derivatives of h, the constant functions are typically in the null space of J, and therefore $\ell_{n,\lambda}(h)$ may not have a unique minimizer. Gu and Qiu (1993) addresses this problem by "dropping the constant functions," achieved by enforcing a side condition such as $\int_B h \; d\nu = 0$ for some measure ν with $\nu(B) > 0$.

Making use of the developments discussed above, we propose a general semiparametric density model

$$f(x) = \frac{\exp\{\eta(x;\theta,h)\}}{\int_{\mathcal{X}} \exp\{\eta(x;\theta,h)\}\,dx},$$

where $\mathcal{X}$ is the domain of the density function f. The detailed specification of the model is presented in Chapter II.

I.4.1 Examples of nonlinear density models

In the following, we present several examples of nonlinear semiparametric density models that have been proposed in the literature. We write each model in notation consistent with our general framework to show how each model can be viewed as a special case of our own.

Olkin and Spiegelman (1987) propose a mixture of a parametric and a nonparametric density function,

$$f(x) = \theta_1 f_1(x;\theta_2) + (1-\theta_1) f_2(x),$$

where $f_1(x;\theta_2)$ is a known density function depending on parameters $\theta_2 \in \mathbb{R}^p$, $\theta_1 \in [0,1]$ is an unknown weight parameter, and $f_2(x)$ is a nonparametric density function. Such a model provides a compromise between the parametric and nonparametric estimates. It is a special case of our model with $\eta(x) = \log\{\theta_1 f_1(x,\theta_2) + (1-\theta_1) f_2(x)\}$, where θ_1 and θ_2 are the parameters and f_2 is the nonparametric function.

Hjort and Glad (1995) propose a density estimation procedure by starting out with a parametric density estimate $f_1(x;\hat{\theta})$, and then multiplying

by a nonparametric kernel type estimate of a correction function $r(x) = f(x)/f_1(x;\hat{\theta})$, producing

$$\hat{f}(x) = f_1(x,\hat{\theta})\hat{r}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)\frac{f_1(x,\hat{\theta})}{f_1(X_i,\hat{\theta})}.$$

Moreover, they show that this model can perform better than a nonparametric density model when the true density is in the neighborhood of the initial parametric density. Their model is a special case of our model with $\eta(x) = \log\{f_1(x,\theta)\} + h(x)$, where θ is the parameter and $h(x) = \log\{r(x)\}$ is the nonparametric function.

Efron and Tibshirani (1996) propose a specially designed exponential family for density estimation,

$$f(x) = f_0(x)\exp\left\{\theta_1 + t^T(x)\theta_2\right\},$$

where $f_0(x)$ is a carrier density and is estimated by kernel density estimation, $t(x)$ is a p-dimensional vector of sufficient statistics, θ_2 is a p-dimensional vector of parameters, and θ_1 is a normalizing parameter ensuring that $f(x)$ integrates to 1 over the sample space $\mathcal{X}$. The proposed method matches the estimated expectation of $t(x)$ with its sample expectation. They also used the exponential family model to investigate density differences in the case of multiple samples, with shared carrier $f_0(x)$ estimated nonparametrically, but with possibly different values of the exponential parameters θ_1 and θ_2. The model is a special case of our model with $\eta(x) = \log\{f_0(x)\} + \theta_1 + t^T(x)\theta_2$, where $f_0(x) = \exp\{h(x)\}/\int \exp\{h(x)\}dx$ is the carrier density given by nonparametric function $h(x)$, and θ_1 and θ_2 are the parameters.

Lenk (2003) proposes a flexible semiparametric model for Bayesian testing,

$$f(x) = \exp\left\{\alpha^T(x)\theta + Z(x)\right\} / \int_{\mathcal{X}} \exp\left\{\alpha^T(x)\theta + Z(x)\right\} dG(x),$$

where $\alpha(x)$ is a vector of m nonconstant functions, Z is a zero mean, second-order Gaussian process with bounded, continuous covariance function, and G is a known dominating measure on the support $\mathcal{X}$. The semiparametric model allows the predictive distribution to deviate from the parametric family. If the parametric family is inadequate, the semiparametric predictive density coherently adapts to the data.

Yang (2009) also uses the partially linear semiparametric logistic density transform as in Lenk (2003), with $Z(x)$ replaced by an unknown smooth function $h(x)$. The model in Yang (2009) is a special case of our model with $\eta(x) = \alpha^T(x)\theta + h(x)$.

Wand et al. (1991) consider density estimation of data with local features, and propose a data transformation technique so that global smoothing is appropriate. This transformation model is a special case of our model with $\eta(x) = h(\alpha(x;\theta)) + \log|\alpha'(x;\theta)|$ where $\alpha(x;\theta)$ is a known transformation.

I.5 Regression models

We now turn to the second class of models that form the core of our work. Regression is one of the fundamental techniques in statistics for estimating the relationship between certain variables of interest. It is often used for exploratory data analysis, forecasting, testing whether a theoretically derived model is consistent with observations, and many other applications.

Given i.i.d. observed data (Y_i, X_i) for $i = 1, \ldots, n$ sampled from variables $(Y, X) \in \mathcal{Y} \times \mathcal{X} \subset \mathbb{R} \times \mathbb{R}^d$, we propose a general semiparametric regression model

$$\mu_0(X) \equiv \mathrm{E}[Y|X] = g[\eta(X; \theta_0, h_0)],$$

where $(\theta_0, h_0) \in \mathbb{R}^p \times \mathcal{H}$ is the true joint parameter to be estimated, η is known, and g is a known *link function*. Together with this assumption, the data can be modeled by $Y \mid X \sim p(Y; \mu_0(X))$ for a conditional distribution p. Alternatively, one can model the data by specifying the relation between the conditional mean and conditional variance (Cheng and Shang, 2015). Note that when $\eta(x; \theta, h) = \theta^T x$, our model is reduced to the well known generalized linear model. For the detailed specification of our regression model, see Chapter II.

I.5.1 Examples of nonlinear regression models

We give several examples of nonlinear semiparametric Gaussian regression models, which are special cases of our general regression framework taking g to be the canonical link function. These examples are taken from Ke and Wang (2008) and include both the original formulation and a proposed reformulation according to Ke and Wang (2008).

Genton and Hall (2007) propose to model the Mira variable R Hydrae declining period and amplitude via the model

$$Y_i = a(x_i)h\left(g(x_i)\right) + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $n = 2315$ observations, y_i is the magnitude (brightness) at time x_i, h is the common periodic shape function with unit period, $a(x)$ is the amplitude function, $g(x)$ is a strictly increase time transformation, and $\epsilon_i \sim N(0, \sigma^2)$ are i.i.d. They suggest modeling h nonparametrically, and modeling $a(x) =$

$1 + \theta_1 x$ and $g(x) = \theta_3^{-1} \log(1 + \theta_2^{-1}\theta_3 x)$ parametrically, so that the model becomes

$$Y_i = (1 + \theta_1 x_i) h\left(\theta_3^{-1} \log(1 + \theta_2^{-1}\theta_3 x_i)\right) + \epsilon_i, \qquad i = 1, \ldots, n.$$

Ke and Wang (2008) give an alternate model of the same data,

$$Y_i = \theta_1 + \exp\{h_1(x_i)\} \sin\{2\pi h_2(x_i)\} + \epsilon_i, \qquad i = 1, \ldots, n,$$

where both the amplitude and period are modeled nonparametrically.

Yu and Ruppert (2002) model the dependence of ozone concentration on wind speed x_1, temperature x_2, and solar radiation x_3. They note that due to the small sample size of $n = 111$ observations, a fully nonparametric model may not be appropriate. Thus, they suggest the single index model

$$Y_i = h\left(\theta_1 x_{1i} + \theta_2 x_{2i} + \sqrt{1 - \theta_1^2 - \theta_2^2}\, x_{3i}\right) + \epsilon_i, \qquad i = 1, \ldots, n,$$

and the partially linear single index model

$$Y_i = h\left(\theta_1 x_{1i} + \sqrt{1 - \theta_1^2}\, x_{2i}\right) + \theta_2 x_{3i} + \epsilon_i, \qquad i = 1, \ldots, n,$$

where y_i is the cube root of the ozone concentration and ϵ_i are i.i.d. Citing a potentially nonlinear effect of radiation, Ke and Wang (2008) propose instead

$$Y_i = h\left(\theta_1 x_{1i} + \sqrt{1 - \theta_1^2}\, x_{2i}\right) + \theta_2 x_{3i} + \epsilon_i, \qquad i = 1, \ldots, n.$$

I.6 Literature review

Below we review the literature on nonparametric and semiparametric penalized likelihood estimation, in the context of both density estimation and

regression. We also review some literature on the regularization approach to the variable selection problem for regression models. However, certain important topics, such as computation, inference, and smoothing / regularization parameter estimation, will be omitted since these topics are not addressed in this dissertation.

I.6.1 Penalized likelihood estimation

The pioneering work of Good and Gaskins (1971) introduces nonparametric penalized log likelihood density estimation, achieved by minimizing a penalized likelihood function

$$\ell_{n,\lambda}(h) = -\sum_{i=1}^{n} \log h(x_i) + \lambda J(h).$$

They take the penalty functional $J(h)$ to be a roughness penalty such as $\int (h'')^2 dx$, and provide a Bayesian justification for this approach.

Silverman (1982) gives a variation of this formulation in which the penalty applies to the logarithm of the density, so that the entire penalized likelihood function is expressed in terms of the logarithm of the density. By estimating $\log f$ instead of f, the positivity requirement for density functions is automatically satisfied. Furthermore, Silverman (1982) proves that if the penalty functional $J(\log h)$ involves only derivatives of $\log h$, then the minimizer of the constrained problem

$$\arg\min_{h \in \mathcal{H}} \left\{ -\sum_{i=1}^{n} \log h(x_i) + \frac{\lambda}{2} J(\log h) \right\} \quad \text{subject to} \quad \int h \, dx = 1$$

is equal to the minimizer of the unconstrained problem

$$\ell_{n,\lambda}(h) = -\sum_{i=1}^{n} \log h(x_i) + \int h \, dx + \frac{\lambda}{2} J(\log h),$$

thereby alleviating the mathematical and computational difficulties of the constraints required for density estimation. Silverman (1982) also gives examples of such penalty functionals with respect to which the limiting estimator as $\lambda \to \infty$ converges to the density of some natural parametric family. For example, take

$$J(h) = \int_{-\infty}^{\infty} \left(\frac{d^3}{dx^3} \log f(x) \right) dx.$$

As $\lambda \to 0$, the resulting estimator $\hat{h}_\lambda$ approaches the sum of spikes at the observations, but as $\lambda \to \infty$, the $\hat{h}_\lambda$ approaches the normal density with the same mean and variance as the data. Silverman (1982) gives conditions for the existence of the general estimator, as well as rates of consistency in various norms and conditions for asymptotic normality.

Cox (1988) introduces the Tikhonov method of regularization estimator as a general estimation technique for nonparametric regression in an abstract linear model of the form $Y_n = T_n f(x_n) + \epsilon_n$. A special case of this approach is the nonparametric smoothing spline estimator. Suppose f lies in the Sobolev space $W_2^m[0,1]$, equipped with inner product $\langle g, h \rangle_{W_2^m} = \langle g_h \rangle_{L^2} + \langle g^{(m)}, h^{(m)} \rangle_{L^2}$. Suppose also $\epsilon_n = (\epsilon_{n1}, \ldots, \epsilon_{nn})$ have mean 0 and covariance $\sigma^2 I$, and that the components of y_n are of the form $y_{ni} = f(x_{ni}) + \epsilon_{ni}$. Then the smoothing spline estimator for f is given by minimizing

$$\ell_{n,\lambda}(h) = \frac{1}{n} \sum_{i=1}^{n} \left(y_{ni} - h\left(x_{ni}\right) \right)^2 + \lambda \int_0^1 \left(h^{(m)}(x) \right)^2 \, dx.$$

Cox (1988) provides an asymptotic analysis for the general case by approximating the discrete problem with a continuous one, and derives rates of convergence in appropriate norms.

Cox and O'Sullivan (1990) study the first-order asymptotic analysis of a general nonparametric penalized likelihood estimator. The estimator minimizes the penalized likelihood

$$\ell_{n,\lambda}(h \mid \text{data}) = \ell_n(h \mid \text{data}) + \frac{\lambda}{2} J(h),$$

where it is assumed that the loss functional $\ell_n(h \mid \text{data})$ approaches a limiting loss functional $\ell(h \mid \text{data})$ as $n \to \infty$. The limiting loss functional $\ell(h \mid \text{data})$ is used to identify the true parameter h_0. Let D_h denote the Fréchet differential operator, and define the score vector $Z_{n\lambda} = D_h \ell_n(h \mid \text{data})$ and the limiting score vector $Z_\lambda = D_h \ell_n(h \mid \text{data}) + \lambda D_h J(h)$. The roots of $Z_{n\lambda}$ and Z_λ are shown to exist and can be approximated using a first-order Taylor series expansion of $Z_{n\lambda}$ and Z_λ about the true parameter h_0, which is the locally unique root of $D_h \ell_n(h \mid \text{data})$. For λ sufficiently small, it is shown that $Z_{n\lambda}$ and Z_λ have locally unique roots in a neighborhood of h_0, and that the roots of $Z_{n\lambda}$ and Z_λ are approximated by the roots of their linearizations. Moreover, these linearizations yield expansions of the estimation errors, and explicit error estimates are derived using spectral analysis.

O'Sullivan (1990) establishes convergence rates for penalized likelihood esimators for an abstract nonlinear regression model of the form

$$y_i = \eta(x; h) + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $\eta(x; h)$ is a nonlinear functional in h for each x, and the errors ϵ_i are uncorrelated mean zero errors with bounded variance. For the corresponding penalized least squares criterion

$$\ell_{n,\lambda}(h \mid \text{data}) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - h\left(x_i\right)\right)^2 + \lambda J(h),$$

the behavior of the estimator $\hat{h}_{n\lambda}$ as a function of n and λ is analyzed using the linearization of the functionals $\eta(x; h)$. Measures of the degree of non-linearity of the functionals $\eta(x_i; h)$ play an essential role in the asymptotic behavior of the estimator.

In her celebrated book, Wahba (1990) introduces reproducing kernel Hilbert space (RKHS) methods and gives a comprehensive analysis of the smoothing spline nonparametric regression model. Some of the many important advances enabled by RKHS methods include:

- The evaluation functionals $f(x_i)$ can be replaced by arbitrary bounded linear functionals $L_i(f)$ so that the regression model becomes

$$y_i = L_i f + \epsilon_i, \qquad i = 1, \ldots, n,$$

  enabling the treatmemt of variational problems where $L_i(f) = f^{(m)}(x_i)$ within the same framework as the standard regression problem.

- The RKHS allows a direct sum decomposition $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0$ is the null space of the penalty functional J, such that J is the orthogonal projection onto $\mathcal{H}_1$. The geometry of the Hilbert space can therefore be used to prove existence and uniqueness of the smoothing spline estimator. In fact, Wahba (1990) obtains a representation of the estimator using the RKHS representer theorem.

- The sample space $\mathcal{X}$, which serves as the domain of the functions in $\mathcal{H}$, is completely arbitrary. The RKHS framework therefore allows the estimation of smoothing splines on the circle, sphere, Euclidean space, and even on discrete domains.

- Wahba (1990) shows that for every positive definite functional K on $\mathcal{H} \times \mathcal{H}$, there exists a zero mean Gaussian process ξ on $\mathcal{X}$ with K

as its covariance. Moreover, the RKHS with reproducing kernel K is isometrically isomorphic to the Hilbert space spanned by $\xi(x_i)$, $x_i \in \mathcal{X}$. Using this result, Wahba (1990) shows that smoothing spline estimates are also Bayes estimates with a certain prior on f.

- Tensor products of reproducing kernels yield a reproducing kernel on the product Hilbert space, paving the way for the study of ANOVA decompositions of multivariate functions.

Gu and Qiu (1993) apply the RKHS approach of Wahba (1990) to the nonparametric density estimation problem using the logistic density transform formulation of Leonard (1978),

$$\ell_{n,\lambda}(h) = -\frac{1}{n}\sum_{i=1}^{n} h(x_i) + \log \int e^h \, dx + \frac{\lambda}{2} J(h).$$

As discussed above, Gu and Qiu (1993) solve the identifiability issue by imposing a side condition $A_x h = 0$ for some averaging operator A_x. This is achieved by eliminating the constant term in the one-way ANOVA decomposition $h = A_x h + (I - A_x)h = h_\emptyset + h_x$. Moreover, they prove existence and uniqueness under mild conditions and establish the rate of convergence of the estimator in the symmetrized Kullback-Leibler (SKL) distance.

Gu (1995) extends the work in Gu and Qiu (1993) to the estimation of conditional probability densities. To estimate $f(y \mid x)$, Gu (1995) employs the logistic conditional density transform $e^{h(x,y)} / \int e^{h(x,y)} \, dy$, enforcing the identifiability of the model by setting the constant term to 0 in a multivariate ANOVA decomposition of h. The penalized likelihood studied is given by

$$\ell_{n,\lambda}(h) = -\frac{1}{n}\sum_{i=1}^{n} \left\{ h(x_i, y_i) + \log \int_{\mathcal{Y}} e^{h(x_i,y)} \, dy \right\} + \frac{\lambda}{2} J(h),$$

where $\mathcal{X}$ and $\mathcal{Y}$ are arbitrary sample spaces, and the procedure is implemented via tensor product splines. This allows the estimation of the whole conditional density $f(y \mid x)$, whereas most prior work considered only certain properties of $f(y \mid x)$ such as the conditional mean or conditional percentiles. Gu (1995) also provides asymptotic theory for the estimator adapted from Gu and Qiu (1993).

Yang (2009) studies penalized likelihood estimation for a partially linear semiparametric density model. Specifically, the density function f is assumed to be of the form

$$f(x) = \frac{\exp\left\{a^T(x)\theta + h(x)\right\}}{\int \exp\left\{a^T(x)\theta + h(x)\right\} dG(x)},$$

where $a(x)$ is a known vector of nonconstant functions on the support $\mathcal{X}$ and $dG(x)$ is a known dominating measure such that f is absolutely continuous with respect to G. The model is analogous to Lenk (2003) (see examples above), but replacing the zero mean second-order Gaussian process $Z(x)$ with bounded continuous covariance function, by the smooth function $h(x)$. The penalized likelihood approach is analogous to Gu and Qiu (1993),

$$\ell_{n,\lambda}(\theta, h) = -\frac{1}{n}\sum_{i=1}^{n}\left(a^T(x_i)\theta + h(x_i)\right) + \log\int e^{a^T(x)\theta + h(x)}\, dx + \frac{\lambda}{2}J(h).$$

This work is one of the first to consider the asymptotic theory for the joint estimator, deriving the consistency and convergence rate in the SKL distance.

Cheng and Shang (2015) make great progress on the joint asymptotic theory for the *generalized partially linear regression model*

$$y_i = g\left(\theta^T x + h(x)\right) + \epsilon_i, \qquad i = 1, \ldots, n,$$

where g is a known link function. They establish joint asymptotic normality and also show that the estimators of the parametric and nonparametric components are asymptotically independent.

I.6.2 Variable selection via regularization

Turning now to the literature on variable selection, Frank and Friedman (1993) introduce the so-called bridge regression approach for the parametric linear regression model

$$L_n(\theta \mid \text{data}) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \theta^T (x_i) \right)^2 + \lambda_1 \|\theta\|_{l^q}^q .$$

They explore different values for q, noting that $q = 2$ corresponds to the well-known ridge regression model, which is commonly used when the covariates exhibit a high degree of multicollinearity. For variable selection, they observe that values of q satisfying $0 < q \leq 1$ yield sparsity.

Tibshirani (1996) analyzes the particular case of $q = 1$, which he calles the *least absolute shrinkage and selection operator* (LASSO). He formulates the estimation as a least squared minimization subject to inequality constraints $\|\theta\|_{l^1} \leq t$. However, the formulation is equivalent to the penalized version of Frank and Friedman (1993). Some useful observations from Tibshirani (1996) include the relative merits of subset selection, lasso regression, and ridge regression. He points out that subset selection is most effective when there are a small number of variables with large effects, the lasso is most effective when there are a small to moderate number of variables with moderate-size effects, and ridge regression is preferred when there is a large number of variables with small effects.

We note that the LASSO estimator exhibits a tension between parameter estimation consistency and model selection consistency, that is, whether the

estimator correctly identifies the significant variables relative to the true parameter. For example, Leng et al. (2006) show that the optimal LASSO estimate does not satisfy model selection consistency. Moreover, Zhao and Yu (2006) show that if irrelevant variables are highly correlated to significant variables, then LASSO estimation may not be able to distinguish between the two with any amount of data and regularization.

Fan and Li (2001) specify three criteria that a good penalized likelihood estimator should satisfy: unbiasedness, sparsity, and continuity. They analyze the bridge regression penalty and find that the three desired properties cannot be simultaneously satisfied for any value of q. They then introduce the *smoothly clipped absolute deviation* (SCAD) penalty function

$$q'_\lambda(\theta) = \lambda \left\{ I(\theta \leq \lambda) + \frac{(a\lambda - \theta)^+}{(a-1)\lambda} I(\theta > \lambda) \right\},$$

where $a > 2$ and $\theta > 0$, and prove that it satisfies the oracle properties. However, the SCAD penalty is not convex.

Zou (2006) proves that the LASSO does not satisfy the oracle properties, but also introduces a variation called the adaptive LASSO,

$$q_\lambda(\theta) = \lambda \sum_{i=1}^{p} w_i \left|\theta^i\right|,$$

where $w = (w^1, \ldots, w^p)$ is a known weights vector. Zou (2006) shows that if the weights are data-dependent and cleverly chosen, then adaptive LASSO can satisfy the oracle properties. Moreover, the adaptive LASSO preserves the simplicity and convexity of the LASSO penalty.

Zou and Hastie (2005) improve on the LASSO in a different direction, specifically when the number of covariates p is much larger than the number

of observations n. The *naïve elastic net* estimation combines the LASSO and ridge regression,

$$L_n(\theta \mid \text{data}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \theta^T(x_i)\right)^2 + \lambda_1 \|\theta\|_{l^1} + \lambda_2 \|\theta\|_{l^2}^2 .$$

However, the resulting estimator has unnecessary extra bias compared to the LASSO or ridge estimators. Zou and Hastie (2005) show that naïve eslatic net estimation is equivalent to the LASSO with a particular choice of regularization parameter, and thus are able to define a "corrected" estimator, called the elastic net estimator.

I.7 Contributions

Many semiparametric models have been proposed and studied, in both the regression and the density estimation settings. However, systematic frameworks for general nonlinear semiparametric density models and regression models are lacking, especially in the context of joint asymptotic theory. Most existing asymptotic results for semiparametric models focus on the Euclidean parameter only, while the functional parameter is treated as a nuisance parameter. Existing work that has been done on joint asymptotics has been restricted to models with partial linear structure. This thesis proposes smoothing spline based frameworks for general nonlinear semiparametric density models and regression models, which include many existing nonparametric and semiparametric models as special cases.

An important contribution of this dissertation is the specification of reasonable assumptions under which our theoretic results are possible. These assumptions can provide insight into why certain results may or may not hold for specific models. In the Chapter II, we emphasize that many of our

assumptions are either mild bounds on the "degree of nonlinearity" of the model, or natural extensions of conditions imposed in the literature on nonlinear nonparametric models, or are otherwise not excessively restrictive. In addition, we introduce a new inner product on the joint parameter space, whose induced norm is suitable for our study of joint asymptotic properties.

Chapter III of this thesis focuses on the local existence and uniqueness of a penalized likelihood estimator for both density estimation and regression models simultaneously. The parameters are jointly estimated by minimizing the penalized likelihood function. We note that prior global existence and uniqueness results for similar estimation methods for nonparametric models can be extended to the semiparametric setting, but require very strong assumptions. In our setting, a theory for local existence and uniqueness is established under much less restrictive conditions.

In Chapter IV, we prove joint consistency for our density model and obtain the rates of convergence of the joint estimator in an appropriate norm. We use these results to get the convergence rate of the parametric component in the standard Euclidean norm and for the overall density function in the symmetric Kullback-Leibler (SKL) distance. We also extend our results on density estimation to the case of multiple samples.

Finally, in Chapter V, we obtain the joint consistency and rates of convergence of our estimators for the regression model. In addition, we prove results on the joint consistency, parameter estimation consistency, and model selection consistency for a doubly penalized likelihood estimator.

CHAPTER II

MODELS AND ASSUMPTIONS

In the first section of this preliminary chapter, we define the notation and terminology that will be used throughout this thesis. We will provide the model assumptions and definition of an inner product and its induced norm needed for our asymptotic analysis in either the density estimation setting or the regression setting in the second section. The properties of the inner product are presented in the last section.

II.1 Notation

We begin by introducing some notation and terminology that will be used throughout the rest of the paper unless specified otherwise.

- Suppose $\mathcal{H}$ is a Hilbert space of functions on domain U. If for all $u \in U$, the evaluation functional $\hat{u}(f) = f(u)$ is continuous in $\mathcal{H}$, then we say $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS).

- Denote the product parameter space as $\mathcal{Q} \equiv \mathbb{R}^p \times \mathcal{H}$, where $\mathcal{H}$ is a RKHS. For simplicity, we sometimes use $\tau \equiv (\theta, h) \in \mathcal{Q}$ to represent an element in $\mathcal{Q}$.

- If X and Y are any (real) Banach spaces, $\mathscr{L}(X, Y)$ represents the space of bounded linear operators from X to Y.

- For $\mathcal{X} \subset \mathbb{R}^p$, $L^q(\mathcal{X})$ denotes the space of functions on $\mathcal{X}$ with finite q-th moment with respect to the probability measure given by the true sample distribution. In addition, define $L_0^q(\mathcal{X}) \equiv L^q(\mathcal{X}) \ominus \{1\}$.

- We use the notation $\mathrm{E}(\cdot)$ to represent the expectation taken over the joint sample distribution, and $\mathrm{E}_X(\cdot)$ represents the expectation taken with respect to the covariate X.

- If there exist positive constants c_1, c_2, such that $c_1 A \leq B \leq c_2 A$, we write $A \sim B$.

- Let $a = \left[a^k\right]_{k=1}^p$ denote any $p \times 1$ vector, whose the kth element is a^k.

- The standard Euclidean norm is denoted $\|a\|_{l^2} = [\sum_{k=1}^p (a^k)^2]^{1/2}$ for $a \in \mathbb{R}^p$.

- Denote any $p \times q$ matrix $M = [M^{i,j}]_{i,j=1}^{p,q}$, where $M^{i,j}$ represent the (i,j)th entry of the matrix. When $p = q$, $M \equiv [M^{i,j}]_{i,j=1}^{p} = [M^{i,j}]_{i,j=1}^{p,p}$.

- Let D_h, D_θ be the Fréchet partial differential operators with respect to h and θ, respectively. Note that for any function $f : \mathcal{Q} \to Y$, where Y is a (real) Banach space, the Fréchet partial derivatives of f are maps $D_\theta f : \mathcal{Q} \to \mathscr{L}(\mathbb{R}^p, Y)$ and $D_h f : \mathcal{Q} \to \mathscr{L}(\mathcal{H}, Y)$ by definition.

- Denote the kth order partial Fréchet derivative operators as $D^k_{a_1 \ldots a_k} = D_{a_1} \ldots D_{a_k}$, where $a_i \in \{\theta, h\}$ for $i = 1, \ldots, k$.

- We say a quadratic functional U is completely continuous with respect to another quadratic functional W if for any $\epsilon > 0$, there exist a finite number of linear functionals $L_1, \ldots, L_k$ such that if $L_j f = 0$ for all $j = 1, \ldots, k$, then $U(f) \leq \epsilon W(f)$. See Weinberger (1974) Section 3.3 for details.

II.2 Models and assumptions

Let $(\theta, h) \in \mathcal{Q}$ be the Euclidean parameter and nonparametric function one wishes to estimate using stochastic data. Consider the semiparametric estimator given by the minimizer of the penalized likelihood

$$\ell_{n,\lambda}(\theta, h \mid \text{data}) = \ell_n(\eta(\theta, h) \mid \text{data}) + \frac{\lambda}{2} J(h) \tag{II.1}$$

with respect to $(\theta, h) \in \mathcal{Q}$. Here, $\eta(\theta, h)$ is a known general function, $\ell_n(\eta(\theta, h))$ is the negative log likelihood of the data, $J(h)$ is the roughness penalty term, which is assumed to be a quadratic functional with a null space $\mathcal{H}_0$ of finite dimension, and λ is the smoothing parameter. We use penalized likelihood estimation for our density estimation and regression models, and denote the true joint parameter by $\tau_0 = (\theta_0, h_0)$.

For density estimation, suppose $X_1, \ldots, X_n$ is an i.i.d. random sample from a common probability density $f(x)$ on a bounded domain $\mathcal{X} \subset \mathbb{R}^d$. We consider the following general semiparametric density model,

$$f(x; \theta, h) = \frac{\exp\{\eta(x; \theta, h)\}}{\int_{\mathcal{X}} \exp\{\eta(x; \theta, h)\} dx},$$

where $\eta : \mathcal{Q} \to L_0^2(\mathcal{X})$ is a known function that is one-to-one in a neighborhood $\mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$ of the true parameter. We specify the precise assumptions on this neighborhood in Section II.2.2. The penalized likelihood (II.1) in this setting has the form

$$\begin{aligned} \ell_{n,\lambda}(\theta, h \mid \text{data}) &= \ell_n(\eta(\theta, h) \mid \text{data}) + \frac{\lambda}{2} J(h) \\ &= -\frac{1}{n} \sum_{i=1}^{n} \eta(X_i; \theta, h) + \log \int_{\mathcal{X}} e^{\eta(x;\theta,h)} dx + \frac{\lambda}{2} J(h). \end{aligned} \tag{II.2}$$

For regression models, given i.i.d. observed data (Y_i, X_i) for $i = 1, \ldots, n$ of the variables $(Y, X) \in \mathcal{Y} \times \mathcal{X} \subset \mathbb{R} \times \mathbb{R}^d$, we consider a general class of semiparametric regression models for which

$$\mu_0(X) \equiv \mathrm{E}[Y|X] = g[\eta(X; \theta_0, h_0)],$$

where $g(\cdot)$ is a known link function, $\eta : \mathcal{Q} \to L^4(\mathcal{X})$ is a known function that represent the relationship between the parametric and nonparametric components of the model. Similarly to Cheng and Shang (2015), we assume that the penalized likelihood (II.1) is given by a general criterion function $l(y; a) : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$,

$$\begin{aligned} \ell_{n,\lambda}(\theta, h \mid \text{data}) &= \ell_n(\eta(\theta, h) \mid \text{data}) + \frac{\lambda}{2} J(h) \\ &= \frac{1}{n} \sum_{i=1}^{n} l(Y_i; \eta(X_i; \theta, h)) + \frac{\lambda}{2} J(h), \end{aligned} \tag{II.3}$$

where $l(y; a)$ can be taken to represent two classes of models. The first class assumes the observed data $Y_i \mid X_i$ follows a conditional distribution $p(y; \mu_0(x))$, and $l(y; \eta(X; \theta, h)) = -\log p(y; g(\eta(X; \theta, h))$ is the negative log likelihood of the conditional density function. The conditional distribution may come from an exponential family distributions, which covers the cases of Gaussian regression, logistic regression, and Poisson regression. When $\eta = X^T\theta$, this is better known as the generalized linear models (McCullagh and Nelder, 1989). The second class assumes a form of conditional variance $\mathrm{Var}(Y \mid X) = V(\mu_0(X))$ in addition to the conditional mean $\mu_0(X)$, where $V(\cdot)$ is some known positive-valued function. The criterion function is then chosen to be the quasi-likelihood $l(y; a) = -Q(y; g(a)) \equiv -\int_y^{g(a)} (y - s)/V(s) ds$ with $a = \eta(x; \theta, h)$ (Wedderburn, 1974). We assume the following basic model conditions on $l(y; a)$.

Let R_0 be the range for the true function $\eta(x;\theta_0,h_0)$, which we assume to be a bounded interval in $\mathbb{R}$. We denote the first- and second-order derivatives of $l(y;a)$ with respect to a by $l_a^{'}(y;a)$ and $l_a^{''}(y;a)$, respectively. We have the following model assumptions.

Assumption 1.

(i). $l(y;a)$ *is three times continuously differentiable and convex with respect to* a. *There exists a bounded open interval* R *that contains* R_0 *and a positive constant* C_0 *such that*

$$\mathrm{E}\left\{\sup_{a\in R}|l_a^{''}(Y;a)|^2 \mid X\right\} \leq C_0 \qquad a.s..$$

(ii). *Let* $I_\tau(X) \equiv \mathrm{E}[l_a^{''}(Y;\eta(X;\theta,h)) \mid X]$. *There exist positive constants* M_0 *such that* $M_0^{-1} \leq I_{\tau_0}(X) \leq M_0$ *a.s.*

(iii). *There exist constant* σ *such that*

$$\begin{aligned} &\mathrm{E}\left[l_a^{'}(Y;\eta(X;\theta_0,h_0)) \mid X\right] = 0, \\ &\mathrm{E}\left[l_a^{'}(Y;\eta(X;\theta_0,h_0))^2 \mid X\right] = \sigma^2 I_{\tau_0}(X) \qquad a.s.. \end{aligned} \tag{II.4}$$

Remark 1. *We note that assumptions such as these are standard in the literature. See Shang and Cheng (2013) for a detailed discussion. The first assumption above is weaker than the standard assumption typically used in semiparametric quasi-likelihood models, see Mammen and van de Geer (1997).*

For the rest of this section, we present a uniform set of assumptions for studying both the density estimation model and the regression model given a model-specific definition of a bilinear form $V_\tau(f,g)$ on $L^2(\mathcal{X})$ for any $\tau=(\theta,h)\in\mathcal{Q}$.

- Density estimation model:

$$V_\tau(f,g) = \mu_\tau(fg) - \mu_\tau(f)\mu_\tau(g),$$

where

$$\mu_\tau(f) = \frac{\int_{\mathcal{X}} f(x) e^{\eta(x;\tau)} dx}{\int_{\mathcal{X}} e^{\eta(x;\tau)} dx}.$$

- Regression model:

$$\begin{aligned} V_\tau(f,g) &= \mathrm{E}\left[l_a''(Y;\eta(X;\theta,h) f(X) g(X) \right] \\ &= \mathrm{E}_X\left[\mathrm{E}[l_a''(Y;\eta(X;\theta,h) \mid X] f(X) g(X) \right]. \end{aligned}$$

II.2.1 Parameter space

We first discuss the assumptions for the parameter space $\mathcal{Q}$, which allows us to define the appropriate metric in which the joint asymptotic properties of our proposed estimators can be studied. For simplicity, we denote $V_\tau(f) = V_\tau(f,f)$.

Assumption 2. *The penalty $J(h)$ is a square seminorm in $\mathcal{H}$ with a finite-dimensional null space $\mathcal{H}_0 \subset \mathcal{H}$. Therefore, $J((\theta,h)) \equiv J(h)$ extends J to a square seminorm on $\mathcal{Q}$, and its null space $\mathbb{R}^p \times \mathcal{H}_0$ is again finite-dimensional. Denote by $J(g,h)$ the semi-inner product associated with the seminorm $J(h)$. We also assume that $J(h_0) < \infty$.*

Assumption 3. *There are bounded linear operators $L_\theta : \mathbb{R}^p \to L_0^2(\mathcal{X})$ and $L_h : \mathcal{H} \to L_0^2(\mathcal{X})$, with zero nullspaces, which satisfy the following conditions:*

(i). Suppose $\mathcal{H}$ is a real Hilbert space of functions, equipped with norm $\|\cdot\|$. For any $g \in \mathcal{H}$, there exist positive constants M_1, M_2, such that

$$M_1 \|g\|^2 \le V_{\tau_0}(L_h g) + \lambda J(g) \le M_2 \|g\|^2.$$

(ii). For any $\zeta \in \mathbb{R}^p$ satisfying $\|\zeta\|_{l^2} = 1$ and for any $g \in \mathcal{H}$, there exists a positive constant c_δ such that

$$V_{\tau_0}(L_\theta\zeta - L_h g) = V_{\tau_0}(L_\theta\zeta - L_h g, L_\theta\zeta - L_h g) > c_\delta. \tag{II.5}$$

By Assumptions 2 and 3(i), we see that

$$\langle g_1, g_2\rangle_{\mathcal{H}} \equiv V_{\tau_0}(L_h g_1, L_h g_2) + \lambda J(g_1, g_2)$$

is an inner product on $\mathcal{H}$, and its induced norm, denoted by $\|\cdot\|_{\mathcal{H}}$, is complete on $\mathcal{H}$. One can also see that L_θ can be represented by the $p \times 1$ vector of $L_0^2(\mathcal{X})$ functions $[L_\theta^k(x)]_{k=1}^p$. We may use L_θ to denote the linear operator or its vector form, i.e., $L_\theta\zeta = \zeta^T L_\theta$. Denote by $V_{\tau_0}(L_\theta, L_\theta)$ the $p \times p$ matrix in which the (i,j)th entry is $V_{\tau_0}[L_\theta^i(x), L_\theta^j(x)]$, and note that one can write $V_{\tau_0}(L_\theta\zeta, L_\theta\zeta) = \zeta^T V_{\tau_0}(L_\theta, L_\theta)\zeta$. When $g = 0$, (II.5) implies that $V_{\tau_0}(L_\theta, L_\theta)$ is positive definite. Therefore,

$$\langle \zeta_1, \zeta_2\rangle_{\mathbb{R}^p} \equiv V_{\tau_0}(L_\theta\zeta_1, L_\theta\zeta_2)$$

is an inner product on $\mathbb{R}^p$ and we use $\|\cdot\|_{\mathbb{R}^p}$ to denote its induced norm.

For any $(\zeta_1, g_1), (\zeta_2, g_2) \in \mathcal{Q}$, we define an inner product by

$$\langle(\zeta_1, g_1), (\zeta_2, g_2)\rangle_{\mathcal{Q}} \equiv V_{\tau_0}(L_\theta\zeta_1 + L_h g_1, L_\theta\zeta_2 + L_h g_2) + \lambda J(g_1, g_2), \tag{II.6}$$

and we denote the norm induced by this inner product by $\|\cdot\|_{\mathcal{Q}}$. In Section II.3, we prove that $\langle\cdot,\cdot\rangle_{\mathcal{Q}}$ as defined above is, in fact, an inner product, and that its induced norm is complete.

Remark 2.

(i). We need to use the auxiliary operators L_θ and L_h to define the norm in which the rate of convergence can be conveniently measured. We point out here that L_θ and L_h are related to the Fréchet partial derivatives of η_l by Assumption 5 in Section II.2.2. This is analogous to the approach in O'Sullivan (1990).

(ii). It is well known that all norms on $\mathbb{R}^p$, including $\|\cdot\|_{\mathbb{R}^p}$ as defined above, are equivalent to the Euclidean norm, namely $\|\cdot\|_{l^2}$ (see Theorem 3.1 in Conway (1990)). Therefore, $\|\cdot\|_{\mathbb{R}^p}$ is complete. Moreover, the convergence of the estimated parameters in $\|\cdot\|_{\mathbb{R}^p}$ implies the convergence in the Euclidean norm.

(iii). Assumption 3(ii) is a regularity condition similar to Assumption A3 in Cheng and Shang (2015). This assumption guarantees that $\langle\cdot,\cdot\rangle_{\mathcal{Q}}$ is indeed an inner product, and its induced norm is complete (see Theorem 1 in Section II.3).

(iv). We note that under Assumption 3, for any $(\zeta, g) \in \mathcal{Q}$, $\|(\zeta, g)\|_{\mathcal{Q},1} \equiv \|\zeta\|_{\mathbb{R}^p} + \|g\|_{\mathcal{H}}$ defines a norm on the product space $\mathcal{Q}$. By a similar argument as in the proof of Theorem 1 in the Appendix, a sequence $\{(\zeta_i, g_i)\}_{i=1}^{\infty} \in \mathcal{Q}$ converges in norm $\|\cdot\|_{\mathcal{Q}}$ if and only if it converges in norm $\|\cdot\|_{\mathcal{Q},1}$. This implies that $\|\cdot\|_{\mathcal{Q}}$ and $\|\cdot\|_{\mathcal{Q},1}$ are equivalent norms on $\mathcal{Q}$.

II.2.2 Properties of $\eta(\theta, h)$ and (θ_0, h_0)

Recall $\tau_0 = (\theta_0, h_0)$ is the true parameter in $\mathcal{Q}$. We assume there are neighborhoods $\mathcal{N}_{\theta_0} \subset \mathbb{R}^p$ of θ_0 and $\mathcal{N}_{h_0} \subset \mathcal{H}$ of h_0 such that the following Assumptions 4 to 7 hold.

Assumption 4. *$\eta(\theta, h)$ is three times continuously Fréchet differentiable with respect to (θ, h) in $\mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$. Moreover, $\tau_0 = (\theta_0, h_0)$ is the unique root of*

$$\mathrm{E}\left[D_\theta \ell_n(\theta, h)\right] = 0 \quad \text{and} \quad \mathrm{E}\left[D_h \ell_n(\theta, h)\right] = 0$$

in $\mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$.

Assumption 5. *For any $\theta_* \in \mathcal{N}_{\theta_0}, h_* \in \mathcal{N}_{h_0}$, there exist positive constants C_1, C_2, such that for all $(\zeta, g) \in \mathcal{Q}$,*

$$C_1 V_{\tau_0}(L_\theta \zeta + L_h g) \leq V_{l,\tau_0}(D_\theta \eta(\theta_*, h_*)\zeta + D_h \eta(\theta_*, h_*) g) \leq C_2 V_{\tau_0}(L_\theta \zeta + L_h g).$$

Assumption 6. *For any $(\theta_1, h_1), (\theta_2, h_2) \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$, there exists a positive constant $C_d < 2C_1$, where C_1 is as defined in Assumption 5, such that for any $(\zeta, g) \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$,*

$$\begin{aligned} &V_{\tau_0}\left[(D_\theta \eta(\theta_1, h_1) - D_\theta \eta(\theta_2, h_2))\zeta + (D_h \eta(\theta_1, h_1) - D_h \eta(\theta_2, h_2)) g\right] \\ &\qquad \leq C_d V_{\tau_0}(L_\theta \zeta + L_h g). \end{aligned}$$

Assumption 7. *$\mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$ is a convex set and there exist $C_3, C_4 > 0$ such that*

$$C_3 V_{\tau_0}(f) \leq V_\tau(f) \leq C_4 V_{\tau_0}(f)$$

holds uniformly for any $\tau = (\theta, h) \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$.

Remark 3. *We compare the assumptions above to the existing literature on nonparametric models and the semiparametric partially linear models. But it is important to note that our general setting is not limited to the existing linear models.*

(i). For the nonparametric model where $\eta(x;\theta,h) = h(x)$, and for the semiparametric additive model where $\eta(x;\theta,h) = \alpha(x;\theta) + h(x)$, we see that $\mathcal{H} \subset L_0^2(\mathcal{X})$ and L_h can be chosen to be the inclusion operator $\iota : \mathcal{H} \to L_0^2(\mathcal{X})$. It is easy to see that $V_{\tau_0}(g_1, g_2)$ defines an inner product on $L_0^2(\mathcal{X})$. Moreover, the norm $\|g\|_{\mathcal{H}} = [V_{\tau_0}(g,g) + \lambda J(g,g)]^{1/2}$ has been widely used in the smoothing spline literature for nonparametric models (e.g. Silverman (1982); Cox (1988); Gu and Qiu (1993); Shang (2010)), with different V_{τ_0} as defined early for either the density estimation setting or the regression setting.

(ii). Consider the linear additive model $\eta(x;\theta,h) = \theta^T a(x) + h(x)$, where $a(x) = [a^k(x)]_{k=1}^p$ is a vector of bounded $L_0^2(\mathcal{X})$ functions. One may choose $L_\theta = D_\theta \eta(\theta,h) = a(x)$ and $L_h = D_h \eta(\theta,h) = \iota$, where ι is the inclusion operator from $\mathcal{H}$ to $L_0^2(\mathcal{X})$. For the density estimation model, since $V_{\tau_0}(\theta^T a - h)$ measures the variance of the difference between the parametric and nonparametric components, identifiability of this model follows from Assumption 3(ii).

(iii). As discussed above, by choosing the appropriate operators L_θ and L_h for either the nonparametric model or the semiparametric linear additive model, our analysis for the density estimation model reduces to that studied in Gu and Qiu (1993) or Yang (2009), respectively. Similarly, for the regression model, by choosing the appropriate L_θ and L_h, our model reduces to the nonparametric model or partially linear model. In all these cases, due to the linearity of η, the neighborhood $\mathcal{N}_\theta \times \mathcal{N}_h$ can be taken to be the whole parameter space. Furthermore, Assumptions 4 to 6 are satisfied automatically. One may see that such assumptions are simply redundant when η is linear in θ and h.

(iv). Note that Assumption 7 is similar to Assumption A.3 in Gu and Qiu (1993), Assumption A.3 in Gu (1995), and Condition 3 in Yang (2009). As mentioned in Gu (1995), $V_\tau(f-g)$ can be viewed as a kind of weighted mean square error between functions f and g with weight function $\exp\{\eta(x;\tau)\}$ in the density estimation setting or $\mathrm{E}[l_a''(Y;\eta(X;\theta,h)\mid X]$ in the regression setting. Since η is continuous in $\mathcal{N}_{\theta_0}\times\mathcal{N}_{h_0}$, a small change in τ yields a small change in the weight function, and Assumption 7 simply guarantees that this also yields a relatively small change in the weighted mean square error.

II.2.3 Spectral decomposition

We now construct an eigensystem for the functionals $g\mapsto\sum_{l=1}^m V_{l,\tau_0}(L_{l,h}g)$ and J, which is sometimes referred to as a "simultaneous diagonalization" of the two bilinear forms. Such assumptions are typical in the smoothing spline literature, see Chapter 9 of Gu (2013) and Cox (1988) for detailed discussions and their connections to RKHS, in particular Sobolev spaces.

Assumption 8. *The quadratic functional $g\mapsto V_{\tau_0}(L_hg)$ is completely continuous with respect to the quadratic functional J.*

Under Assumption 8, Theorem 3.1 of Weinberger (1974) yields a sequence $\{\phi_\nu:\nu=1,2,\ldots\}$ of eigenfunctions and a sequence $\{\rho_\nu:\nu=1,2,\ldots\}$ of eigenvalues such that

$$V_{\tau_0}(L_h\phi_\nu,L_h\phi_\mu)=\delta_{\nu\mu}\quad\text{ and }\quad J(\phi_\mu,\phi_\nu)=\rho_\nu\delta_{\nu\mu}$$

for all pairs ν,μ of positive integers, where $\delta_{\nu\mu}$ is the Kronecker delta and $0\le\rho_\nu\to\infty$.

Assumption 9. *$\rho_\nu=\kappa_\nu\nu^r$, where $r>1$ and $\kappa_\nu\in(\beta_1,\beta_2)\subset(0,\infty)$.*

By Assumptions 5 and 8, for any $(\theta_*, h_*) \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$, there exist sequences of eigenfunctions $\{\phi_{*,\nu} : \nu = 1, 2, \ldots\}$ and eigenvalues $\{\rho_{*,\nu} : \nu = 1, 2, \ldots\}$ such that

$$V_{\tau_0}[D_h\eta(\theta_*, h_*)\phi_{*,\nu}, D_h\eta(\theta_*, h_*)\phi_{*,\mu}] = \delta_{\nu\mu} \quad \text{and} \quad J(\phi_{*,\mu}, \phi_{*,\nu}) = \rho_{*,\nu}\delta_{\nu\mu}$$

for all pairs ν, μ of positive integers, where $0 \leq \rho_{*,\nu} \to \infty$. Assumption 5 implies that there exist positive constants c_1, c_2 such that for ν large enough, $c_1\rho_\nu \leq \rho_{*,\nu} \leq c_2\rho_\nu$. By Assumption 9, $\rho_{*,\nu} \sim \nu^r$ for large enough ν, where $r > 1$. Furthermore, for every $h \in \mathcal{H}$ and any $(\theta_*, h_*) \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$, we have a Fourier expansion

$$h = \sum_{\nu=1}^{\infty} V_{\tau_0}(D_h\eta(\theta_*, h_*)h, D_h\eta(\theta_*, h_*)\phi_{*,\nu})\phi_{*,\nu}.$$

II.3 Properties of inner product

We now turn to the discussion of the validity of (II.6). For any $p_1 \times 1$ and $p_2 \times 1$ vectors of functions in $\mathcal{H}$, say $G_1 = [G_1^k]_{k=1}^{p_1}$ and $G_2 = [G_2^k]_{k=1}^{p_2}$, we use the vector form of the inner product $\langle G_1, G_2\rangle_{\mathcal{H}}$ to denote a $p_1 \times p_2$ matrix in which the (i, j)th entry is $\langle G_1^i, G_2^j\rangle_{\mathcal{H}}$. For any $g \in \mathcal{H}$, let $\mathcal{F}^k g = V_{\tau_0}[L_\theta^k, L_h g]$. Since

$$\left|\mathcal{F}^k g\right| \leq \left[V_{\tau_0}(L_\theta^k)\right]^{\frac{1}{2}} \left[V_{\tau_0}(L_h g)\right]^{\frac{1}{2}} \leq C \left\|g\right\|_{\mathcal{H}}$$

for some positive constant C, $\mathcal{F}^k$ is a bounded linear functional on $\mathcal{H}$. By the Riesz representation theorem, there exists a $F^k \in \mathcal{H}$ such that for any $g \in \mathcal{H}$, $\mathcal{F}^k g = \langle F^k, g\rangle_{\mathcal{H}}$. Let $F = \left[F^k\right]_{k=1}^{p}$. We define $V_{\tau_0}(L_\theta, L_h g)$, $V_{\tau_0}(L_h F, L_h g)$, and $J(F, g)$ to be p-dimensional vectors whose kth entries are $V_{\tau_0}(L_\theta^k, L_h g)$, $V_{\tau_0}(L_h F^k, L_h g)$, and $J(F^k, g)$, respectively. Therefore,

$$V_{\tau_0}(L_\theta, L_h g) = V_{\tau_0}(L_h F, L_h g) + \lambda J(F, g) = \langle F, g\rangle_{\mathcal{H}}.$$

We also define the $p \times p$ matrix $\Omega_F = V_{\tau_0}(L_\theta - L_h F, L_\theta - L_h F)$, whose (i,j)th entry is $V_{\tau_0}(L_\theta^i - L_h F^i, L_\theta^j - L_h F^j)$.

Lemma 1. *Under Assumption 3, Ω_F is positive definite and the eigenvalues of Ω_F are greater than c_δ, which is as defined in* (II.5).

Proof. Fix a non-zero vector $\zeta \in \mathbb{R}^p$ and write $\zeta_* = \zeta / \|\zeta\|_{l^2}$. We have

$$\begin{aligned}
\zeta^T \Omega_F \zeta &= \zeta^T V_{\tau_0}(L_\theta - L_h F, L_\theta - L_h F)\zeta \\
&= \|\zeta\|_{l^2}^2 \, V_{\tau_0}[L_\theta \zeta_* - L_h(\zeta_*^T F), L_\theta \zeta_* - L_h(\zeta_*^T F)] \\
&> c_\delta \|\zeta\|_{l^2}^2 ,
\end{aligned}$$

where the last inequality holds by Assumption 3(ii) because $\|\zeta_*\|_{l^2} = 1$ and $\zeta_*^T F \in \mathcal{H}$. Therefore, Ω_F is positive definite.

Let δ be any eigenvalue of Ω_F, and let $\zeta_\delta \in \mathbb{R}^p$ be a unit eigenvector associated with δ. By definition, we have $\Omega_F \zeta_\delta = \delta \zeta_\delta$. We have

$$\begin{aligned}
\delta = \zeta_\delta^T \delta \zeta_\delta &= \zeta_\delta^T \Omega_F \zeta_\delta \\
&= V_{l,\tau_0}[L_\theta \zeta_\delta - L_h(\zeta_\delta^T F), L_\theta \zeta_* - L_h(\zeta_*^T F)] > c_\delta.
\end{aligned}$$

□

Theorem 1. *Suppose Assumption 3 holds. Then $\langle \cdot, \cdot \rangle_{\mathcal{Q}}$ given by* (II.6) *is a well-defined inner product on $\mathcal{Q}$, and $\mathcal{Q}$ is complete with respect to the norm $\|\cdot\|_{\mathcal{Q}}$ induced by this inner product. Hence, $\mathcal{Q}$ is a Hilbert space.*

Proof. It is easy to check that (II.6) satisfies symmetry, linearity and positive semi-definiteness for an inner product. If $(\zeta, g) = 0$, $\langle (\zeta, g), (\zeta, g) \rangle_{\mathcal{Q}} = 0$ is

obvious. We will now show that $\langle(\zeta, g), (\zeta, g)\rangle_{\mathcal{Q}} = 0$ implies $(\zeta, g) = 0$. We see that

$$\begin{aligned}
\langle(\zeta, g),(\zeta, g)\rangle_{\mathcal{Q}} &= V_{\tau_0}(L_\theta\zeta + L_h g, L_\theta\zeta + L_h g) + \lambda J(g, g) \\
&= \left[\zeta^T V_{\tau_0}(L_\theta, L_\theta)\zeta + 2\zeta^T V_{\tau_0}(L_\theta, L_h g) + V_{\tau_0}(L_h g, L_h g)\right] + \lambda J(g, g) \\
&= \zeta^T V_{\tau_0}(L_\theta - L_h F, L_\theta - L_h F)\zeta - \left[V_{\tau_0}(L_h F, L_h F) - 2V_{\tau_0}(L_\theta, L_h F)\right]\zeta \\
&\quad + 2\zeta^T V_{l,\tau_0}(L_\theta, L_h g) + \langle g, g\rangle_{\mathcal{H}} \\
&= \zeta^T \Omega_F \zeta + \langle\zeta^T F + g, \zeta^T F + g\rangle_{\mathcal{H}} + \lambda J(\zeta^T F, \zeta^T F),
\end{aligned} \tag{II.7}$$

and every term in (II.7) is non-negative. If $\langle(\zeta, g), (\zeta, g)\rangle = 0$, the first term in (II.7) implies $\zeta = 0$ by Lemma 1. This further implies that

$$\langle\zeta^T F + g, \zeta^T F + g\rangle_{\mathcal{H}} = \langle g, g\rangle_{\mathcal{H}} = 0.$$

Therefore, $g = 0$ because $\langle\cdot, \cdot\rangle_{\mathcal{H}}$ is an inner product on $\mathcal{H}$. Hence, $\langle\cdot, \cdot\rangle_{\mathcal{Q}}$ is a well-defined inner product on $\mathcal{Q}$.

Next, we want to show that $\mathcal{Q}$ is complete with respect to the norm $\|\cdot\|_{\mathcal{Q}}$. Let $\{(\zeta_i, g_i)\}_{i=1}^{\infty} \subset \mathcal{Q}$ be a Cauchy sequence. For any $\epsilon > 0$, there exist a positive integer M such that for all $i, j > M$, we have

$$\|(\zeta_i, g_i) - (\zeta_j, g_j)\|_{\mathcal{Q}}^2 = V_{\tau_0}[L_\theta(\zeta_i - \zeta_j) + L_h(g_i - g_j)] + \lambda J(g_i - g_j) \leq \epsilon.$$

This implies that

$$\begin{aligned}
&V_{\tau_0}[L_\theta(\zeta_i - \zeta_j) + L_h(g_i - g_j)] \\
&\qquad = \|\zeta_i - \zeta_j\|_{l^2}^2 \, V_{\tau_0}[L_\theta(\zeta_i - \zeta_j)^* + L_h(g_i - g_j)^*] \leq \epsilon,
\end{aligned}$$

where $(\zeta_i - \zeta_j)^* = (\zeta_i - \zeta_j)/\left\|\zeta_i - \zeta_j\right\|_{l^2}$, and $(g_i - g_j)^* = (g_i - g_j)/\left\|\zeta_i - \zeta_j\right\|_{l^2}$. By Assumption 3($ii$), $V_{\tau_0}[L_\theta(\zeta_i - \zeta_j)^* + L_h(g_i - g_j)^*] > c_\delta$ for some positive constant c_δ as defined in (II.5). Therefore,

$$\left\|\zeta_i - \zeta_j\right\|_{l^2}^2 \leq \frac{\epsilon}{c_\delta}, \tag{II.8}$$

and $\{\zeta_i\}_{i=1}^\infty$ is a Cauchy sequence in $\mathbb{R}^p$ under the Euclidean norm, which therefore converges to some limit $\zeta_\infty \in \mathbb{R}^p$.

To find a limit for the sequence $\{g_i\}_{i=1}^\infty$ in $\mathcal{H}$, we consider

$$\begin{aligned}
&\left\|\zeta_i - \zeta_j\right\|_{\mathbb{R}^p}^2 + \left\|g_i - g_j\right\|_{\mathcal{H}}^2 - 2\left\|\zeta_i - \zeta_j\right\|_{\mathbb{R}^p}\left\|g_i - g_j\right\|_{\mathcal{H}} \\
&\leq \left\|\zeta_i - \zeta_j\right\|_{\mathbb{R}^p}^2 + \left\|g_i - g_j\right\|_{\mathcal{H}}^2 - 2\left|V_{\tau_0}[L_\theta(\zeta_i - \zeta_j), L_h(g_i - g_j)]\right| \\
&\leq \left\|\zeta_i - \zeta_j\right\|_{\mathbb{R}^p}^2 + \left\|g_i - g_j\right\|_{\mathcal{H}}^2 + 2V_{\tau_0}[L_\theta(\zeta_i - \zeta_j), L_h(g_i - g_j)] \\
&= V_{\tau_0}[L_\theta(\zeta_i - \zeta_j) + L_h(g_i - g_j)] + \lambda J(g_i - g_j) \\
&= \left\|(\zeta_i, g_i) - (\zeta_j, g_j)\right\|_{\mathcal{Q}}^2 \leq \epsilon
\end{aligned} \tag{II.9}$$

where the first inequality follows from the Cauchy-Schwarz inequality, and the second inequality follows from the triangle inequality. For $a > 0, b > 0$, we have $(1/4)a + b - a^{1/2}b^{1/2} = [(1/2)a^{1/2} - b^{1/2}]^2 \geq 0$, and it follows that $2a^{1/2}b^{1/2} \leq (1/2)a + 2b$. For $a = \left\|g_i - g_j\right\|_{\mathcal{H}}^2$ and $b = \left\|\zeta_i - \zeta_j\right\|_{\mathbb{R}^p}^2$, (II.9) becomes

$$\begin{aligned}
\left\|\zeta_i - \zeta_j\right\|_{\mathbb{R}^p}^2 + \left\|g_i - g_j\right\|_{\mathcal{H}}^2 &\leq \epsilon + 2\left\|\zeta_i - \zeta_j\right\|_{\mathbb{R}^p}\left\|g_i - g_j\right\|_{\mathcal{H}} \\
&\leq \epsilon + \frac{1}{2}\left\|g_i - g_j\right\|_{\mathcal{H}}^2 + 2\left\|\zeta_i - \zeta_j\right\|_{\mathbb{R}^p}^2.
\end{aligned} \tag{II.10}$$

Since $\left\|\cdot\right\|_{\mathbb{R}^p}$ is equivalent to $\left\|\cdot\right\|_{l^2}$ on $\mathbb{R}^p$, $\left\|\zeta_i - \zeta_j\right\|_{\mathbb{R}^p}^2 \leq C\epsilon$ for some positive constant C by (II.8). Therefore, after rearranging (II.10), we get $\left\|g_i - g_j\right\|_{\mathcal{H}}^2 \leq (2 + C)\epsilon$. Hence, $\{g_i\}_{i=1}^\infty$ is a Cauchy sequence in $\mathcal{H}$ under the norm $\left\|\cdot\right\|_{\mathcal{H}}$. By Assumption 3($i$), this sequence converges to some limit $g_\infty \in \mathcal{H}$.

Lastly, we show that (ζ_i, g_i) converges to (ζ_∞, g_∞) in $\|\cdot\|_{\mathcal{Q}}$. By the Cauchy-Schwarz inequality and triangle inequality, as $i \to \infty$, we have

$$\begin{aligned}
&\|(\zeta_i, g_i) - (\zeta_\infty, g_\infty)\|_{\mathcal{Q}}^2 \\
&\quad = \|\zeta_i - \zeta_\infty\|_{\mathbb{R}^p}^2 + \|g_i - g_\infty\|_{\mathcal{H}}^2 + 2V_{\tau_0}[L_\theta(\zeta_i - \zeta_\infty), L_h(g_i - g_\infty)] \\
&\quad \leq \|\zeta_i - \zeta_\infty\|_{\mathbb{R}^p}^2 + \|g_i - g_\infty\|_{\mathcal{H}}^2 + 2\,\|\zeta_i - \zeta_\infty\|_{\mathbb{R}^p}\,\|g_i - g_\infty\|_{\mathcal{H}} \to 0.
\end{aligned}$$

Therefore, we conclude that $\mathcal{Q}$ is a Hilbert space with respect to the inner product $\langle \cdot, \cdot \rangle_{\mathcal{Q}}$. □

CHAPTER III

LOCAL EXISTENCE AND UNIQUENESS OF PENALIZED LIKELIHOOD ESTIMATORS

In this chapter, under Assumptions 1 to 7, we establish a theory for local existence and uniqueness of the penalized likelihood estimator in the neighborhood $\mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$.

When $\eta(\theta, h) = h$, our models reduce to the nonparametric models, for which the existence of penalized likelihood estimators have been studied by many authors Cox (1988); Cox and O'Sullivan (1990); Tapia and Thompson (1978); Silverman (1982); Gu and Qiu (1993). In particular, the existence result stated in Theorem 4.1 of Gu and Qiu (1993) can be extended to the semiparametric case when $\ell_n(\eta(\theta, h))$ is continuous, convex with respect to (θ, h), and has a unique minimizer in $\mathcal{Q}$. In general, the functions $\eta(\theta, h)$ and $\ell_n(\eta(\theta, h))$ need not satisfy these conditions, and hence, we are motivated to establish a local existence theory for our penalized likelihood estimator in a neighborhood $\mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$ of the true parameter (θ_0, h_0).

Denote the local penalized likelihood estimator as

$$(\hat{\theta}, \hat{h}) = \underset{(\theta,h)\in\mathcal{N}_{\theta_0}\times\mathcal{N}_{h_0}}{\arg\min} \ell_{n,\lambda}(\theta, h),$$

which satisfies

$$\begin{cases} D_\theta \ell_{n,\lambda}(\hat{\theta}, \hat{h}) = 0 \\ D_h \ell_{n,\lambda}(\hat{\theta}, \hat{h}) = 0. \end{cases}$$

We also denote $\ell_\lambda(\theta, h) = \mathrm{E}[\ell_{n,\lambda}(\theta, h)]$, and it is easy to verify that the true parameter $\tau_0 = (\theta_0, h_0)$ is the solution for

$$\begin{cases} D_\theta \ell_0(\theta, h) = 0 \\ D_h \ell_0(\theta, h) = 0. \end{cases}$$

We assume $\ell_n(\theta, h)$ is locally convex in $\mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$. Lastly, as defined in Chapter II, we use $\langle \cdot, \cdot \rangle_{\mathbb{R}^p}, \langle \cdot, \cdot \rangle_{\mathcal{H}}$ to denote the inner products on $\mathbb{R}^p$ and $\mathcal{H}$, respectively, and $\|\cdot\|_{\mathbb{R}^p}$, $\|\cdot\|_{\mathcal{H}}$ denote the corresponding induced norms.

III.1 Linearization

In this section, we extend the linearization technique used to approximate the systematic and stochastic components of the estimation error as in Cox and O'Sullivan (1990) and O'Sullivan (1990) to our semiparametric setting by using the bivariate Taylor series expansions for nonlinear operators. We first state the following proposition, whose proof is provided in Ke and Wang (2004).

Proposition 1. *Let $f : Domain(f) \subset X \times Y \to Z$, where X, Y and Z are Banach spaces. If f'' exists at (x, y), then the partial Fréchet derivatives f_{xx}, f_{xy}, f_{yx} and f_{yy} exist at (x, y). For any $h, a \in X$, $k, b \in Y$,*

$$f''(x, y)(h, k)(a, b) = f_{xx}(x, y)ha + f_{xy}(x, y)ka + f_{yx}(x, y)hb + f_{yy}(x, y)kb.$$

By the above theorem and the Taylor formula given in Chapter 1 Section 4 in Ambrosetti and Prodi (1995), we can write the first order Taylor series expansion of $f(x, y)$ as

$$f(x + h, y + k) = f(x, y) + f_x(x, y)h + f_y(x, y)k + R,$$

where R is the remainder, given by

$$\begin{aligned} R &= \int_0^1 (1-t) f''(x+th, y+tk)(h,k)(h,k)dt \\ &= \int_0^1 (1-t) \left[f_{xx}(x+th, y+tk)hh + f_{xy}(x+th, y+tk)kh \right. \\ &\qquad \left. + f_{yx}(x+th, y+tk)hk + f_{yy}(x+th, y+tk)kk \right] dt. \end{aligned}$$

III.1.1 Linear expansions

Since $D_\theta \ell_\lambda(\theta, h)$ is a bounded linear functional on $\mathbb{R}^p$, by the Riesz representation theorem, there exists $Z_\theta(\theta, h) \in \mathbb{R}^p$ such that for any $a \in \mathbb{R}^p$,

$$D_\theta \ell_\lambda(\theta, h)a = \langle Z_\theta(\theta, h), a \rangle_{\mathbb{R}^p}.$$

Similarly, we can denote the Riesz representer of $D_h \ell_\lambda(\theta, h)$ in $\mathcal{H}$ by $Z_h(\theta, h)$. For convenience, we use

$$Z_\theta(\theta, h) = D_\theta \ell_\lambda(\theta, h) \quad \text{and} \quad Z_h(\theta, h) = D_h \ell_\lambda(\theta, h)$$

to represent either the functionals or their Riesz representers in $\mathbb{R}^p$ and $\mathcal{H}$, respectively. For any $\theta_0 + a \in \mathcal{N}_{\theta_0}$ and $h_0 + g \in \mathcal{N}_{h_0}$, the first order Taylor series expansions of Z_h, Z_θ at the true parameter (θ_0, h_0) are

$$\begin{aligned} Z_h(\theta_0 + a, h_0 + g) &= Z_h(\theta_0, h_0) + D_\theta Z_h(\theta_0, h_0)a \\ &\qquad + D_h Z_h(\theta_0, h_0)g + R_h(\theta_0, h_0)ag, \\ Z_\theta(\theta_0 + a, h_0 + g) &= Z_\theta(\theta_0, h_0) + D_\theta Z_\theta(\theta_0, h_0)a \\ &\qquad + D_h Z_\theta(\theta_0, h_0)g + R_\theta(\theta_0, h_0)ag, \end{aligned}$$

where

$$R_h(\theta,h)ag = \int_0^1 (1-t)\left[D^2_{\theta\theta}Z_h(\theta+ta,h+tg)aa + D^2_{h\theta}Z_h(\theta+ta,h+tg)ag\right.$$
$$\left. + D^2_{\theta h}Z_h(\theta+ta,h+tg)ga + D^2_{hh}Z_h(\theta+ta,h+tg)gg\right]\ dt,$$
$$R_\theta(\theta,h)ag = \int_0^1 (1-t)\left[D^2_{\theta\theta}Z_\theta(\theta+ta,h+tg)aa + D^2_{h\theta}Z_\theta(\theta+ta,h+tg)ag\right.$$
$$\left. + D^2_{\theta h}Z_\theta(\theta+ta,h+tg)ga + D^2_{hh}Z_\theta(\theta+ta,h+tg)gg\right]\ dt.$$

We next gives the precise forms of Z_θ, Z_h and their Fréchet partial derivatives in the density estimation setting and regression setting separately.

(I) Density estimation setting

Recall that for any $\tau = (\theta, h) \in \mathcal{Q}$ and any function $f, g \in L^2(\mathcal{X})$, we denote

$$\mu_\tau(f) = \frac{\int_{\mathcal{X}} f(x)e^{\eta(x;\tau)}dx}{\int_{\mathcal{X}} e^{\eta(x;\tau)}dx},$$
$$V_\tau(f,g) = \mu_\tau(fg) - \mu_\tau(f)\mu_\tau(g).$$

For $u, v \in \mathcal{H}$, $a, b \in \mathbb{R}^p$, direct calculation gives

$$Z_h(\theta,h)u = -\mu_{\tau_0}[D_h\eta(\theta,h)u] + \mu_\tau[D_h\eta(\theta,h)u] + \lambda J(h,u),$$
$$D_hZ_h(\theta,h)uv = -\left\{\mu_{\tau_0}[D^2_{hh}\eta(\theta,h)uv] - \mu_\tau[D^2_{hh}\eta(\theta,h)uv]\right\}$$
$$+ V_\tau[D_h\eta(\theta,h)v, D_h\eta(\theta,h)u] + \lambda J(v,u),$$
$$D_\theta Z_h(\theta,h)ua = -\left\{\mu_{\tau_0}[D^2_{\theta h}\eta(\theta,h)ua] - \mu_\tau[D^2_{\theta h}\eta(\theta,h)ua]\right\}$$
$$+ V_\tau[D_\theta\eta(\theta,h)a, D_h\eta(\theta,h)u].$$

$$
\begin{aligned}
Z_\theta(\theta,h)a &= -\mu_{\tau_0}[D_\theta\eta(\theta,h)a] + \mu_\tau[D_\theta\eta(\theta,h)a],\\
D_hZ_\theta(\theta,h)au &= -\left\{\mu_{\tau_0}[D^2_{h\theta}\eta(\theta,h)au] - \mu_\tau[D^2_{h\theta}\eta(\theta,h)au]\right\}\\
&\quad + V_\tau[D_\theta\eta(\theta,h)a, D_h\eta(\theta,h)u],\\
D_\theta Z_\theta(\theta,h)ab &= -\left\{\mu_{\tau_0}[D^2_{\theta\theta}\eta(\theta,h)ab] - \mu_\tau[D^2_{\theta\theta}\eta(\theta,h)ab]\right\}\\
&\quad + V_\tau[D_\theta\eta(\theta,h)a, D_\theta\eta(\theta,h)b].
\end{aligned}
$$

(II) Regression setting

Recall that for any $\tau = (\theta, h) \in \mathcal{Q}$ and any function $f, g \in L^2(\mathcal{X})$, we denote

$$
\begin{aligned}
V_\tau(f,g) &= \mathrm{E}\left[l''_a(Y;\eta(X;\theta,h)f(X)g(X)\right]\\
&= \mathrm{E}_X\left[\mathrm{E}[l''_a(Y;\eta(X;\theta,h) \mid X]f(X)g(X)\right].
\end{aligned}
$$

For $u, v \in \mathcal{H}$, $a, b \in \mathbb{R}^p$, direct calculation gives

$$
\begin{aligned}
Z_h(\theta,h)u &= \mathrm{E}\left[l'_a(Y;\eta(X;\theta,h))D_h\eta(\theta,h)u\right] + \lambda J(h,u),\\
D_hZ_h(\theta,h)uv &= \mathrm{E}\left[l'_a(Y;\eta(X;\theta,h))D^2_{hh}\eta(\theta,h)uv\right]\\
&\quad + V_\tau[D_h\eta(\theta,h)v, D_h\eta(\theta,h)u] + \lambda J(v,u),\\
D_\theta Z_h(\theta,h)ua &= \mathrm{E}\left[l'_a(Y;\eta(X;\theta,h))D^2_{\theta h}\eta(\theta,h)ua\right]\\
&\quad + V_\tau[D_\theta\eta(\theta,h)a, D_h\eta(\theta,h)u].
\end{aligned}
$$

$$
\begin{aligned}
Z_\theta(\theta,h)a &= \mathrm{E}\left[l_a'(Y;\eta(X;\theta,h))D_\theta\eta(\theta,h)a\right], \\
D_hZ_\theta(\theta,h)au &= \mathrm{E}\left[l_a'(Y;\eta(X;\theta,h))D_{h\theta}^2\eta(\theta,h)au\right] \\
&\quad + V_\tau[D_\theta\eta(\theta,h)a, D_h\eta(\theta,h)u], \\
D_\theta Z_\theta(\theta,h)ab &= \mathrm{E}\left[l_a'(Y;\eta(X;\theta,h))D_{\theta\theta}^2\eta(\theta,h)ab\right] \\
&\quad + V_\tau[D_\theta\eta(\theta,h)a, D_\theta\eta(\theta,h)b].
\end{aligned}
$$

Using the representations in either setting as given above, for any $\tau = (\theta,h) \in \mathcal{Q}$, $u,v \in \mathcal{H}$, and $a,b \in \mathbb{R}^p$, define the operators $U_\theta(\theta,h)$ and $U_h(\theta,h)$ on $\mathbb{R}^p$ and $\mathcal{H}$, respectively, such that

$$
\begin{aligned}
\langle u, U_h(\theta,h)v\rangle_{\mathcal{H}} &= V_\tau[D_h\eta(\theta,h)u, D_h\eta(\theta,h)v], \\
\langle a, U_\theta(\theta,h)b\rangle_{\mathbb{R}^p} &= V_\tau[D_\theta\eta(\theta,h)a, D_\theta\eta(\theta,h)b].
\end{aligned}
$$

Note that these operators are well-defined by the Riesz representation theorem applied to the linear functionals

$$
v \mapsto V_\tau[D_h\eta(\theta,h)u, D_h\eta(\theta,h)v], \qquad b \mapsto V_\tau[D_\theta\eta(\theta,h)a, D_\theta\eta(\theta,h)b],
$$

which are bounded in the corresponding norms on $\mathcal{H}$ and $\mathbb{R}^p$, respectively. Similarly, we also define $U_{h\theta}(\theta,h) : \mathcal{H} \to \mathbb{R}^p$ and $U_{\theta h}(\theta,h) : \mathbb{R}^p \to \mathcal{H}$ by

$$
\langle a, U_{h\theta}(\theta,h)u\rangle_{\mathbb{R}^p} = V_\tau[D_\theta\eta(\theta,h)a, D_h\eta(\theta,h)u] = \langle u, U_{\theta h}(\theta,h)a\rangle_{\mathcal{H}}.
$$

By the same arguement as Lemma S.2 in the supplement of Cheng and Shang (2015), there exists a bounded linear operator W_λ on $\mathcal{H}$ such that

$$
\langle u, W_\lambda v\rangle_{\mathcal{H}} = \lambda J(u,v).
$$

Therefore, we have

$$
\begin{aligned}
Z_h(\theta_0 + a, h_0 + g) &= Z_h(\theta_0, h_0) + G_h(\theta_0, h_0)g + U_{\theta h}(\theta_0, h_0)a + R_h(\theta_0, h_0)ag, \\
Z_\theta(\theta_0 + a, h_0 + g) &= Z_\theta(\theta_0, h_0) + U_\theta(\theta_0, h_0)a + U_{h\theta}(\theta_0, h_0)g + R_\theta(\theta_0, h_0)ag,
\end{aligned} \tag{III.1}
$$

where $G_h(\theta, h) = U_h(\theta, h) + W_\lambda$. We provide the presentations of the remainder terms $R_h(\theta, h)ag$ and $R_\theta(\theta, h)ag$ in Section III.1.3.

Suppose $(\theta_\lambda, h_\lambda)$ is a solution for $Z_\theta(\theta, h) = Z_h(\theta, h) = 0$. We define the systematic error as $(\theta_\lambda - \theta_0, h_\lambda - h_0)$. Ignoring the remainder terms, we get an approximation to the systematic error by setting the system of equations (III.1) to 0 and solving for $\bar{\theta}_\lambda - \theta_0$, and $\bar{h}_\lambda - h_0$, i.e.,

$$
\begin{aligned}
Z_h(\theta_0, h_0) + G_h(\theta_0, h_0)(\bar{h}_\lambda - h_0) + U_{\theta h}(\theta_0, h_0)(\bar{\theta}_\lambda - \theta_0) = 0, \\
Z_\theta(\theta_0, h_0) + U_\theta(\theta_0, h_0)(\bar{\theta}_\lambda - \theta_0) + U_{h\theta}(\theta_0, h_0)(\bar{h}_\lambda - h_0) = 0.
\end{aligned}
$$

By the Lax-Milgram theorem (Section 3.6 of Aubin (1979)) and Assumptions 3 and 5, for any $(\theta, h) \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$, the operators $G_h(\theta, h)$, $U_\theta(\theta, h)$ have bounded inverses on $\mathcal{H}$ and $\mathbb{R}^p$, respectively. Let

$$
\begin{aligned}
G_{hh}(\theta, h) &= (G_h - U_{\theta h}U_\theta^{-1}U_{h\theta})(\theta, h) : \mathcal{H} \to \mathcal{H}, \\
G_{\theta\theta}(\theta, h) &= (U_\theta - U_{h\theta}G_h^{-1}U_{\theta h})(\theta, h) : \mathbb{R}^p \to \mathbb{R}^p.
\end{aligned}
$$

Assuming both operators above have bounded inverses for any $(\theta, h) \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$, we get

$$
\begin{aligned}
\bar{h}_\lambda - h_0 &= -G_{hh}^{-1}(\theta_0, h_0)\left[Z_h(\theta_0, h_0) - U_{\theta h}(\theta_0, h_0)U_\theta^{-1}(\theta_0, h_0)Z_\theta(\theta_0, h_0)\right], \\
\bar{\theta}_\lambda - \theta_0 &= -G_{\theta\theta}^{-1}(\theta_0, h_0)\left[Z_\theta(\theta_0, h_0) - U_{h\theta}(\theta_0, h_0)G_h^{-1}(\theta_0, h_0)Z_h(\theta_0, h_0)\right].
\end{aligned}
$$

Note that with Assumptions 8 and 9, $G_{hh}(\theta, h)$ can be shown to have a bounded inverse by the analysis of Ω_λ in the proof for Lemma 2 in Chapter IV. $G_{\theta\theta}(\theta, h)$ can be shown to have a bounded inverse in a similar manner.

Next, we define the stochastic error as $(\hat{\theta} - \theta_\lambda, \hat{h} - h_\lambda)$. Let

$$Z_{n\theta}(\theta, h) = D_\theta \ell_{n,\lambda}(\theta, h) \quad \text{and} \quad Z_{nh}(\theta, h) = D_h \ell_{n,\lambda}(\theta, h).$$

The approximation of the stochastic errors can be obtained by the linearizations of $Z_{n\theta}$ and Z_{nh}. The precise forms of $Z_{n\theta}$, Z_{nh} and their Fréchet partial derivatives are given below for the density estimation and regression separately.

(I) Density estimation setting

For $u, v \in \mathcal{H}$ and $a, b \in \mathbb{R}^p$, we have

$$\begin{aligned}
Z_{n\theta}(\theta, h)a &= -\frac{1}{n}\sum_{i=1}^{n} D_\theta \eta(x_i; \theta, h)a + \mu_\tau \left[D_\theta \eta(\theta, h)a \right], \\
D_\theta Z_{n\theta}(\theta, h)ab &= e_{\theta\theta}(\theta, h)ab + V_\tau \left[D_\theta \eta(\theta, h)a, D_\theta \eta(\theta, h)b \right], \\
D_h Z_{n\theta}(\theta, h)au &= e_{h\theta}(\theta, h)au + V_\tau \left[D_\theta \eta(\theta, h)a, D_h \eta(\theta, h)u \right], \\
Z_{nh}(\theta, h)u &= -\frac{1}{n}\sum_{i=1}^{n} D_h \eta(x_i; \theta, h)u + \mu_\tau \left[D_h \eta(\theta, h)u \right] + \lambda J(h, u), \\
D_\theta Z_{nh}(\theta, h)ua &= e_{\theta h}(\theta, h)ua + V_\tau \left[D_h \eta(\theta, h)u, D_\theta \eta(\theta, h)a \right], \\
D_h Z_{n\theta}(\theta, h)uv &= e_{hh}(\theta, h)uv + V_\tau \left[D_h \eta(\theta, h)u, D_h \eta(\theta, h)v \right] + \lambda J(v, u),
\end{aligned}$$

where

$$e_{\theta\theta}(\theta,h)ab = -\frac{1}{n}\sum_{i=1}^{n} D^2_{\theta\theta}\eta(X_i;\theta,h)ab + \mu_\tau\left[D^2_{\theta\theta}\eta(\theta,h)ab\right],$$

$$e_{h\theta}(\theta,h)au = -\frac{1}{n}\sum_{i=1}^{n} D^2_{h\theta}\eta(X_i;\theta,h)au + \mu_\tau\left[D^2_{h\theta}\eta(\theta,h)au\right],$$

$$e_{\theta h}(\theta,h)ua = -\frac{1}{n}\sum_{i=1}^{n} D^2_{\theta h}\eta(X_i;\theta,h)ua + \mu_\tau\left[D^2_{\theta h}\eta(\theta,h)ua\right],$$

$$e_{hh}(\theta,h)uv = -\frac{1}{n}\sum_{i=1}^{n} D^2_{hh}\eta(X_i;\theta,h)uv + \mu_\tau\left[D^2_{hh}\eta(\theta,h)uv\right].$$

(II) Regression setting

For $u, v \in \mathcal{H}$ and $a, b \in \mathbb{R}^p$, we have

$$Z_{n\theta}(\theta,h)a = \frac{1}{n}\sum_{i=1}^{n} l'_a(Y_i;\eta(X_i;\theta,h))D_\theta\eta(X_i;\theta,h)a,$$

$$D_\theta Z_{n\theta}(\theta,h)ab = e_{\theta\theta}(\theta,h)ab + V_\tau\left[D_\theta\eta(\theta,h)a, D_\theta\eta(\theta,h)b\right],$$

$$D_h Z_{n\theta}(\theta,h)au = e_{h\theta}(\theta,h)au + V_\tau\left[D_\theta\eta(\theta,h)a, D_h\eta(\theta,h)u\right],$$

$$Z_{nh}(\theta,h)u = \frac{1}{n}\sum_{i=1}^{n} l'_a(Y_i;\eta(X_i;\theta,h))D_h\eta(X_i;\theta,h)u, +\lambda J(h,u),$$

$$D_\theta Z_{nh}(\theta,h)ua = e_{\theta h}(\theta,h)ua + V_\tau\left[D_h\eta(\theta,h)u, D_\theta\eta(\theta,h)a\right],$$

$$D_h Z_{n\theta}(\theta,h)uv = e_{hh}(\theta,h)uv + V_\tau\left[D_h\eta(\theta,h)u, D_h\eta(\theta,h)v\right] + \lambda J(v,u),$$

where

$$\begin{aligned} e_{\theta\theta}(\theta,h)ab = \frac{1}{n}\sum_{i=1}^{n} & l'_a(Y_i;\eta(X_i;\theta,h))D^2_{\theta\theta}\eta(X_i;\theta,h)ab \\ & - V_\tau\left[D_\theta\eta(\theta,h)a, D_\theta\eta(\theta,h)b\right] \\ & + \frac{1}{n}\sum_{i=1}^{n} l''_a(Y_i;\eta(X_i;\theta,h))[D_\theta\eta(X_i;\theta,h)a][D_\theta\eta(X_i;\theta,h)a], \end{aligned}$$

$$
\begin{aligned}
e_{h\theta}(\theta,h)au &= \frac{1}{n}\sum_{i=1}^{n} l_a'(Y_i;\eta(X_i;\theta,h))D^2_{h\theta}\eta(X_i;\theta,h)au \\
&\quad - V_\tau\left[D_\theta\eta(\theta,h)a, D_h\eta(\theta,h)u\right] \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} l_a''(Y_i;\eta(X_i;\theta,h))[D_\theta\eta(X_i;\theta,h)a][D_h\eta(X_i;\theta,h)u], \\
e_{\theta h}(\theta,h)ua &= \frac{1}{n}\sum_{i=1}^{n} l_a'(Y_i;\eta(X_i;\theta,h))D^2_{\theta h}\eta(X_i;\theta,h)ua \\
&\quad - V_\tau\left[D_h\eta(\theta,h)u, D_\theta\eta(\theta,h)a\right] \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} l_a''(Y_i;\eta(X_i;\theta,h))[D_h\eta(X_i;\theta,h)u][D_\theta\eta(X_i;\theta,h)a], \\
e_{hh}(\theta,h)uv &= \frac{1}{n}\sum_{i=1}^{n} l_a'(Y_i;\eta(X_i;\theta,h))D^2_{hh}\eta(X_i;\theta,h)uv \\
&\quad - V_\tau\left[D_h\eta(\theta,h)u, D_h\eta(\theta,h)v\right] \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} l_a''(Y_i;\eta(X_i;\theta,h))[D_h\eta(X_i;\theta,h)u][D_h\eta(X_i;\theta,h)v].
\end{aligned}
$$

For both derivations above, we see that $e_{h\theta}(\theta,h)au = e_{\theta h}(\theta,h)ua$. Thus, for any $\theta_\lambda + a \in \mathcal{N}_{\theta_0}, h_\lambda + g \in \mathcal{N}_{h_0}$, the first order Taylor series expansions of $Z_{n\theta}$ and Z_{nh} at $(\theta_\lambda, h_\lambda) \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$ can be written as

$$
\begin{aligned}
Z_{n\theta}(\theta_\lambda + a, h_\lambda + g) &= Z_{n\theta}(\theta_\lambda, h_\lambda) + U_\theta(\theta_\lambda, h_\lambda)a + U_{h\theta}(\theta_\lambda, h_\lambda)g \\
&\quad + e_\theta(\theta_\lambda, h_\lambda)ag + R_{n\theta}(\theta_\lambda, h_\lambda)ag, \\
Z_{nh}(\theta_\lambda + a, h_\lambda + g) &= Z_{nh}(\theta_\lambda, h_\lambda) + G_h(\theta_\lambda, h_\lambda)g + U_{\theta h}(\theta_\lambda, h_\lambda)a \\
&\quad + e_h(\theta_\lambda, h_\lambda)ag + R_{nh}(\theta_\lambda, h_\lambda)ag,
\end{aligned}
\tag{III.2}
$$

where the error terms are given by

$$
\begin{aligned}
e_\theta(\theta_\lambda, h_\lambda)ag &= e_{\theta\theta}(\theta_\lambda, h_\lambda)a + e_{\theta h}(\theta_\lambda, h_\lambda)g, \\
e_h(\theta_\lambda, h_\lambda)ag &= e_{h\theta}(\theta_\lambda, h_\lambda)a + e_{hh}(\theta_\lambda, h_\lambda)g,
\end{aligned}
$$

$R_{nh}, R_{n\theta}$ are defined similarly to R_h, R_θ by replacing Z_h, Z_θ with Z_{nh}, $Z_{n\theta}$, respectively, whose representations are given in Section III.1.3.

Recall that $(\hat{\theta}, \hat{h})$ is the solution for $Z_{n\theta}(\theta, h) = Z_{nh}(\theta, h) = 0$. Dropping the error terms and the remainder terms, we get an approximation to the stochastic error $(\hat{\theta} - \theta_\lambda, \hat{h} - h_\lambda)$ by setting the linearizations (III.2) to 0 and solving for $\bar{\theta}_{n\lambda} - \theta_\lambda$ and $\bar{h}_{n\lambda} - h_\lambda$. We get

$$\bar{h}_{n\lambda} - h_\lambda = -G_{hh}^{-1}(\theta_\lambda, h_\lambda)[Z_{nh}(\theta_\lambda, h_\lambda) - U_{\theta h}(\theta_\lambda, h_\lambda)U_\theta^{-1}(\theta_\lambda, h_\lambda)Z_{n\theta}(\theta_\lambda, h_\lambda)],$$
$$\bar{\theta}_{n\lambda} - \theta_\lambda = -G_{\theta\theta}^{-1}(\theta_\lambda, h_\lambda)[Z_{n\theta}(\theta_\lambda, h_\lambda) - U_{h\theta}(\theta_\lambda, h_\lambda)G_h^{-1}(\theta_\lambda, h_\lambda)Z_{nh}(\theta_\lambda, h_\lambda)].$$

III.1.2 Bounds for the remainders

We see that the magnitude of the remainder terms R_θ, R_h, $R_{n\theta}$, R_{nh}, e_θ, and e_h determine how accurate $(\bar{\theta}_\lambda - \theta_0, \bar{h}_\lambda - h_0)$ and $(\bar{\theta}_{n\lambda} - \theta_\lambda, \bar{h}_{n\lambda} - h_\lambda)$ are as approximations of the systematic error and the stochastic error, respectively. To obtain bounds of these terms, we first define for $\lambda > 0$, $\tau_1 = (\theta_1, h_1), \tau_2 = (\theta_2, h_2) \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$, and unit elements $u_1, u_2 \in \mathbb{R}^p$ and $v_1, v_2 \in \mathcal{H}$,

$$K_h^1 = \sup_{\tau_1,\tau_2} \sup_{v_1,v_2} \left\| G_{hh}^{-1}(\tau_1)[D_{hh}^2 Z_h(\tau_2)v_1v_2 - U_{\theta h}(\tau_1)U_\theta^{-1}(\tau_1)D_{hh}^2 Z_\theta(\tau_2)v_1v_2] \right\|_{\mathcal{H}},$$
$$K_h^2 = \sup_{\tau_1,\tau_2} \sup_{v_1,u_1} \left\| G_{hh}^{-1}(\tau_1)[D_{\theta h}^2 Z_h(\tau_2)v_1u_1 - U_{\theta h}(\tau_1)U_\theta^{-1}(\tau_1)D_{\theta h}^2 Z_\theta(\tau_2)v_1u_1] \right\|_{\mathcal{H}},$$
$$K_h^3 = \sup_{\tau_1,\tau_2} \sup_{u_1,v_1} \left\| G_{hh}^{-1}(\tau_1)[D_{h\theta}^2 Z_h(\tau_2)u_1v_1 - U_{\theta h}(\tau_1)U_\theta^{-1}(\tau_1)D_{h\theta}^2 Z_\theta(\tau_2)u_1v_1] \right\|_{\mathcal{H}},$$
$$K_h^4 = \sup_{\tau_1,\tau_2} \sup_{u_1,u_2} \left\| G_{hh}^{-1}(\tau_1)[D_{\theta\theta}^2 Z_h(\tau_2)u_1u_2 - U_{\theta h}(\tau_1)U_\theta^{-1}(\tau_1)D_{\theta\theta}^2 Z_\theta(\tau_2)u_1u_2] \right\|_{\mathcal{H}},$$
$$K_\theta^1 = \sup_{\tau_1,\tau_2} \sup_{v_1,v_2} \left\| G_{\theta\theta}^{-1}(\tau_1)[D_{hh}^2 Z_\theta(\tau_2)v_1v_2 - U_{h\theta}(\tau_1)G_h^{-1}(\tau_1)D_{hh}^2 Z_h(\tau_2)v_1v_2] \right\|_{\mathbb{R}^p},$$
$$K_\theta^2 = \sup_{\tau_1,\tau_2} \sup_{v_1,u_1} \left\| G_{\theta\theta}^{-1}(\tau_1)[D_{\theta h}^2 Z_\theta(\tau_2)v_1u_1 - U_{h\theta}(\tau_1)G_h^{-1}(\tau_1)D_{\theta h}^2 Z_h(\tau_2)v_1u_1] \right\|_{\mathbb{R}^p},$$
$$K_\theta^3 = \sup_{\tau_1,\tau_2} \sup_{u_1,v_1} \left\| G_{\theta\theta}^{-1}(\tau_1)[D_{h\theta}^2 Z_\theta(\tau_2)u_1v_1 - U_{h\theta}(\tau_1)G_h^{-1}(\tau_1)D_{h\theta}^2 Z_h(\tau_2)u_1v_1] \right\|_{\mathbb{R}^p},$$
$$K_\theta^4 = \sup_{\tau_1,\tau_2} \sup_{u_1,u_2} \left\| G_{\theta\theta}^{-1}(\tau_1)[D_{\theta\theta}^2 Z_\theta(\tau_2)u_1u_2 - U_{h\theta}(\tau_1)G_h^{-1}(\tau_1)D_{\theta\theta}^2 Z_h(\tau_2)u_1u_2] \right\|_{\mathbb{R}^p}.$$

For $i = 1, 2, 3, 4$, we also define $K^i_{nh}, K^i_{n\theta}$ by replacing Z_θ, Z_h with $Z_{n\theta}, Z_{nh}$ in K^i_h and K^i_θ, respectively. In addition, we define

$$
\begin{aligned}
E^{12}_{nh} &= \sup_{\theta_1,h_1} \sup_{u_1} \left\| G^{-1}_{hh}(\theta_1,h_1) e_{h\theta}(\theta_1,h_1) u_1 \right\|_{\mathcal{H}}, \\
E^{11}_{nh} &= \sup_{\theta_1,h_1} \sup_{v_1} \left\| G^{-1}_{hh}(\theta_1,h_1) e_{hh}(\theta_1,h_1) v_1 \right\|_{\mathcal{H}}, \\
E^{22}_{nh} &= \sup_{\theta_1,h_1} \sup_{u_1} \left\| G^{-1}_{hh}(\theta_1,h_1) U_{\theta h}(\theta_1,h_1) U^{-1}_{\theta}(\theta_1,h_1) e_{\theta\theta}(\theta_1,h_1) u_1 \right\|_{\mathcal{H}}, \\
E^{21}_{nh} &= \sup_{\theta_1,h_1} \sup_{v_1} \left\| G^{-1}_{hh}(\theta_1,h_1) U_{\theta h}(\theta_1,h_1) U^{-1}_{\theta}(\theta_1,h_1) e_{\theta h}(\theta_1,h_1) v_1 \right\|_{\mathcal{H}}, \\
E^{22}_{n\theta} &= \sup_{\theta_1,h_1} \sup_{u_1} \left\| G^{-1}_{\theta\theta}(\theta_1,h_1) e_{\theta\theta}(\theta_1,h_1) u_1 \right\|_{\mathbb{R}^p}, \\
E^{21}_{n\theta} &= \sup_{\theta_1,h_1} \sup_{v_1} \left\| G^{-1}_{\theta\theta}(\theta_1,h_1) e_{\theta h}(\theta_1,h_1) v_1 \right\|_{\mathbb{R}^p}, \\
E^{12}_{n\theta} &= \sup_{\theta_1,h_1} \sup_{u_1} \left\| G^{-1}_{\theta\theta}(\theta_1,h_1) U_{h\theta}(\theta_1,h_1) G^{-1}_{h}(\theta_1,h_1) e_{h\theta}(\theta_1,h_1) u_1 \right\|_{\mathbb{R}^p}, \\
E^{11}_{n\theta} &= \sup_{\theta_1,h_1} \sup_{v_1} \left\| G^{-1}_{\theta\theta}(\theta_1,h_1) U_{h\theta}(\theta_1,h_1) G^{-1}_{h}(\theta_1,h_1) e_{hh}(\theta_1,h_1) v_1 \right\|_{\mathbb{R}^p}.
\end{aligned}
$$

Therefore, for any $a \in \mathbb{R}^p$ and $g \in \mathcal{H}$, standard analysis yields the following bounds for the remainder terms for the systematic error and the stochastic error,

$$
\begin{aligned}
&\left\| G^{-1}_{hh}(\theta_0,h_0) \left[R_h(\theta_0,h_0) ag - U_{\theta h}(\theta_0,h_0) U^{-1}_{\theta}(\theta_0,h_0) R_\theta(\theta_0,h_0) ag \right] \right\|_{\mathcal{H}} \\
&\quad \le \frac{1}{2} \left[\left(K^1_h \|g\|_{\mathcal{H}} + K^2_h \|a\|_{\mathbb{R}^p} \right) \|g\|_{\mathcal{H}} + \left(K^3_h \|g\|_{\mathcal{H}} + K^4_h \|a\|_{\mathbb{R}^p} \right) \|a\|_{\mathbb{R}^p} \right],
\end{aligned} \tag{III.3}
$$

$$
\begin{aligned}
&\left\| G^{-1}_{\theta\theta}(\theta_0,h_0) \left[R_\theta(\theta_0,h_0) ag - U_{h\theta}(\theta_0,h_0) G^{-1}_{h}(\theta_0,h_0) R_h(\theta_0,h_0) ag \right] \right\|_{\mathbb{R}^p} \\
&\quad \le \frac{1}{2} \left[\left(K^1_\theta \|g\|_{\mathcal{H}} + K^2_\theta \|a\|_{\mathbb{R}^p} \right) \|g\|_{\mathcal{H}} + \left(K^3_\theta \|g\|_{\mathcal{H}} + K^4_\theta \|a\|_{\mathbb{R}^p} \right) \|a\|_{\mathbb{R}^p} \right],
\end{aligned} \tag{III.4}
$$

$$
\begin{aligned}
&\left\|G_{hh}^{-1}(\theta_\lambda, h_\lambda)\left\{\left[e_h(\theta_\lambda, h_\lambda) + R_{nh}(\theta_\lambda, h_\lambda)\right] ag\right.\right. \\
&\quad \left.\left. - U_{\theta h}(\theta_\lambda, h_\lambda) U_\theta^{-1}(\theta_\lambda, h_\lambda)\left[e_\theta(\theta_\lambda, h_\lambda) + R_{n\theta}(\theta_\lambda, h_\lambda)\right] ag\right\}\right\|_{\mathcal{H}} \\
&\quad \leq \frac{1}{2}\left[\left(K_{nh}^1 \|g\|_{\mathcal{H}} + K_{nh}^2 \|a\|_{\mathbb{R}^p}\right)\|g\|_{\mathcal{H}} + \left(K_{nh}^3 \|g\|_{\mathcal{H}} + K_{nh}^4 \|a\|_{\mathbb{R}^p}\right)\|a\|_{\mathbb{R}^p}\right] \\
&\qquad + E_{nh}^1 \|g\|_{\mathcal{H}} + E_{nh}^2 \|a\|_{\mathbb{R}^p},
\end{aligned}
\tag{III.5}
$$

$$
\begin{aligned}
&\left\|G_{\theta\theta}^{-1}(\theta_\lambda, h_\lambda)\left\{\left[e_\theta(\theta_\lambda, h_\lambda) + R_{n\theta}(\theta_\lambda, h_\lambda)\right] ag\right.\right. \\
&\quad \left.\left. - U_{h\theta}(\theta_\lambda, h_\lambda) G_h^{-1}(\theta_\lambda, h_\lambda)\left[e_h(\theta_\lambda, h_\lambda) + R_{nh}(\theta_\lambda, h_\lambda)\right] ag\right\}\right\|_{\mathbb{R}^p} \\
&\quad \leq \frac{1}{2}\left[\left(K_{n\theta}^1 \|g\|_{\mathcal{H}} + K_{n\theta}^2 \|a\|_{\mathbb{R}^p}\right)\|g\|_{\mathcal{H}} + \left(K_{n\theta}^3 \|g\|_{\mathcal{H}} + K_{n\theta}^4 \|a\|_{\mathbb{R}^p}\right)\|a\|_{\mathbb{R}^p}\right] \\
&\qquad + E_{n\theta}^1 \|g\|_{\mathcal{H}} + E_{n\theta}^2 \|a\|_{\mathbb{R}^p},
\end{aligned}
\tag{III.6}
$$

where $E_{nh}^1 = E_{nh}^{11} + E_{nh}^{21}$, $E_{nh}^2 = E_{nh}^{12} + E_{nh}^{22}$, $E_{n\theta}^1 = E_{n\theta}^{21} + E_{n\theta}^{11}$ and $E_{n\theta}^2 = E_{n\theta}^{22} + E_{n\theta}^{12}$.

III.1.3 Remainder terms

In this section, we provide the precise forms of the second partial Fréchet derivatives of $Z_\theta(\theta, h)$ and $Z_h(\theta, h)$. In practice, for a given form of $\eta(\theta, h)$, these can be used to determine the bounds for the remainder terms R_h and R_θ given in Section III.1.2. The second partial Fréchet derivatives of $Z_{n\theta}$ and Z_{nh} can be found in a similar manner.

(I) Density estimation setting

For $u, v, w \in \mathcal{H}$, $a, b, c \in \mathbb{R}^p$, we have

$$
\begin{aligned}
&D^2_{hh} Z_h(\theta, h) uvw \\
&= -\left\{ \mu_{\tau_0} \left[D^2_{hhh} \eta(\theta, h) uvw \right] - \mu_\tau \left[D^2_{hhh} \eta(\theta, h) uvw \right] \right\} \\
&\quad + V_\tau \left[D^2_{hh} \eta(\theta, h) uv, D_h \eta(\theta, h) w \right] + V_\tau \left[D^2_{hh} \eta(\theta, h) vw, D_h \eta(\theta, h) u \right] \\
&\quad + V_\tau \left[D^2_{hh} \eta(\theta, h) uw, D_h \eta(\theta, h) v \right] + V_\tau \left[D_h \eta(\theta, h) v, D_h \eta(\theta, h) u \cdot D_h \eta(\theta, h) w \right] \\
&\quad - \mu_\tau \left[D_h \eta(\theta, h) u \right] V_\tau \left[D_h \eta(\theta, h) v, D_h \eta(\theta, h) w \right] \\
&\quad - \mu_\tau \left[D_h \eta(\theta, h) w \right] V_\tau \left[D_h \eta(\theta, h) v, D_h \eta(\theta, h) u \right],
\end{aligned}
$$

$$
\begin{aligned}
&D^2_{\theta h} Z_h(\theta, h) uva = D^2_{h\theta} Z_h(\theta, h) uav = D^2_{hh} Z_\theta(\theta, h) auv \\
&= -\left\{ \mu_{\tau_0} \left[D_\theta D^2_{hh} \eta(\theta, h) uva \right] - \mu_\tau \left[D_\theta D^2_{hh} \eta(\theta, h) uva \right] \right\} \\
&\quad + V_\tau \left[D^2_{hh} \eta(\theta, h) uv, D_\theta \eta(\theta, h) a \right] + V_\tau \left[D^2_{\theta h} \eta(\theta, h) va, D_h \eta(\theta, h) u \right] \\
&\quad + V_\tau \left[D^2_{\theta h} \eta(\theta, h) ua, D_h \eta(\theta, h) v \right] + V_\tau \left[D_h \eta(\theta, h) u \cdot D_h \eta(\theta, h) v, D_\theta \eta(\theta, h) a \right] \\
&\quad - \mu_\tau \left[D_h \eta(\theta, h) u \right] V_\tau \left[D_h \eta(\theta, h) v, D_\theta \eta(\theta, h) a \right] \\
&\quad - \mu_\tau \left[D_h \eta(\theta, h) v \right] V_\tau \left[D_\theta \eta(\theta, h) a, D_h \eta(\theta, h) u \right],
\end{aligned}
$$

$$
\begin{aligned}
&D^2_{\theta\theta} Z_h(\theta, h) uab = D^2_{\theta h} Z_\theta(\theta, h) aub = D^2_{h\theta} Z_\theta(\theta, h) abu \\
&= -\left\{ \mu_{\tau_0} \left[D^2_{\theta\theta} D_h \eta(\theta, h) uab \right] - \mu_\tau \left[D^2_{\theta\theta} D_h \eta(\theta, h) uab \right] \right\} \\
&\quad + V_\tau \left[D^2_{\theta h} \eta(\theta, h) ua, D_\theta \eta(\theta, h) b \right] + V_\tau \left[D^2_{\theta\theta} \eta(\theta, h) ab, D_h \eta(\theta, h) u \right] \\
&\quad + V_\tau \left[D^2_{\theta h} \eta(\theta, h) ub, D_\theta \eta(\theta, h) a \right] + V_\tau \left[D_\theta \eta(\theta, h) a, D_\theta \eta(\theta, h) b \cdot D_h \eta(\theta, h) u \right] \\
&\quad - \mu_\tau \left[D_\theta \eta(\theta, h) b \right] V_\tau \left[D_\theta \eta(\theta, h) a, D_h \eta(\theta, h) u \right] \\
&\quad - \mu_\tau \left[D_h \eta(\theta, h) u \right] V_\tau \left[D_\theta \eta(\theta, h) a, D_\theta \eta(\theta, h) b \right],
\end{aligned}
$$

$$
\begin{aligned}
&D^2_{\theta\theta} Z_\theta(\theta, h)abc \\
&= -\left\{\mu_{\tau_0}\left[D^3_{\theta\theta\theta}\eta(\theta, h)abc\right] - \mu_\tau\left[D^3_{\theta\theta\theta}\eta(\theta, h)abc\right]\right\} \\
&\quad + V_\tau\left[D^2_{\theta\theta}\eta(\theta, h)ab, D_\theta\eta(\theta, h)c\right] + V_\tau\left[D^2_{\theta\theta}\eta(\theta, h)bc, D_\theta\eta(\theta, h)a\right] \\
&\quad + V_\tau\left[D^2_{\theta\theta}\eta(\theta, h)ac, D_\theta\eta(\theta, h)b\right] + V_\tau\left[D_\theta\eta(\theta, h)a, D_\theta\eta(\theta, h)b \cdot D_\theta\eta(\theta, h)c\right] \\
&\quad - \mu_\tau\left[D_\theta\eta(\theta, h)b\right] V_\tau\left[D_\theta\eta(\theta, h)a, D_\theta\eta(\theta, h)c\right] \\
&\quad - \mu_\tau\left[D_\theta\eta(\theta, h)c\right] V_\tau\left[D_\theta\eta(\theta, h)a, D_\theta\eta(\theta, h)b\right].
\end{aligned}
$$

By replacing the terms $\mu_{\tau_0}[\cdot(x)]$ with $\frac{1}{n}\sum_{i=1}^n \cdot(x_i)$ in each term above, we have the second partial Fréchet derivatives of $Z_{n\theta}(\theta, h)$ and $Z_{nh}(\theta, h)$ for the remainder terms R_{nh} and $R_{n\theta}$.

(II) Regression setting

For $u, v, w \in \mathcal{H}$, $a, b, c \in \mathbb{R}^p$, we have

$$
\begin{aligned}
&D^2_{hh} Z_h(\theta, h)uvw \\
&= \mathrm{E}\left[l'''_a(Y; \eta(X; \theta, h))[D_h\eta(X; \theta, h)u][D_h\eta(X; \theta, h)v][D_h\eta(X; \theta, h)w]\right] \\
&\quad + V_\tau[D^2_{hh}\eta(\theta, h)uv, D_h\eta(\theta, h)w] + V_\tau\left[D^2_{hh}\eta(\theta, h)vw, D_h\eta(\theta, h)u\right] \\
&\quad + V_\tau\left[D^2_{hh}\eta(\theta, h)uw, D_h\eta(\theta, h)v\right] + \mathrm{E}\left[l'_a(Y; \eta(X; \theta, h))D^3_{hhh}\eta(X; \theta, h)uvs\right],
\end{aligned}
$$

$$
\begin{aligned}
&D^2_{\theta h} Z_h(\theta, h)uva = D^2_{h\theta} Z_h(\theta, h)uav = D^2_{hh} Z_\theta(\theta, h)auv \\
&= \mathrm{E}\left[l'''_a(Y; \eta(X; \theta, h))[D_h\eta(X; \theta, h)u][D_h\eta(X; \theta, h)v][D_\theta\eta(X; \theta, h)a]\right] \\
&\quad + V_\tau\left[D^2_{hh}\eta(\theta, h)uv, D_\theta\eta(\theta, h)a\right] + V_\tau\left[D^2_{\theta h}\eta(\theta, h)va, D_h\eta(\theta, h)u\right] \\
&\quad + V_\tau\left[D^2_{\theta h}\eta(\theta, h)ua, D_h\eta(\theta, h)v\right] + \mathrm{E}\left[l'_a(Y; \eta(X; \theta, h))D^3_{\theta hh}\eta(X; \theta, h)uva\right],
\end{aligned}
$$

$$
\begin{aligned}
&D^2_{\theta\theta} Z_h(\theta,h)uab = D^2_{\theta h} Z_\theta(\theta,h)aub = D^2_{h\theta} Z_\theta(\theta,h)abu \\
&= \mathrm{E}\left[l'''_a(Y;\eta(X;\theta,h))[D_h\eta(X;\theta,h)u][D_\theta\eta(X;\theta,h)a][D_\theta\eta(X;\theta,h)b]\right] \\
&\quad + V_\tau\left[D^2_{\theta h}\eta(\theta,h)ua, D_\theta\eta(\theta,h)b\right] + V_\tau\left[D^2_{\theta\theta}\eta(\theta,h)ab, D_h\eta(\theta,h)u\right] \\
&\quad + V_\tau\left[D^2_{\theta h}\eta(\theta,h)ub, D_\theta\eta(\theta,h)a\right] + \mathrm{E}\left[l'_a(Y;\eta(X;\theta,h))D^3_{\theta\theta h}\eta(X;\theta,h)abu\right],
\end{aligned}
$$

$$
\begin{aligned}
&D^2_{\theta\theta} Z_\theta(\theta,h)abc \\
&= \mathrm{E}\left[l'''_a(Y;\eta(X;\theta,h))[D_\theta\eta(X;\theta,h)a][D_\theta\eta(X;\theta,h)b][D_\theta\eta(X;\theta,h)c]\right] \\
&\quad + V_\tau\left[D^2_{\theta\theta}\eta(\theta,h)ab, D_\theta\eta(\theta,h)c\right] + V_\tau\left[D^2_{\theta\theta}\eta(\theta,h)bc, D_\theta\eta(\theta,h)a\right] \\
&\quad + V_\tau\left[D^2_{\theta\theta}\eta(\theta,h)ac, D_\theta\eta(\theta,h)b\right] + \mathrm{E}\left[l'_a(Y;\eta(X;\theta,h))D^3_{\theta\theta\theta}\eta(X;\theta,h)abc\right],
\end{aligned}
$$

By replacing all expectation $\mathrm{E}[\cdot(Y,X)]$ with $\frac{1}{n}\sum_{i=1}^n \cdot(Y_i,X_i)$ in each term above, we have the second partial Fréchet derivatives of $Z_{n\theta}(\theta,h)$ and $Z_{nh}(\theta,h)$ for the remainder terms R_{nh} and $R_{n\theta}$.

III.2 Proof of existence and uniqueness

We are now ready to show the local existence and uniqueness of $(\theta_\lambda, h_\lambda)$ and $(\hat{\theta},\hat{h})$ in the neighborhood $\mathcal{N}_{\theta_0}\times\mathcal{N}_{h_0}$. Let

$$
\begin{aligned}
d_\theta(\lambda) &= \left\|\bar{\theta}_\lambda - \theta_0\right\|_{\mathbb{R}^p}, \\
d_h(\lambda) &= \left\|\bar{h}_\lambda - h_0\right\|_{\mathcal{H}}, \\
r_\theta(\lambda) &= (K_h^3 + K_\theta^3) d_h(\lambda) + (K_h^4 + K_\theta^4) d_\theta(\lambda), \\
r_h(\lambda) &= (K_h^1 + K_\theta^1) d_h(\lambda) + (K_h^2 + K_\theta^2) d_\theta(\lambda), \\
S_{\theta,\theta_1}(\gamma) &= \{a \in \mathbb{R}^p : \|a - \theta_1\|_{\mathbb{R}^p} \le \gamma\} \text{ for } \theta_1 \in \mathbb{R}^p, \\
S_{h,h_1}(\gamma) &= \{u \in \mathcal{H} : \|g - h_1\|_{\mathcal{H}} \le \gamma\} \text{ for } h_1 \in \mathcal{H}, \\
S_\theta(\gamma) &= S_{\theta,0}(\gamma), \\
S_h(\gamma) &= S_{h,0}(\gamma).
\end{aligned}
$$

One can get the following theorem for the existence and uniqueness of $(\theta_\lambda, h_\lambda)$ via a contraction mapping argument.

Theorem 2. *If $d_\theta(\lambda) \to 0, d_h(\lambda) \to 0, r_\theta(\lambda) \to 0, r_h(\lambda) \to 0$ as $\lambda \to 0$, there exists $\lambda_0 > 0$ such that for $\lambda \in [0, \lambda_0]$, there are unique $\theta_\lambda \in S_{\theta,\theta_0}(2d_\theta(\lambda))$ and $h_\lambda \in S_{h,h_0}(2d_h(\lambda))$ satisfying $Z_\theta(\theta_\lambda, h_\lambda) = 0, Z_h(\theta_\lambda, h_\lambda) = 0$, and $(\theta_\lambda, h_\lambda) \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$. In addition, as $\lambda \to 0$,*

$$
\left\|\bar{\theta}_\lambda - \theta_\lambda\right\|_{\mathbb{R}^p} + \left\|\bar{h}_\lambda - h_\lambda\right\|_{\mathcal{H}} \le 4\left[r_h(\lambda) d_h(\lambda) + r_\theta(\lambda) d_\theta(\lambda)\right].
$$

Proof. Let $t_{\theta\lambda} = 2d_\theta(\lambda)$, $t_{h\lambda} = 2d_h(\lambda)$. Define

$$
\begin{aligned}
F_\theta(\zeta, g) = \zeta - G_{\theta\theta}^{-1}(\theta_0, h_0) \big[& Z_\theta(\theta_0 + \zeta, h_0 + g) \\
& - U_{h\theta}(\theta_0, h_0) G_h^{-1}(\theta_0, h_0) Z_h(\theta_0 + \zeta, h_0 + g) \big], \\
F_h(\zeta, g) = g - G_{hh}^{-1}(\theta_0, h_0) \big[& Z_h(\theta_0 + \zeta, h_0 + g) \\
& - U_{\theta h}(\theta_0, h_0) U_\theta^{-1}(\theta_0, h_0) Z_\theta(\theta_0 + \zeta, h_0 + g) \big].
\end{aligned}
$$

Let $\vec{F}(\zeta, g) = (F_\theta(\zeta, g), F_h(\zeta, g))$ be a function on $\mathcal{Q} = \mathbb{R}^p \times \mathcal{H}$, and for any subset $\mathcal{Q}_1 \subset \mathcal{Q}$, denote by $\vec{F}(\mathcal{Q}_1)$ the image of $\mathcal{Q}_1$ under $\vec{F}$. The proof has three steps:

1. $\vec{F}(S_\theta(t_{\theta\lambda}) \times S_h(t_{h\lambda})) \subset S_\theta(t_{\theta\lambda}) \times S_h(t_{h\lambda})$.

2. $\vec{F}$ is a contraction map on $S_\theta(t_{\theta\lambda}) \times S_h(t_{h\lambda})$.

3. Obtaining the bound for $\|\bar{\theta}_\lambda - \theta_\lambda\|_{\mathbb{R}^p} + \|\bar{h}_\lambda - h_\lambda\|_{\mathcal{H}}$.

For step 1, by our assumption, we can choose λ_0 small enough that $S_{\theta,\theta_0}(t_{\theta\lambda}) \subset \mathcal{N}_{\theta_0}$, $S_{h,h_0}(t_{h\lambda}) \subset \mathcal{N}_{h_0}$, and $r_\theta(\lambda) < 1/2$ for all $\lambda \in (0, \lambda_0]$. Recall that for every $(\theta, h) \in \mathcal{Q}$, we denote $\|(\theta, h)\|_{\mathcal{Q},1} = \|\theta\|_{\mathbb{R}^p} + \|h\|_{\mathcal{H}}$. For $(\zeta, g) \in S_\theta(t_{\theta\lambda}) \times S_h(t_{h\lambda})$, we have

$$\left\|\vec{F}(\zeta, g)\right\|_{\mathcal{Q},1} = \|F_\theta(\zeta, g)\|_{\mathbb{R}^p} + \|F_h(\zeta, g)\|_{\mathcal{H}}\,.$$

For $\|F_\theta(\zeta, g)\|_{\mathbb{R}^p}$, by the triangle inequality, we have

$$\begin{aligned}\|F_\theta(\zeta, g)\|_{\mathbb{R}^p} \leq \Big\|\zeta - G_{\theta\theta}^{-1}(\theta_0, h_0)\,[Z_\theta(\theta_0 + \zeta, h_0 + g) \\ - U_{h\theta}(\theta_0, h_0) G_h^{-1}(\theta_0, h_0) Z_h(\theta_0 + \zeta, h_0 + g)] - (\bar{\theta}_\lambda - \theta_0)\Big\|_{\mathbb{R}^p} \\ + \left\|\bar{\theta}_\lambda - \theta_0\right\|_{\mathbb{R}^p}.\end{aligned}$$

By the definition of $\bar{\theta}_\lambda - \theta_0$ and $G_{\theta\theta}(\theta, h)$, the Taylor series expansions of $Z_\theta(\theta_0 + \zeta, h_0 + g)$ and $Z_h(\theta_0 + \zeta, h_0 + g)$, and the remainder bound (III.4), we get

$$
\begin{aligned}
&\big\|\zeta - G_{\theta\theta}^{-1}(\theta_0,h_0)\left[Z_\theta(\theta_0+\zeta,h_0+g)\right.\\
&\qquad \left.-U_{h\theta}(\theta_0,h_0)G_h^{-1}(\theta_0,h_0)Z_h(\theta_0+\zeta,h_0+g)\right] - (\bar{\theta}_\lambda-\theta_0)\big\|_{\mathbb{R}^p}\\
&= \big\|\zeta - G_{\theta\theta}^{-1}(\theta_0,h_0)\left\{\left[Z_\theta(\theta_0+\zeta,h_0+g) - Z_\theta(\theta_0,h_0)\right]\right.\\
&\qquad \left.-U_{h\theta}(\theta_0,h_0)G_h^{-1}(\theta_0,h_0)\left[Z_h(\theta_0+\zeta,h_0+g) - Z_h(\theta_0,h_0)\right]\right\}\big\|_{\mathbb{R}^p}\\
&= \big\|\zeta - G_{\theta\theta}^{-1}(\theta_0,h_0)\left\{\left[U_\theta(\theta_0,h_0)\zeta + R_\theta(\theta_0,h_0)\zeta g\right]\right.\\
&\qquad \left.-U_{h\theta}(\theta_0,h_0)G_h^{-1}(\theta_0,h_0)\left[U_{\theta h}(\theta_0,h_0)\zeta + R_h(\theta_0,h_0)\zeta g\right]\right\}\big\|_{\mathbb{R}^p}\\
&= \big\|\zeta - G_{\theta\theta}^{-1}(\theta_0,h_0)\left\{\left[U_\theta(\theta_0,h_0) - U_{h\theta}(\theta_0,h_0)G_h^{-1}(\theta_0,h_0)U_{\theta h}(\theta_0,h_0)\right]\zeta\right.\\
&\qquad \left.+\left[R_\theta(\theta_0,h_0) - U_{h\theta}(\theta_0,h_0)G_h^{-1}(\theta_0,h_0)R_h(\theta_0,h_0)\right]\zeta g\right\}\big\|_{\mathbb{R}^p}\\
&= \big\|G_{\theta\theta}^{-1}(\theta_0,h_0)\left[R_\theta(\theta_0,h_0)\zeta g - U_{h\theta}(\theta_0,h_0)G_h^{-1}(\theta_0,h_0)R_h(\theta_0,h_0)\zeta g\right]\big\|_{\mathbb{R}^p}\\
&\le \frac{1}{2}\left(K_\theta^1\|g\|_{\mathcal{H}} + K_\theta^2\|\zeta\|_{\mathbb{R}^p}\right)\|g\|_{\mathcal{H}} + \frac{1}{2}\left(K_\theta^3\|g\|_{\mathcal{H}} + K_\theta^4\|\zeta\|_{\mathbb{R}^p}\right)\|\zeta\|_{\mathbb{R}^p}.
\end{aligned}
$$

Similarly, by the definition of $\bar{h}_\lambda - h_0$ and $G_{hh}(\theta,h)$, the Taylor series expansions of $Z_\theta(\theta_0+\zeta,h_0+g)$ and $Z_h(\theta_0+\zeta,h_0+g)$, and the remainder bound (III.3), we also have

$$
\begin{aligned}
\|F_h(\zeta,g)\|_{\mathcal{H}} \le \big\|&g - G_{hh}^{-1}(\theta_0,h_0)\left[Z_h(\theta_0+\zeta,h_0+g)\right.\\
&\left.-U_{\theta h}(\theta_0,h_0)U_\theta^{-1}(\theta_0,h_0)Z_\theta(\theta_0+\zeta,h_0+g)\right] - (\bar{h}_\lambda-h_0)\big\|_{\mathcal{H}}\\
&+\big\|\bar{h}_\lambda - h_0\big\|_{\mathcal{H}},
\end{aligned}
$$

and

$$
\begin{aligned}
&\big\|g - G_{hh}^{-1}(\theta_0,h_0)\left[Z_h(\theta_0+\zeta,h_0+g)\right.\\
&\qquad \left.-U_{\theta h}(\theta_0,h_0)U_\theta^{-1}(\theta_0,h_0)Z_\theta(\theta_0+\zeta,h_0+g)\right] - (\bar{h}_\lambda-h_0)\big\|_{\mathcal{H}}\\
&= \big\|G_{hh}^{-1}(\theta_0,h_0)\left[R_h(\theta_0,h_0)\zeta g - U_{\theta h}(\theta_0,h_0)U_\theta^{-1}(\theta_0,h_0)R_\theta(\theta_0,h_0)\zeta g\right]\big\|_{\mathcal{H}}\\
&\le \frac{1}{2}\left(K_h^1\|g\|_{\mathcal{H}} + K_h^2\|\zeta\|_{\mathbb{R}^p}\right)\|g\|_{\mathcal{H}} + \frac{1}{2}\left(K_h^3\|g\|_{\mathcal{H}} + K_h^4\|\zeta\|_{\mathbb{R}^p}\right)\|\zeta\|_{\mathbb{R}^p}.
\end{aligned}
$$

Since $t_{\theta\lambda} = 2\left\|\bar{\theta}_\lambda - \theta_0\right\|_{\mathbb{R}^p}$, $t_{h\lambda} = 2\left\|\bar{h}_\lambda - h_0\right\|_{\mathcal{H}}$, $r_h(\lambda) < 1/2$, and $r_\theta(\lambda) < 1/2$, for $(\zeta, g) \in S_\theta(t_{\theta\lambda}) \times S_h(t_{h\lambda})$, we have

$$\begin{aligned}
\left\|\vec{F}(\zeta, g)\right\|_{\mathcal{Q},1} &\leq \frac{1}{2}\left[\left(K_h^1 + K_\theta^1\right)\|g\|_{\mathcal{H}} + \left(K_h^2 + K_\theta^2\right)\|\zeta\|_{\mathbb{R}^p}\right]\|g\|_{\mathcal{H}} \\
&\quad + \frac{1}{2}\left[\left(K_h^3 + K_\theta^3\right)\|g\|_{\mathcal{H}} + \left(K_h^4 + K_\theta^4\right)\|\zeta\|_{\mathbb{R}^p}\right]\|\zeta\|_{\mathbb{R}^p} \\
&\quad + \left\|\bar{\theta}_\lambda - \theta_0\right\|_{\mathbb{R}^p} + \left\|\bar{h}_\lambda - h_0\right\|_{\mathcal{H}} \\
&\leq \frac{1}{2}\left[\left(K_h^1 + K_\theta^1\right)t_{h\lambda} + \left(K_h^2 + K_\theta^2\right)t_{\theta\lambda}\right]t_{h\lambda} \\
&\quad + \frac{1}{2}\left[\left(K_h^3 + K_\theta^3\right)t_{h\lambda} + \left(K_h^4 + K_\theta^4\right)t_{\theta\lambda}\right]t_{\theta\lambda} \\
&\quad + \frac{1}{2}t_{\theta\lambda} + \frac{1}{2}t_{h\lambda} \\
&= r_h(\lambda)t_{h\lambda} + r_\theta(\lambda)t_{\theta\lambda} + \frac{1}{2}t_{\theta\lambda} + \frac{1}{2}t_{h\lambda} \\
&= \left(r_h(\lambda) + \frac{1}{2}\right)t_{h\lambda} + \left(r_\theta(\lambda) + \frac{1}{2}\right)t_{\theta\lambda} \\
&< t_{h\alpha} + t_{\theta\alpha}.
\end{aligned}$$

Now for step 2, by Taylor expansion, we get that for $\zeta_1, \zeta_2 \in S_\theta(t_{\theta\lambda})$, $g_1, g_2 \in S_h(t_{h\lambda})$,

$$\begin{aligned}
&Z_\theta(\theta_0 + \zeta_2, h_0 + g_2) = Z_\theta(\theta_0 + \zeta_1, h_0 + g_1) \\
&\quad + \int_0^1 D_\theta Z_\theta\left[\theta_0 + \zeta_1 + t(\zeta_2 - \zeta_1), h_0 + g_1 + t(g_2 - g_1)\right](\zeta_2 - \zeta_1) \\
&\qquad + D_h Z_\theta\left[\theta_0 + \zeta_1 + t(\zeta_2 - \zeta_1), h_0 + g_1 + t(g_2 - g_1)\right](g_2 - g_1)\, dt.
\end{aligned}$$

Applying Taylor expansion again to the terms inside the integral and letting $\zeta^* = \zeta_1 + t(\zeta_2 - \zeta_1)$, $g^* = g_1 + t(g_2 - g_1)$, we have

$$
\begin{aligned}
&Z_\theta(\theta_0+\zeta_2,h_0+g_2)-Z_\theta(\theta_0+\zeta_1,h_0+g_1)\\
&\quad=U_\theta(\theta_0,h_0)(\zeta_2-\zeta_1)+U_{h\theta}(\theta_0,h_0)(g_2-g_1)\\
&\qquad+\int_0^1\int_0^1\big[D^2_{\theta\theta}Z_\theta(\theta_0+t'\zeta^*,h_0+t'g^*)\zeta^*\\
&\qquad\qquad+D^2_{h\theta}Z_\theta(\theta_0+t'\zeta^*,h_0+t'g^*)g^*\big](\zeta_2-\zeta_1)\,dt'dt\\
&\qquad+\int_0^1\int_0^1\big[D^2_{\theta h}Z_\theta(\theta_0+t'\zeta^*,h_0+t'g^*)\zeta^*\\
&\qquad\qquad+D^2_{hh}Z_\theta(\theta_0+t'\zeta^*,h_0+t'g^*)g^*\big](g_2-g_1)\,dt'dt.
\end{aligned}
$$

Note that for $0\le t\le 1$, $\zeta^*=\zeta_1+t(\zeta_2-\zeta_1)\in S_\theta(t_{\theta\lambda})$, $g^*=g_1+t(g_2-g_1)\in S_h(t_{h\lambda})$ by convexity of $S_\theta(t_{\theta\lambda})$ and $S_h(t_{h\lambda})$. Since

$$
\begin{aligned}
&F_\theta(\zeta_1,g_1)-F_\theta(\zeta_2,g_2)\\
&=(\zeta_1-\zeta_2)-G^{-1}_{\theta\theta}(\theta_0,h_0)\left\{[Z_\theta(\theta_0+\zeta_1,h_0+g_1)-Z_\theta(\theta_0+\zeta_2,h_0+g_2)]\right.\\
&\qquad\left.-U_{h\theta}(\theta_0,h_0)G_h^{-1}(\theta_0,h_0)\left[Z_h(\theta_0+\zeta_1,h_0+g_1)-Z_h(\theta_0+\zeta_2,h_0+g_2)\right]\right\},
\end{aligned}
$$

similar algebraic manipulations as in the proof of step 1 show that

$$
\begin{aligned}
\|F_\theta(\zeta_1,g_1)-F_\theta(\zeta_2,g_2)\|_{\mathbb{R}^p}&\le\left(K^3_\theta\|g^*\|_{\mathcal{H}}+K^4_\theta\|\zeta^*\|_{\mathbb{R}^p}\right)\|\zeta_2-\zeta_1\|_{\mathbb{R}^p}\\
&\quad+\left(K^2_\theta\|\zeta^*\|_{\mathbb{R}^p}+K^1_\theta\|g^*\|_{\mathcal{H}}\right)\|g_2-g_1\|_{\mathcal{H}}.
\end{aligned}
$$

Similarly for F_h, we get

$$
\begin{aligned}
\|F_h(\zeta_1,g_1)-F_h(\zeta_2,g_2)\|_{\mathcal{H}}&\le\left(K^3_h\|g^*\|_{\mathcal{H}}+K^4_h\|\zeta^*\|_{\mathbb{R}^p}\right)\|\zeta_2-\zeta_1\|_{\mathbb{R}^p}\\
&\quad+\left(K^2_h\|\zeta^*\|_{\mathbb{R}^p}+K^1_h\|g^*\|_{\mathcal{H}}\right)\|g_2-g_1\|_{\mathcal{H}}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\left\|\vec{F}(\zeta_1, g_1) - \vec{F}(\zeta_2, g_2)\right\|_{\mathcal{Q},1} \\
&\quad \leq (K_h^1 + K_\theta^1) \|g^*\|_{\mathcal{H}} \|g_2 - g_1\|_{\mathcal{H}} + (K_h^2 + K_\theta^2) \|\zeta^*\|_{\mathbb{R}^p} \|g_2 - g_1\|_{\mathcal{H}} \\
&\qquad + (K_h^3 + K_\theta^3) \|g^*\|_{\mathcal{H}} \|\zeta_2 - \zeta_1\|_{\mathbb{R}^p} + (K_h^4 + K_\theta^4) \|\zeta^*\|_{\mathbb{R}^p} \|\zeta_2 - \zeta_1\|_{\mathbb{R}^p} \\
&\quad = 2\left[(K_h^1 + K_\theta^1) d_h(\lambda) + (K_h^2 + K_\theta^2) d_\theta(\lambda)\right] \|g_2 - g_1\|_{\mathcal{H}} \\
&\qquad + 2\left[(K_h^3 + K_\theta^3) d_h(\lambda) + (K_h^4 + K_\theta^4) d_\theta(\lambda)\right] \|\zeta_2 - \zeta_1\|_{\mathbb{R}^p} \\
&\quad = 2 r_h(\lambda) \|g_2 - g_1\|_{\mathcal{H}} + 2 r_\theta(\lambda) \|\zeta_2 - \zeta_1\|_{\mathbb{R}^p} \\
&\quad \leq C_1 \|g_2 - g_1\|_{\mathcal{H}} + C_2 \|\zeta_2 - \zeta_1\|_{\mathbb{R}^p},
\end{aligned}
$$

where $0 < C_1 < 1$, $0 < C_2 < 1$, so $\vec{F} = (F_\theta, F_h)$ is a contraction map on $S_\theta(t_{\theta\lambda}) \times S_h(t_{h\lambda})$. By the contraction mapping theorem (Theorem 9.23 in Rudin (1976)), there exists a unique $(\zeta_\lambda, g_\lambda) \in S_\theta(t_{\theta\lambda}) \times S_h(t_{h\lambda})$ such that $\vec{F}(\zeta_\lambda, g_\lambda) = (\zeta_\lambda, g_\lambda)$. Let $\theta_\lambda = \theta_0 + \zeta_\lambda$, $h_\lambda = h_0 + g_\lambda$. Then $\theta_\lambda \in S_{\theta,\theta_0}(t_{\theta\lambda})$, $h_\lambda \in S_{h,h_0}(t_{\theta\lambda})$, and $(\theta_\lambda, h_\lambda)$ are the unique solutions to $Z_\theta(\theta_\lambda, h_\lambda) = 0$, $Z_h(\theta_\lambda, h_\lambda) = 0$.

For step 3, note that

$$
\begin{aligned}
(\bar{\theta}_\lambda - \theta_\lambda, \bar{h}_\lambda - h_\lambda) &= (\bar{\theta}_\lambda - \theta_0, \bar{h}_\lambda - h_0) - (\theta_\lambda - \theta_0, h_\lambda - h_0) \\
&= \vec{F}(0,0) - \vec{F}(\zeta_\lambda, g_\lambda).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\left\|\bar{\theta}_\lambda - \theta_\lambda\right\|_{\mathbb{R}^p} + \left\|\bar{h}_\lambda - h_\lambda\right\|_{\mathcal{H}} &= \left\|\vec{F}(\zeta_\lambda, g_\lambda) - \vec{F}(0,0)\right\|_{\mathbb{R}^p \times \mathcal{H}} \\
&\leq 2 r_h(\lambda) \|g_\lambda\|_{\mathcal{H}} + 2 r_\theta(\lambda) \|\zeta_\lambda\|_{\mathbb{R}^p} \\
&\leq 4\left[r_h(\lambda) d_h(\lambda) + r_\theta(\lambda) d_\theta(\lambda)\right].
\end{aligned}
$$

This completes the proof of Theorem 2. □

Next, we consider the existence of $(\hat{\theta}, \hat{h}) \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$. Define

$$
\begin{aligned}
d_{n\theta}(\lambda) &= \left\| \bar{\theta}_{n\lambda} - \theta_\lambda \right\|_{\mathbb{R}^p}, \\
d_{nh}(\lambda) &= \left\| \bar{h}_{n\lambda} - h_\lambda \right\|_{\mathcal{H}}, \\
r_{n\theta}(\lambda) &= E^2_{n\theta} + E^2_{nh} + (K^3_{n\theta} + K^3_{nh}) d_{nh}(\lambda) + (K^4_{n\theta} + K^4_{nh}) d_{n\theta}(\lambda), \\
r_{nh}(\lambda) &= E^1_{n\theta} + E^1_{nh} + (K^1_{n\theta} + K^1_{nh}) d_{nh}(\lambda) + (K^2_{n\theta} + K^2_{nh}) d_{n\theta}(\lambda).
\end{aligned}
$$

We get the following existence theorem for $(\hat{\theta}, \hat{h}) \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$.

Theorem 3. *Suppose λ_n is a sequence such that for all n sufficiently large, $\theta_{\lambda_n} \in \mathcal{N}_{\theta_0}$, $h_{\lambda_n} \in \mathcal{N}_{h_0}$, and*

$$
\begin{aligned}
d_{n\theta}(\lambda_n) &\xrightarrow{P} 0, & d_{nh}(\lambda_n) &\xrightarrow{P} 0, \\
r_{n\theta}(\lambda_n) &\xrightarrow{P} 0, & r_{nh}(\lambda_n) &\xrightarrow{P} 0.
\end{aligned}
$$

Then, with probability tending to unity as $n \to \infty$, there is a unique root $(\hat{\theta}, \hat{h})$ of $Z_{n\theta}(\hat{\theta}, \hat{h}) = 0$, $Z_{nh}(\hat{\theta}, \hat{h}) = 0$ in $S_{\theta, \theta_{\lambda_n}}(2 d_{n\theta}(\lambda_n)) \times S_{h, h_{\lambda_n}}(2 d_{nh}(\lambda_n)) \subset \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$. In addition, as $n \to \infty$ and $\lambda_n \to 0$,

$$
\left\| \hat{\theta} - \bar{\theta}_{n\lambda_n} \right\|_{\mathbb{R}^p} + \left\| \hat{h} - \bar{h}_{n\lambda_n} \right\|_{\mathcal{H}} \leq 4 r_{n\theta}(\lambda_n) d_{n\theta}(\lambda_n) + 4 r_{nh}(\lambda_n) d_{nh}(\lambda_n).
$$

Proof. For convenience, we drop the subscript on λ_n and let $t_{n\theta\lambda} = 2 d_{n\theta}(\lambda)$, $t_{nh\lambda} = 2 d_{nh}(\lambda)$. Let

$$
\begin{aligned}
F_{n\theta}(\zeta, g) = \zeta - G^{-1}_{\theta\theta}(\theta_\lambda, h_\lambda) \big[& Z_{n\theta}(\theta_\lambda + \zeta, h_\lambda + g) \\
& - U_{h\theta}(\theta_\lambda, h_\lambda) G^{-1}_h(\theta_\lambda, h_\lambda) Z_{nh}(\theta_\lambda + \zeta, h_\lambda + g) \big], \\
F_{nh}(\zeta, g) = g - G^{-1}_{hh}(\theta_\lambda, h_\lambda) \big[& Z_{nh}(\theta_\lambda + \zeta, h_\lambda + g) \\
& - U_{\theta h}(\theta_\lambda, h_\lambda) U^{-1}_\theta(\theta_\lambda, h_\lambda) Z_{n\theta}(\theta_\lambda + \zeta, h_\lambda + g) \big].
\end{aligned}
$$

The proof proceeds in three steps, similar to the proof of Theorem 2, with additional terms introduced in approximating $D_\theta Z_{n\theta}$ and $D_h Z_{nh}$ by $D_\theta Z_\theta$ and $D_h Z_h$, respectively. Take n large enough so that $S_{\theta,\theta_\lambda}(t_{n\theta\lambda}) \subset \mathcal{N}_{\theta_0}$, $S_{h,h_\lambda}(t_{nh\lambda}) \subset \mathcal{N}_{h_0}$ and $r_{n\theta}(\lambda) < \frac{1}{2}$, $r_{nh}(\lambda) < \frac{1}{2}$.

First, we show that $\vec{F}_n(\zeta, g) = (F_{n\theta}(\zeta, g), F_{nh}(\zeta, g))$ maps $S_\theta(t_{n\theta\lambda}) \times S_h(t_{nh\lambda})$ to itself, i.e., $\vec{F}_n\left(S_\theta(t_{n\theta\lambda}) \times S_h(t_{nh\lambda})\right) \subset S_\theta(t_{n\theta\lambda}) \times S_h(t_{nh\lambda})$. By definition, for $(\zeta, g) \in S_\theta(t_{n\theta\lambda}) \times S_h(t_{nh\lambda})$, we have

$$\left\|\vec{F}_n(\zeta, g)\right\|_{\mathcal{Q},1} = \|F_{n\theta}(\zeta, g)\|_{\mathbb{R}^p} + \|F_{nh}(\zeta, g)\|_{\mathcal{H}}\,.$$

For $F_{n\theta}$, by the triangle inequality, we get

$$\begin{aligned}\|F_{n\theta}(\zeta, g)\|_{\mathbb{R}^p} \leq \big\|\zeta - G_{\theta\theta}^{-1}(\theta_\lambda, h_\lambda)\,[&Z_{n\theta}(\theta_\lambda + \zeta, h_\lambda + g) \\ &- U_{h\theta}(\theta_\lambda, h_\lambda) G_h^{-1}(\theta_\lambda, h_\lambda) Z_{nh}(\theta_\lambda + \zeta, h_\lambda + g)\big] \\ &- (\bar{\theta}_{n\lambda} - \theta_\lambda)\big\|_{\mathbb{R}^p} + \big\|\bar{\theta}_{n\lambda} - \theta_\lambda\big\|_{\mathbb{R}^p}\end{aligned}$$

Using the definition of $\bar{\theta}_{n\lambda} - \theta_\lambda$, $G_{\theta\theta}(\theta, h)$, the Taylor expansions of $Z_{n\theta}(\theta_\lambda + \zeta, h_\lambda + g)$ and $Z_{nh}(\theta_\lambda + \zeta, h_\lambda + g)$, and the remainder bound (III.6), we get

$$
\begin{aligned}
&\big\|\zeta - G_{\theta\theta}^{-1}((\theta_\lambda, h_\lambda)\left[Z_{n\theta}(\theta_\lambda+\zeta, h_\lambda+g)\right.\\
&\qquad \left.-U_{h\theta}(\theta_\lambda, h_\lambda)G_h^{-1}(\theta_\lambda, h_\lambda)Z_{nh}(\theta_\lambda+\zeta, h_\lambda+g)\right] - (\bar{\theta}_{n\lambda}-\theta_\lambda)\big\|_{\mathbb{R}^p}\\
&= \big\|\zeta - G_{\theta\theta}^{-1}(\theta_\lambda, h_\lambda)\left\{\left[Z_{n\theta}(\theta_\lambda+\zeta, h_\lambda+g) - Z_{n\theta}(\theta_\lambda, h_\lambda)\right]\right.\\
&\qquad \left.-U_{h\theta}(\theta_\lambda, h_\lambda)G_h^{-1}(\theta_\lambda, h_\lambda)\left[Z_{nh}(\theta_\lambda+\zeta, h_\lambda+g) - Z_{nh}(\theta_\lambda, h_\lambda)\right]\right\}\big\|_{\mathbb{R}^p}\\
&= \big\|\zeta - G_{\theta\theta}^{-1}(\theta_\lambda, h_\lambda)\left[U_\theta(\theta_\lambda, h_\lambda) - (U_{h\theta}(\theta_\lambda, h_\lambda)G_h^{-1}(\theta_\lambda, h_\lambda)U_{h\theta}(\theta_\lambda, h_\lambda)\right]\zeta\\
&\qquad - G_{\theta\theta}^{-1}(\theta_\lambda, h_\lambda)\left[e_\theta(\theta_\lambda, h_\lambda) + R_{n\theta}(\theta_\lambda, h_\lambda)\right]\zeta g\\
&\qquad +G_{\theta\theta}^{-1}(\theta_\lambda, h_\lambda)U_{h\theta}(\theta_\lambda, h_\lambda)G_h^{-1}(\theta_\lambda, h_\lambda)\left[e_h(\theta_\lambda, h_\lambda) + R_{nh}(\theta_\lambda, h_\lambda)\right]\zeta g\big\|_{\mathbb{R}^p}\\
&= \big\|G_{\theta\theta}^{-1}(\theta_\lambda, h_\lambda)\left\{\left[e_\theta(\theta_\lambda, h_\lambda) + R_{n\theta}(\theta_\lambda, h_\lambda)\right]\zeta g\right.\\
&\qquad \left.-U_{h\theta}(\theta_\lambda, h_\lambda)G_h^{-1}(\theta_\lambda, h_\lambda)\left[e_h(\theta_\lambda, h_\lambda) + R_{nh}(\theta_\lambda, h_\lambda)\right]\zeta g\right\}\big\|_{\mathbb{R}^p}\\
&\le E_{n\theta}^1\|g\|_{\mathcal{H}} + E_{n\theta}^2\|\zeta\|_{\mathbb{R}^p} + \frac{1}{2}\left(K_{n\theta}^1\|g\|_{\mathcal{H}} + K_{n\theta}^2\|\zeta\|_{\mathbb{R}^p}\right)\|g\|_{\mathcal{H}}\\
&\qquad + \frac{1}{2}\left(K_{n\theta}^3\|g\|_{\mathcal{H}} + K_{n\theta}^4\|\zeta\|_{\mathbb{R}^p}\right)\|\zeta\|_{\mathbb{R}^p}.
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
\|F_{nh}(\zeta, g)\|_{\mathcal{H}} \le \|g - G_{hh}^{-1}(\theta_\lambda, h_\lambda)&\left[Z_h(\theta_\lambda+\zeta, h_\lambda+g)\right.\\
&\left.- U_{\theta h}(\theta_\lambda, h_\lambda)U_\theta^{-1}(\theta_\lambda, h_\lambda)Z_\theta(\theta_\lambda+\zeta, h_\lambda+g)\right]\\
&- (\bar{h}_{n\lambda}-h_\lambda)\|_{\mathcal{H}} + \big\|\bar{h}_{n\lambda}-h_\lambda\big\|_{\mathcal{H}},
\end{aligned}
$$

and

$$
\begin{aligned}
&\|g - G_{hh}^{-1}(\theta_\lambda, h_\lambda)\left[Z_h(\theta_\lambda+\zeta, h_\lambda+g)\right.\\
&\quad \left.-U_{\theta h}(\theta_\lambda, h_\lambda)U_\theta^{-1}(\theta_\lambda, h_\lambda)Z_\theta(\theta_\lambda+\zeta, h_\lambda+g)\right] - (\bar{h}_{n\lambda}-h_\lambda)\|_{\mathcal{H}}\\
&\le E_{nh}^1\|g\|_{\mathcal{H}} + E_{nh}^2\|\zeta\|_{\mathbb{R}^p} + \frac{1}{2}(K_{nh}^1\|g\|_{\mathcal{H}} + K_{nh}^2\|\zeta\|_{\mathbb{R}^p})\|g\|_{\mathcal{H}}\\
&\quad + \frac{1}{2}(K_{nh}^3\|g\|_{\mathcal{H}} + K_{nh}^4\|\zeta\|_{\mathbb{R}^p})\|\zeta\|_{\mathbb{R}^p}.
\end{aligned}
$$

Thus, for $(\zeta, g) \in S_\theta(t_{n\theta\lambda}) \times S_h(t_{nh\lambda})$,

$$
\begin{aligned}
&\left\| \vec{F}(\zeta, g) \right\|_{\mathcal{Q},1} \\
&\leq \left[(E^1_{n\theta} + E^1_{nh}) + \frac{1}{2}(K^1_{n\theta} + K^1_{nh}) \|g\|_{\mathcal{H}} + \frac{1}{2}(K^2_{n\theta} + K^2_{nh}) \|\zeta\|_{\mathbb{R}^p} \right] \|g\|_{\mathcal{H}} \\
&\quad + \left[(E^2_{n\theta} + E^2_{nh}) + \frac{1}{2}(K^3_{n\theta} + K^3_{nh}) \|g\|_{\mathcal{H}} + \frac{1}{2}(K^4_{n\theta} + K^4_{nh}) \|\zeta\|_{\mathbb{R}^p} \right] \|\zeta\|_{\mathbb{R}^p} \\
&\quad + \left\| \bar{\theta}_{n\lambda} - \theta_\lambda \right\|_{\mathbb{R}^p} + \left\| \bar{h}_{n\lambda} - h_\lambda \right\|_{\mathcal{H}} \\
&\leq r_{nh}(\lambda) t_{nh\lambda} + r_{n\theta}(\lambda) t_{n\theta\lambda} + \frac{1}{2} t_{n\theta\lambda} + \frac{1}{2} t_{nh\lambda} \\
&= \left[r_{nh}(\lambda) + \frac{1}{2} \right] t_{nh\lambda} + \left[r_{n\theta}(\lambda) + \frac{1}{2} \right] t_{n\theta\lambda} \\
&< t_{nh\lambda} + t_{n\theta\lambda}.
\end{aligned}
$$

Therefore, we have shown that $\vec{F}_n(S_\theta(t_{n\theta\lambda}) \times S_h(t_{nh\lambda})) \subset S_\theta(t_{n\theta\lambda}) \times S_h(t_{nh\lambda})$.

Next, we show that $\vec{F}_n$ is a contraction map. By similar calculations as in the proof for Theorem 2, after applying Taylor expansion twice, for $\zeta_1, \zeta_2 \in S_\theta(t_{n\theta\lambda})$, $g_1, g_2 \in S_h(t_{nh\lambda})$, we get

$$
\begin{aligned}
\|F_{n\theta}(\zeta_1, g_1) - F_{n\theta}(\zeta_2, g_2)\|_{\mathbb{R}^p} \leq &\left(E^2_{n\theta} + K^3_{n\theta} t_{nh\lambda} + K^4_{n\theta} t_{n\theta\lambda} \right) \|\zeta_1 - \zeta_2\|_{\mathbb{R}^p} \\
&+ \left(E^1_{n\theta} + K^2_{n\theta} t_{n\theta\lambda} + K^1_{n\theta} t_{nh\lambda} \right) \|g_1 - g_2\|_{\mathcal{H}},
\end{aligned}
$$

$$
\begin{aligned}
\|F_{nh}(\zeta_1, g_1) - F_{nh}(\zeta_2, g_2)\|_{\mathcal{H}} \leq &\left(E^2_{nh} + K^3_{nh} t_{nh\lambda} + K^4_{nh} t_{n\theta\lambda} \right) \|\zeta_1 - \zeta_2\|_{\mathbb{R}^p} \\
&+ \left(E^1_{nh} + K^2_{nh} t_{n\theta\lambda} + K^1_{nh} t_{nh\lambda} \right) \|g_1 - g_2\|_{\mathcal{H}}.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\left\|\vec{F}_n(\zeta_1, g_1) - \vec{F}_n(\zeta_2, g_2)\right\|_{\mathcal{Q},1} \\
&\quad \le \left[E^2_{n\theta} + E^2_{nh} + (K^3_{n\theta} + K^3_{nh})t_{nh\lambda} + (K^4_{n\theta} + K^4_{nh})t_{n\theta\lambda}\right] \|\zeta_1 - \zeta_2\|_{\mathbb{R}^p} \\
&\qquad + \left[E^1_{n\theta} + E^1_{nh} + (K^1_{n\theta} + K^1_{nh})t_{nh\lambda} + (K^2_{n\theta} + K^2_{nh})t_{n\theta\lambda}\right] \|g_1 - g_2\|_{\mathcal{H}} \\
&\quad \le 2r_{n\theta}(\lambda) \|\zeta_1 - \zeta_2\|_{\mathbb{R}^p} + 2r_{nh}(\lambda) \|g_1 - g_2\|_{\mathcal{H}} .
\end{aligned}
$$

Since $r_{n\lambda}(\lambda) < \frac{1}{2}$, $r_{nh}(\lambda) < \frac{1}{2}$, we have shown that $\vec{F}(\zeta, g)$ is a contraction map on $S_\theta(t_{n\theta\lambda}) \times S_h(t_{nh\lambda})$. By the contraction mapping theorem, there exists a unique $(\zeta_{n\lambda}, g_{n\lambda}) \in S_\theta(t_{n\theta\lambda}, \alpha) \times S_h(t_{nh\lambda}, \alpha)$ such that $\vec{F}_n(\zeta_{n\lambda}, g_{n\lambda}) = (\zeta_{n\lambda}, g_{n\lambda})$. Let $\hat{\theta} = \theta_\lambda + \zeta_{n\lambda} \in S_{\theta,\theta_\lambda}(2d_{n\theta}(\lambda))$ and $\hat{h} = h_\lambda + g_{n\lambda} \in S_{h,h_\lambda}(2d_{nh}(\lambda))$. Then $(\hat{\theta}, \hat{h})$ is the unique root of $Z_{n\theta}(\hat{\theta}, \hat{h}) = 0$ and $Z_{nh}(\hat{\theta}, \hat{h}) = 0$.

To get the upper bound, we observe that

$$
\begin{aligned}
(\bar{\theta}_{n\lambda} - \hat{\theta}, \bar{h}_{n\lambda} - \hat{h}) &= (\bar{\theta}_{n\lambda} - \theta_\lambda, \bar{h}_{n\lambda} - h_\lambda) - (\hat{\theta} - \theta_\lambda, \hat{h} - h_\lambda) \\
&= \vec{F}_n(0, 0) - \vec{F}_n(\zeta_{n\lambda}, g_{n\lambda}).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\left\|\bar{\theta}_{n\lambda} - \theta_{n\lambda}\right\|_{\mathbb{R}^p} + \left\|\bar{h}_{n\lambda} - h_{n\lambda}\right\|_{\mathcal{H}} &= \left\|\vec{F}_n(\zeta_{n\lambda}, g_{n\lambda}) - \vec{F}_n(0, 0)\right\|_{\mathcal{Q},1} \\
&\le 2r_{n\theta}(\lambda) \|\zeta_{n\lambda}\|_{\mathbb{R}^p} + 2r_{nh}(\lambda) \|g_{n\lambda}\|_{\mathcal{H}} \\
&\le 4 \left[r_{n\theta}(\lambda) d_{n\theta}(\lambda) + r_{nh}(\lambda) d_{nh}(\lambda)\right] .
\end{aligned}
$$

This completes the proof of Theorem 2. □

CHAPTER IV

DENSITY MODELS

In this chapter, under Assumptions 2 to 9, we develop a joint asymptotic theory for the penalized likelihood estimators for the density estimation model.

Recall that for an i.i.d. random sample $X_1, \ldots, X_n$ with a common probability density $f(x)$ on a bounded domain $\mathcal{X} \subset \mathbb{R}^d$, we consider the following general semiparametric density model

$$f(x;\theta,h) = \frac{\exp\{\eta(x;\theta,h)\}}{\int_{\mathcal{X}} \exp\{\eta(x;\theta,h)\}dx},$$

where $\eta : \mathcal{Q} = \mathbb{R}^p \times \mathcal{H} \to L_0^2(\mathcal{X})$ is a known function that is one-to-one in a neighborhood $\mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$ of the true parameter. We consider the semiparametric penalized likelihood estimator

$$(\hat{\theta}, \hat{h}) = \underset{(\theta,h)\in\mathcal{N}_{\theta_0}\times\mathcal{N}_{h_0}}{\arg\min} \ell_{n,\lambda}(\theta,h),$$

where

$$\begin{aligned} \ell_{n,\lambda}(\theta,h \mid \text{data}) &= \ell_n(\eta(\theta,h) \mid \text{data}) + \frac{\lambda}{2}J(h) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\eta(X_i;\theta,h) + \log\int_{\mathcal{X}} e^{\eta(x;\theta,h)}dx + \frac{\lambda}{2}J(h). \end{aligned} \tag{IV.1}$$

Assuming the local existence and uniqueness of $\hat{\tau} \equiv (\hat{\theta}, \hat{h})$, we study the consistency of this estimator and obtain the rate of convergence of $(\hat{\theta}, \hat{h})$ to the true parameters (θ_0, h_0) as $n \to \infty$ and $\lambda \to 0$.

IV.1 Outline of the proof of consistency

It is easy to see that the estimator $\hat{\tau} = (\hat{\theta}, \hat{h})$ is taken to be the unique solution of the system $D_\theta \ell_{n,\lambda}(\theta, h) = 0$, $D_h \ell_{n,\lambda}(\theta, h) = 0$ in $\mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$. To study the asymptotic behavior of the estimator $\hat{\tau} = (\hat{\theta}, \hat{h})$, we first introduce an approximation of $\hat{\tau}$, denoted $\tilde{\tau} \equiv (\tilde{\theta}, \tilde{h})$, and establish the rate of convergence for $\tilde{\tau} - \tau_0$. Let $\tilde{\tau}$ minimize $\tilde{\ell}_{n,\lambda}(\theta, h) \equiv \tilde{\ell}_n(\theta, h) + \frac{\lambda}{2} J(h)$, where

$$\begin{aligned}\tilde{\ell}_n(\theta, h) \equiv -\frac{1}{n} \sum_{i=1}^{n} [D_\theta \eta(X_i; \tau_0)\theta + D_h \eta(X_i; \tau_0) h] + \mu_{\tau_0}[D_\theta \eta(\tau_0)\theta] \\ + \mu_{\tau_0}[D_h \eta(\tau_0) h] + \frac{1}{2} V_{\tau_0}[D_\theta \eta(\tau_0)(\theta - \theta_0) + D_h \eta(\tau_0)(h - h_0)].\end{aligned} \tag{IV.2}$$

Recall that for any funcitons $f(x), g(x) \in L^2(\mathcal{X})$,

$$\mu_\tau(f) = \frac{\int f(t) e^{\eta(t;\tau))} dt}{\int e^{\eta(t;\tau)} dt},$$

$$V_\tau(f, g) = \mu_\tau(fg) - \mu_\tau(f)\mu_\tau(g),$$

and $V_\tau(f) = V_\tau(f, f)$. We see that $\tilde{\ell}_n(\theta, h)$ is almost like a quadratic approximation of $\ell_n(\theta, h)$ at $\tau_0 = (\theta_0, h_0)$, ignoring terms independent of (θ, h) and terms involving second derivatives of $\eta(\theta, h)$. Since $V_{l,\tau_0}(\cdot)$ and $J(\cdot)$ are quadratic functionals, one can check that $\tilde{\ell}_{n,\lambda}(\theta, h)$ is convex with respect to (θ, h), and attains its minimum at $\tilde{\tau} = (\tilde{\theta}, \tilde{h})$ if and only if $\tilde{\tau}$ is the unique solution for the system $D_\theta \tilde{\ell}_{n,\lambda}(\theta, h) = 0$, $D_h \tilde{\ell}_{n,\lambda}(\theta, h) = 0$, see Proposition 2.

Using a quadratic approximation of the penalized likelihood to attain an approximation of the estimator is a common intermediate step in the study of convergence of smoothing spline estimators (see Silverman (1982); Cox and O'Sullivan (1990); O'Sullivan (1990); Gu and Qiu (1993)). In particular, our definition of $\tilde{\ell}_{n,\lambda}(\theta, h)$ can be seen as a generalized semiparametric version of the quadratic approximation given by equation (5.2) in Gu and Qiu (1993).

We establish the consistency and rate of convergence for $\tilde{\tau} - \tau_0$ measured in $\|\cdot\|_{\mathcal{Q}}$ in Section IV.2. Together with a bound for the approximation error $\hat{\tau} - \tilde{\tau}$, the rate of convergence for $\hat{\tau} - \tau_0$ then follows by the triangle inequality (Section IV.3). In addition, we also establish the rate of convergence of the estimate $\hat{\tau}$ in terms of the symmetrized Kullback-Leibler distance, which is defined as

$$\begin{aligned} \mathrm{SKL}(\tau_0, \hat{\tau}) &= \mathrm{KL}(\tau_0, \hat{\tau}) + \mathrm{KL}(\hat{\tau}, \tau_0) = \mathrm{E}_{f_{\tau_0}} \log \frac{f_{\tau_0}}{f_{\hat{\tau}}} + \mathrm{E}_{f_{\hat{\tau}}} \log \frac{f_{\hat{\tau}}}{f_{\tau_0}} \\ &= \mu_{\tau_0}[\eta(\theta_0, h_0) - \eta(\hat{\theta}, \hat{h})] + \mu_{\hat{\tau}}[\eta(\hat{\theta}, \hat{h}) - \eta(\theta_0, h_0)]. \end{aligned} \tag{IV.3}$$

In Section IV.4 we extend our analysis to a density estimation model with multiple samples. Since the proofs are straightforward extensions of those in Sections IV.2 and IV.3, we will only provide the model setup and state the results without proofs .

IV.2 Linear approximation

In this section, we derive the rate of convergence for $\tilde{\tau} - \tau_0$. We first prove the following proposition, which guarantees the existence of $\tilde{\tau}$.

Proposition 2. *$\tilde{l}_{n,\lambda}(\theta, h)$ attains its minimum at $\tilde{\tau} = (\tilde{\theta}, \tilde{h})$ if and only if $\tilde{\tau}$ is the solution for the system $D_\theta \tilde{l}_{n,\lambda}(\theta, h) = 0$, $D_h \tilde{l}_{n,\lambda}(\theta, h) = 0$.*

Proof. Let

$$\begin{aligned} B^{\theta}_{\tau_1,\tau_2}(\alpha) &\equiv \tilde{\ell}_n(\theta_1 + \alpha\theta_2, h_1) + \frac{\lambda}{2} J(h_1), \\ B^{h}_{\tau_1,\tau_2}(\alpha) &\equiv \tilde{\ell}_n(\theta_1, h_1 + \alpha h_2) + \frac{\lambda}{2} J(h_1 + \alpha h_2), \end{aligned}$$

and

$$
\begin{aligned}
\dot{B}^{\theta}_{\tau_1,\tau_2}(0) \equiv \frac{dB^{\theta}_{\tau_1,\tau_2}}{d\alpha}(0) = &-\frac{1}{n}\sum_{i=1}^{n} D_\theta\eta(X_i;\tau_0)\theta_2 + \mu_{\tau_0}[D_\theta\eta(\tau_0)\theta_2] \\
&+ V_{\tau_0}[D_\theta\eta(\tau_0)(\theta_1-\theta_0), D_\theta\eta(\tau_0)\theta_2] \\
&+ V_{\tau_0}[D_h\eta(\tau_0)(h_1-h_0), D_\theta\eta(\tau_0)\theta_2], \\
\dot{B}^{h}_{\tau_1,\tau_2}(0) \equiv \frac{dB^{h}_{\tau_1,\tau_2}}{d\alpha}(0) = &-\frac{1}{n}\sum_{i=1}^{n} D_h\eta(X_i;\tau_0)h_2 + \mu_{\tau_0}[D_h\eta(\tau_0)h_2] \\
&+ V_{\tau_0}[D_h\eta(\tau_0)(h_1-h_0), D_h\eta(\tau_0)h_2] \\
&+ V_{\tau_0}[D_\theta\eta(\tau_0)(\theta_1-\theta_0), D_h\eta(\tau_0)h_2] + \lambda J(h_1,h_2).
\end{aligned}
\tag{IV.4}
$$

Suppose $\tilde{\tau} = (\tilde{\theta}, \tilde{h})$ is the solution for the system $D_\theta \tilde{l}_{n,\lambda}(\theta, h) = 0$, $D_h \tilde{l}_{n,\lambda}(\theta, h) = 0$. By the definition of Fréchet partial differentiation, we have

$$
\dot{B}^{\theta}_{\tilde{\tau},\tau_2}(0) = \dot{B}^{h}_{\tilde{\tau},\tau_2}(0) = 0, \text{ for any } \tau_2 \in \mathcal{H}.
$$

For any $\tau_* = (\theta_*, h_*) \neq \tilde{\tau}$, since $V(\cdot)$, and $J(\cdot)$ are quadratic functionals (nonnegative definite), we see that after rearrangements

$$
\begin{aligned}
\tilde{l}_{n,\lambda}(\tau_*) &= \tilde{l}_{n,\lambda}[\tilde{\tau} + (\tau_* - \tilde{\tau})] \\
&= \tilde{l}_{n,\lambda}(\tilde{\tau}) + \dot{B}^{\theta}_{\tilde{\tau},\tau_*-\tilde{\tau}}(0) + \dot{B}^{h}_{\tilde{\tau},\tau_*-\tilde{\tau}}(0) \\
&\quad + \frac{1}{2}V_{\tau_0}[D_\theta\eta(\tau_0)(\theta_* - \tilde{\theta})] + \frac{1}{2}V_{\tau_0}[D_h\eta(\tau_0)(h_* - \tilde{h})] \\
&\quad + V_{\tau_0}[D_\theta\eta(\tau_0)(\theta_* - \tilde{\theta}), D_h\eta(\tau_0)(h_* - \tilde{h})] + \frac{\lambda}{2}J(h_* - \tilde{h}) \\
&= \tilde{l}_{n,\lambda}(\tilde{\tau}) + \frac{1}{2}V_{\tau_0}[D_\theta\eta(\tau_0)(\theta_* - \tilde{\theta}) + D_h\eta(\tau_0)(h_* - \tilde{h})] + \frac{\lambda}{2}J(h_* - \tilde{h}) \\
&\geq \tilde{l}_{n,\lambda}(\tilde{\tau}).
\end{aligned}
$$

Therefore, $\tilde{\tau}$ is a minimizer of $\tilde{l}_{n,\lambda}(\theta, h)$. Next, suppose $\tilde{\tau} = (\tilde{\theta}, \tilde{h})$ is a minimizer of $\tilde{l}_{n,\lambda}(\theta, h)$, but

$$
\dot{B}^{\theta}_{\tilde{\tau},\tau_2}(0) \neq 0, \text{ or } \dot{B}^{h}_{\tilde{\tau},\tau_2}(0) \neq 0, \text{ for some } \tau_2 \in \mathcal{H}.
$$

For some constant α, consider

$$\begin{aligned}\tilde{l}_{n,\lambda}(\tilde{\tau}+\alpha\tau_2) &= \tilde{l}_{n,\lambda}(\tilde{\tau}) + \alpha\dot{B}^{\theta}_{\tilde{\tau},\tau_2}(0) + \alpha\dot{B}^{h}_{\tilde{\tau},\tau_2}(0) \\ &\quad + \frac{\alpha^2}{2}V_{\tau_0}[D_\theta\eta(\tau_0)\theta_2] + \frac{\alpha^2}{2}V_{\tau_0}[D_h\eta(\tau_0)h_2] \\ &\quad + \alpha^2 V_{\tau_0}[D_\theta\eta(\tau_0)\theta_2, D_h\eta(\tau_0)h_2] + \frac{\lambda\alpha^2}{2}J(h_2) \\ &= \tilde{l}_{n,\lambda}(\tilde{\tau}) + \alpha\left[\dot{B}^{\theta}_{\tilde{\tau},\tau_2}(0) + \dot{B}^{h}_{\tilde{\tau},\tau_2}(0)\right] \\ &\quad + \frac{\alpha^2}{2}V_{\tau_0}[D_\theta\eta(\tau_0)\theta_2 + D_h\eta(\tau_0)h_2] + \frac{\lambda\alpha^2}{2}J(h_2).\end{aligned}$$

Notice that if $[\dot{B}^{\theta}_{\tilde{\tau},\tau_2}(0) + \dot{B}^{h}_{\tilde{\tau},\tau_2}(0)]$ greater than 0 (or less than 0), for $\alpha < 0$ (or $\alpha > 0$) and $|\alpha|$ is sufficiently small, the last expression will be strictly less than $\tilde{l}_{n,\lambda}(\tilde{\tau})$. Therefore, we have shown that $\tilde{l}_{n,\lambda}(\tilde{\tau}+\alpha\tau_2) < \tilde{l}_{n,\lambda}(\tilde{\tau})$, which contradicts with the fact that $\tilde{\tau}$ is a minimizer of $\tilde{l}_{n,\lambda}(\theta,h)$. Hence, if $\tilde{\tau}$ is a minimizer of $\tilde{l}_{n,\lambda}(\theta,h)$, it must be the solution for the system $D_\theta\tilde{l}_{n,\lambda}(\theta,h)=0$, $D_h\tilde{l}_{n,\lambda}(\theta,h)=0$. □

Next, we find the solution $\tilde{\tau}$ using the eigensystem discussed in Section II.2.3. Let

$$h_\nu = V_{\tau_0}(D_h\eta(\tau_0)h, D_h\eta(\tau_0)\phi_{0,\nu}) \quad \text{and} \quad h_{0,\nu} = V_{\tau_0}(D_h\eta(\tau_0)h_0, D_h\eta(\tau_0)\phi_{0,\nu}).$$

We have the Fourier expansions $h = \sum_\nu h_\nu\phi_{0,\nu}$ and $h_0 = \sum_\nu h_{0,\nu}\phi_{0,\nu}$ of h and h_0 with respect to the base $\phi_{0,\nu}$. Plugging these into equation (IV.2), we get

$$
\begin{aligned}
\tilde{\ell}_{n,\lambda}(\theta, h) = &-\theta^T \left\{ \frac{1}{n} \sum_{i=1}^{n} D_\theta \eta(X_i; \tau_0) - \mu_{\tau_0} \left[D_\theta \eta(\tau_0) \right] \right\} \\
&- \sum_\nu h_\nu \left\{ \frac{1}{n} \sum_{i=1}^{n} D_h \eta(X_i; \tau_0) \phi_{0,\nu} - \mu_{\tau_0} \left[D_h \eta(\tau_0) \phi_{0,\nu} \right] \right\} \\
&+ \frac{1}{2} (\theta - \theta_0)^T V_{\tau_0} \left[D_\theta \eta(\tau_0) \right] (\theta - \theta_0) + \frac{1}{2} \sum_\nu (h_\nu - h_{0,\nu})^2 \qquad \text{(IV.5)} \\
&+ \sum_\nu (h_\nu - h_{0,\nu}) (\theta - \theta_0)^T V_{\tau_0} \left[D_\theta \eta(\tau_0), D_h \eta(\tau_0) \phi_{0,\nu} \right] \\
&+ \frac{\lambda}{2} \sum_\nu \rho_{0,\nu} h_\nu^2.
\end{aligned}
$$

In equation (IV.5), using the fact that $D_\theta \eta(x; \tau_0)$ can be represented by a $p \times 1$ vector whose kth entry is $D_{\theta^k} \eta(x; \tau_0)$, we denote by $D_\theta \eta(x; \tau_0)$ both the linear operator and its vector form, i.e., $D_\theta \eta(x; \tau_0)\theta = \theta^T D_\theta \eta(x; \tau_0)$. We also have $\mu_{\tau_0}[D_\theta \eta(\tau_0)]$ and $V_{\tau_0}[D_\theta \eta(\tau_0), D_h \eta(\tau_0)\phi_{0,\nu}]$ the p-dimentional vectors whose ith entries are $\mu_{\tau_0}[D_{\theta^i} \eta(\tau_0)]$ and $V_{\tau_0}[D_{\theta^i} \eta(\tau_0), D_h \eta(\tau_0)\phi_{0,\nu}]$, respectively, and $V_{\tau_0}[D_\theta \eta(\tau_0)]$ is the $p \times p$ covariance matrix whose (i, j)th entry is $V_{\tau_0}[D_{\theta^i} \eta(\tau_0), D_{\theta^j} \eta(\tau_0)]$. The Fourier coefficients $\tilde{h}_\nu$ and $\tilde{\theta}$ that minimize equation (IV.5) are therefore given by the roots of the Fréchet partial derivatives of equation (IV.5). Let $\bar{\alpha}_n = \frac{1}{n} \sum_{i=1}^{n} D_\theta \eta(X_i; \tau_0) - \mu_{\tau_0}[D_\theta \eta(\tau_0)]$ and $\bar{\beta}_{\nu,n} = \frac{1}{n} \sum_{i=1}^{n} D_h \eta(X_i; \tau_0)\phi_{0,\nu} - \mu_{\tau_0}[D_h \eta(\tau_0)\phi_{0,\nu}]$, we get

$$
\begin{aligned}
\tilde{\theta} &= \theta_0 + \Omega_\lambda^{-1} \left\{ \bar{\alpha}_n - \sum_\nu \frac{\bar{\beta}_{\nu,n} - \lambda \rho_{0,\nu} h_{0,\nu}}{1 + \lambda \rho_{0,\nu}} V_{\tau_0}[D_\theta \eta(\tau_0), D_h \eta(\tau_0)\phi_{0,\nu}] \right\}, \\
\tilde{h}_\nu &= \frac{\bar{\beta}_{\nu,n} + h_{0,\nu}}{1 + \lambda \rho_{0,\nu}} - (\tilde{\theta} - \theta_0)^T \frac{V_{\tau_0}[D_\theta \eta(\tau_0), D_h \eta(\tau_0)\phi_{0,\nu}]}{1 + \lambda \rho_{0,\nu}},
\end{aligned}
$$

where

$$
\Omega_\lambda = V_{\tau_0}[D_\theta \eta(\tau_0)] - \sum_\nu \frac{V_{\tau_0}[D_\theta \eta(\tau_0), D_h \eta(\tau_0)\phi_{0,\nu}]^{\otimes 2}}{1 + \lambda \rho_{0,\nu}},
$$

and $a^{\otimes 2} = aa^T$ (a is a vector or matrix). Therefore, $(\tilde{\theta}, \tilde{h})$, where $\tilde{h} = \sum_\nu \tilde{h}_\nu \phi_{0,\nu}$, is the minimizer of equation (IV.2). Using the terminology from

the nonparametric setting Gu (2013), $\tilde{h}$ can be called a linear approximation of $\hat{h}$ since $\tilde{h}_\nu$ is linear in $\phi_{0,\nu}(X_i)$ given $\tilde{\theta}$. We will slightly abuse the terminology for our nonlinear case here and also call $\tilde{\tau} = (\tilde{\theta}, \tilde{h})$ the linear approximation of $\hat{\tau} = (\hat{\theta}, \hat{h})$.

Using the facts that $\mathrm{E}(\bar{\alpha}_n) = \mathrm{E}(\bar{\beta}_{\nu,n}) = 0$, $\mathrm{E}(\|\bar{\alpha}_n\|_{l^2}^2) = O(n^{-1})$, $\mathrm{E}(\bar{\beta}_{\nu,n}^2) = n^{-1}$, and after tedious calculation, we get the following lemma and theorem. The proof of the lemma is given in Section IV.2.1.

Lemma 2. *Under Assumptions 2 to 4, 8, and 9, as $n \to \infty$ and $\lambda \to 0$,*

$$\mathrm{E}\left\{V_{\tau_0}[D_\theta \eta(\tau_0)(\tilde{\theta} - \theta_0)]\right\} \le c\,\mathrm{E}\left[(\tilde{\theta} - \theta_0)^T(\tilde{\theta} - \theta_0)\right] = O(n^{-1}\lambda^{-\frac{1}{r}}),$$
$$\mathrm{E}\left\{V_{\tau_0}[D_h \eta(\tau_0)(\tilde{h} - h_0)] + \lambda J(\tilde{h} - h_0)\right\} = O(n^{-1}\lambda^{-\frac{1}{r}} + \lambda).$$

Theorem 4. *Under Assumptions 2 to 5, 8, and 9, as $n \to \infty$ and $\lambda \to 0$,*

$$V_{\tau_0}[L_\theta(\tilde{\theta} - \theta_0) + L_h(\tilde{h} - h_0)] + \lambda J(\tilde{h} - h_0) = O_p(n^{-1}\lambda^{-\frac{1}{r}} + \lambda).$$

Proof. By Assumption 5, Lemma 2 implies that

$$\mathrm{E}\left\{V_{\tau_0}[L_\theta(\tilde{\theta} - \theta_0)]\right\} = O(n^{-1}\lambda^{-\frac{1}{r}}),$$
$$\mathrm{E}\left\{V_{\tau_0}[L_h(\tilde{h} - h_0)] + \lambda J(\tilde{h} - h_0)\right\} = O(n^{-1}\lambda^{-\frac{1}{r}} + \lambda).$$

By the Cauchy-Schwarz inequality and completing the square, we have

$$\begin{aligned}
&V_{\tau_0}[L_\theta(\tilde{\theta} - \theta_0) + L_h(\tilde{h} - h_0)] \\
&\quad = V_{\tau_0}[L_\theta(\tilde{\theta} - \theta_0)] + V_{\tau_0}[L_h(\tilde{h} - h_0)] + 2V_{\tau_0}[L_\theta(\tilde{\theta} - \theta_0), L_h(\tilde{h} - h_0)] \\
&\quad \le V_{\tau_0}[L_\theta(\tilde{\theta} - \theta_0)] + V_{\tau_0}[L_h(\tilde{h} - h_0)] + 2V_{\tau_0}^{\frac{1}{2}}[L_\theta(\tilde{\theta} - \theta_0)]V_{\tau_0}^{\frac{1}{2}}[L_h(\tilde{h} - h_0)] \\
&\quad = \left[V_{\tau_0}^{\frac{1}{2}}[L_\theta(\tilde{\theta} - \theta_0)] + V_{\tau_0}^{\frac{1}{2}}[L_h(\tilde{h} - h_0)]\right]^2 \\
&\quad = O_p(n^{-1}\lambda^{-\frac{1}{r}} + \lambda).
\end{aligned}$$

Together with $\lambda J(\tilde{h} - h_0) = O_p(n^{-1}\lambda^{-\frac{1}{r}} + \lambda)$, we have the desired result. $\square$

Note that in addition to $n \to \infty$ and $\lambda \to 0$, if also $n^{-1}\lambda^{-\frac{1}{r}} \to 0$, then the probability that $\tilde{\tau} \to \tau_0$ tends to 1, which means $\tilde{\tau} \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$. We can restrict our attention to this event for the rest of our analysis, or for simplicity, we assume that $\tilde{\tau} \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$.

IV.2.1 Proof of Lemma 2

We now give detailed calculations and a proof of Lemma 2. Let $\mathcal{G}^k g = V_{\tau_0}[D_{\theta^k}\eta(\tau_0), D_h\eta(\tau_0)g]$ for any $g \in \mathcal{H}$. Since

$$\left|\mathcal{G}^k g\right| \leq V_{\tau_0}^{\frac{1}{2}}[D_{\theta^k}\eta(\tau_0)] V_{\tau_0}^{\frac{1}{2}}[D_h\eta(\tau_0)g] \leq C \left\|g\right\|_{\mathcal{H}}$$

for some positive constant C, $\mathcal{G}^k$ is a bounded linear functional on $\mathcal{H}$. By the Riesz representation theorem, there exists $G^k \in \mathcal{H}$ such that for any $g \in \mathcal{H}$,

$$\mathcal{G}^k g = V_{\tau_0}[D_{\theta^k}\eta(\tau_0), D_h\eta(\tau_0)g] = V_{\tau_0}[D_h\eta(\tau_0)G^k, D_h\eta(\tau_0)g] + \lambda J(G^k, g).$$

Let $G = \left[G^k\right]_{k=1}^p$. Define $V_{\tau_0}[D_\theta\eta(\tau_0), D_h\eta(\tau_0)g]$, $V_{\tau_0}[D_h\eta(\tau_0)G, D_h\eta(\tau_0)g]$, and $J(G, g)$ to be $p \times 1$ vectors whose kth entries are $V_{\tau_0}[D_{\theta^k}\eta(\tau_0), D_h\eta(\tau_0)g]$, $V_{\tau_0}[D_h\eta(\tau_0)G^k, D_h\eta(\tau_0)g]$, and $J(G^k, g)$, respectively. Therefore,

$$V_{\tau_0}[D_\theta\eta(\tau_0), D_h\eta(\tau_0)g] = V_{\tau_0}[D_h\eta(\tau_0)G, D_h\eta(\tau_0)g] + \lambda J(G, g).$$

The Fourier expansion of G^k with respect to the eigensystem discussed in Section II.2.3 is

$$G^k = \sum_\nu V_{\tau_0}[D_h\eta(\tau_0)G^k, D_h\eta(\tau_0)\phi_{0,\nu}]\phi_{0,\nu}.$$

A simple calculation shows that

$$V_{\tau_0}[D_h\eta(\tau_0)G^k, D_h\eta(\tau_0)\phi_{0,\nu}] = (1+\lambda\rho_{0,\nu})^{-1}V_{\tau_0}[D_{\theta^k}\eta(\tau_0), D_h\eta(\tau_0)\phi_{0,\nu}],$$

and hence,

$$G^k = \sum_\nu \frac{1}{1+\lambda\rho_{0,\nu}} V_{\tau_0}[D_{\theta^k}\eta(\tau_0), D_h\eta(\tau_0)\phi_{0,\nu}]\phi_{0,\nu}. \tag{IV.6}$$

Recall from Section IV.2 that

$$\tilde{\theta} - \theta_0 = \Omega_\lambda^{-1}\left\{\bar{\alpha}_n - \sum_\nu \frac{\bar{\beta}_{\nu,n} - \lambda\rho_{0,\nu}h_{0,\nu}}{1+\lambda\rho_{0,\nu}} V_{\tau_0}[D_\theta\eta(\tau_0), D_h\eta(\tau_0)\phi_{0,\nu}]\right\},$$

where

$$\Omega_\lambda = V_{\tau_0}[D_\theta\eta(\tau_0)] - \sum_\nu \frac{V_{\tau_0}[D_\theta\eta(\tau_0), D_h\eta(\tau_0)\phi_{0,\nu}]^{\otimes 2}}{1+\lambda\rho_{0,\nu}}.$$

The (i,j)th entry of this $p\times p$ matrix Ω_λ can be written as

$$\begin{aligned}\Omega_\lambda^{i,j} &= V_{\tau_0}[D_{\theta^i}\eta(\tau_0), D_{\theta^j}\eta(\tau_0)] - V_{\tau_0}[D_{\theta^i}\eta(\tau_0), D_h\eta(\tau_0)G^j] \\ &= V_{\tau_0}[D_{\theta^i}\eta(\tau_0) - D_h\eta(\tau_0)G^i, D_{\theta^j}\eta(\tau_0) - D_h\eta(\tau_0)G^j] + \lambda J(G^i, G^j).\end{aligned}$$

Let $\Omega = V_{\tau_0}[D_\theta\eta(\tau_0) - D_h\eta(\tau_0)G]$ and $\Sigma_\lambda = \lambda J(G)$ be the matrices such that

$$\Omega^{i,j} = V_{\tau_0}[D_{\theta^i}\eta(\tau_0) - D_h\eta(\tau_0)G^i, D_{\theta^j}\eta(\tau_0) - D_h\eta(\tau_0)G^j],$$

and $\Sigma_\lambda^{i,j} = \lambda J(G^i, G^j)$. Thus, $\Omega_\lambda = \Omega + \Sigma_\lambda$.

We now prove some properties of Ω and Σ_λ, which will be used to establish the bound for $\mathrm{E}[(\tilde{\theta}-\theta_0)^T(\tilde{\theta}-\theta_0)]$.

Lemma 3. *$\Sigma_\lambda \to 0$ as $\lambda \to 0$.*

Proof. By (IV.6), we get that the (i,j)th entry of Σ_λ is

$$\sum_\nu \frac{\lambda\rho_{0,\nu}}{(1+\lambda\rho_{0,\nu})^2} V_{\tau_0}[D_{\theta^i}\eta(\tau_0), D_h\eta(\tau_0)\phi_{0,\nu}] V_{\tau_0}[D_{\theta^j}\eta(\tau_0), D_h\eta(\tau_0)\phi_{0,\nu}].$$

By square summability of $\{V_{\tau_0}[D_{\theta^i}\eta(\tau_0), D_h\eta(\tau_0)\phi_{0,\nu}]\}_{\nu\in\mathbb{N}}$ and the dominated convergence theorem, the above sum converges to 0 as $\lambda \to 0$. □

Lemma 4. *Under Assumption 3, Ω is positive definite. Let c_δ and C_1 be constants defined in Assumptions 3 and 5 respectively, and let δ be any eigenvalue of Ω. Then $\delta > C_1 c_\delta = \tilde{c}_\delta$.*

Proof. The proof is similar to the one for Lemma 1. □

Note that by Lemma 4, the eigenvalues of Ω have a uniform lower bound independent of λ. Then as $\lambda \to 0$, we have

$$\begin{aligned}\mathrm{E}\left[(\tilde{\theta}-\theta_0)^T(\tilde{\theta}-\theta_0)\right] &= \mathrm{E}\left[a_n^T(\Omega+\Sigma_\lambda)^{-2}a_n\right] \\ &\to \mathrm{E}\left[a_n^T\Omega^{-2}a_n\right] \le \tilde{c}_\delta^{-2}\,\mathrm{E}\left[a_n^T a_n\right] = \tilde{c}_\delta^{-2}\sum_{i=1}^p \mathrm{E}\left[(a_n^i)^2\right],\end{aligned}$$

where

$$a_n = \bar{\alpha}_n - \sum_\nu \frac{\bar{\beta}_{\nu,n} - \lambda\rho_{0,\nu}h_{0,\nu}}{1+\lambda\rho_{0,\nu}} V_{\tau_0}[D_\theta\eta(\tau_0), D_h\eta(\tau_0)\phi_{0,\nu}],$$

and a_n^i is the ith entry of a_n.

Before we proceed to derive the bound for $\mathrm{E}[(a_n^i)^2]$, we also need the following lemma, whose proof is given by Lemma 5.2 in Gu and Qiu (1993).

Lemma 5. *Under Assumption 9, as* $\lambda \to 0$,

$$\sum_{\nu} \frac{\lambda\rho_{0,\nu}}{(1+\lambda\rho_{0,\nu})^2} = O(\lambda^{-\frac{1}{r}}),$$
$$\sum_{\nu} \frac{1}{(1+\lambda\rho_{0,\nu})^2} = O(\lambda^{-\frac{1}{r}}),$$
$$\sum_{\nu} \frac{1}{1+\lambda\rho_{0,\nu}} = O(\lambda^{-\frac{1}{r}}).$$

We now ready to establish the upper bound for $\mathrm{E}\left[(a_n^i)^2\right]$. We have

$$\begin{aligned}
&\mathrm{E}\left[(a_n^i)^2\right] \\
&= \mathrm{E}\left\{\left[\bar{\alpha}_n^i - \sum_{\nu} \frac{\bar{\beta}_{\nu,n} - \lambda\rho_{0,\nu}h_{0,\nu}}{1+\lambda\rho_{0,\nu}} V_{\tau_0}[D_{\theta^i}\eta(\tau_0), D_h\eta(\tau_0)\phi_{0,\nu}]\right]^2\right\} \\
&\leq 2\,\mathrm{E}\left[(\bar{\alpha}_n^i)^2\right] + 2\,\mathrm{E}\left[\left(\sum_{\nu} \frac{\bar{\beta}_{\nu,n} - \lambda\rho_{0,\nu}h_{0,\nu}}{1+\lambda\rho_{0,\nu}} V_{\tau_0}[D_{\theta^i}\eta(\tau_0), D_h\eta(\tau_0)\phi_{0,\nu}]\right)^2\right].
\end{aligned}$$

Note that $\mathrm{E}\left[(\bar{\alpha}_n^i)^2\right] = O(n^{-1})$, and by square summability of $\{h_{0,\nu}\}_{\nu\in\mathbb{N}}$ and $\{V_{\tau_0}[D_{\theta^i}\eta(\tau_0), D_h\eta(\tau_0)\phi_{0,\nu}]\}_{\nu\in\mathbb{N}}$, the dominated convergence theorem, the Cauchy-Schwarz inequality, $\mathrm{E}(\bar{\beta}_{\nu,n}^2) = n^{-1}$, and Lemma 5, we have

$$\begin{aligned}
&\mathrm{E}\left[\left(\sum_{\nu} \frac{\bar{\beta}_{\nu,n} - \lambda\rho_{0,\nu}h_{0,\nu}}{1+\lambda\rho_{0,\nu}} V_{\tau_0}[D_{\theta^i}\eta(\tau_0), D_h\eta(\tau_0)\phi_{0,\nu}]\right)^2\right] \\
&\to \mathrm{E}\left[\left(\sum_{\nu} \frac{\bar{\beta}_{\nu,n}}{1+\lambda\rho_{0,\nu}} V_{\tau_0}[D_{\theta^i}\eta(\tau_0), D_h\eta(\tau_0)\phi_{0,\nu}]\right)^2\right] \\
&\leq C\left[\sum_{\nu} \frac{\mathrm{E}(\bar{\beta}_{\nu,n}^2)}{(1+\lambda\rho_{0,\nu})^2}\right] = O(n^{-1}\lambda^{-\frac{1}{r}}).
\end{aligned}$$

Therefore, we conclude that as $\lambda \to 0$, $\mathrm{E}\left[(a_n^i)^2\right] = O(n^{-1}\lambda^{-\frac{1}{r}})$, which implies that

$$\mathrm{E}\left\{V_{\tau_0}[D_\theta\eta(\tau_0)(\tilde{\theta}-\theta_0)]\right\} \leq c\,\mathrm{E}\left[(\tilde{\theta}-\theta_0)^T(\tilde{\theta}-\theta_0)\right] = O(n^{-1}\lambda^{-\frac{1}{r}}).$$

This concludes the proof for the first bound in Lemma 2.

Now for the second bound in Lemma 2, we see that

$$V_{\tau_0}[D_h\eta(\tau_0)(\tilde{h}-h_0)] = \sum_\nu \left(\tilde{h}_\nu - h_{0,\nu}\right)^2, \lambda J(\tilde{h}-h_0) = \sum_\nu \lambda\rho_\nu^0 \left(\tilde{h}_\nu - h_{0,\nu}\right)^2.$$

Plugging in the formula of $\tilde{h}_\nu$ given in Section IV.2, we get

$$\begin{aligned}
&\sum_\nu \left(\tilde{h}_\nu - h_{0,\nu}\right)^2 \\
&\quad = \sum_\nu \left\{ \frac{\bar{\beta}_{\nu,n} - \lambda\rho_{0,\nu}h_{0,\nu}}{1+\lambda\rho_{0,\nu}} - (\tilde{\theta}-\theta_0)^T \frac{V_{\tau_0}[D_\theta\eta(\tau_0), D_h\eta(\tau_0)\phi_{0,\nu}]}{1+\lambda\rho_{0,\nu}} \right\}^2 \\
&\quad \le C\left[(I_h) + (II_h)\right],
\end{aligned}$$

where

$$\begin{aligned}
(I_h) &= \sum_\nu \left(\frac{\bar{\beta}_{\nu,n} - \lambda\rho_{0,\nu}h_{0,\nu}}{1+\lambda\rho_{0,\nu}} \right)^2, \\
(II_h) &= \sum_\nu \left[(\tilde{\theta}-\theta_0)^T \frac{V_{\tau_0}[D_\theta\eta(\tau_0), D_h\eta(\tau_0)\phi_{0,\nu}]}{1+\lambda\rho_{0,\nu}} \right]^2.
\end{aligned}$$

Since $\mathrm{E}(\bar{\beta}_{\nu,n}) = 0$, $\mathrm{E}(\bar{\beta}_{\nu,n}^2) = \frac{1}{n}$, and $\sum_\nu \rho_{0,\nu}h_{0,\nu}^2 = J(h_0) < \infty$, by Lemma 5, we have

$$\mathrm{E}[(I)] = \frac{1}{n}\sum_\nu \frac{1}{(1+\lambda\rho_{0,\nu})^2} + \lambda \sum_\nu \frac{\lambda\rho_{0,\nu}}{(1+\lambda\rho_{0,\nu})^2}\rho_{0,\nu}h_{0,\nu}^2 = O(n^{-1}\lambda^{-\frac{1}{r}} + \lambda).$$

If a, b are $p \times 1$ vectors, then a^Tb is 1×1, which implies that $a^Tb = b^Ta$ and $(a^Tb)^2 = b^Taa^Tb$. Using this fact and the bound for $\mathrm{E}[(\tilde{\theta}-\theta_0)^T(\tilde{\theta}-\theta_0)]$, we have

$$
\begin{aligned}
&\mathrm{E}[(II)] \\
&= \sum_{i,j=1}^{p} \Big\{ \mathrm{E}\left[(\tilde{\theta}-\theta_0)^i (\tilde{\theta}-\theta_0)^j \right] \\
&\qquad \cdot \sum_{\nu} \frac{1}{(1+\lambda\rho_{0,\nu})^2} V_{\tau_0}[D_{\theta^i}\eta(\tau_0), D_h\eta(\tau_0)\phi_{0,\nu}] V_{\tau_0}[D_{\theta^j}\eta(\tau_0), D_h\eta(\tau_0)\phi_{0,\nu}] \Big\} \\
&\le C \sum_{i,j=1}^{p} \mathrm{E}\left[(\tilde{\theta}-\theta_0)^i (\tilde{\theta}-\theta_0)^j \right] \\
&\le C \sum_{i,j=1}^{p} \left\{ \mathrm{E}\left[(\tilde{\theta}^i-\theta_0^i)^2 \right] \right\}^{\frac{1}{2}} \left\{ \mathrm{E}\left[(\tilde{\theta}^j-\theta_0^j)^2 \right] \right\}^{\frac{1}{2}} = O(n^{-1}\lambda^{-\frac{1}{r}}).
\end{aligned}
$$

Therefore, $\mathrm{E}\{V_{\tau_0}[D_h\eta(\tau_0)(\tilde{h}-h_0)]\} = O(n^{-1}\lambda^{-\frac{1}{r}}+\lambda)$. Similar analysis shows that

$$
\mathrm{E}\left[\lambda J(\tilde{h}-h_0) \right] = \mathrm{E}\left[\sum_{\nu} \lambda\rho_{0,\nu} \left(\tilde{h}_\nu - h_{0,\nu} \right)^2 \right] = O(n^{-1}\lambda^{-\frac{1}{r}}+\lambda).
$$

Hence, the second bound in Lemma 2 is established.

IV.3 Approximation error and main results

We next find a bound for the approximation error $\hat{\tau}-\tilde{\tau} = (\hat{\theta}-\tilde{\theta}, \hat{h}-\tilde{h})$, which will then imply the convergence of $\hat{\tau}-\tau_0 = (\hat{\theta}-\theta_0, \hat{h}-h_0)$. Define

$$
\begin{aligned}
A^{\theta}_{\tau_1,\tau_2}(\alpha) &\equiv \ell_n(\theta_1+\alpha\theta_2, h_1) + \frac{\lambda}{2} J(h_1), \\
A^{h}_{\tau_1,\tau_2}(\alpha) &\equiv \ell_n(\theta_1, h_1+\alpha h_2) + \frac{\lambda}{2} J(h_1+\alpha h_2),
\end{aligned}
$$

and

$$
\begin{aligned}
\dot{A}^{\theta}_{\tau_1,\tau_2}(0) &\equiv \frac{dA^{\theta}_{\tau_1,\tau_2}}{d\alpha}(0) = -\frac{1}{n}\sum_{i=1}^{n} D_\theta\eta(X_i;\tau_1)\theta_2 + \mu_{\tau_1}[D_\theta\eta(\tau_1)\theta_2], \\
\dot{A}^{h}_{\tau_1,\tau_2}(0) &\equiv \frac{dA^{h}_{\tau_1,\tau_2}}{d\alpha}(0) = -\frac{1}{n}\sum_{i=1}^{n} D_h\eta(X_i;\tau_1)h_2 + \mu_{\tau_1}[D_h\eta(\tau_1)h_2] + \lambda J(h_1,h_2).
\end{aligned} \tag{IV.7}
$$

One can see that $\dot{A}^{\theta}_{\tau_1,\tau_2}(0)$ (or $\dot{A}^{h}_{\tau_1,\tau_2}(0)$) can be understood as the Fréchet partial derivative of $\ell_{n,\lambda}(\theta,h)$ with respect to θ (or h) at $\tau_1=(\theta_1,h_1)$ in the direction of θ_2 (or h_2). Likewise, $\dot{B}^{\theta}_{\tau_1,\tau_2}(0)$ (or $\dot{B}^{h}_{\tau_1,\tau_2}(0)$) as given in (IV.4) is the Fréchet partial derivative of $\tilde{\ell}_{n,\lambda}(\theta,h)$ with respect to θ (or h) at $\tau_1=(\theta_1,h_1)$ in the direction of θ_2 (or h_2). Therefore, for any τ_2, we note that $\dot{A}^{\theta}_{\hat{\tau},\tau_2}(0)=\dot{A}^{h}_{\hat{\tau},\tau_2}(0)=0$, and $\dot{B}^{\theta}_{\tilde{\tau},\tau_2}(0)=\dot{B}^{h}_{\tilde{\tau},\tau_2}(0)=0$. Let $\tau_2=\hat{\tau}-\tilde{\tau}$, we get

$$\begin{aligned}
&\dot{A}^{\theta}_{\hat{\tau},\tau_2}(0)+\dot{A}^{h}_{\hat{\tau},\tau_2}(0)\\
&\quad=\mu_{\hat{\tau}}[D_\theta\eta(\hat{\tau})(\hat{\theta}-\tilde{\theta})+D_h\eta(\hat{\tau})(\hat{h}-\tilde{h})]+\lambda J(\hat{h},\hat{h}-\tilde{h})\\
&\qquad-\frac{1}{n}\sum_{i=1}^{n}\Big[D_\theta\eta(X_i;\hat{\tau})(\hat{\theta}-\tilde{\theta})+D_h\eta(X_i;\hat{\tau})(\hat{h}-\tilde{h})\Big]=0,
\end{aligned}\tag{IV.8}$$

$$\begin{aligned}
&\dot{B}^{\theta}_{\tilde{\tau},\tau_2}(0)+\dot{B}^{h}_{\tilde{\tau},\tau_2}(0)\\
&=V_{\tau_0}[D_\theta\eta(\tau_0)(\tilde{\theta}-\theta_0)+D_h\eta(\tau_0)(\tilde{h}-h_0),D_\theta\eta(\tau_0)(\hat{\theta}-\tilde{\theta})+D_h\eta(\tau_0)(\hat{h}-\tilde{h})]\\
&\qquad+\mu_{\tau_0}[D_\theta\eta(\tau_0)(\hat{\theta}-\tilde{\theta})+D_h\eta(\tau_0)(\hat{h}-\tilde{h})]+\lambda J(\tilde{h},\hat{h}-\tilde{h})\\
&\qquad-\frac{1}{n}\sum_{i=1}^{n}\Big[D_\theta\eta(X_i;\tau_0)(\hat{\theta}-\tilde{\theta})+D_h\eta(X_i;\tau_0)(\hat{h}-\tilde{h})\Big]=0.
\end{aligned}\tag{IV.9}$$

Equating (IV.8) and (IV.9) and subtracting $\mu_{\tilde{\tau}}[D_\theta\eta(\hat{\tau})(\hat{\theta}-\tilde{\theta})+D_h\eta(\hat{\tau})(\hat{h}-\tilde{h})]$ on both sides. After some rearrangements, we have

$$\begin{aligned}
&\mu_{\hat{\tau}}[D_\theta\eta(\hat{\tau})(\hat{\theta}-\tilde{\theta})+D_h\eta(\hat{\tau})(\hat{h}-\tilde{h})]-\mu_{\tilde{\tau}}[D_\theta\eta(\hat{\tau})(\hat{\theta}-\tilde{\theta})+D_h\eta(\hat{\tau})(\hat{h}-\tilde{h})]\\
&\qquad+\lambda J(\hat{h}-\tilde{h},\hat{h}-\tilde{h})\\
&=V_{\tau_0}[D_\theta\eta(\tau_0)(\tilde{\theta}-\theta_0)+D_h\eta(\tau_0)(\tilde{h}-h_0),D_\theta\eta(\tau_0)(\hat{\theta}-\tilde{\theta})+D_h\eta(\tau_0)(\hat{h}-\tilde{h})]\\
&\qquad+(I)+(II)+(III),
\end{aligned}\tag{IV.10}$$

where

$$
\begin{aligned}
(I) &= \mu_{\tau_0}[D_\theta\eta(\hat{\tau})(\hat{\theta}-\tilde{\theta}) + D_h\eta(\hat{\tau})(\hat{h}-\tilde{h})] \\
&\quad - \mu_{\tilde{\tau}}[D_\theta\eta(\hat{\tau})(\hat{\theta}-\tilde{\theta}) + D_h\eta(\hat{\tau})(\hat{h}-\tilde{h})], \\
(II) &= \frac{1}{n}\sum_{i=1}^{n}\left[D_\theta\eta(X_i;\hat{\tau})(\hat{\theta}-\tilde{\theta}) + D_h\eta(X_i;\hat{\tau})(\hat{h}-\tilde{h})\right] \\
&\quad - \mu_{\tau_0}[D_\theta\eta(\hat{\tau})(\hat{\theta}-\tilde{\theta}) + D_h\eta(\hat{\tau})(\hat{h}-\tilde{h})], \\
(III) &= -\frac{1}{n}\sum_{i=1}^{n}\left[D_\theta\eta(X_i;\tau_0)(\hat{\theta}-\tilde{\theta}) + D_h\eta(X_i;\tau_0)(\hat{h}-\tilde{h})\right] \\
&\quad + \mu_{\tau_0}[D_\theta\eta(\tau_0)(\hat{\theta}-\tilde{\theta}) + D_h\eta(\tau_0)(\hat{h}-\tilde{h})].
\end{aligned}
$$

Note that for any function $f \in L_0^2(\mathcal{X})$ and $\tau_1, \tau_2 \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$, it is easy to show using the mean value theorem that

$$
\mu_{\tau_2}[f] - \mu_{\tau_1}[f] = V_{\tau_*}[f, D_\theta\eta(\tau_*)(\theta_2-\theta_1) + D_h\eta(\tau_*)(h_2-h_1)], \tag{IV.11}
$$

where $\tau_* = \tau_1 + \alpha(\tau_2 - \tau_1)$ for some $0 \le \alpha \le 1$. We are now ready to prove the following theorem.

Theorem 5. *Under Assumptions 2 to 9, as $n \to \infty$, $\lambda \to 0$, and $n^{-1}\lambda^{-\frac{1}{r}} \to 0$,*

$$
V_{\tau_0}[L_\theta(\hat{\theta}-\tilde{\theta}) + L_h(\hat{h}-\tilde{h})] + \lambda J(\hat{h}-\tilde{h}) = O_p\left(n^{-1}\lambda^{-\frac{1}{r}} + \lambda\right).
$$

Therefore,

$$
V_{\tau_0}[L_\theta(\hat{\theta}-\theta_0) + L_h(\hat{h}-h_0)] + \lambda J(\hat{h}-h_0) = O_p\left(n^{-1}\lambda^{-\frac{1}{r}} + \lambda\right).
$$

Proof. We first obtain a lower bound for the first two terms in the LHS of (IV.10) and an upper bound for (I) in the RHS of (IV.10). For some

$0 \leq \alpha \leq 1$, let $\tau_* = \tilde{\tau} + \alpha(\hat{\tau} - \tilde{\tau})$. By Assumptions 5, 6, and 7, and equation (IV.11), the first two terms in the LHS of (IV.10) yield

$$
\begin{aligned}
&\mu_{\hat{\tau}}[D_\theta\eta(\hat{\tau})(\hat{\theta} - \tilde{\theta}) + D_h\eta(\hat{\tau})(\hat{h} - \tilde{h})] - \mu_{\tilde{\tau}}[D_\theta\eta(\hat{\tau})(\hat{\theta} - \tilde{\theta}) + D_h\eta(\hat{\tau})(\hat{h} - \tilde{h})] \\
&= V_{\tau_*}[D_\theta\eta(\hat{\tau})(\hat{\theta} - \tilde{\theta}) + D_h\eta(\hat{\tau})(\hat{h} - \tilde{h}), D_\theta\eta(\tau_*)(\hat{\theta} - \tilde{\theta}) + D_h\eta(\tau_*)(\hat{h} - \tilde{h})] \\
&= \frac{1}{2}\Big\{V_{\tau_*}[D_\theta\eta(\hat{\tau})(\hat{\theta} - \tilde{\theta}) + D_h\eta(\hat{\tau})(\hat{h} - \tilde{h})] \\
&\quad + V_{\tau_*}[D_\theta\eta(\tau_*)(\hat{\theta} - \tilde{\theta}) + D_h\eta(\tau_*)(\hat{h} - \tilde{h})] \\
&\quad - V_{\tau_*}[(D_\theta\eta(\hat{\tau}) - D_\theta\eta(\tau_*))(\hat{\theta} - \tilde{\theta}) + (D_h\eta(\hat{\tau}) - D_h\eta(\tau_*))(\hat{h} - \tilde{h})]\Big\} \\
&\geq \frac{C_3}{2}\Big\{2C_1 V_{\tau_0}[L_\theta(\hat{\theta} - \tilde{\theta}) + L_h(\hat{h} - \tilde{h})] - C_d V_{\tau_0}[L_\theta(\hat{\theta} - \tilde{\theta}) + L_h(\hat{h} - \tilde{h})]\Big\} \\
&\geq \frac{c_3}{2} V_{\tau_0}[L_\theta(\hat{\theta} - \tilde{\theta}) + L_h(\hat{h} - \tilde{h})],
\end{aligned}
$$

for some $0 \leq c_3 \leq C_3(2C_1 - C_d)$. To get an upper bound for (I), let $\tau_u = \tilde{\tau} + u(\tau_0 - \tilde{\tau})$ for some $0 \leq u \leq 1$. By equation (IV.11), the Cauchy-Schwarz inequality, and Assumptions 5 and 7, we have

$$
\begin{aligned}
&\mu_{\tau_0}[D_\theta\eta(\hat{\tau})(\hat{\theta} - \tilde{\theta}) + D_h\eta(\hat{\tau})(\hat{h} - \tilde{h})] - \mu_{\tilde{\tau}}[D_\theta\eta(\hat{\tau})(\hat{\theta} - \tilde{\theta}) + D_h\eta(\hat{\tau})(\hat{h} - \tilde{h})] \\
&= V_{\tau_u}[D_\theta\eta(\hat{\tau})(\hat{\theta} - \tilde{\theta}) + D_h\eta(\hat{\tau})(\hat{h} - \tilde{h}), D_\theta\eta(\tau_u)(\theta_0 - \tilde{\theta}) + D_h\eta(\tau_u)(h_0 - \tilde{h})] \\
&\leq C_2 C_4 V_{\tau_0}^{\frac{1}{2}}[L_\theta(\hat{\theta} - \tilde{\theta}) + L_h(\hat{h} - \tilde{h})] V_{\tau_0}^{\frac{1}{2}}[L_\theta(\theta_0 - \tilde{\theta}) + L_h(h_0 - \tilde{h})].
\end{aligned}
$$

Next, for any random sample $X_1, \ldots, X_n$ and any function $f \in L_0^2(\mathcal{X})$, it is well known that $\mathrm{E}\{|n^{-1}\sum_{i=1}^n f(X_i) - \mathrm{E}(f(X))|^2\} = n^{-1}\,\mathrm{Var}[f(X)]$, which implies $n^{-1}\sum_{i=1}^n f(X_i) - \mathrm{E}[f(X)] = O_p(n^{-\frac{1}{2}})\,\mathrm{Var}^{\frac{1}{2}}[f(X)]$. Together with Assumption 5, (II) and (III) in (IV.10) are both bounded above by $O_p(n^{-\frac{1}{2}})V_{\tau_0}^{\frac{1}{2}}[L_\theta(\hat{\theta} - \tilde{\theta}) + L_h(\hat{h} - \tilde{h})]$.

Putting everything together, as $\lambda \to 0$, $n \to \infty$, and $n^{-1}\lambda^{-\frac{1}{r}} \to 0$, by all assumptions and Theorem 4, we get

$$
\begin{aligned}
&\frac{c_3}{2} V_{\tau_0}[L_\theta(\hat{\theta}-\tilde{\theta}) + L_h(\hat{h}-\tilde{h})] + \lambda J(\hat{h}-\tilde{h}) \\
&\leq V_{\tau_0}[D_\theta\eta(\tau_0)(\tilde{\theta}-\theta_0) + D_h\eta(\tau_0)(\tilde{h}-h_0), D_\theta\eta(\tau_0)(\hat{\theta}-\tilde{\theta}) + D_h\eta(\tau_0)(\hat{h}-\tilde{h})] \\
&\quad + C_2C_4 V_{\tau_0}^{\frac{1}{2}}[L_\theta(\hat{\theta}-\tilde{\theta}) + L_h(\hat{h}-\tilde{h})] V_{\tau_0}^{\frac{1}{2}}[L_\theta(\theta_0-\tilde{\theta}) + L_h(h_0-\tilde{h})] \\
&\quad + O_p\left(n^{-\frac{1}{2}}\right) V_{\tau_0}^{\frac{1}{2}}[L_\theta(\hat{\theta}-\tilde{\theta}) + L_h(\hat{h}-\tilde{h})] \\
&\leq C_2(1+C_4) V_{\tau_0}^{\frac{1}{2}}[L_\theta(\tilde{\theta}-\theta_0) + L_h(\tilde{h}-h_0)] V_{\tau_0}^{\frac{1}{2}}[L_\theta(\hat{\theta}-\tilde{\theta}) + L_h(\hat{h}-\tilde{h})] \\
&\quad + O_p\left(n^{-\frac{1}{2}}\right) V_{\tau_0}^{\frac{1}{2}}[L_\theta(\hat{\theta}-\tilde{\theta}) + L_h(\hat{h}-\tilde{h})] \\
&\leq O_p\left(n^{-\frac{1}{2}}\lambda^{-\frac{1}{2r}} + \lambda^{\frac{1}{2}}\right) V_{\tau_0}^{\frac{1}{2}}[L_\theta(\hat{\theta}-\tilde{\theta}) + L_h(\hat{h}-\tilde{h})].
\end{aligned}
$$

The result then follows after trivial manipulation. □

Since $\|\cdot\|_{\mathbb{R}^p}$ is equivalent to $\|\cdot\|_{l^2}$ on $\mathbb{R}^p$, and $\|\cdot\|_{\mathcal{Q}}$ is equivalent to the product norm $\|\cdot\|_{\mathcal{Q},1}$ on the joint parameter space $\mathcal{Q}$ (see Remark 2(ii) and (iv)). The following corollaries are direct consequences of Theorem 5.

Corollary 1. *Under Assumptions 2 to 9, as $n \to \infty$, $\lambda \to 0$, and $n^{-1}\lambda^{-\frac{1}{r}} \to 0$, we have*

$$
\|\hat{\theta}-\theta_0\|_{l^2} \sim \|\hat{\theta}-\theta_0\|_{\mathbb{R}^p} = V_{\tau_0}^{\frac{1}{2}}[L_\theta(\hat{\theta}-\theta_0)] = O_p\left(n^{-\frac{1}{2}}\lambda^{-\frac{1}{2r}} + \lambda^{\frac{1}{2}}\right).
$$

Corollary 2. *If $\eta(\theta, h) = \alpha(x;\theta) + h(x)$, L_h can be chosen to be the inclusion operator from $\mathcal{H}$ to $L_0^2(\mathcal{X})$. Under the same conditions as in Theorem 5, we have*

$$
\|\hat{h}-h_0\|_{\mathcal{H}} = [V_{\tau_0}(\hat{h}-h_0) + \lambda J(\hat{h}-h_0)]^{1/2} = O_p\left(n^{-\frac{1}{2}}\lambda^{-\frac{1}{2r}} + \lambda^{\frac{1}{2}}\right).
$$

We now derive a convergence rate of the overall density function, measured by the symmetrized Kullback-Leibler distance, defined in (IV.3).

Theorem 6. *Under Assumptions 2 to 9, as* $n \to \infty$, $\lambda \to 0$, *and* $n^{-1}\lambda^{-\frac{1}{r}} \to 0$,

$$\mathrm{SKL}(\tau_0, \hat{\tau}) = O_p\left(n^{-1}\lambda^{-\frac{1}{r}} + \lambda\right).$$

Proof. The definition of the Fréchet derivative of $\eta(\theta, h)$ at (θ_0, h_0) gives

$$\begin{aligned}
&\eta(\hat{\theta}, \hat{h}) - \eta(\theta_0, h_0) \\
&\quad = D_\theta \eta(\theta_0, h_0)(\hat{\theta} - \theta_0) + D_h \eta(\theta_0, h_0)(\hat{h} - h_0) + o(\|(\hat{\theta} - \theta_0, \hat{h} - h_0)\|_{\mathcal{Q}}).
\end{aligned}$$

By Theorem 5, $\|(\hat{\theta} - \theta_0, \hat{h} - h_0)\|_{\mathcal{Q}} \to 0$ with probability tending to 1 as $\lambda \to 0$, and $n^{-1}\lambda^{-\frac{1}{r}} \to 0$. Therefore,

$$\begin{aligned}
\mathrm{SKL}(\tau_0, \hat{\tau}) &= \mu_{\tau_0}[\eta(\theta_0, h_0) - \eta(\hat{\theta}, \hat{h})] + \mu_{\hat{\tau}}[\eta(\hat{\theta}, \hat{h}) - \eta(\theta_0, h_0)] \\
&\to \Big\{\mu_{\tau_0}[D_\theta \eta(\tau_0)(\theta_0 - \hat{\theta}) + D_h \eta(\tau_0)(h_0 - \hat{h})] \\
&\qquad -\mu_{\hat{\tau}}[D_\theta \eta(\tau_0)(\theta_0 - \hat{\theta}) + D_h \eta(\tau_0)(h_0 - \hat{h})]\Big\} \qquad \text{(IV.12)} \\
&= V_{\tau_*}[D_\theta \eta(\tau_0)(\theta_0 - \hat{\theta}) + D_h \eta(\tau_0)(h_0 - \hat{h}), \\
&\qquad\qquad D_\theta \eta(\tau_*)(\theta_0 - \hat{\theta}) + D_h \eta(\tau_*)(h_0 - \hat{h})],
\end{aligned}$$

where $\tau_* = \hat{\tau} + u(\tau_0 - \hat{\tau})$ for some $0 \le u \le 1$. The last equality holds by (IV.11). By the Cauchy-Schwarz inequality, Assumptions 5 and 7, and Theorem 5, (IV.12) is bounded above by

$$C_2 C_4 V_{\tau_0}[L_\theta(\theta_0 - \hat{\theta}) + L_h(h_0 - \hat{h})] = O_p\left(n^{-1}\lambda^{-\frac{1}{r}} + \lambda\right).$$

□

Remark 4. *Note that results in Corollary 1, Corollary 2, and Theorem 6 are independent of the linear operators* L_θ *and* L_h.

IV.4 Extension to the multiple sample case

In the case of multiple samples, assume there are m independent groups, and in each group $l = 1, \ldots, m$, there are n_l i.i.d. observations such that $X_{l1}, \ldots, X_{ln_l} \overset{iid}{\sim} f_l(x; \theta, h)$ on domain $\mathcal{X}_l$. We consider the following general semiparametric density model

$$f_l(x, \theta, h) = \frac{\exp\{\eta_l(x; \theta, h)\}}{\int_{\mathcal{X}_l} \exp\{\eta_l(x; \theta, h)\}\, dx}, \tag{IV.13}$$

where $\eta_l : \mathcal{Q} \to L_0^2(\mathcal{X}_l)$, the logistic transformation of f_l, is a known function $(\theta, h) \in \mathcal{Q}$. We are interested in the estimation of θ, h, and ultimately the overall density function $f_l(x, \theta, h)$ from the observations. The estimator is the local minimizer of the following penalized likelihood

$$\begin{aligned} \sum_{l=1}^{m} \ell_{n_l,\lambda}(\theta, h) &= \sum_{l=1}^{m} \ell_{n_l}(\theta, h) + \frac{\lambda}{2} J(h) \\ &= \sum_{l=1}^{m} \left\{ -\frac{1}{n_l} \sum_{i=1}^{n_l} \eta_l(X_{li}; \theta, h) + \log \int_{\mathcal{X}_l} e^{\eta_l(x;\theta,h)} dx \right\} + \frac{\lambda}{2} J(h), \end{aligned} \tag{IV.14}$$

where $\ell_{n_l}(\theta, h)$ is the negative log likelihood of f_l, $J(h)$ is the roughness penalty term (assumed to be a quadratic functional), and λ is the smoothing parameter. Assumptions 2 to 9 can be adjusted to this multiple sample setting by replacing V_τ with $\sum_{l=1}^{m} V_{l,\tau}$, which is defined as

$$V_{l,\tau}(\varphi_1, \varphi_2) = \mu_{l,\tau}(\varphi_1\varphi_2) - \mu_{l,\tau}(\varphi_1)\mu_{l,\tau}(\varphi_2),$$

where $\mu_{l,\tau}(\varphi_1) = \int \varphi_1(t) e^{\eta_l(t;\tau))} dt / \int e^{\eta_l(t;\tau)} dt$, for any funcitons $\varphi_1(x), \varphi_2(x)$ over the appropriate sample space. Denote $V_{l,\tau}(\varphi_1) = V_{l,\tau}(\varphi_1, \varphi_1)$. The existence of bounded linear operators L_θ and L_h also needs to be extended

to the existence of bounded linear operators $L_{l,\theta} : \mathbb{R}^p \to L_0^2(\mathcal{X}_l)$ and $L_{l,h} : \mathcal{H} \to L_0^2(\mathcal{X}_l)$ for all $l = 1, \ldots, m$. The precise statements of the assumptions are given in the Appendix. We next state our results for density estimation with multiple samples in parallel to those for the single sample case in the previous sections.

Denote $n = \min_{1 \le l \le m}\{n_l\}$ for $m > 1$. The linear approximation estimate $\tilde{\tau} = (\tilde{\theta}, \tilde{h})$ of $\hat{\tau} = (\hat{\theta}, \hat{h})$, where $\tilde{h} = \sum_\nu \tilde{h}_\nu \phi_{0,\nu}$, can be obtained by minimizing $\sum_{l=1}^m \tilde{\ell}_{n_l,\lambda}(\theta, h)$. The calculation is similar to that in Section IV.2. With all assumptions as stated in the Appendix, one can get the following results, which are similar to Lemma 2 and Theorems 4 to 6. Note that one also has corollaries to Theorem 8 which are analogous to Corollaries 1 and 2, but for $m > 1$. They are omitted to save space.

Lemma 6. *As $n \to \infty$ and $\lambda \to 0$,*

$$\mathrm{E}\left\{\sum_{l=1}^m V_{l,\tau_0}[D_\theta \eta_l(\tau_0)(\tilde{\theta} - \theta_0)]\right\} \le c\,\mathrm{E}\left[(\tilde{\theta} - \theta_0)^T(\tilde{\theta} - \theta_0)\right] = O(n^{-1}\lambda^{-\frac{1}{r}}),$$
$$\mathrm{E}\left\{\sum_{l=1}^m V_{l,\tau_0}[D_h \eta_l(\tau_0)(\tilde{h} - h_0)] + \lambda J(\tilde{h} - h_0)\right\} = O(n^{-1}\lambda^{-\frac{1}{r}} + \lambda).$$

Theorem 7. *As $n \to \infty$, $\lambda \to 0$, and $n^{-1}\lambda^{-\frac{1}{r}} \to 0$,*

$$\sum_{l=1}^m V_{l,\tau_0}[L_{l,\theta}(\tilde{\theta} - \theta_0) + L_{l,h}(\tilde{h} - h_0)] + \lambda J(\tilde{h} - h_0) = O_p(n^{-1}\lambda^{-\frac{1}{r}} + \lambda).$$

Theorem 8. *As $n \to \infty$, $\lambda \to 0$, and $n^{-1}\lambda^{-\frac{1}{r}} \to 0$,*

$$\sum_{l=1}^m V_{l,\tau_0}[L_{l,\theta}(\hat{\theta} - \tilde{\theta}) + L_{l,h}(\hat{h} - \tilde{h})] + \lambda J(\hat{h} - \tilde{h}) = O_p\left(n^{-1}\lambda^{-\frac{1}{r}} + \lambda\right).$$

Therefore,

$$\sum_{l=1}^m V_{l,\tau_0}[L_{l,\theta}(\hat{\theta} - \theta_0) + L_{l,h}(\hat{h} - h_0)] + \lambda J(\hat{h} - h_0) = O_p\left(n^{-1}\lambda^{-\frac{1}{r}} + \lambda\right).$$

Theorem 9. *As $n \to \infty$, $\lambda \to 0$, and $n^{-1}\lambda^{-\frac{1}{r}} \to 0$,*

$$\begin{aligned}\mathrm{SKL}(\tau_0, \hat{\tau}) &= \sum_{l=1}^{m} \left\{ \mu_{l,\tau_0}[\eta(\theta_0, h_0) - \eta(\hat{\theta}, \hat{h})] + \mu_{l,\hat{\tau}}[\eta(\hat{\theta}, \hat{h}) - \eta(\theta_0, h_0)] \right\} \\ &= O_p\left(n^{-1}\lambda^{-\frac{1}{r}} + \lambda\right).\end{aligned}$$

CHAPTER V

REGRESSION MODELS

In this chapter, under Assumptions 1 to 9, we study the asymptotic properties for the penalized likelihood estimator and the doubly penalized likelihood estimator for the semiparametric regression model.

Recall that given i.i.d. observed data (Y_i, X_i) for $i = 1, \ldots, n$ of the variables $(Y, X) \in \mathcal{Y} \times \mathcal{X} \subset \mathbb{R} \times \mathbb{R}^d$, we consider a general class of semiparametric regression models for which

$$\mu_0(X) \equiv \mathrm{E}[Y|X] = g[\eta(X; \theta_0, h_0)],$$

where $g(\cdot)$ is a known link function, $\eta : \mathcal{Q} \to L^4(\mathcal{X})$ is a known function that represent the relationship between the parametric and nonparametric components of the model, and (θ_0, h_0) is the true parameter.

In Section V.1, we first consider the semiparametric penalized likelihood estimator

$$(\hat{\theta}, \hat{h}) = \underset{(\theta,h)\in\mathcal{N}_{\theta_0}\times\mathcal{N}_{h_0}}{\arg\min} \ell_{n,\lambda}(\theta, h),$$

where

$$\begin{aligned} \ell_{n,\lambda}(\theta, h \mid \text{data}) &= \ell_n(\eta(\theta, h) \mid \text{data}) + \frac{\lambda}{2} J(h) \\ &= \frac{1}{n} \sum_{i=1}^{n} l(Y_i; \eta(X_i; \theta, h)) + \frac{\lambda}{2} J(h), \end{aligned} \tag{V.1}$$

and $l(y; a) : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ is a general criterion function representing either $-\log p(y; g(a))$, the negative log likelihood given by the conditional distribution p, or $-Q(y; g(a))$, the negative quasi-likelihood, which were discussed

in Chapter II. Recall the open interval R that contains the range R_0 of $\eta(x;\theta_0,h_0)$ as defined in Assumption 1. We choose $\mathcal{N}_{\theta_0}\times\mathcal{N}_{h_0}$ such that for any $(\theta,h)\in\mathcal{N}_{\theta_0}\times\mathcal{N}_{h_0}$, the range of $\eta(x;\theta,h)$ is contained in R as well.

The outline of the proof of consistency in terms of parameter estimation given by (V.1) is similar to the proof for density estimation in Chapter IV. We will first define an "quadratic" approximation $\tilde{\ell}_{n,\lambda}(\theta,h)$ of the penalized likelihood and derive the rate of convergence for $\tilde{\tau}-\tau_0$, where $\tilde{\tau}$ is the minimizer of $\tilde{\ell}_{n,\lambda}(\theta,h)$ (Section V.1.1). An upper bound for the approximation error $\hat{\tau}-\tilde{\tau}$ is then derived, and the rate of convergence for $\hat{\tau}-\tau_0$ follows by the triangle inequality (Section V.1.2).

If the dimension of parameter $\theta\in\mathbb{R}^p$ is large and the underlying true parameter θ_0 is sparse, one may be interested in a parameter reduction process to identify the significant parameters. In Section V.2, we will consider the asymptotic properties of a doubly penalized likelihood estimator, that is, with penalties on both the nonparametric function h and the parameter θ. Our proposed estimation method is motivated by the class of sparsity-encouraged regularization methods developed for variable selection problems for models with partially linear structures. In particular, we consider the following doubly penalized likelihood function

$$\begin{aligned} L_n(\theta,h) &= \ell_{n,\lambda}(\theta,h)+\sum_{j=1}^{p} q_{\lambda_j}(\left|\theta^j\right|) \\ &= \frac{1}{n}\sum_{i=1}^{n} l(Y_i;\eta(X_i;\theta,h))+\frac{\lambda}{2}J(h)+\sum_{j=1}^{p} q_{\lambda_j}(|\theta^j|), \end{aligned} \tag{V.2}$$

where $\ell_{n,\lambda}(\theta,h)$ is as defined in equation (V.1), $q_{\lambda_j}(\cdot)$ are specified penalty functions with regularization parameters $\lambda_j\geq 0$, and $(\theta,h)\in\mathcal{Q}=\mathbb{R}^p\times\mathcal{H}$ are the parameter to be estimated.

We denote the estimator minimizing (V.2) above by

$$\hat{\tau}_D = (\hat{\theta}_D, \hat{h}_D) = \underset{(\theta,h)\in\mathcal{Q}}{\arg\min}\, L_n(\theta, h).$$

We note that when $\sum_{j=1}^{p} q_{\lambda_j}(|\theta^j|)$ is continuous in θ and locally convex in $\mathcal{N}_{\theta_0}$, the local existence and uniqueness of $\hat{\tau}_D$ can be established following the same ideas as in the existence and uniqueness proof discussed previously. Since nonconvex penalities such as SCAD (Fan and Li, 2001) and bridge penalties for $0 < q < 1$ (Frank and Friedman, 1993) could be of interest, we do not assume convexity of $\sum_{j=1}^{p} q_{\lambda_j}(|\theta^j|)$ for our analysis in this section, and we will establish local existence of a joint consistent estimator in Theorem 12. Since the estimation and dimension reduction of the parameters θ are of major interest in this model, conditioning on the joint consistency result, we further prove the rate of convergence for $\hat{\theta}_D$ in terms of parameter estimation in Theorem 13 and model selection consistency in Theorem 14.

Throughout this chapter, we will also need the following assumption for a uniform bound on the fourth moments of $\phi_{*,\nu}$, which are the basis functions from the spectral decomposition as discussed in Section II.2.3. As noted in Gu (2013), such a condition appears mild as $\phi_{*,\nu}$ typically grow in roughness but not necessarily in magnitude. When $\eta(\theta, h)$ is linear additive in h, a stronger condition is usually assumed, such as a uniform bound on the L^∞ norm on $\phi_{*,\nu}$ (see, for example, Cheng and Shang (2015)).

Assumption 10. *For a fixed $(\theta_*, h_*) \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$, let $\{\phi_{*,\nu} : \nu = 1, 2, \ldots\}$ be the eigenfunctions as defined above. There exists a constant $C_{\tau_*} < \infty$ such that*

$$\|D_h \eta(\theta_*, h_*)\phi_{*,\nu}\|_{L^4} < C_{\tau_*},$$

for all $\nu \in \mathbb{N}$.

V.1 Consistency of the penalized likelihood estimator

In this section, assuming the existence of $\hat{\tau} = (\hat{\theta}, \hat{h}) \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$, we study the consistency and rate of convergence of the estimator $\hat{\tau}$ given by the penalized likelihood (V.1).

V.1.1 Linear approximation

Consider the quadratic functional

$$\begin{aligned}\tilde{\ell}_{n,\lambda}(\theta, h) = \frac{1}{n}\sum_{i=1}^{n} l_a'(Y_i; \eta(X_i;\tau_0))\left[D_h(X_i,\tau_0)h + D_\theta(X_i,\tau_0)\theta\right] \\ + \frac{1}{2}V_{\tau_0}\left[D_h(X_i,\tau_0)(h-h_0) + D_\theta(X_i,\tau_0)(\theta-\theta_0)\right] + \frac{\lambda}{2}J(h).\end{aligned} \tag{V.3}$$

Let $h_\nu = V_{\tau_0}[D_h\eta(\tau_0)h, D_h\eta(\tau_0)\phi_{0,\nu}]$ and $h_{0,\nu} = V_{\tau_0}[D_h\eta(\tau_0)h_0, D_h\eta(\tau_0)\phi_{0,\nu}]$, and plug the Fourier series expansions $h = \sum_\nu h_\nu\phi_{0,\nu}$ and $h_0 = \sum_\nu h_{0,\nu}\phi_{0,\nu}$ into equation (V.3),

$$\begin{aligned}\tilde{\ell}_{n,\lambda}(\theta, h) = \theta^T\left\{\frac{1}{n}\sum_{i=1}^{n} l_a'(Y_i;\eta(X_i;\tau_0))D_\theta\eta(X_i;\tau_0)\right\} \\ + \sum_\nu h_\nu\left\{\frac{1}{n}\sum_{i=1}^{n} l_a'(Y_i;\eta(X_i;\tau_0))D_h\eta(X_i;\tau_0)\phi_{0,\nu}\right\} \\ + \frac{1}{2}(\theta-\theta_0)^T V_{\tau_0}\left[D_\theta\eta(\tau_0)\right](\theta-\theta_0) + \frac{1}{2}\sum_\nu (h_\nu - h_{0,\nu})^2 \\ + \sum_\nu (h_\nu - h_{0,\nu})(\theta-\theta_0)^T V_{\tau_0}\left[D_\theta\eta(\tau_0), D_h\eta(\tau_0)\phi_{0,\nu}\right] \\ + \frac{\lambda}{2}\sum_\nu \rho_{0,\nu}h_\nu^2.\end{aligned} \tag{V.4}$$

Similar to the density estimation case, we denote by $D_\theta\eta(x;\tau_0)$ both the linear operator and its vector form $[D_{\theta^k}\eta(x;\tau_0)]_{k=1}^p$. Moreover, $V_{\tau_0}[D_\theta\eta(\tau_0)]$ is the $p \times p$ covariance matrix whose (i,j)th entry is

$$V_{\tau_0}[D_{\theta^i}\eta(\tau_0), D_{\theta^j}\eta(\tau_0)] = \mathrm{E}_X\left[I_{\tau_0}(X)D_{\theta^i}\eta(\tau_0)D_{\theta^j}\eta(\tau_0)\right],$$

and $V_{\tau_0}[D_\theta\eta(\tau_0), D_h\eta(\tau_0)\phi_{0,\nu}]$ is the $p \times 1$ vector whose ith entry is

$$V_{\tau_0}[D_{\theta^i}\eta(\tau_0), D_h\eta(\tau_0)\phi_{0,\nu}] = \mathrm{E}_X\left[I_{\tau_0}(X) D_{\theta^i} D_h\eta(\tau_0)\phi_{0,\nu}\right].$$

Let

$$\bar{\alpha}_n = -\frac{1}{n}\sum_{i=1}^{n} l_a'(Y_i; \eta(X_i;\tau_0)) D_\theta\eta(X_i;\tau_0),$$
$$\bar{\beta}_{\nu,n} = -\frac{1}{n}\sum_{i=1}^{n} l_a'(Y_i; \eta(X_i;\tau_0)) D_h\eta(X_i;\tau_0)\phi_{0,\nu}.$$

Solving the system $D_\theta\tilde{\ell}_{n,\lambda}(\theta,h) = 0$, $D_{h_\nu}\tilde{\ell}_{n,\lambda}(\theta,h) = 0$, we get

$$\tilde{\theta} = \theta_0 + \Omega_\lambda^{-1}\left\{\bar{\alpha}_n - \sum_\nu \frac{\bar{\beta}_{\nu,n} - \lambda\rho_{0,\nu}h_{0,\nu}}{1+\lambda\rho_{0,\nu}} V_{\tau_0}[D_\theta\eta(\tau_0), D_h\eta(\tau_0)\phi_{0,\nu}]\right\},$$
$$\tilde{h}_\nu = \frac{\bar{\beta}_{\nu,n} + h_{0,\nu}}{1+\lambda\rho_{0,\nu}} - (\tilde{\theta}-\theta_0)^T \frac{V_{\tau_0}[D_\theta\eta(\tau_0), D_h\eta(\tau_0)\phi_{0,\nu}]}{1+\lambda\rho_{0,\nu}},$$

where

$$\Omega_\lambda = V_{\tau_0}[D_\theta\eta(\tau_0)] - \sum_\nu \frac{V_{\tau_0}[D_\theta\eta(\tau_0), D_h\eta(\tau_0)\phi_{0,\nu}]^{\otimes 2}}{1+\lambda\rho_{0,\nu}}.$$

Hence, $\tilde{\tau} = (\tilde{\theta}, \tilde{h})$, where $\tilde{h} = \sum_\nu \tilde{h}_\nu\phi_{0,\nu}$, is the minimizer of (V.3). Since for fixed $\tilde{\theta}$, $\tilde{h}_\nu$ is linear in $\phi_{0,\nu}$, analogously to the density estimation proof, we refer $\tilde{\tau}$ as the linear approximation of $\hat{\tau}$. By Assumption 1, we note that

$$\mathrm{E}(\bar{\beta}_{\nu,n}) = -\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_X\left[\mathrm{E}[l_a'(Y_i;\eta(X_i;\tau_0))|X_i] D_h\eta(X_i;\tau_0)\phi_{0,\nu}\right] = 0,$$
$$\begin{aligned}\mathrm{E}(\bar{\beta}_{\nu,n}^2) &= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{E}_X\left[\mathrm{E}[l_a'(Y_i;\eta(X_i;\tau_0))^2|X_i]\left(D_h\eta(X_i;\tau_0)\phi_{0,\nu}\right)^2\right] \\ &= \frac{\sigma^2}{n}\mathrm{E}_X\left[I_{\tau_0}(X)\left(D_h\eta(X;\tau_0)\phi_{0,\nu}\right)^2\right] \\ &= \frac{\sigma^2}{n} = O(n^{-1}),\end{aligned}$$

$$
\begin{aligned}
\mathrm{E}(\bar{\alpha}_n) &= -\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_X\left[\mathrm{E}[l_a'(Y_i;\eta(X_i;\tau_0))|X_i] D_\theta \eta(X_i;\tau_0)\right] = 0, \\
\mathrm{E}(\|\bar{\alpha}_n\|_{l^2}^2) &= \sum_{k=1}^{p}\frac{1}{n^2}\sum_{i=1}^{n} \mathrm{E}_X\left[\mathrm{E}[l_a'(Y_i;\eta(X_i;\tau_0))^2|X_i]\left(D_{\theta^k}\eta(X_i;\tau_0)\right)^2\right] \\
&= \sum_{k=1}^{p}\frac{\sigma^2}{n}\mathrm{E}_X\left[I_{\tau_0}(X)\left(D_{\theta^k}\eta(X_i;\tau_0)\right)^2\right] \\
&= O(\frac{p\sigma^2}{n}) = O(n^{-1}).
\end{aligned}
$$

The following lemma and theorem can be easily proven analogously to Lemma 2 and Theorem 4 in Chapter IV.

Lemma 7. *Under Assumptions 1 to 4, 8, and 9, as $n \to \infty$ and $\lambda \to 0$,*

$$
\begin{aligned}
&\mathrm{E}\left\{V_{\tau_0}[D_\theta\eta(\tau_0)(\tilde{\theta}-\theta_0)]\right\} \le c\,\mathrm{E}\left[(\tilde{\theta}-\theta_0)^T(\tilde{\theta}-\theta_0)\right] = O(n^{-1}\lambda^{-\frac{1}{r}}), \\
&\mathrm{E}\left\{V_{\tau_0}[D_h\eta(\tau_0)(\tilde{h}-h_0)] + \lambda J(\tilde{h}-h_0)\right\} = O(n^{-1}\lambda^{-\frac{1}{r}}+\lambda).
\end{aligned}
$$

Theorem 10. *Under Assumptions 1 to 5, 8, and 9, as $n \to \infty$ and $\lambda \to 0$,*

$$
V_{\tau_0}[L_\theta(\tilde{\theta}-\theta_0) + L_h(\tilde{h}-h_0)] + \lambda J(\tilde{h}-h_0) = O_p(n^{-1}\lambda^{-\frac{1}{r}}+\lambda).
$$

Note that as $n^{-1}\lambda^{-\frac{1}{r}} \to 0$, Theorem 10 implies that $\tilde{\tau} \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$ with probability tending to one. We restrict our attention to this event for the rest of the analysis, or for simplicity, we assume $\tilde{\tau} \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$.

V.1.2 Approximation error and main results

We now proceed to the analysis for the approximation error $\hat{\tau} - \tilde{\tau}$. For any $\tau_1 = (\theta_1, h_1)$ and $\tau_2 = (\theta_2, h_2)$, define

$$
\begin{aligned}
A^\theta_{\tau_1,\tau_2}(\alpha) &= \ell_{n,\lambda}(\theta_1 + \alpha\theta_2, h_1), \quad A^h_{\tau_1,\tau_2}(\alpha) = \ell_{n,\lambda}(\theta_1, h_1 + \alpha h_2), \\
B^\theta_{\tau_1,\tau_2}(\alpha) &= \tilde{\ell}_{n,\lambda}(\theta_1 + \alpha\theta_2, h_1), \quad B^h_{\tau_1,\tau_2}(\alpha) = \tilde{\ell}_{n,\lambda}(\theta_1, h_1 + \alpha h_2).
\end{aligned}
$$

It can be easily shown that the derivatives with respect to α of the above equations evaluated at $\alpha = 0$ are

$$\begin{aligned}
\dot{A}^{\theta}_{\tau_1,\tau_2}(0) &= \frac{dA^{\theta}_{\tau_1,\tau_2}}{d\alpha}(0) = \frac{1}{n}\sum_{i=1}^{n} l_a^{'}(Y_i;\eta(X_i;\tau_1))D_\theta\eta(X_i;\tau_1)\theta_2, \\
\dot{A}^{h}_{\tau_1,\tau_2}(0) &= \frac{dA^{h}_{\tau_1,\tau_2}}{d\alpha}(0) = \frac{1}{n}\sum_{i=1}^{n} l_a^{'}(Y_i;\eta(X_i;\tau_1))D_h\eta(X_i;\tau_1)h_2 + \lambda J(h_1,h_2), \\
\dot{B}^{\theta}_{\tau_1,\tau_2}(0) &= \frac{dB^{\theta}_{\tau_1,\tau_2}}{d\alpha}(0) = \frac{1}{n}\sum_{i=1}^{n} l_a^{'}(Y_i;\eta(X_i;\tau_0))D_\theta\eta(X_i;\tau_0)\theta_2 \\
&\quad + V_{\tau_0}[D_\theta\eta(\tau_0)(\theta_1-\theta_0), D_\theta\eta(\tau_0)\theta_2] \\
&\quad + V_{\tau_0}[D_h\eta(\tau_0)(h_1-h_0), D_\theta\eta(\tau_0)\theta_2], \\
\dot{B}^{h}_{\tau_1,\tau_2}(0) &= \frac{dB^{h}_{\tau_1,\tau_2}}{d\alpha}(0) = \frac{1}{n}\sum_{i=1}^{n} l_a^{'}(Y_i;\eta(X_i;\tau_0))D_h\eta(X_i;\tau_0)h_2 \\
&\quad + V_{\tau_0}[D_h\eta(\tau_0)(h_1-h_0), D_h\eta(\tau_0)h_2] \\
&\quad + V_{\tau_0}[D_\theta\eta(\tau_0)(\theta_1-\theta_0), D_h\eta(\tau_0)h_2] + \lambda J(h_1,h_2).
\end{aligned}$$

By definition, $\dot{A}^{\theta}_{\hat{\tau},\tau_2}(0) + \dot{A}^{h}_{\hat{\tau},\tau_2}(0) = 0$ and $\dot{B}^{\theta}_{\tilde{\tau},\tau_2}(0) + \dot{B}^{h}_{\tilde{\tau},\tau_2}(0) = 0$ for any $\tau_2 \in \mathcal{Q}$. Setting $\tau_2 = \hat{\tau} - \tilde{\tau}$, we get

$$\begin{aligned}
&\dot{A}^{\theta}_{\hat{\tau},\tau_2}(0) + \dot{A}^{h}_{\hat{\tau},\tau_2}(0) \\
&= \frac{1}{n}\sum_{i=1}^{n} l_a^{'}(Y_i;\eta(X_i;\hat{\tau}))\left[D_\theta\eta(X_i;\hat{\tau})(\hat{\theta}-\tilde{\theta}) + D_h\eta(X_i;\hat{\tau})(\hat{h}-\tilde{h})\right] \qquad \text{(V.5)} \\
&\quad + \lambda J(\hat{h}, \hat{h}-\tilde{h}) = 0,
\end{aligned}$$

$$
\begin{aligned}
&\dot{B}^{\theta}_{\tilde{\tau},\tau_2}(0) + \dot{B}^{h}_{\tilde{\tau},\tau_2}(0) \\
&= \frac{1}{n}\sum_{i=1}^{n} l_a^{'}(Y_i;\eta(X_i;\tau_0))\left[D_\theta\eta(X_i;\tau_0)(\hat{\theta}-\tilde{\theta}) + D_h\eta(X_i;\tau_0)(\hat{h}-\tilde{h})\right] \\
&\quad + V_{\tau_0}[D_\theta\eta(\tau_0)(\tilde{\theta}-\theta_0) + D_h\eta(\tau_0)(\tilde{h}-h_0), D_\theta\eta(\tau_0)(\hat{\theta}-\tilde{\theta}) + D_h\eta(\tau_0)(\hat{h}-\tilde{h})] \\
&\quad + \lambda J(\tilde{h}, \hat{h}-\tilde{h}) = 0.
\end{aligned}
\tag{V.6}
$$

Subtracting (V.6) from (V.5), and after some algebraic manipulation, one gets

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[l_a^{'}(Y_i;\eta(X_i;\hat{\tau})) - l_a^{'}(Y_i;\eta(X_i;\tilde{\tau}))\right] \\
&\qquad\qquad \cdot \left[D_\theta\eta(X_i;\hat{\tau})(\hat{\theta}-\tilde{\theta}) + D_h\eta(X_i;\hat{\tau})(\hat{h}-\tilde{h})\right] + \lambda J(\hat{h}-\tilde{h}) \\
&= V_{\tau_0}[D_\theta\eta(\tau_0)(\tilde{\theta}-\theta_0) + D_h\eta(\tau_0)(\tilde{h}-h_0), D_\theta\eta(\tau_0)(\hat{\theta}-\tilde{\theta}) + D_h\eta(\tau_0)(\hat{h}-\tilde{h})] \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\left[l_a^{'}(Y_i;\eta(X_i;\tau_0)) - l_a^{'}(Y_i;\eta(X_i;\tilde{\tau}))\right] \\
&\qquad\qquad \cdot \left[D_\theta\eta(X_i;\hat{\tau})(\hat{\theta}-\tilde{\theta}) + D_h\eta(X_i;\hat{\tau})(\hat{h}-\tilde{h})\right] \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} l_a^{'}(Y_i;\eta(X_i;\tau_0))\left[D_\theta\eta(X_i;\tau_0)(\hat{\theta}-\tilde{\theta}) + D_h\eta(X_i;\tau_0)(\hat{h}-\tilde{h})\right] \\
&\quad - \frac{1}{n}\sum_{i=1}^{n} l_a^{'}(Y_i;\eta(X_i;\tau_0))\left[D_\theta\eta(X_i;\hat{\tau})(\hat{\theta}-\tilde{\theta}) + D_h\eta(X_i;\hat{\tau})(\hat{h}-\tilde{h})\right].
\end{aligned}
\tag{V.7}
$$

For any $\tau_1, \tau_2 \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$, let

$$
C(u) = l_a^{'}\{Y_i;\eta[X_i;\tau_1 + u(\tau_2-\tau_1)]\}.
$$

By the mean value theorem, for some $0 < u < 1$, we have

$$
\begin{aligned}
&l_a^{'}(Y_i;\eta(X_i;\tau_2)) - l_a^{'}(Y_i;\eta(X_i;\tau_1)) \\
&\quad = C(1) - C(0) = C'(u) \\
&\quad = l_a^{''}(Y_i;\eta(X_i;\tau_u))\left[D_\theta\eta(X_i;\tau_u)(\theta_2-\theta_1) + D_h\eta(X_i;\tau_u)(h_2-h_1)\right],
\end{aligned}
$$

where $\tau_u = \tau_1 + u(\tau_2 - \tau_1)$. Using this fact, we see that the first term on the left hand side and second term on the right hand side of (V.7) become

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[l_a'(Y_i;\eta(X_i;\hat{\tau})) - l_a'(Y_i;\eta(X_i;\tilde{\tau}))\right] \\
&\qquad\qquad \cdot \left[D_\theta\eta(X_i;\hat{\tau})(\hat{\theta}-\tilde{\theta}) + D_h\eta(X_i;\hat{\tau})(\hat{h}-\tilde{h})\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} l_a''(Y_i;\eta(X_i;\tau_*))\left[D_\theta\eta(X_i;\tau_*)(\hat{\theta}-\tilde{\theta}) + D_h\eta(X_i;\tau_*)(\hat{h}-\tilde{h})\right] \\
&\qquad\qquad \cdot \left[D_\theta\eta(X_i;\hat{\tau})(\hat{\theta}-\tilde{\theta}) + D_h\eta(X_i;\hat{\tau})(\hat{h}-\tilde{h})\right],
\end{aligned} \tag{V.8}$$

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[l_a'(Y_i;\eta(X_i;\tau_0)) - l_a'(Y_i;\eta(X_i;\tilde{\tau}))\right] \\
&\qquad\qquad \cdot \left[D_\theta\eta(X_i;\hat{\tau})(\hat{\theta}-\tilde{\theta}) + D_h\eta(X_i;\hat{\tau})(\hat{h}-\tilde{h})\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} l_a''(Y_i;\eta(X_i;\tau_{**}))\left[D_\theta\eta(X_i;\tau_{**})(\theta_0-\tilde{\theta}) + D_h\eta(X_i;\tau_{**})(h_0-\tilde{h})\right] \\
&\qquad\qquad \cdot \left[D_\theta\eta(X_i;\hat{\tau})(\hat{\theta}-\tilde{\theta}) + D_h\eta(X_i;\hat{\tau})(\hat{h}-\tilde{h})\right],
\end{aligned} \tag{V.9}$$

respectively, where $\tau_* = \tilde{\tau} + u(\hat{\tau} - \tilde{\tau})$ and $\tau_{**} = \tilde{\tau} + u(\tau_0 - \tilde{\tau})$. Before we state and prove the main theorem for the approximation error, we first establish the following lemma.

Lemma 8. *Under Assumptions 1 to 10, for any* $\tau, \tau_1, \tau_2 \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$ *and* $\tau_3, \tau_4 \in \mathcal{Q}$,

1. *As* $\lambda \to 0$, $n \to \infty$, *and* $n\lambda^{\frac{1}{r}} \to \infty$,

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n} l_a'(Y_i;\eta(X_i;\tau_0))\left[D_\theta\eta(X_i;\tau_1)\theta_3 + D_h(X_i;\tau_1)h_3\right]\right| \\
&\qquad = o_p\left\{\left[V_{\tau_0}(L_\theta\theta_3 + L_h h_3) + \lambda J(h_3)\right]^{\frac{1}{2}}\right\}.
\end{aligned} \tag{V.10}$$

2. As $\lambda \to 0$, $n \to \infty$, and $n\lambda^{\frac{2}{r}} \to \infty$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} l_a''(Y_i;\eta(X_i;\tau))\left[D_\theta\eta(X_i;\tau_1)\theta_3 + D_h\eta(X_i;\tau_1)h_3\right] \\
&\qquad\qquad \cdot \left[D_\theta\eta(X_i;\tau_2)\theta_4 + D_h\eta(X_i;\tau_2)h_4\right] \\
&= V_\tau\left[D_\theta\eta(\tau_1)\theta_3 + D_h\eta(\tau_1)h_3, D_\theta\eta(\tau_2)\theta_4 + D_h\eta(\tau_2)h_4\right] \\
&\quad + o_p\left\{\left[V_{\tau_0}(L_\theta\theta_3 + L_h h_3) + \lambda J(h_3)\right]^{\frac{1}{2}}\left[V_{\tau_0}(L_\theta\theta_4 + L_h h_4) + \lambda J(h_4)\right]^{\frac{1}{2}}\right\}
\end{aligned} \tag{V.11}$$

Proof. Let $\phi_{1,\nu}, \phi_{2,\nu}$ be the eigenfunctions and $\rho_{1,\nu}, \rho_{2,\nu}$ be the eigenvalues, for $\nu \in \mathbb{N}$, corresponding to τ_1, τ_2, respectively. We also denote

$$h_{3,\nu} = V_{\tau_0}[D_h\eta(\tau_1)h_3, D_h\eta(\tau_1)\phi_{1,\nu}] \quad \text{and} \quad h_{4,\nu} = V_{\tau_0}[D_h\eta(\tau_2)h_4, D_h\eta(\tau_2)\phi_{2,\nu}].$$

For (V.10), since

$$\begin{aligned}
&\mathrm{E}\left\{\left[\frac{1}{n}\sum_{i=1}^{n} l_a'(Y_i;\eta(X_i;\tau_0))D_{\theta^k}\eta(X_i;\tau_1)\right]^2\right\} \\
&= \frac{1}{n}\,\mathrm{E}_X\left\{\mathrm{E}\left[l_a'(Y;\eta(X;\tau_0))^2|X\right] D_{\theta^k}\eta(X_i;\tau_1)^2\right\} \\
&= \frac{\sigma}{n}\,\mathrm{E}_X\left\{I_{\tau_0}(X)D_{\theta^k}\eta(X_i;\tau_1)^2\right\} \\
&= O(n^{-1}),
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{E}\left\{\left[\frac{1}{n}\sum_{i=1}^{n} l_a'(Y_i;\eta(X_i;\tau_0))D_h\eta(X_i;\tau_1)\phi_{1,\nu}\right]^2\right\} \\
&= \frac{1}{n}\,\mathrm{E}_X\left\{\mathrm{E}\left[l_a'(Y;\eta(X;\tau_0))^2|X\right]\left(D_h\eta(X_i;\tau_1)\phi_{1,\nu}\right)^2\right\} \\
&= \frac{\sigma}{n}\,\mathrm{E}_X\left\{I_{\tau_0}(X)\left(D_h\eta(X_i;\tau_1)\phi_{1,\nu}\right)^2\right\} \\
&= \sigma(n^{-1}),
\end{aligned}$$

by the Cauchy-Schwarz inequality and Lemma 9.1 in Gu (2013), we have

$$
\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n} l_a^{'}(Y_i;\eta(X_i;\tau_0))\left[D_\theta\eta(X_i;\tau_1)\theta_3 + D_h(X_i;\tau_1)h_3\right]\right| \\
&\leq \left|\sum_{k=1}^{p}\theta_3^k\left[\frac{1}{n}\sum_{i=1}^{n} l_a^{'}(Y_i;\eta(X_i;\tau_0))D_{\theta^k}\eta(X_i;\tau_1)\right]\right| \\
&\quad + \left|\sum_{\nu=1}^{\infty} h_{3,\nu}\left[\frac{1}{n}\sum_{i=1}^{n} l_a^{'}(Y_i;\eta(X_i;\tau_0))D_h\eta(X_i;\tau_1)\phi_{1,\nu}\right]\right| \\
&\leq \left\{\sum_{k=1}^{p}\left(\theta_3^k\right)^2\right\}^{\frac{1}{2}}\left\{\sum_{k=1}^{p}\left[\frac{1}{n}\sum_{i=1}^{n} l_a^{'}(Y_i;\eta(X_i;\tau_0))D_{\theta^k}\eta(X_i;\tau_1)\right]^2\right\}^{\frac{1}{2}} \\
&\quad + \left\{\sum_{\nu=1}^{\infty}\frac{1}{1+\lambda\rho_{1,\nu}}\left[\frac{1}{n}\sum_{i=1}^{n} l_a^{'}(Y_i;\eta(X_i;\tau_0))D_h\eta(X_i;\tau_1)\phi_{1,\nu}\right]^2\right\}^{\frac{1}{2}} \\
&\qquad \cdot\left\{\sum_{\nu=1}^{\infty}(1+\lambda\rho_{1,\nu})h_{3,\nu}^2\right\}^{\frac{1}{2}} \\
&= O_p\left(n^{-\frac{1}{2}}\right)\|\theta_3\|_{l^2} + O_p\left(n^{-\frac{1}{2}}\lambda^{-\frac{1}{2r}}\right)\|h\|_{\mathcal{H}} \\
&= o_p\left\{\left[V_{\tau_0}(L_\theta\theta_3 + L_h h_3) + \lambda J(h_3)\right]^{\frac{1}{2}}\right\}.
\end{aligned}
$$

The last equality holds because all norms on $\mathbb{R}^p$ are equivalent to the Euclidean norm and $\|\cdot\|_{\mathcal{Q},1}$ is equivalent to $\|\cdot\|_{\mathcal{Q}}$.

For (V.11), we first write

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} l_a''(Y_i;\eta(X_i;\tau))\left[D_\theta\eta(X_i;\tau_1)\theta_3 + D_h\eta(X_i;\tau_1)h_3\right] \\
&\qquad\qquad \cdot \left[D_\theta\eta(X_i;\tau_2)\theta_4 + D_h\eta(X_i;\tau_2)h_4\right] \\
&\quad - \mathrm{E}\left\{ l_a''(Y;\eta(X;\tau))\left[D_\theta\eta(X;\tau_1)\theta_3 + D_h\eta(X;\tau_1)h_3\right]\right. \\
&\qquad\qquad \left. \cdot \left[D_\theta\eta(X;\tau_2)\theta_4 + D_h\eta(X;\tau_2)h_4\right]\right\} \\
&= \left\{ \frac{1}{n}\sum_{i=1}^{n} l_a''(Y_i;\eta(X_i;\tau))D_\theta\eta(X_i;\tau_1)\theta_3 D_\theta\eta(X;\tau_2)\theta_4 \right. \\
&\qquad \left. - \mathrm{E}\left[l_a''(Y;\eta(X;\tau))D_\theta\eta(X;\tau_1)\theta_3 D_\theta\eta(X;\tau_2)\theta_4\right]\right\} \\
&+ \left\{ \frac{1}{n}\sum_{i=1}^{n} l_a''(Y_i;\eta(X_i;\tau))D_h\eta(X_i;\tau_1)h_3 D_h\eta(X;\tau_2)h_4 \right. \\
&\qquad \left. - \mathrm{E}\left[l_a''(Y;\eta(X;\tau))D_h\eta(X;\tau_1)h_3 D_h\eta(X;\tau_2)h_4\right]\right\} \\
&+ \left\{ \frac{1}{n}\sum_{i=1}^{n} l_a''(Y_i;\eta(X_i;\tau))D_\theta\eta(X_i;\tau_1)\theta_3 D_h\eta(X;\tau_2)h_4 \right. \\
&\qquad \left. - \mathrm{E}\left[l_a''(Y;\eta(X;\tau))D_\theta\eta(X;\tau_1)\theta_3 D_h\eta(X;\tau_2)h_4\right]\right\} \\
&+ \left\{ \frac{1}{n}\sum_{i=1}^{n} l_a''(Y_i;\eta(X_i;\tau))D_h\eta(X_i;\tau_1)h_3 D_\theta\eta(X;\tau_2)\theta_4 \right. \\
&\qquad \left. - \mathrm{E}\left[l_a''(Y;\eta(X;\tau))D_h\eta(X;\tau_1)h_3 D_\theta\eta(X;\tau_2)\theta_4\right]\right\} \\
&= (I) + (II) + (III) + (IV),
\end{aligned}
\tag{V.12}
$$

and we will derive a bound for each term on the right hand side of the above equation.

For (I), since for any $1 \le k, j \le p$,

$$
\begin{aligned}
&\operatorname{Var}\left[l_a''(Y;\eta(X;\tau))D_{\theta^k}\eta(X;\tau_1)D_{\theta^j}\eta(X;\tau_2)\right] \\
&\leq \mathrm{E}\left[l_a''(Y;\eta(X;\tau))^2 D_{\theta^k}\eta(X;\tau_1)^2 D_{\theta^j}\eta(X;\tau_2)^2\right] \\
&\leq \left\{\mathrm{E}\left[l_a''(Y;\eta(X;\tau))^2 D_{\theta^k}\eta(X;\tau_1)^4\right]\right\}^{\frac{1}{2}} \left\{\mathrm{E}\left[l_a''(Y;\eta(X;\tau))^2 D_{\theta^j}\eta(X;\tau_2)^4\right]\right\}^{\frac{1}{2}} \\
&= \left\{\mathrm{E}_X\left[\mathrm{E}\left(l_a''(Y;\eta(X;\tau))^2|X\right) D_{\theta^k}\eta(X;\tau_1)^4\right]\right\}^{\frac{1}{2}} \\
&\qquad \cdot \left\{\mathrm{E}_X\left[\mathrm{E}\left(l_a''(Y;\eta(X;\tau))^2|X\right) D_{\theta^j}\eta(X;\tau_2)^4\right]\right\}^{\frac{1}{2}} \\
&\leq C_0 \left\{\mathrm{E}_X\left[D_{\theta^k}\eta(X;\tau_1)^4\right]\right\}^{\frac{1}{2}} \left\{\mathrm{E}_X\left[D_{\theta^j}\eta(X;\tau_2)^4\right]\right\}^{\frac{1}{2}} \\
&\leq C C_k^2 C_j^2 < \infty
\end{aligned}
$$

where C_0 is as defined in Assumption 1,

$$
\begin{aligned}
C_k &= \sup_k \left\| D_{\theta^k}\eta(X;\tau_1) \right\|_{L^4}, \\
C_j &= \sup_j \left\| D_{\theta^j}\eta(X;\tau_2) \right\|_{L^4},
\end{aligned}
$$

and C is some positive constant. The last inequality follows from $D_{\theta^k}\eta(x;\tau) \in L^4(\mathcal{X})$ by definition. Hence,

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} l_a''(Y_i;\eta(X_i;\tau))D_{\theta^k}\eta(X_i;\tau_1)D_{\theta^j}\eta(X;\tau_2) \\
&\qquad - \mathrm{E}\left[l_a''(Y;\eta(X;\tau))D_{\theta^k}\eta(X;\tau_1)D_{\theta^j}\eta(X;\tau_2)\right] \\
&\qquad = O_p\left(n^{-1}\operatorname{Var}\left[l_a''(Y;\eta(X;\tau))D_{\theta^k}\eta(X;\tau_1)D_{\theta^j}\eta(X;\tau_2)\right]\right)^{\frac{1}{2}} \\
&\qquad = O_p(n^{-\frac{1}{2}}),
\end{aligned}
$$

and as $n \to \infty$, by Cauchy-Schwarz inequality, Assumptions 1 and 3, we have

$$
\begin{aligned}
|(I)| &= \left| \sum_{k=1}^{p} \sum_{j=1}^{p} \theta_3^k \theta_4^j \left\{ \frac{1}{n} \sum_{i=1}^{n} l_a''(Y; \eta(X;\tau)) D_{\theta^k} \eta(X;\tau_1) D_{\theta^j} \eta(X;\tau_2) \right.\right. \\
&\qquad \left.\left. - \mathrm{E}\left[l_a''(Y; \eta(X;\tau)) D_{\theta^k} \eta(X;\tau_1) D_{\theta^j} \eta(X;\tau_2) \right] \right\} \right| \\
&\leq \left\{ \sum_{k=1}^{p} \sum_{j=1}^{p} \left(\frac{1}{n} \sum_{i=1^n} l_a''(Y; \eta(X;\tau)) D_{\theta^k} \eta(X;\tau_1) D_{\theta^j} \eta(X;\tau_2) \right.\right. \\
&\qquad \left.\left. - \mathrm{E}\left[l_a''(Y; \eta(X;\tau)) D_{\theta^k} \eta(X;\tau_1) D_{\theta^j} \eta(X;\tau_2) \right] \right)^2 \right\}^{\frac{1}{2}} \\
&\quad \cdot \left\{ \sum_{k=1}^{p} (\theta_3^k)^2 \sum_{j=1}^{p} (\theta_4^j)^2 \right\}^{\frac{1}{2}} \\
&= O_p\left(p n^{-\frac{1}{2}} \right) \|\theta_3\|_{l^2} \|\theta_4\|_{l^2} \\
&= o_p\left(\left[V_{\tau_0}(L_\theta \theta_3) \right]^{\frac{1}{2}} \left[V_{\tau_0}(L_\theta \theta_4) \right]^{\frac{1}{2}} \right).
\end{aligned}
$$

For (II), since for any $\nu, \mu \in \mathcal{N}$,

$$
\begin{aligned}
&\mathrm{Var}\left[l_a''(Y; \eta(X;\tau)) \left(D_h \eta(X;\tau_1) \phi_{1,\nu} \right) \left(D_h \eta(X;\tau_2) \phi_{2,\mu} \right) \right] \\
&\quad \leq \mathrm{E}\left[l_a''(Y; \eta(X;\tau))^2 \left(D_h \eta(X;\tau_1) \phi_{1,\nu} \right)^2 \left(D_h \eta(X;\tau_2) \phi_{2,\mu} \right)^2 \right] \\
&\quad \leq \left\{ \mathrm{E}\left[l_a''(Y; \eta(X;\tau))^2 \left(D_h \eta(X;\tau_1) \phi_{1,\nu} \right)^4 \right] \right\}^{\frac{1}{2}} \\
&\qquad \cdot \left\{ \mathrm{E}\left[l_a''(Y; \eta(X;\tau))^2 \left(D_h \eta(X;\tau_2) \phi_{2,\mu} \right)^4 \right] \right\}^{\frac{1}{2}} \\
&\quad \leq C_0 \left\{ \mathrm{E}_X\left[\left(D_h \eta(X;\tau_1) \phi_{1,\nu} \right)^4 \right] \right\}^{\frac{1}{2}} \left\{ \mathrm{E}_X\left[\left(D_h \eta(X;\tau_2) \phi_{2,\mu} \right)^4 \right] \right\}^{\frac{1}{2}} \\
&\quad \leq C \left\| D_h \eta(\cdot;\tau_1) \phi_{1,\nu} \right\|_{L^4}^2 \left\| D_h \eta(\cdot;\tau_2) \phi_{2,\mu} \right\|_{L^4}^2 \\
&\quad \leq C C_{\tau_1}^2 C_{\tau_2}^2,
\end{aligned}
$$

where C is some positive constant and $C_{\tau_i} = \sup_\nu \left\| D_h \eta(\cdot;\tau_i) \phi_{i,\nu} \right\|_{L^4}$, for $i = 1, 2$, are the uniform upper bounds as define in Assumption 10. By Lemma 9.1 in Gu (2013) and Assumption 5, as $n\lambda^{\frac{2}{r}} \to \infty$, we have

$$
\begin{aligned}
|(II)| &= \left| \sum_{\nu} \sum_{\mu} h_{3,\nu} h_{4,\mu} \left\{ \frac{1}{n} \sum_{i=1}^{n} l_a''(Y; \eta(X;\tau)) \left(D_h \eta(X;\tau_1)\phi_{1,\nu}\right) \left(D_h \eta(X;\tau_2)\phi_{2,\mu}\right) \right.\right. \\
&\qquad \left.\left. - \mathrm{E}\left[l_a''(Y; \eta(X;\tau)) \left(D_h \eta(X;\tau_1)\phi_{1,\nu}\right) \left(D_h \eta(X;\tau_2)\phi_{2,\mu}\right) \right] \right\} \right| \\
&\leq \left\{ \sum_{\nu} \sum_{\mu} \frac{1}{1+\lambda\rho_{1,\nu}} \frac{1}{1+\lambda\rho_{2,\mu}} \right. \\
&\qquad \cdot \left(\frac{1}{n} \sum_{i=1}^{n} l_a''(Y; \eta(X;\tau)) \left(D_h \eta(X;\tau_1)\phi_{1,\nu}\right) \left(D_h \eta(X;\tau_2)\phi_{2,\mu}\right) \right. \\
&\qquad\qquad \left. \left. - \mathrm{E}\left[l_a''(Y; \eta(X;\tau)) \left(D_h \eta(X;\tau_1)\phi_{1,\nu}\right) \left(D_h \eta(X;\tau_2)\phi_{2,\mu}\right) \right] \right)^2 \right\}^{\frac{1}{2}} \\
&\quad \cdot \left\{ \sum_{\nu} \sum_{\mu} (1+\lambda\rho_{1,\nu})(1+\lambda\rho_{2,\mu}) h_{3,\nu}^2 h_{4,\mu}^2 \right\}^{\frac{1}{2}} \\
&= O_p\left(n^{-\frac{1}{2}} \lambda^{-\frac{1}{r}} \right) \left(V_{\tau_0}[D_h \eta(\tau_1) h_3] + \lambda J(h_3) \right)^{\frac{1}{2}} \left(V_{\tau_0}[D_h \eta(\tau_2) h_4] + \lambda J(h_4) \right)^{\frac{1}{2}} \\
&= o_p\left(\left[V_{\tau_0}(L_h h_3) + \lambda J(h_3) \right]^{\frac{1}{2}} \left[V_{\tau_0}(L_h h_4) + \lambda J(h_4) \right]^{\frac{1}{2}} \right).
\end{aligned}
$$

The bound for (III) follows a similar approach. We see that for any $1 \leq k \leq p$ and $\mu \in \mathbb{N}$,

$$
\begin{aligned}
&\mathrm{Var}\left[l_a''(Y; \eta(X;\tau)) D_{\theta^k} \eta(X;\tau_1) \left(D_h \eta(X;\tau_2)\phi_{2,\mu}\right) \right] \\
&\quad \leq \mathrm{E}\left[l_a''(Y; \eta(X;\tau))^2 D_{\theta^k} \eta(X;\tau_1)^2 \left(D_h \eta(X;\tau_2)\phi_{2,\mu}\right)^2 \right] \\
&\quad \leq \left\{ \mathrm{E}\left[l_a''(Y; \eta(X;\tau))^2 D_{\theta^k} \eta(X;\tau_1)^4 \right] \right\}^{\frac{1}{2}} \\
&\qquad \cdot \left\{ \mathrm{E}\left[l_a''(Y; \eta(X;\tau))^2 \left(D_h \eta(X;\tau_2)\phi_{2,\mu}\right)^4 \right] \right\}^{\frac{1}{2}} \\
&\quad \leq C_0 \left\{ \mathrm{E}_X\left[D_{\theta^k} \eta(X;\tau_1)^4 \right] \right\}^{\frac{1}{2}} \left\{ \mathrm{E}_X\left[\left(D_h \eta(X;\tau_2)\phi_{2,\mu}\right)^4 \right] \right\}^{\frac{1}{2}} \\
&\quad \leq C \left\| D_{\theta^k} \eta(X;\tau_1) \right\|_{L^4}^2 \left\| D_h \eta(\cdot;\tau_2)\phi_{2,\mu} \right\|_{L^4}^2 \\
&\quad \leq C C_k^2 C_{\tau_2}^2,
\end{aligned}
$$

where C, C_k, C_{τ_2} are positive constants as defined earlier. As $n\lambda^{\frac{1}{r}} \to \infty$, we have

$$
\begin{aligned}
|(III)| = \Bigg| \sum_{k=1}^{p} \sum_{\mu=1}^{\infty} \theta_3^k h_{4,\mu} \Bigg\{ \frac{1}{n} \sum_{i=1}^{n} l_a''(Y; \eta(X;\tau)) D_{\theta^k} \eta(X;\tau_1) \left(D_h \eta(X;\tau_2) \phi_{2,\mu}\right) \\
- \mathrm{E}\left[l_a''(Y; \eta(X;\tau)) D_{\theta^k} \eta(X;\tau_1) \left(D_h \eta(X;\tau_2) \phi_{2,\mu}\right) \right] \Bigg\} \Bigg| \\
\leq \left\{ \sum_{k=1}^{p} \sum_{\mu=1}^{\infty} \frac{1}{1+\lambda\rho_{2,\mu}} \left(\frac{1}{n} \sum_{i=1}^{n} l_a''(Y; \eta(X;\tau)) D_{\theta^k} \eta(X;\tau_1) \left(D_h \eta(X;\tau_2) \phi_{2,\mu}\right) \right.\right. \\
\left.\left. - \mathrm{E}\left[l_a''(Y; \eta(X;\tau)) D_{\theta^k} \eta(X;\tau_1) \left(D_h \eta(X;\tau_2) \phi_{2,\mu}\right) \right] \right)^2 \right\}^{\frac{1}{2}} \\
\cdot \left\{ \sum_{k=1}^{p} (\theta_3^k)^2 \sum_{\mu=1}^{\infty} (1+\lambda\rho_{2,\mu}) h_{4,\mu}^2 \right\}^{\frac{1}{2}} \\
= O_p\left(p n^{-\frac{1}{2}} \lambda^{-\frac{1}{2r}} \right) \left(V_{\tau_0}[D_\theta \eta(\tau_1)\theta_3]) \right)^{\frac{1}{2}} \left(V_{\tau_0}[D_h \eta(\tau_2) h_4] + \lambda J(h_4) \right)^{\frac{1}{2}} \\
= o_p\left(\left[V_{\tau_0}(L_\theta \theta_3) \right]^{\frac{1}{2}} \left[V_{\tau_0}(L_h h_4) + \lambda J(h_4) \right]^{\frac{1}{2}} \right).
\end{aligned}
$$

The analysis for (IV) is identical to that for (III) after exchanging all terms associated with τ_1 and τ_2 and replacing θ_3 by θ_4 and h_4 by h_3. Hence, as $n\lambda^{\frac{1}{r}} \to \infty$,

$$
|(IV)| = o_p\left(\left[V_{\tau_0}(L_h h_3) + \lambda J(h_3) \right]^{\frac{1}{2}} \left[V_{\tau_0}(L_\theta \theta_4) \right]^{\frac{1}{2}} \right).
$$

Equation (V.12) then becomes

$$
\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} l_a''(Y_i; \eta(X_i;\tau)) \left[D_\theta \eta(X_i;\tau_1)\theta_3 + D_h \eta(X_i;\tau_1) h_3 \right] \\
\cdot \left[D_\theta \eta(X_i;\tau_2)\theta_4 + D_h \eta(X_i;\tau_2) h_4 \right] \\
= V_\tau \left[D_\theta \eta(\tau_1)\theta_3 + D_h \eta(\tau_1) h_3, D_\theta \eta(\tau_2)\theta_4 + D_h \eta(\tau_2) h_4 \right] \\
+ O_p\left(n^{-\frac{1}{2}} \lambda^{-\frac{1}{r}} \right) \left(\left[V_{\tau_0}(L_\theta \theta_3) \right]^{\frac{1}{2}} + \left[V_{\tau_0}(L_h h_3) + \lambda J(h_3) \right]^{\frac{1}{2}} \right) \\
\cdot \left(\left[V_{\tau_0}(L_\theta \theta_4) \right]^{\frac{1}{2}} + \left[V_{\tau_0}(L_h h_4) + \lambda J(h_4) \right]^{\frac{1}{2}} \right) \\
= V_\tau \left[D_\theta \eta(\tau_1)\theta_3 + D_h \eta(\tau_1) h_3, D_\theta \eta(\tau_2)\theta_4 + D_h \eta(\tau_2) h_4 \right] \\
+ o_p\left\{ \left[V_{\tau_0}(L_\theta \theta_3 + L_h h_3) + \lambda J(h_3) \right]^{\frac{1}{2}} \left[V_{\tau_0}(L_\theta \theta_4 + L_h h_4) + \lambda J(h_4) \right]^{\frac{1}{2}} \right\}.
\end{aligned}
$$

The last equality holds because $\|\cdot\|_{\mathcal{Q}}$ is equivalent to $\|\cdot\|_{\mathcal{Q},1}$. □

We are now ready to prove the following theorem.

Theorem 11. *Under Assumptions 1 - 10, as* $n \to \infty$, $\lambda \to 0$, *and* $n\lambda^{\frac{2}{r}} \to \infty$,

$$V_{\tau_0}[L_\theta(\hat{\theta} - \tilde{\theta}) + L_h(\hat{h} - \tilde{h})] + \lambda J(\hat{h} - \tilde{h}) = O_p\left(n^{-1}\lambda^{-\frac{1}{r}} + \lambda\right).$$

Therefore,

$$V_{\tau_0}[L_\theta(\hat{\theta} - \theta_0) + L_h(\hat{h} - h_0)] + \lambda J(\hat{h} - h_0) = O_p\left(n^{-1}\lambda^{-\frac{1}{r}} + \lambda\right).$$

Proof. Apply Lemma 8 to equation (V.8), the first term on the left hand side of (V.7) becomes

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} l_a''(Y_i; \eta(X_i;\tau_*)) \left[D_\theta\eta(X_i;\tau_*)(\hat{\theta} - \tilde{\theta}) + D_h\eta(X_i;\tau_*)(\hat{h} - \tilde{h})\right] \\
&\qquad\qquad \cdot \left[D_\theta\eta(X_i;\hat{\tau})(\hat{\theta} - \tilde{\theta}) + D_h\eta(X_i;\hat{\tau})(\hat{h} - \tilde{h})\right] \\
&= V_{\tau_*}[D_\theta\eta(\hat{\tau})(\hat{\theta} - \tilde{\theta}) + D_h\eta(\hat{\tau})(\hat{h} - \tilde{h}), D_\theta\eta(\tau_*)(\hat{\theta} - \tilde{\theta}) + D_h\eta(\tau_*)(\hat{h} - \tilde{h})] \\
&\quad + o_p\left(V_{\tau_0}[L_\theta(\hat{\theta} - \tilde{\theta}) + L_h(\hat{h} - \tilde{h})] + \lambda J(\hat{h} - \tilde{h})\right),
\end{aligned}$$

where $\tau_* = \tilde{\tau} + u(\hat{\tau} - \tilde{\tau})$. By Assumptions 6, we have

$$\begin{aligned}
&2V_{\tau_*}[D_\theta\eta(\hat{\tau})(\hat{\theta} - \tilde{\theta}) + D_h\eta(\hat{\tau})(\hat{h} - \tilde{h}), D_\theta\eta(\tau_*)(\hat{\theta} - \tilde{\theta}) + D_h\eta(\tau_*)(\hat{h} - \tilde{h})] \\
&= V_{\tau_*}[D_\theta\eta(\hat{\tau})(\hat{\theta} - \tilde{\theta}) + D_h\eta(\hat{\tau})(\hat{h} - \tilde{h})] + V_{\tau_*}[D_\theta\eta(\tau_*)(\hat{\theta} - \tilde{\theta}) + D_h\eta(\tau_*)(\hat{h} - \tilde{h})] \\
&\quad - V_{\tau_*}[(D_\theta\eta(\hat{\tau}) - D_\theta\eta(\tau_*))(\hat{\theta} - \tilde{\theta}) + (D_h\eta(\hat{\tau}) - D_h\eta(\tau_*))(\hat{h} - \tilde{h})] \\
&\geq C_3\left\{2C_1 V_{\tau_0}[L_\theta(\hat{\theta} - \tilde{\theta}) + L_h(\hat{h} - \tilde{h})] - C_d V_{\tau_0}[L_\theta(\hat{\theta} - \tilde{\theta}) + L_h(\hat{h} - \tilde{h})]\right\} \\
&\geq mV_{\tau_0}[L_\theta(\hat{\theta} - \tilde{\theta}) + L_h(\hat{h} - \tilde{h})],
\end{aligned}$$

for some $0 \leq m \leq C_3(2C_1 - C_d)$. Recall $\tau_{**} = \tilde{\tau} + u(\tau_0 - \tilde{\tau})$. By Lemma 8 and Assumptions 5 and 7, (V.7) gives

$$
\begin{aligned}
&\left\{ \frac{m}{2} V_{\tau_0}[L_\theta(\hat{\theta} - \tilde{\theta}) + L_h(\hat{h} - \tilde{h})] + \lambda J(\hat{h} - \tilde{h}) \right\} (1 + o_p(1)) \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \left[l_a'(Y_i; \eta(X_i; \hat{\tau})) - l_a'(Y_i; \eta(X_i; \tilde{\tau})) \right] \\
&\qquad \cdot \left[D_\theta \eta(X_i; \hat{\tau})(\hat{\theta} - \tilde{\theta}) + D_h \eta(X_i; \hat{\tau})(\hat{h} - \tilde{h}) \right] \\
&\quad + \lambda J(\hat{h} - \tilde{h}) \\
&\leq V_{\tau_0}^{\frac{1}{2}}[D_\theta \eta(\tau_0)(\tilde{\theta} - \theta_0) + D_h \eta(\tau_0)(\tilde{h} - h_0)] \\
&\qquad \cdot V_{\tau_0}^{\frac{1}{2}}[D_\theta \eta(\tau_0)(\hat{\theta} - \tilde{\theta}) + D_h \eta(\tau_0)(\hat{h} - \tilde{h})] \\
&\quad + C_4 V_{\tau_0}^{\frac{1}{2}} \left[D_\theta \eta(\tau_{**})(\tilde{\theta} - \theta_0) + D_h \eta(\tau_{**})(\tilde{h} - h_0) \right] \\
&\qquad \cdot V_{\tau_0}^{\frac{1}{2}} \left[D_\theta \eta(\hat{\tau})(\hat{\theta} - \tilde{\theta}) + D_h \eta(\hat{\tau})(\hat{h} - \tilde{h}) \right] \\
&\quad + O_p \left\{ \left[V_{\tau_0}[L_\theta(\tilde{\theta} - \theta_0) + L_h(\tilde{h} - h_0)] + \lambda J(\tilde{h} - h_0) \right]^{\frac{1}{2}} \right. \\
&\qquad \left. \cdot \left[V_{\tau_0}[L_\theta(\hat{\theta} - \tilde{\theta}) + L_h(\hat{h} - \tilde{h})] + \lambda J(\hat{h} - \tilde{h}) \right]^{\frac{1}{2}} \right\} \\
&\quad + o_p(\left\{ \left[V_{\tau_0}[D_\theta \eta(\tau_0)(\hat{\theta} - \tilde{\theta}) + D_h \eta(\tau_0)(\hat{h} - \tilde{h})] + \lambda J(\hat{h} - \tilde{h}) \right]^{\frac{1}{2}} \right\} \\
&\leq O_p \left\{ \left[(1 + C_4) V_{\tau_0}[L_\theta(\tilde{\theta} - \theta_0) + L_h(\tilde{h} - h_0)] + \lambda J(\tilde{h} - h_0) \right]^{\frac{1}{2}} \right\} \\
&\qquad \cdot \left[(1 + C_4) V_{\tau_0}[L_\theta(\hat{\theta} - \tilde{\theta}) + L_h(\hat{h} - \tilde{h})] + \lambda J(\hat{h} - \tilde{h}) \right]^{\frac{1}{2}} \\
&\quad + o_p(\left\{ \left[V_{\tau_0}[D_\theta \eta(\tau_0)(\hat{\theta} - \tilde{\theta}) + D_h \eta(\tau_0)(\hat{h} - \tilde{h})] + \lambda J(\hat{h} - \tilde{h}) \right]^{\frac{1}{2}} \right\} \\
&\leq O_p \left(n^{-1} \lambda^{-\frac{1}{r}} + \lambda \right)^{\frac{1}{2}} \left[V_{\tau_0}[L_\theta(\hat{\theta} - \tilde{\theta}) + L_h(\hat{h} - \tilde{h})] + \lambda J(\hat{h} - \tilde{h}) \right]^{\frac{1}{2}}
\end{aligned}
$$

The result follows from Theorem 10 after trivial manipulations. □

Since $\|\cdot\|_{\mathbb{R}^p}$ is equivalent to $\|\cdot\|_{l^2}$ on $\mathbb{R}^p$ and $\|\cdot\|_{\mathcal{Q}}$ is equivalent to the product norm $\|\cdot\|_{\mathcal{Q},1}$ on the joint parameter space $\mathcal{Q}$, the following corollaries are direct consequences of Theorem 11.

Corollary 3. *Under Assumptions 1 to 10, as* $n \to \infty$, $\lambda \to 0$, *and* $n\lambda^{\frac{2}{r}} \to \infty$, *we have*

$$\|\hat{\theta} - \theta_0\|_{l^2} \sim \|\hat{\theta} - \theta_0\|_{\mathbb{R}^p} = V_{\tau_0}^{\frac{1}{2}}[L_\theta(\hat{\theta} - \theta_0)] = O_p\left(n^{-\frac{1}{2}}\lambda^{-\frac{1}{2r}} + \lambda^{\frac{1}{2}}\right).$$

Corollary 4. *If* $\eta(\theta, h) = \alpha(x;\theta) + h(x)$, L_h *can be chosen to be the inclusion operator from* $\mathcal{H}$ *to* $L_0^2(\mathcal{X})$. *Under the same conditions as in Theorem 11, we have*

$$\|\hat{h} - h_0\|_{\mathcal{H}} = [V_{\tau_0}(\hat{h} - h_0) + \lambda J(\hat{h} - h_0)]^{1/2} = O_p\left(n^{-\frac{1}{2}}\lambda^{-\frac{1}{2r}} + \lambda^{\frac{1}{2}}\right).$$

We note that the rate of convergence for the nonparametric component in our semiparametric setting coincides with the rate of convergence for the nonparametric models studied by Gu and Qiu (1994) and the rate of convergence for the nonlinear nonparametric model studied by O'Sullivan (1990) (for the penalized least squares estimator only). Given that the nonparametric estimator $\hat{h}$ converges to the true nonparametric parameter h_0, that is, $\left\|\hat{h} - h_0\right\|_{\mathcal{H}} = o_p(1)$, we show in the following proposition that one can obtain a better rate of convergence for the parametric component $\hat{\theta} - \theta_0$. This shows that the joint rate of convergence is dominated by the convergence of the nonparametric component.

Proposition 3. *Under Assumptions 1 to 10, as* $n \to \infty$, $\lambda \to 0$, *and* $n\lambda^{\frac{2}{r}} \to \infty$,

$$\left\|\hat{\theta} - \theta_0\right\|_{l^2} \sim \left\|\hat{\theta} - \theta_0\right\|_{\mathbb{R}^p} = O_p(n^{-\frac{1}{2}}).$$

Proof. Equipped with our joint consistency result in Theorem 11, we have $\left\|\hat{h} - h_0\right\|_{\mathcal{H}} = o_p(1)$ when $n \to \infty$, $\lambda \to 0$, and $n\lambda^{\frac{2}{r}} \to \infty$. Similar to the idea of proving the joint consistency, we first consider an approximation of $\hat{\theta}$,

$$\bar{\theta} = \arg\min_{\theta} \bar{\ell}_n(\theta),$$

where

$$\bar{\ell}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} l_a'(Y_i; \eta(X_i, \tau_0)) D_\theta \eta(X_i; \tau_0)\theta + \frac{1}{2} V_{\tau_0}[D_\theta \eta(X; \tau_0)(\theta - \theta_0)].$$

It is easy to solve $D_\theta \bar{\ell}_n(\theta) = 0$ and get

$$\bar{\theta} = \theta_0 + V_{\tau_0}[D_\theta \eta(X; \tau_0)]^{-1} \left[-\frac{1}{n}\sum_{i=1}^{n} l_a'(Y_i; \eta(X_i, \tau_0)) D_\theta \eta(X_i; \tau_0) \right].$$

Since $V_{\tau_0}[D_\theta \eta(X; \tau_0)]$ is invertible by Assumption 3 and its smallest eigenvalue is bounded below by $C_1 c_\delta$, which are as defined in equation (II.5) and Assumption 5. By similar analysis as in the proofs of Lemma 7 and Theorem 10, we get

$$\begin{aligned} \mathrm{E}\left[(\bar{\theta} - \theta_0)^T (\bar{\theta} - \theta_0)\right] &= O(n^{-1}), \\ V_{\tau_0}[L_\theta(\bar{\theta} - \theta_0)] &= O_p(n^{-1}). \end{aligned}$$

Since

$$\begin{aligned} D_\theta \ell_{n,\lambda}(\hat{\theta}, \hat{h})(\hat{\theta} - \bar{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} l_a'(Y_i; \eta(X_i; \hat{\theta}, \hat{h})) D_\theta \eta(X_i; \hat{\theta}, \hat{h})(\hat{\theta} - \bar{\theta}) = 0, \\ D_\theta \bar{\ell}_n(\bar{\theta})(\hat{\theta} - \bar{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} l_a'(Y_i; \eta(X_i, \tau_0)) D_\theta \eta(X_i; \tau_0)(\hat{\theta} - \bar{\theta}) \\ &\quad + V_{\tau_0}[D_\theta \eta(X; \tau_0)(\bar{\theta} - \theta_0), D_\theta \eta(X; \tau_0)(\hat{\theta} - \bar{\theta})] = 0, \end{aligned}$$

it follows that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[l_a'(Y_i;\eta(X_i;\hat{\theta},\hat{h}))-l_a'(Y_i;\eta(X_i;\bar{\theta},\hat{h}))\right]D_\theta\eta(X_i;\hat{\theta},\hat{h})(\hat{\theta}-\bar{\theta})\\
&=V_{\tau_0}[D_\theta\eta(X;\tau_0)(\bar{\theta}-\theta_0),D_\theta\eta(X;\tau_0)(\hat{\theta}-\bar{\theta})]\\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\left[l_a'(Y_i;\eta(X_i;\theta_0,h_0))-l_a'(Y_i;\eta(X_i;\bar{\theta},\hat{h}))\right]D_\theta\eta(X_i;\hat{\theta},\hat{h})(\hat{\theta}-\bar{\theta})\\
&\quad+\frac{1}{n}\sum_{i=1}^{n}l_a'(Y_i;\eta(X_i;\theta_0,h_0))\left[D_\theta\eta(X_i;\theta_0,h_0)(\hat{\theta}-\bar{\theta})-D_\theta\eta(X_i;\hat{\theta},\hat{h})(\hat{\theta}-\bar{\theta})\right].
\end{aligned}\tag{V.13}$$

Apply the mean value theorem to the left hand side and the second term in the right hand side of equation (V.13) and Lemma 8 to get

$$\begin{aligned}
&V_{\tau_u}[D_\theta\eta(\tau_u)(\hat{\theta}-\bar{\theta}),D_\theta\eta(\hat{\tau})(\hat{\theta}-\bar{\theta})]+o_p\left(V_{\tau_0}^{\frac{1}{2}}[L_\theta(\hat{\theta}-\bar{\theta})]\right)\\
&\le\left|V_{\tau_0}[D_\theta\eta(X;\tau_0)(\bar{\theta}-\theta_0),D_\theta\eta(X;\tau_0)(\hat{\theta}-\bar{\theta})]\right|\\
&\quad+\left|V_{\tau_a}[D_\theta\eta(\tau_a)(\bar{\theta}-\theta_0),D_\theta\eta(\hat{\tau})(\hat{\theta}-\bar{\theta})]\right|\\
&\quad+\left|V_{\tau_a}[D_h\eta(\tau_a)(\hat{h}-h_0),D_\theta\eta(\hat{\tau})(\hat{\theta}-\bar{\theta})]\right|\\
&\quad+o_p\left(V_{\tau_0}^{\frac{1}{2}}[L_\theta(\bar{\theta}-\theta_0)]V_{\tau_0}^{\frac{1}{2}}[L_\theta(\hat{\theta}-\bar{\theta})]\right)\\
&\quad+o_p\left(\left\{V_{\tau_0}[L_h(\hat{h}-h_0)]+\lambda J(\hat{h}-h_0)\right\}^{\frac{1}{2}}V_{\tau_0}^{\frac{1}{2}}[L_\theta(\hat{\theta}-\bar{\theta})]\right)\\
&\quad+o_p\left(V_{\tau_0}^{\frac{1}{2}}[L_\theta(\hat{\theta}-\bar{\theta})]\right),
\end{aligned}$$

where $\tau_u=(\bar{\theta},\hat{h})+u[(\hat{\theta},\hat{h})-(\bar{\theta},\hat{h})]$ and $\tau_a=(\bar{\theta},\hat{h})+a[(\theta_0,h_0)-(\bar{\theta},\hat{h})]$ for some $0<u,a<1$. By Assumption 6, the Cauchy-Schwarz inequality, and the convergence of $\hat{h}-h_0$, similar analysis as in Theorem 11 gives

$$V_{\tau_0}[L_\theta(\bar{\theta}-\theta_0)]=O_p\left(n^{-1}\right).$$

The theorem follows from the triangle inequality. □

V.2 Parameter selection via doubly penalized likelihood

In this section, we will consider the asymptotic properties of a doubly penalized likelihood estimator given by (V.2).

Recall that $\tau_0 = (\theta_0, h_0)$ is the true parameter. We assume that θ_0 is sparse, and without loss of generality, $\theta_0 = (\theta_{10}^T, \theta_{20}^T)^T$, where $\theta_{10} = (\theta_0^1, \ldots, \theta_0^s)^T$ consists of all nonzero components and $\theta_{20} = (\theta_0^{s+1}, \ldots, \theta_0^p)^T = 0$ is a 0 vector of dimension $p-s$. For our analysis, we do not assume a specific form of $q_{\lambda_j}(\cdot)$, but rather we make the following assumptions.

Assumption 11. *For any $j = 1, \ldots, p$,*

1. *$q_{\lambda_j}(|a|) \geq 0$ for $a \in \mathbb{R}$, and $q_{\lambda_j}(0) = 0$.*
2. *$q'_{\lambda_j}(|a|)$ and $q''_{\lambda_j}(|a|)$ exist when $a \neq 0$.*

Let

$$a_n = \max_{1 \leq j \leq s} \left\{ |q'_{\lambda_j}(|\theta_0^j|)| \right\},$$
$$b_n = \max_{1 \leq j \leq s} \left\{ |q''_{\lambda_j}(|\theta_0^j|)| \right\}.$$

We show the existence and joint consistency of a local estimator $\hat{\tau}_D = (\hat{\theta}_D, \hat{h}_D)$ in the following theorem. For notational simplicity, we use $\|\cdot\|$ to denote the norm $\|\cdot\|_{\mathcal{Q}}$ for the rest of the section, unless otherwise specified.

Theorem 12. *(Joint consistency) Under Assumptions 1 to 11, as $n \to \infty$, $n^{-\frac{1}{2}}\lambda^{-\frac{1}{r}} \to 0$, $a_n \to 0$ and $b_n \to 0$, with probability approaching unity, there exists a local minimizer of $L_n(\theta, h)$ such that*

$$\|\hat{\tau}_D - \tau_0\|_{\mathcal{Q}} = O_p\left(n^{-\frac{1}{2}}\lambda^{-\frac{1}{2r}} + \lambda^{\frac{1}{2}} + a_n\right).$$

Proof. Let $r_n = n^{-\frac{1}{2}}\lambda^{-\frac{1}{2r}} + \lambda^{\frac{1}{2}} + a_n$. It suffices to show that for any $\epsilon > 0$, there exists $C > 0$ large enough such that

$$\mathbb{P}\left\{\inf_{\|\Delta\tau\|=Cr_n} L_n(\tau_0 + \Delta\tau) > L_n(\tau_0)\right\} \geq 1 - \epsilon. \tag{V.14}$$

This means that with probability approaching 1, there exists a local minimizer of $L_n(\theta, h)$, say $\hat{\tau}_D = (\hat{\theta}_D, \hat{h}_D)$, such that

$$\|\hat{\tau}_D - \tau_0\|_{\mathcal{Q}} \leq Cr_n.$$

For any $\Delta\tau$ with $\|\Delta\tau\| = Cr_n$, by Assumption 11, $q_{\lambda_j}(|\theta_0^j|) = 0$ for $j = s+1, \ldots, p$, together with the first order Taylor expansion, we have

$$\begin{aligned}
&L_n(\tau_0 + \Delta\tau) - L_n(\tau_0) \\
&= \ell_{n,\lambda}(\tau_0 + \Delta\tau) - \ell_{n,\lambda}(\tau_0) + \sum_{j=1}^{p} q_{\lambda_j}(|\theta_0^j - \Delta\theta^j|) - \sum_{j=1}^{p} q_{\lambda_j}(|\theta_0^j|) \\
&\geq \ell_{n,\lambda}(\tau_0 + \Delta\tau) - \ell_{n,\lambda}(\tau_0) + \sum_{j=1}^{s} q_{\lambda_j}(|\theta_0^j - \Delta\theta^j|) - \sum_{j=1}^{s} q_{\lambda_j}(|\theta_0^j|) \\
&= D\ell_{n,\lambda}(\tau_0)\Delta\tau + \frac{1}{2}D^2\ell_{n,\lambda}(\tau_0)\Delta\tau\Delta\tau + o\left(\|\Delta\tau\|^2\right) \\
&\quad + \sum_{j=1}^{s} q'_{\lambda_j}(|\theta_0^j|)\operatorname{sgn}(\theta_0^j)\Delta\theta^j + \frac{1}{2}\sum_{j=1}^{s} q''_{\lambda_j}(|\theta_0^j|)\left(\Delta\theta^j\right)^2 + o\left(\|\Delta\theta\|_{\mathbb{R}^p}^2\right) \\
&= \left\{\frac{1}{n}\sum_{i=1}^{n} l'_a(Y_i; \eta(X_i; \tau_0))\left[D_\theta\eta(X_i; \tau_0)\Delta\theta + D_h(X_i; \tau_0)\Delta h\right] + \lambda J(h_0, \Delta h)\right\} \\
&\quad + \left\{\frac{1}{2n}\sum_{i=1}^{n} l''_a(Y_i; \eta(X_i; \tau_0))\left[D_\theta\eta(X_i; \tau_0)\Delta\theta + D_h(X_i; \tau_0)\Delta h\right]^2 + \lambda J(\Delta h, \Delta h)\right\} \\
&\quad + \left\{\frac{1}{2n}\sum_{i=1}^{n} l'_a(Y_i; \eta(X_i; \tau_0))\left[D^2\eta(X_i; \tau_0)\Delta\tau\Delta\tau\right]\right\} \\
&\quad + \left\{\sum_{j=1}^{s} q'_{\lambda_j}(|\theta_0^j|)\operatorname{sgn}(\theta_0^j)\Delta\theta^j + \frac{1}{2}\sum_{j=1}^{s} q''_{\lambda_j}(|\theta_0^j|)\left(\Delta\theta^j\right)^2\right\} + o\left(\|\Delta\tau\|^2\right) \\
&= I + II + III + IV + o\left(\|\Delta\tau\|^2\right),
\end{aligned} \tag{V.15}$$

where sgn$(\cdot)$ is the sign function, D and D^2 are the first- and second-order Fréchet differential operators, i.e., for any twice Fréchet differentiable function $f(\theta, h)$ and any $\tau_1 = (\theta_1, h_1), \tau_2 = (\theta_2, h_2) \in \mathcal{Q}$,

$$\begin{aligned} Df(\tau)\tau_1 &= D_\theta f(\tau)\theta_1 + D_h f(\tau)h_1, \\ D^2 f(\tau)\tau_1\tau_2 &= D^2_{\theta\theta} f(\tau)\theta_1\theta_2 + D^2_{hh} f(\tau)h_1 h_2 + D^2_{h\theta} f(\tau)\theta_1 h_2 + D^2_{\theta h} f(\tau)h_1\theta_2. \end{aligned}$$

It is easy to see that by Lemma 8(i) and

$$|\lambda J(h_0, \Delta h)| \leq \sqrt{\lambda J(h_0)}\sqrt{\lambda J(\Delta h)} \leq O\left(\lambda^{\frac{1}{2}}\right) \|\Delta\tau\|,$$

we have

$$I = O_p\left(n^{-\frac{1}{2}}\lambda^{-\frac{1}{2r}} + \lambda^{\frac{1}{2}}\right) \|\Delta\tau\| = O_p\left(Cr_n^2\right).$$

By Lemma 8(ii), we have

$$\begin{aligned} II &= \left\{\frac{c}{2} V_{\tau_0}[L_\theta \Delta\theta + L_h \Delta h] + \lambda J(\Delta h)\right\}[1 + o_p(1)] \\ &\geq \tilde{C}\,[1 + o_p(1)]\,\|\Delta\tau\|^2 \\ &= \left[\tilde{C} + o_p(1)\right]\left(C^2 r_n^2\right), \end{aligned}$$

for some $c \in [C_1, C_2]$ and $0 < \tilde{C} \leq 1$. For (III), since $\eta(x;\tau)$ is three times continuously Fréchet differentiable by Assumption 4, $D^2\eta(X;\tau_0) \in \mathscr{L}(\mathcal{Q}, \mathscr{L}(\mathcal{Q}, L_0^4(X))) \cong \mathscr{L}(\mathcal{Q}\times\mathcal{Q}, L_0^4(X))$ is a bounded bilinear operator by definition. Since

$$
\begin{aligned}
&\mathrm{E}\left\{\left|\frac{1}{2n}\sum_{i=1}^{n} l_a^{'}(Y_i;\eta(X_i;\tau_0))\left[D^2\eta(X_i;\tau_0)\Delta\tau\Delta\tau\right]\right|^2\right\} \\
&= \frac{\sigma}{4n}\,\mathrm{E}_X\left[I_{\tau_0}(X)\left|D^2\eta(X_i;\tau_0)\Delta\tau\Delta\tau\right|^2\right] \\
&= O(n^{-1})\left\|D^2\eta(X_i;\tau_0)\Delta\tau\Delta\tau\right\|_{L^2}^2 \\
&\le O(n^{-1})\left\|D^2\eta(X_i;\tau_0)\Delta\tau\Delta\tau\right\|_{L^4}^2 \\
&\le O(n^{-1})\left\|\Delta\tau\right\|^2\left\|\Delta\tau\right\|^2,
\end{aligned}
$$

we have

$$
III = O_p\left(n^{-\frac{1}{2}}\right)\left\|\Delta\tau\right\|^2 = o_p\left(C^2 r_n^2\right).
$$

For (IV), by the Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
&\left|\sum_{j=1}^{s} q_{\lambda_j}^{'}(|\theta_0^j|)\,\mathrm{sgn}(\theta_0^j)\Delta\theta^j + \frac{1}{2}\sum_{j=1}^{s} q_{\lambda_j}^{''}(|\theta_0^j|)\left(\Delta\theta^j\right)^2\right| \\
&\le \left\{\sum_{j=1}^{s}\left[q_{\lambda_j}^{'}(|\theta_0^j|)\right]^2\sum_{j=1}^{s}\left(\Delta\theta^j\right)^2\right\}^{\frac{1}{2}} + \frac{1}{2}\sum_{j=1}^{s}\left|q_{\lambda_j}^{''}(|\theta_0^j|)\right|\left(\Delta\theta^j\right)^2 \\
&\le \sqrt{s}a_n\left\|\Delta\theta\right\|_{l^2} + \frac{1}{2}sb_n\left\|\Delta\theta\right\|_{l^2}^2 \\
&= O\left(a_n\left\|\Delta\theta\right\|_{\mathbb{R}^p}\right) + o_p\left(\left\|\Delta\theta\right\|_{\mathbb{R}^p}^2\right) \\
&= O\left(Cr_n^2\right) + o_p\left(C^2 r_n^2\right).
\end{aligned}
$$

It can be seen that when C is large enough, terms (I), (III), and (IV) are dominated by term (II), which is greater than 0. Therefore, equation (V.15) is greater than 0 for sufficiently large C and arbitrary $\Delta\tau$, hence we have proved (V.14). □

We note that Theorem 12 implies the probability that $\hat{\tau}_D \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$ tends to 1. We will restrict our attention to this event, or simply assume

that $\hat{\tau}_D \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$ for the rest of our analysis. Suppose we replace h with its estimator $\hat{h}_D$ in $L_n(\theta, h)$, then

$$\hat{\theta}_D = \underset{\theta \in \mathcal{N}_{\theta_0}}{\arg\min}\, L_n(\theta, \hat{h}_D).$$

Similar to the idea behind Proposition V.1.2, we can establish the following rate of convergence for $\hat{\theta}_D - \theta_0$.

Theorem 13. *(Parameter estimation consistency) Under the same assumptions as in Theorem 12, as $n \to \infty$, $\lambda \to 0$, $n^{-\frac{1}{2}}\lambda^{-\frac{1}{r}} \to 0$, $a_n \to 0$ and $b_n \to 0$,*

$$\left\|\hat{\theta}_D - \theta_0\right\|_{l^2} \sim \left\|\hat{\theta}_D - \theta_0\right\|_{\mathbb{R}^p} = O_p\left(n^{-\frac{1}{2}} + a_n\right).$$

Proof. Let $t_n = n^{-\frac{1}{2}} + a_n$. It suffices to show that for any $\epsilon > 0$, there exists $C > 0$ large enough such that

$$\mathbb{P}\left\{\inf_{\|\Delta\theta\|_{\mathbb{R}^n} = Ct_n} L_n(\theta_0 + \Delta\theta, \hat{h}_D) > L_n(\theta_0, \hat{h}_D)\right\} \geq 1 - \epsilon. \tag{V.16}$$

By Assumption 11, $q_{\lambda_j}(|\theta_0^j|) = 0$ for $j = s+1, \ldots, p$, together with the first-order Taylor expansion, we have

$$
\begin{aligned}
&L_n(\theta_0+\Delta\theta,\hat{h}_D)-L_n(\theta_0,\hat{h}_D)\\
&\geq \ell_{n,\lambda}(\theta_0+\Delta\theta,\hat{h}_D)-\ell_{n,\lambda}(\theta_0,\hat{h}_D)+\sum_{j=1}^{s}q_{\lambda_j}(|\theta_0^j-\Delta\theta^j|)-\sum_{j=1}^{s}q_{\lambda_j}(|\theta_0^j|)\\
&=D_\theta\ell_{n,\lambda}(\theta_0,\hat{h}_D)\Delta\theta+\frac{1}{2}D^2_{\theta\theta}\ell_{n,\lambda}(\theta_0,\hat{h}_D)\Delta\theta\Delta\theta+o\left(\|\Delta\theta\|^2_{\mathbb{R}^p}\right)\\
&\quad+\sum_{j=1}^{s}q'_{\lambda_j}(|\theta_0^j|)\operatorname{sgn}(\theta_0^j)\Delta\theta^j+\frac{1}{2}\sum_{j=1}^{s}q''_{\lambda_j}(|\theta_0^j|)\left(\Delta\theta^j\right)^2+o\left(\|\Delta\theta\|^2_{\mathbb{R}^p}\right)\\
&=\left\{\frac{1}{n}\sum_{i=1}^{n}l'_a(Y_i;\eta(X_i;\theta_0,\hat{h}_D))D_\theta\eta(X_i;\theta_0,\hat{h}_D)\Delta\theta\right\}\\
&\quad+\left\{\frac{1}{2n}\sum_{i=1}^{n}l''_a(Y_i;\eta(X_i;\theta_0,\hat{h}_D))\left[D_\theta\eta(X_i;\theta_0,\hat{h}_D)\Delta\theta\right]^2\right\}\\
&\quad+\left\{\frac{1}{2n}\sum_{i=1}^{n}l'_a(Y_i;\eta(X_i;\theta_0,\hat{h}_D))\left[D^2_{\theta\theta}\eta(X_i;\theta_0,\hat{h}_D)\Delta\theta\Delta\theta\right]\right\}\\
&\quad+\left\{\sum_{j=1}^{s}q'_{\lambda_j}(|\theta_0^j|)\operatorname{sgn}(\theta_0^j)\Delta\theta^j+\frac{1}{2}\sum_{j=1}^{s}q''_{\lambda_j}(|\theta_0^j|)\left(\Delta\theta^j\right)^2\right\}+o\left(\|\Delta\theta\|^2_{\mathbb{R}^p}\right)\\
&=I+II+III+IV+o\left(\|\Delta\theta\|^2_{\mathbb{R}^p}\right).
\end{aligned}
\tag{V.17}
$$

For (I), we have

$$
\begin{aligned}
|I|&\leq\left|\frac{1}{n}\sum_{i=1}^{n}\left[l'_a(Y_i;\eta(X_i;\theta_0,\hat{h}_D))-l'_a(Y_i;\eta(X_i;\theta_0,h_0))\right]D_\theta\eta(X_i;\theta_0,\hat{h}_D)\Delta\theta\right|\\
&\quad+\left|\frac{1}{n}\sum_{i=1}^{n}l'_a(Y_i;\eta(X_i;\theta_0,h_0))D_\theta\eta(X_i;\theta_0,\hat{h}_D)\Delta\theta\right|\\
&=I_1+I_2.
\end{aligned}
$$

By the mean value theorem, an argument similar to the analysis of term (III) in the proof of Lemma 8 shows that

$$
\begin{aligned}
I_1 &= \left| \frac{1}{n} \sum_{i=1}^{n} l_a''(Y_i; \eta(X_i; \tau_u)) \left[D_h \eta(X_i; \tau_u)(\hat{h}_D - h_0) \right] \left[D_\theta \eta(X_i; \theta_0, \hat{h}_D) \Delta\theta \right] \right| \\
&= O_p(n^{-\frac{1}{2}} \lambda^{-\frac{1}{2r}}) \left\| \hat{h}_D - h_0 \right\|_{\mathcal{H}} \|\Delta\theta\|_{\mathbb{R}^p} \\
&= O_p\left(n^{-1} \lambda^{-\frac{1}{r}} + n^{-\frac{1}{2}} \lambda^{\frac{1}{2} - \frac{1}{2r}} + n^{-\frac{1}{2}} \lambda^{-\frac{1}{2r}} a_n \right) \|\Delta\theta\|_{\mathbb{R}^p} \\
&= O_p\left\{ n^{-\frac{1}{2}} \left[n^{-\frac{1}{2}} \lambda^{-\frac{1}{r}} + \lambda^{\frac{1}{2} - \frac{1}{2r}} \right] + a_n \left[n^{-\frac{1}{2}} \lambda^{-\frac{1}{2r}} \right] \right\} \|\Delta\theta\|_{\mathbb{R}^p} \\
&\leq O_p\left\{ M \left(n^{-\frac{1}{2}} + a_n \right) \|\Delta\theta\|_{\mathbb{R}^p} \right\} = O_p\left(C t_n^2 \right),
\end{aligned}
$$

where M is some large positive constant and $\tau_u = (\theta_0, h_0) + u[(\theta_0, \hat{h}_D) - (\theta_0, h_0)]$ for some $0 < u < 1$. The last inequality holds because $n^{-\frac{1}{2}} \lambda^{-\frac{1}{r}} + \lambda^{\frac{1}{2} - \frac{1}{2r}} \to 0$ and $n^{-\frac{1}{2}} \lambda^{-\frac{1}{2r}} \to 0$, and hence both are bounded. Equation (V.10) implies

$$
I_2 = O_p\left(n^{-\frac{1}{2}} \|\Delta\theta\|_{\mathbb{R}^p} \right) = O_p\left(C t_n^2 \right).
$$

Hence, $I = O_p(C t_n^2)$. For (II), by Lemma 8, we have

$$
II \geq (C_1 + o_p(1)) \|\Delta\theta\|_{\mathbb{R}^p}^2 = (C_1 + o_p(1)) (C^2 t_n^2) > 0.
$$

For (III), since $\eta(x; \tau)$ is three times continuously Fréchet differentiable by Assumption 4, by definition $D_{\theta\theta}^2 \eta(X; \theta_0, \hat{h}_D) \in \mathscr{L}(\mathbb{R}^p, \mathscr{L}(\mathbb{R}^p, L_0^4(X)) \cong \mathscr{L}(\mathbb{R}^p \times \mathbb{R}^p, L_0^4(X))$ is a bounded bilinear operator. By the mean value theorem,

$$
\begin{aligned}
&|III| \\
&\leq \left| \frac{1}{2n} \sum_{i=1}^{n} l_a''(Y_i; \eta(X_i; \tau_u)) \left[D_h \eta(X_i; \tau_u)(\hat{h}_D - h_0) \right] \left[D_{\theta\theta}^2 \eta(X_i; \theta_0, \hat{h}_D) \Delta\theta \Delta\theta \right] \right| \\
&\quad + \left| \frac{1}{2n} \sum_{i=1}^{n} l_a'(Y_i; \eta(X_i; \tau_0)) \left[D_{\theta\theta}^2 \eta(X_i; \theta_0, \hat{h}_D) \Delta\theta \Delta\theta \right] \right|,
\end{aligned}
$$

where $\tau_u = (\theta_0, h_0) + u[(\theta_0, \hat{h}_D) - (\theta_0, h_0)]$ for some $0 < u < 1$. And since

$$
\begin{aligned}
&\mathrm{E}\left\{\left|\frac{1}{2n}\sum_{i=1}^{n} l_a''(Y_i;\eta(X_i;\tau_u))\left[D_h\eta(X_i;\tau_u)(\hat{h}_D-h_0)\right]\left[D^2_{\theta\theta}\eta(X_i;\theta_0,\hat{h}_D)\Delta\theta\Delta\theta\right]\right|^2\right\}\\
&\quad\le C_0\frac{1}{4n}\left\|D_h\eta(X_i;\tau_u)(\hat{h}_D-h_0)\right\|^2_{L^4}\left\|D^2_{\theta\theta}\eta(X_i;\theta_0,\hat{h}_D)\Delta\theta\Delta\theta\right\|^2_{L^4}\\
&\quad\le O\left(n^{-1}\right)\left\|\hat{h}_D-h_0\right\|^2_{\mathcal{H}}\|\Delta\theta\|^2_{\mathbb{R}^p}\|\Delta\theta\|^2_{\mathbb{R}^p},
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathrm{E}\left\{\left|\frac{1}{2n}\sum_{i=1}^{n} l_a'(Y_i;\eta(X_i;\tau_0))\left[D^2_{\theta\theta}\eta(X_i;\theta_0,\hat{h}_D)\Delta\theta\Delta\theta\right]\right|^2\right\}\\
&\quad=\frac{\sigma}{4n}\,\mathrm{E}_X\left[I_{\tau_0}(X)\left|D^2_{\theta\theta}\eta(X_i;\theta_0,\hat{h}_D)\Delta\theta\Delta\theta\right|^2\right]\\
&\quad=O(n^{-1})\left\|D^2_{\theta\theta}\eta(X_i;\theta_0,\hat{h}_D)\Delta\theta\Delta\theta\right\|^2_{L^2}\\
&\quad\le O(n^{-1})\left\|D^2_{\theta\theta}\eta(X_i;\theta_0,\hat{h}_D)\Delta\theta\Delta\theta\right\|^2_{L^4}\\
&\quad\le O(n^{-1})\|\Delta\theta\|^2_{\mathbb{R}^p}\|\Delta\theta\|^2_{\mathbb{R}^p},
\end{aligned}
$$

we have

$$
III=O_p\left(n^{-1}\lambda^{-\frac{1}{2r}}+n^{-\frac{1}{2}}\lambda^{\frac{1}{2}}+n^{-\frac{1}{2}}a_n\right)\|\Delta\theta\|^2_{\mathbb{R}^p}=o_p\left(C^2t_n^2\right).
$$

The analysis for (IV) is identical to the argument given for term (IV) in the proof of Theorem 12, which shows

$$
IV=O\left(a_n\|\Delta\theta\|_{\mathbb{R}^p}\right)+o_p\left(\|\Delta\theta\|^2_{\mathbb{R}^p}\right)=O_p\left(Ct_n^2\right).
$$

It is easy to see that by choosing $C>0$ large enough, (I), (III), and (IV) are dominated by (II), which is greater than 0. We have shown (V.16), which implies our desired result. □

When the true parameter θ_0 is sparse, it is often of great interest whether a proposed estimator achieves model selection consistency, i.e.,

$$
\mathbb{P}\left(\left\{j:\hat{\theta}_D^j\neq 0\right\}=\left\{j:\theta_0^j\neq 0\right\}\right)\to 1\quad\text{ as } n\to\infty.
$$

In the next theorem, we discuss the model selection consistency for our proposed model under some conditions. Adopting similar notation to what we used for θ_0, we write $\hat{\theta}_D = (\hat{\theta}_{1D}^T, \hat{\theta}_{2D}^T)^T$, where $\hat{\theta}_{1D} = (\hat{\theta}_D^1, \ldots, \hat{\theta}_D^s)^T$ and $\hat{\theta}_{2D} = (\hat{\theta}_D^{s+1}, \ldots, \hat{\theta}_D^p)^T$.

Theorem 14. *(Model selection consistency) Under the same conditions as in Theorem 13, if* $a_n = O(n^{-\frac{1}{2}})$, $n^{\frac{1}{2}}\lambda_j \to \infty$, *and*

$$\liminf_{n\to\infty} \liminf_{\theta^j \to 0^+} \lambda_j^{-1} q'_{\lambda_j}(|\theta^j|) > 0,$$

for $j = 1, \ldots, p$, *then*

$$\mathbb{P}\left(\left\{j : \hat{\theta}_D^j \neq 0\right\} = \left\{j : \theta_0^j \neq 0\right\}\right) = \mathbb{P}\left(\hat{\theta}_{2D} = 0\right) \to 1 \quad \text{as } n \to \infty.$$

Proof. We first show that for any $\theta = (\theta_1^T, \theta_2^T)^T$, if θ_1 satisfies $\|\theta_1 - \theta_{10}\|_{\mathbb{R}^p} = O_p(n^{-\frac{1}{2}})$, for some small $\gamma_n = Cn^{-\frac{1}{2}}$ and $j = s+1, \ldots, p$,

$$\begin{aligned} D_{\theta^j} L_n(\theta, \hat{h}_D) &< 0 \quad \text{if} \quad -\gamma_n < \theta^j < 0, \\ D_{\theta^j} L_n(\theta, \hat{h}_D) &> 0 \quad \text{if} \quad 0 < \theta^j < \gamma_n, \end{aligned} \tag{V.18}$$

with probability tending to 1 as $n \to \infty$. For $j = s+1, \ldots, p$, by Taylor expansion, we have

$$
\begin{aligned}
&D_{\theta^j} L_n(\theta, \hat{h}_D) \\
&= D_{\theta^j} \ell_{n,\lambda}(\theta, \hat{h}_D) + q'_{\lambda_j}(\left|\theta^j\right|) \operatorname{sgn}(\theta^j) \\
&= D_{\theta^j} \ell_{n,\lambda}(\theta_0, \hat{h}_D) + D_\theta D_{\theta^j} \ell_{n,\lambda}(\theta_0, \hat{h}_D)(\theta - \theta_0) + q'_{\lambda_j}(\left|\theta^j\right|) \operatorname{sgn}(\theta^j) + o(n^{-\frac{1}{2}}) \\
&= \left\{ \frac{1}{n} \sum_{i=1}^{n} l'_a(Y_i; \eta(X_i; \theta_0, \hat{h}_D)) D_{\theta^j} \eta(X_i; \theta_0, \hat{h}_D) \right\} \\
&\quad + \left\{ \frac{1}{2n} \sum_{i=1}^{n} l''_a(Y_i; \eta(X_i; \theta_0, \hat{h}_D)) \left[D_\theta \eta(X_i; \theta_0, \hat{h}_D)(\theta - \theta_0) \right] D_{\theta^j} \eta(X_i; \theta_0, \hat{h}_D) \right\} \\
&\quad + \left\{ \frac{1}{2n} \sum_{i=1}^{n} l'_a(Y_i; \eta(X_i; \theta_0, \hat{h}_D)) \left[D_\theta D_{\theta^j} \eta(X_i; \theta_0, \hat{h}_D)(\theta - \theta_0) \right] \right\} \\
&\quad + q'_{\lambda_j}(\left|\theta^j\right|) \operatorname{sgn}(\theta^j) + o(n^{-\frac{1}{2}}) \\
&= I + II + III + q'_{\lambda_j}(\left|\theta^j\right|) \operatorname{sgn}(\theta^j) + o(n^{-\frac{1}{2}}).
\end{aligned}
$$

By similar analysis as in Theorem 13, one can show that

$$
\begin{aligned}
I &= O_p\left(n^{-\frac{1}{2}}\right), \\
II &= \left(\tilde{C} + o_p(1)\right) \|\theta - \theta_0\|_{\mathbb{R}^p} \\
&\sim \left(\tilde{C} + o_p(1)\right) \left(\|\theta_1 - \theta_{10}\|_{l^2}^2 + \|\theta_2\|_{l^2}^2\right)^{\frac{1}{2}} = O_p\left(n^{-\frac{1}{2}}\right), \\
III &= O_p\left(n^{-\frac{1}{2}}\right) \|\theta - \theta_0\|_{\mathbb{R}^p} \\
&\sim O_p\left(n^{-\frac{1}{2}}\right) \left(\|\theta_1 - \theta_{10}\|_{l^2}^2 + \|\theta_2\|_{l^2}^2\right)^{\frac{1}{2}} = o_p\left(n^{-\frac{1}{2}}\right),
\end{aligned}
$$

for some positive constant $\tilde{C}$. Therefore, as $n \to \infty$, if $n^{\frac{1}{2}} \lambda_j \to \infty$ and

$$
\liminf_{n \to \infty} \liminf_{\theta^j \to 0^+} \lambda_j^{-1} q'_{\lambda_j}(|\theta^j|) > 0,
$$

we have

$$
D_{\theta^j} L_n(\theta, \hat{h}_D) = \lambda_j \left[\lambda_j^{-1} q'_{\lambda_j}(\left|\theta^j\right|) \operatorname{sgn}(\theta^j) + O_p(n^{-\frac{1}{2}} \lambda_j^{-1}) \right],
$$

whose sign is determined by θ^j. Hence, (V.18) holds.

It is not difficult to see that when $a_n = O(n^{-\frac{1}{2}})$, the local minimizer $\hat{\theta}_D$ as in Theorem 13 satisfying $\left\|\hat{\theta}_D - \theta_0\right\|_{\mathbb{R}^p} = O_p(n^{-\frac{1}{2}})$ must have $\hat{\theta}_{2D} = 0$ with probability tending to 1. □

Remark 5.

1. *We note that when the LASSO penalty is considered, i.e.,* $\sum_{j=1}^{p} q_{\lambda_j}(|\theta^j|) = \lambda_n \sum_{j=1}^{p} |\theta^j|$, $a_n = \lambda_n$ *and* $\sqrt{n}$ *consistency in terms of parameter estimation requires that* $\lambda_n = O(n^{-\frac{1}{2}})$. *Hence, the condition* $n^{\frac{1}{2}}\lambda_n \to \infty$ *in Theorem 14 cannot be satisfied. In fact, it is well-known for the parametric linear model that the LASSO estimator that achieves optimal consistency in terms of parameter estimation can be inconsistent in model selection, see Leng et al. (2006), Zhao and Yu (2006), and Zou (2006).*

2. *Penalized likelihood estimators given by the SCAD penalty and bridge penalty for some* $0 < q < 1$ *can achieve both estimation consistency and model selection consistency when a suitable regularization parameter is chosen in our setting (Fan and Li, 2001).*

APPENDIX

ASSUMPTIONS FOR DENSITY ESTIMATION WITH MULTIPLE SAMPLES

In this section, we present the precise statements for the assumptions needed for the study of the density estimation model with multiple samples.

Assumption A.1. *The penalty $J(h)$ is a square seminorm in $\mathcal{H}$ with a finite-dimensional null space $\mathcal{H}_0 \subset \mathcal{H}$. Therefore, $J((\theta, h)) \equiv J(h)$ extends J to a square seminorm on $\mathcal{Q}$, and its null space $\mathbb{R}^p \times \mathcal{H}_0$ is again finite-dimensional. Denote by $J(g, h)$ the semi-inner product associated with the seminorm $J(h)$. We also assume that $J(h_0) < \infty$.*

Assumption A.2. *For $l = 1, \ldots, m$, there are bounded linear operators $L_{l,\theta} : \mathbb{R}^p \to L_0^2(\mathcal{X}_l)$ and $L_{l,h} : \mathcal{H} \to L_0^2(\mathcal{X}_l)$, with zero nullspaces, which satisfy the following conditions:*

(i) Suppose $\mathcal{H}$ is a reproducing kernel Hilbert space, equipped with norm $\|\cdot\|$. For any $g \in \mathcal{H}$, there exist positive constants M_1, M_2, such that

$$M_1 \|g\|^2 \leq \sum_{l=1}^{m} V_{l,\tau_0}(L_{l,h}g) + \lambda J(g) \leq M_2 \|g\|^2 .$$

(ii) For any $\zeta \in \mathbb{R}^p$ satisfying $\|\zeta\|_{l^2} = 1$ and for any $g \in \mathcal{H}$, there exists a positive constant c_δ such that

$$\sum_{l=1}^{m} V_{l,\tau_0}(L_{l,\theta}\zeta - L_{l,h}g) > c_\delta. \tag{A.1}$$

By Assumptions A.1 and A.2, we have $\langle g_1, g_2\rangle_{\mathcal{H}} \equiv \sum_{l=1}^{m} V_{l,\tau_0}(L_{l,h}g_1, L_{l,h}g_2) + \lambda J(g_1, g_2)$ an inner product on $\mathcal{H}$, and its induced norm, denoted by $\|\cdot\|_{\mathcal{H}}$, is complete on $\mathcal{H}$. One can also see that $L_{l,\theta}$ can be represented by the $p \times 1$ vector of $L_0^2(\mathcal{X}_l)$ functions $[L_{l,\theta}^k(x)]_{k=1}^p$. We may use $L_{l,\theta}$ to denote the linear operator or its vector form, i.e., $L_{l,\theta}\zeta = \zeta^T L_{l,\theta}$. Denote by $V_{l,\tau_0}(L_{l,\theta}, L_{l,\theta})$ the $p \times p$ matrix in which the (i,j)th entry is $V_{l,\tau_0}(L_{l,\theta}^i(x), L_{l,\theta}^j(x))$, and note that one can write $\sum_{l=1}^{m} V_{l,\tau_0}(L_{l,\theta}\zeta, L_{l,\theta}\zeta) = \zeta^T \sum_{l=1}^{m} V_{l,\tau_0}(L_{l,\theta}, L_{l,\theta})\zeta$. When $g = 0$, (A.1) implies that $\sum_{l=1}^{m} V_{l,\tau_0}(L_{l,\theta}, L_{l,\theta})$ is positive definite. Therefore, $\langle \zeta_1, \zeta_2\rangle_{\mathbb{R}^p} \equiv \sum_{l=1}^{m} V_{l,\tau_0}(L_{l,\theta}\zeta_1, L_{l,\theta}\zeta_2)$ is an inner product on $\mathbb{R}^p$ and we use $\|\cdot\|_{\mathbb{R}^p}$ to denote its induced norm.

For any $(\zeta_1, g_1), (\zeta_2, g_2) \in \mathcal{Q}$, we define an inner product as follows,

$$\langle (\zeta_1, g_1), (\zeta_2, g_2)\rangle_{\mathcal{Q}} \equiv \sum_{l=1}^{m} V_{l,\tau_0}(L_{l,\theta}\zeta_1 + L_{l,h}g_1, L_{l,\theta}\zeta_2 + L_{l,h}g_2) + \lambda J(g_1, g_2), \tag{A.2}$$

and we denote the norm induced by this inner product by $\|\cdot\|_{\mathcal{Q}}$. Following a similar analysis as in Section II.3, it can be shown that $\langle\cdot,\cdot\rangle_{\mathcal{Q}}$ as defined above is, in fact, an inner product, and that its induced norm is complete.

Assumption A.3. *For all $l = 1, \ldots, m$, $\eta_l(\theta, h)$ is three times continuously Fréchet differentiable with respect to (θ, h) in $\mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$. Moreover, $\tau_0 = (\theta_0, h_0)$ is the unique root of*

$$\mathrm{E}\left[D_\theta \sum_{l=1}^{m} \ell_{n_l}(\theta, h)\right] = 0 \quad \textit{and} \quad \mathrm{E}\left[D_h \sum_{l=1}^{m} \ell_{n_l}(\theta, h)\right] = 0$$

in $\mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$.

Assumption A.4. *For any $\theta_* \in \mathcal{N}_{\theta_0}, h_* \in \mathcal{N}_{h_0}, l \in \{1, \ldots, m\}$, $D_\theta\eta_l(\theta_*, h_*)$ is a bounded linear operator from $\mathbb{R}^p$ to $L_0^2(\mathcal{X}_l)$, $D_h\eta_l(\theta_*, h_*)$ is a bounded*

linear operator from $\mathcal{H}$ *to* $L_0^2(\mathcal{X}_l)$*, and there exist positive constants* C_1, C_2*, such that for all* $(\zeta, g) \in \mathcal{Q}$*,*

$$C_1 \sum_{l=1}^{m} V_{l,\tau_0}(L_{l,\theta}\zeta + L_{l,h}g) \leq \sum_{l=1}^{m} V_{l,\tau_0}(D_\theta \eta_l(\theta_*, h_*)\zeta + D_h \eta_l(\theta_*, h_*)g) \leq C_2 \sum_{l=1}^{m} V_{l,\tau_0}(L_{l,\theta}\zeta + L_{l,h}g).$$

Assumption A.5. *For any* $(\theta_1, h_1), (\theta_2, h_2) \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$*, there exists a positive constant* $C_d < 2C_1$*, where* C_1 *is as defined in Assumption A.4, such that for any* $(\zeta, g) \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$*,*

$$\sum_{l=1}^{m} V_{l,\tau_0} \left[(D_\theta \eta_l(\theta_1, h_1) - D_\theta \eta_l(\theta_2, h_2))\zeta + (D_h \eta_l(\theta_1, h_1) - D_h \eta_l(\theta_2, h_2))g \right] \leq C_d \sum_{l=1}^{m} V_{l,\tau_0}(L_{l,\theta}\zeta + L_{l,h}g).$$

Assumption A.6. $\mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$ *is a convex set that contains* $\hat{\tau}$*, and there exist* $C_3, C_4 > 0$ *such that* $C_3 \sum_{l=1}^{m} V_{l,\tau_0}(f) \leq \sum_{l=1}^{m} V_{l,\tau}(f) \leq C_4 \sum_{l=1}^{m} V_{l,\tau_0}(f)$ *holds uniformly for any* $\tau = (\theta, h) \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$ *and* $f \in L_0^2(\mathcal{X}_l)$*.*

Assumption A.7. *The quadratic functional* $g \mapsto \sum_{l=1}^{m} V_{l,\tau_0}(L_{l,h}g)$ *is completely continuous with respect to the quadratic functional* J*.*

Under Assumption A.7, Theorem 3.1 of Weinberger (1974) yields a sequence $\{\phi_\nu : \nu = 1, 2, \ldots\}$ of eigenfunctions and a sequence $\{\rho_\nu : \nu = 1, 2, \ldots\}$ of eigenvalues such that $\sum_{l=1}^{m} V_{l,\tau_0}(L_{l,h}\phi_\nu, L_{l,h}\phi_\mu) = \delta_{\nu\mu}$ and $J(\phi_\mu, \phi_\nu) = \rho_\nu \delta_{\nu\mu}$ for all pairs ν, μ of positive integers, where $\delta_{\nu\mu}$ is the Kronecker delta and $0 \leq \rho_\nu \to \infty$.

Assumption A.8. $\rho_\nu = \kappa_\nu \nu^r$*, where* $r > 1$ *and* $\kappa_\nu \in (\beta_1, \beta_2) \subset (0, \infty)$*.*

By Assumptions A.4 and A.7, for any $(\theta_*, h_*) \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$, there exist sequences of eigenfunctions $\{\phi_{*,\nu} : \nu = 1, 2, \ldots\}$ and eigenvalues $\{\rho_{*,\nu} : \nu = 1, 2, \ldots\}$ such that $\sum_{l=1}^{m} V_{l,\tau_0}(D_h\eta_l(\theta_*, h_*)\phi_{*,\nu}, D_h\eta_l(\theta_*, h_*)\phi_{*,\mu}) = \delta_{\nu\mu}$ and $J(\phi_{*,\mu}, \phi_{*,\nu}) = \rho_{*,\nu}\delta_{\nu\mu}$ for all pairs ν, μ of positive integers, where $0 \leq \rho_{*,\nu} \to \infty$. Assumptions A.4 and A.8 imply that $\rho_{*,\nu} \sim \nu^r$ for large enough ν, where $r > 1$. Furthermore, for every $h \in \mathcal{H}$ and any $(\theta_*, h_*) \in \mathcal{N}_{\theta_0} \times \mathcal{N}_{h_0}$, we have a Fourier expansion $h = \sum_{\nu=1}^{\infty} \sum_{l=1}^{m} V_{l,\tau_0}(D_h\eta_l(\theta_*, h_*)h, D_h\eta_l(\theta_*, h_*)\phi_{*,\nu})\phi_{*,\nu}$.

BIBLIOGRAPHY

Ambrosetti, A. and G. Prodi (1995). *A Primer of Nonlinear Analysis*, Volume 34. Cambridge University Press, Cambridge.

Aubin, J.-P. (1979). *Applied Functional Analysis*. John Wiley & Sons, New York-Chichester-Brisbane.

Breiman, L. (1996). Nonparametric roughness penalties for probability densities. *Machine Learning 24*(2), 123–140.

Cheng, G. and Z. Shang (2015). Joint asymptotics for semi-nonparametric regression models with partially linear structure. *Ann. Statist. 43*(3), 1351–1390.

Conway, J. B. (1990). *A Course in Functional Analysis* (Second ed.), Volume 96. Springer-Verlag, New York.

Cox, D. D. (1988). Approximation of method of regularization estimators. *Ann. Statist. 16*(2), 694–712.

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

Efron, B. and R. Tibshirani (1996). Using specially designed exponential families for density estimation. *Ann. Statist. 24*(6), 2431–2461.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc. 96*(456), 1348–1360.

Frank, L. E. and J. H. Friedman (1993). A statistical view of some chemometrics regression tools. *Technometrics 35*(2), 109–135.

Genton, M. G. and P. Hall (2007). Statistical inference for evolving periodic functions. *J. R. Stat. Soc. Ser. B. Stat. Methodol 69*(4), 643–657.

Good, I. and R. A. Gaskins (1971). Nonparametric roughness penalties for probability densities. *Biometrika 58*(2), 255–277.

Gu, C. (1995). Smoothing spline density estimation: conditional distribution. *Statist. Sinica*, 709–726.

Gu, C. (2013). *Smoothing Spline ANOVA Models* (Second ed.). Springer-Verlag, New York.

Gu, C. and C. Qiu (1993). Smoothing spline density estimation: theory. *Ann. Statist. 21*(1), 217–234.

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

Hjort, N. L. and I. K. Glad (1995). Nonparametric density estimation with a parametric start. *Ann. Statist. 23*(3), 882–904.

Ke, C. and Y. Wang (2004). Existence and uniqueness of penalized least square estimation for smoothing spline nonlinear nonparametric regression models. Unpublished Manuscript.

Ke, C. and Y. Wang (2008). Smoothing spline semi-parametric nonlinear regression models. Unpublished Manuscript.

Leng, C., Y. Lin, and G. Wahba (2006). A note on the lasso and related procedures in model selection. *Statist. Sinica 16*(4), 1273.

Lenk, P. J. (2003). Bayesian semiparametric density estimation and model verification using a logistic-Gaussian process. *J. Comput. Graph. Statist. 12*(3), 548–565.

Leonard, T. (1978). Density estimation, stochastic processes and prior information. *J. R. Stat. Soc. Ser. B. Stat. Methodol 40*(2), 113–132.

Mammen, E. and S. van de Geer (1997). Penalized quasi-likelihood estimation in partial linear models. *Ann. Statist. 25*(3), 1014–1035.

McCullagh, P. and J. A. Nelder (1989). *Generalized linear models* (Second ed.). Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Olkin, I. and C. H. Spiegelman (1987). A semiparametric approach to density estimation. *J. Amer. Statist. Assoc. 82*(399), 858–865.

O'Sullivan, F. (1990). Convergence characteristics of methods of regularization estimators for nonlinear operator equations. *SIAM J. Numer. Anal. 27*(6), 1635–1649.

Rudin, W. (1976). *Principles of Mathematical Analysis* (Third ed.). McGraw-Hill Book Co., New York-Auckland-Düsseldorf.

Shang, Z. (2010). Convergence rate and Bahadur type representation of general smoothing spline M-estimates. *Electron. J. Stat. 4*, 1411–1442.

Shang, Z. and G. Cheng (2013). Local and global asymptotic inference in smoothing spline models. *Ann. Statist. 41*(5), 2608–2638.

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist. 10*(3), 795–810.

Tapia, R. A. and J. R. Thompson (1978). Nonparametric probability density estimation.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol 58*(1), 267–288.

Wahba, G. (1990). *Spline Models for Observational Data*, Volume 59. SIAM, Philadelphia, PA.

Wand, M. P., J. S. Marron, and D. Ruppert (1991). Transformations in density estimation. *J. Amer. Statist. Assoc. 86*(414), 343–361.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the gaussnewton method. *Biometrika 61*(3), 439–447.

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

Yang, Y. (2009). Penalized semiparametric density estimation. *Stat. Comput. 19*(4), 355–366.

Yu, Y. and D. Ruppert (2002). Penalized spline estimation for partially linear single-index models. *J. Amer. Statist. Assoc. 97*(460), 1042–1054.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *Journal of Machine Learning Research 7*, 2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc. 101*(476), 1418–1429.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol 67*(2), 301–320.