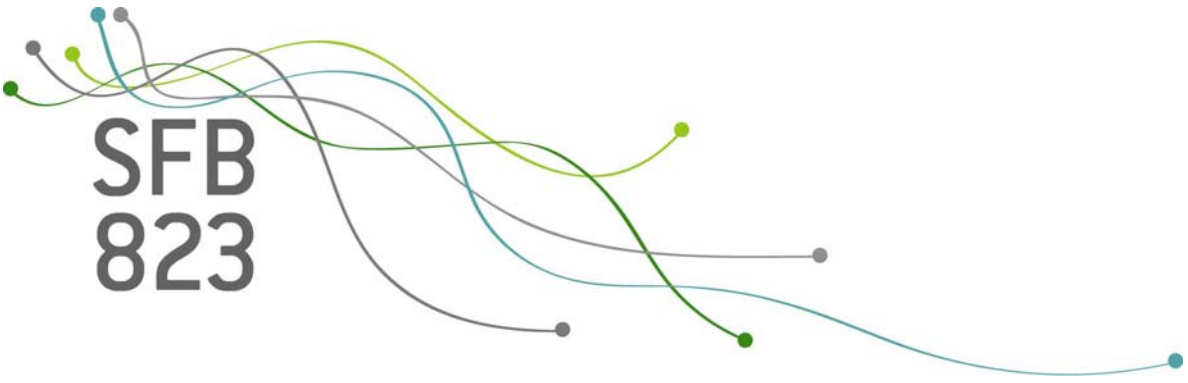
SFB 823

Sieve maximum likelihood estimation in a semi-parametric regression model with errors in variables

Denis Belomestny, Egor Klochkov, Vladimir Spokoiny

Nr. 15/2016

Discussion Paper



Sieve maximum likelihood estimation in a semi-parametric regression model with errors in variables

Denis Belomestny*

Duisburg-Essen University, Germany
and National Research University Higher School of Economics, Russian Federation
denis.belomestny@uni-due.de

Egor Klochkov
Humboldt University Berlin,
IRTG 1792
eklochkov@gmail.com

Vladimir Spokoiny
Weierstrass-Institute,
Humboldt University Berlin and IITP RAS,
Mohrenstr. 39, 10117 Berlin, Germany,
spokoiny@wias-berlin.de

March 29, 2016

Abstract

The paper deals with a semi-parametric regression problem under deterministic and regular design which is observed with errors. We first linearise the problem using a sieve approach and then apply the total penalised maximum likelihood estimator to the linearised model. Sufficient conditions for $\sqrt{n}$-consistency and efficiency under parametric assumption are derived and a possible misspecification bias under different smoothness assumptions on the design is analysed. The Monte Carlo simulations show the performance of the estimator with simulated data.

AMS 2000 Subject Classification: Primary 62F10. Secondary
Keywords: errors-in-varaibles model, regression, $\sqrt{n}$-consistency

*The author's research is supported by the Deutsche Forschungsgemeinschaft through the SFB 823 "Statistical modelling of nonlinear dynamic processes"

1 Introduction

Consider a classical regression model

$$Y_i = f(X_i) + \varepsilon_i\,, \qquad i = 1, \dots, n, \tag{1.1}$$

which describes the relation between the response variable Y and the independent variable $X \in \mathbb{R}^d$ Here Y_i, $i = 1, \dots, n$, are independent observations of Y and ε_i, $i = 1, \dots, n$, stand for the measurement errors, which are assumed to be i.i.d. random variables. We consider the situation when the regressors X_i are observed with some errors as well. Measurement error (or errors-in-variables) models have been extensively studied over the last decades, see, e.g., the monographs of Schneeweiß and Mittag (1986), Fuller (2009), Cheng and Van Ness (1999), Wansbeek and Meijer (2000), and Carroll et al. (1995). In particular, the last book deals almost exclusively with nonlinear regression models. According to the usual interpretation of measurement error models, Y is an observable variable depending on the regressor vector X, here via a nonlinear regression function f. The variable X, however, is not directly observable. Instead a perturbed vector $Z \in \mathbb{R}^d$ is observed, which is related to X via

$$Z_i = X_i + \sigma \boldsymbol{\nu}_i, \quad i = 1, \dots, n,$$

where σ^2 stands for the noise variance in the regressors Z_i, and ν_i are standardised random errors. The aim is to estimate the regression function f from the observations (Y_i, Z_i). The models (1.1) obviously faces an identification problem: if the distribution of the errors-in-regressors $\boldsymbol{\nu}_i$ is unknown, the function f cannot be recovered consistently. Some information about the distribution of the $\boldsymbol{\nu}_i$ can helpful in this context. Indeed, taking the expectation of (1.1) reveals that the conditional mean of Y_i is a convolution of the regression function f with some kernel corresponding to the distribution of the $\boldsymbol{\nu}_i$:

$$\mathbb{E}\big[Y \,\big|\, X\big] = \int f(X + \boldsymbol{x})\varphi_\sigma(\boldsymbol{x})d\boldsymbol{x}$$

Here $\varphi(\boldsymbol{x})$ is the density of the $\boldsymbol{\nu}_i$'s and $\varphi_\sigma(\boldsymbol{x}) = \sigma^{-1}\varphi(\boldsymbol{x}/\sigma)$. To recover f, one has to make a deconvolution which leads to an ill-posed inverse problem (see, e.g. Fan et al. (1991)). Note that the classical assumption that $(\boldsymbol{\nu}_i)$ are normally distributed is in some sense the worst-case and the function f can only be recovered with a $\log n$ accuracy in this case. We refer to Butucea and Taupin (2008) for the detailed discussion and the overview of the literature on this problem. In this paper we suggest a new approach to this problem which treats the unobserved regressors (X_i) as a high-dimensional nuisance parameter. Such a model is characterised by an increasing number of nuisance parameters

and a fixed number of parameters of interest. It is well known that in models with an increasing number of nuisance parameters usual estimation procedures may fail to be consistent. The main problem tackled here is the elimination of the nuisance parameters in such a way that the estimating procedure delivers $\sqrt{n}$-consistent estimators. While in linear models the total least squares approach gives, under some conditions on the design, consistent and efficient estimators for the parameter of interest, the situation is much more involved in the case of nonlinear models.

For simplicity we assume below that $d = 1$ and the function f is univariate. We apply the linear sieve approach which approximates the target function f as a linear combination of the fixed basis functions $\Psi(x) = (\psi_1(x), \ldots, \psi_p(x))$:

$$f(x) \approx f(x, \boldsymbol{\theta}) = \Psi(x)^\top \boldsymbol{\theta} = \sum_{m=1}^{p} \theta_m \psi_m(x).$$

The function $f(x, \boldsymbol{\theta})$ is described by the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ and the estimation problem for the function f is equivalent to estimation of $\boldsymbol{\theta}$. Suppose for a moment that the errors ε_i are i.i.d. standard normal and the errors ν_i are i.i.d. zero mean normal with the variance σ^2. Then the log-likelihood for the full parameter $\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{X})$ with $\boldsymbol{X} = (X_1, \ldots, X_n)$ reads as follows:

$$L(\boldsymbol{\upsilon}) = L(\boldsymbol{\theta}, \boldsymbol{X}) = -\frac{1}{2}\|\boldsymbol{Y} - \Psi(\boldsymbol{X})^\top \boldsymbol{\theta}\|^2 - \frac{1}{2\sigma^2}\|\boldsymbol{Z} - \boldsymbol{X}\|^2$$

Here $\Psi(\boldsymbol{X})$ is the $p \times n$-matrix with entries $\psi_m(X_i)$ for $m = 1, \ldots, p$ and $i = 1, \ldots, n$. Even if $d = 1$, the parameter dimension is $p+n$ and it is larger than the sample size. So we need to reduce the complexity of estimation of the high-dimensional nuisance vector $\boldsymbol{X}$. Let $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_n$ are n orthonormal vectors in $\mathbb{R}^n$, i.e., it holds $\boldsymbol{\phi}_i^\top \boldsymbol{\phi}_j = \delta_{ij}$, then we can always represent the vector $\boldsymbol{X}$ in the form

$$\boldsymbol{X} = \eta_1 \boldsymbol{\phi}_1 + \cdots + \eta_n \boldsymbol{\phi}_n$$

for an unknown vector $\boldsymbol{\eta} \in \mathbb{R}^n$. We assume that $\boldsymbol{X}$ can be approximately spanned by first $q \ll n$ of $\boldsymbol{\phi}_j$, i.e.

$$\boldsymbol{X} \approx \eta_1 \boldsymbol{\phi}_1 + \cdots + \eta_q \boldsymbol{\phi}_q = \Phi\boldsymbol{\eta}, \qquad \boldsymbol{\eta} \in I\!R^q, \qquad \Phi = \big(\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_q\big). \tag{1.2}$$

This leads to the following quasi-log-likelihood parametrized by $\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta})$

$$L_\Phi(\boldsymbol{\upsilon}) = L_\Phi(\boldsymbol{\theta}, \boldsymbol{\eta}) \stackrel{\text{def}}{=} L(\boldsymbol{\theta}, \Phi\boldsymbol{\eta}) = -\frac{1}{2}\|\boldsymbol{Y} - \Psi(\Phi\boldsymbol{\eta})^\top \boldsymbol{\theta}\|^2 - \frac{1}{2\sigma^2}\|\boldsymbol{Z} - \Phi\boldsymbol{\eta}\|^2, \tag{1.3}$$

with the total dimension $p + q$ and dimension of the target parameter p. Note, for example, that the equality $\boldsymbol{X} = \Phi\boldsymbol{\eta}$ with $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_q$ being the first q vectors of Haar

basis means the aggregation of points, which is equivalent to *repeated measurements* model Fuller (2009). In general $\boldsymbol{X} \neq \Phi\boldsymbol{\eta}$ and we need some conditions that allow to control bias caused by dimension reduction.

2 Main results

Let us consider the following auxiliary semiparametric model

$$\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\varepsilon}, \quad \boldsymbol{Z} = \boldsymbol{X}^* + \sigma\boldsymbol{\nu}, \tag{2.1}$$

with $\boldsymbol{f}^* = (f(X_1^*), \ldots, f(X_n^*))^\top$, where $\boldsymbol{X}^* = (X_1, \ldots, X_n)$ is the true (unknown) design. Let us approximate f via $f(x, \boldsymbol{\theta}) = \sum_{m=1}^p \theta_m \psi_m(x)$ with $\boldsymbol{\theta} \in I\!R^p$. The model (2.1) can be viewed as a sieve approximation for the original model (1.1). In the next sections we study the properties of this sieve maximum likelihood approach.

Note that non-concavity of the log-likelihood (1.3) is a problem from both optimization and statistical points of view. In Section 3 we suggest to conduct an iterative optimization procedure starting from the *plug-in* estimator:

$$\begin{aligned}
\widetilde{\boldsymbol{\upsilon}}^{(\mathsf{pl})} &= (\widetilde{\boldsymbol{\theta}}^{(\mathsf{pl})}, \widetilde{\boldsymbol{\eta}}^{(\mathsf{pl})}) \\
\widetilde{\boldsymbol{\theta}}^{(\mathsf{pl})} &= \bigl(\widetilde{\Psi}\widetilde{\Psi}^\top\bigr)^{-1}\widetilde{\Psi}\boldsymbol{Y}, \qquad \widetilde{\Psi} = \Psi(\Phi\widetilde{\boldsymbol{\eta}}^{(\mathsf{pl})}), \\
\widetilde{\boldsymbol{\eta}}^{(\mathsf{pl})} &= \Phi^\top \boldsymbol{Z}.
\end{aligned}$$

This is typically a good initial estimator of the full parameter $\boldsymbol{\upsilon}^* = (\boldsymbol{\theta}^*, \boldsymbol{\eta}^*)$. However, in order to proceed with semiparametric theory, we need to have at least one stationary point of the likelihood, which is close to the true point with high probability, see Section 4.7. Since $\widetilde{\boldsymbol{\upsilon}}^{(\mathsf{pl})}$ is usually already close to $\boldsymbol{\upsilon}^*$, we suggest to define the estimator as the closest stationary point of the likelihood to plug-in $\widetilde{\boldsymbol{\upsilon}}^{(\mathsf{pl})}$:

$$\widetilde{\boldsymbol{\upsilon}} = \arg \min_{\substack{\boldsymbol{\upsilon}=(\boldsymbol{\theta},\boldsymbol{\eta}) \in I\!R^{p+q}: \\ \nabla L_\Phi(\boldsymbol{\upsilon})=0}} \bigl\|\widetilde{\Psi}^\top(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}^{(\mathsf{pl})})\bigr\|^2 + \sigma^{-2}\|\boldsymbol{\eta} - \widetilde{\boldsymbol{\eta}}^{(\mathsf{pl})}\|^2, \tag{2.2}$$

and, similarly,

$$\widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*} = \arg \min_{\substack{\boldsymbol{\eta} \in I\!R^q: \\ \nabla_{\boldsymbol{\eta}} L_\Phi(\boldsymbol{\theta}^*, \boldsymbol{\eta})=0}} \|\boldsymbol{\eta} - \widetilde{\boldsymbol{\eta}}^{(\mathsf{pl})}\|.$$

2.1 Consistency and efficiency

Suppose that $\boldsymbol{\theta}^* \in I\!R^p$ $\boldsymbol{\eta}^* \in I\!R^q$ are two vectors and (ψ_m), (ϕ_m) are two sieves such that the following conditions hold.

(i) Suppose that for each $i = 1, \dots, n,$ and each $s = 0, 1, 2$ it holds

$$|f^{(s)}(X_i)| = \left|\Psi^{(s)}(X_i)^\top \boldsymbol{\theta}^*\right| = \Big|\sum_{m=1}^{p} \theta^*_m \psi^{(s)}_m(X_i)\Big| \le \mathtt{C}_{f,s}.$$

(ii) Assume that the eigenvalues of the matrix

$$\frac{\Psi(\Phi\boldsymbol{\eta}^*)\Psi(\Phi\boldsymbol{\eta}^*)^\top}{n}$$

belong to the interval $[\mathtt{f}, \mathtt{F}]$ for fixed positive constants $\mathtt{f}$ and $\mathtt{F}$;

(iii) For each $s = 0, 1, 2$, there are some fixed constants $\mu_{1,s} \le \mu_{2,s} \le \dots \mu_{p,s}$ such that each function $\psi_m(x)$ fulfills for all x

$$|\psi^{(s)}_m(x)| \le \mu_{m,s}\,. \tag{2.3}$$

(iv) The orthogonal vectors $\boldsymbol{\phi}_j$ fulfil $\|\boldsymbol{\phi}_j\| = 1$ and

$$\max_{i=1,\dots,n} \sum_{j=1}^{q} \phi^2_{ij} \le \frac{\mathtt{v}^2_q}{n},$$

where $\mathtt{v}_q$, as we will see, typically depends on q.

(v) The error of approximating $\boldsymbol{f}^*$ via $\Psi(\Phi\boldsymbol{\eta}^*)^\top\boldsymbol{\theta}^*$ is not large, in the sense that

$$\square \stackrel{\text{def}}{=} \|\boldsymbol{f}^* - \Psi(\Phi\boldsymbol{\eta}^*)^\top\boldsymbol{\theta}^*\| \le \sqrt{n}.$$

(vi) Errors $(\varepsilon_i), (\nu_i)$ are zero mean mutually independent and satisfy for some $0 < \nu_\varepsilon \le 1$, $i = 1, \dots, n,$ and each λ

$$\log I\!E \exp(\lambda\varepsilon_i) \le \frac{\lambda^2\nu^2_\varepsilon}{2}, \qquad \log I\!E \exp(\lambda\nu_i) \le \frac{\lambda^2\nu^2_\varepsilon}{2},$$

which implies $\max\{\mathrm{Var}(\boldsymbol{\varepsilon}), \mathrm{Var}(\boldsymbol{\nu})\} \le \nu^2_\varepsilon \boldsymbol{I}_n$.

(vii) With $\mathtt{f}$ from (ii) and $\mu_{m,s}$ from (iii), it holds for some $\mathtt{w}_{p,2} \ge \mathtt{w}_{p,1} \ge \mathtt{w}_{p,0} > 0$,

$$\mathtt{f}^{-1} \sum_{m=1}^{p} \mu^2_{m,s} \le \mathtt{w}^2_{p,s}, \qquad s = 0, 1, 2, \quad p \in \mathbb{N}.$$

Moreover,

$$\sigma\mathtt{w}_{p,1}\sqrt{p + q + \mathtt{x}} \lesssim \sqrt{n}, \qquad \square \lesssim \mathtt{w}^{-1}_{p,1}\sqrt{n},$$
$$\square \le (\mathtt{w}_{p,2}\mathtt{v}_q)^{-1/2}\sqrt{n}, \qquad \square \le \mathtt{v}^{-1}_q\sqrt{n},$$

where expression "$\lesssim$" means inequality up to a small constant depending on σ, $\mathtt{C}_{f,1}$, $\mathtt{C}_{f,2}$, $\mathtt{f}$ and ν_ε only. It is worth to mention, that in typical situations $\mathtt{w}_{p,s} \asymp p^{s+\frac{1}{2}}$ and the two last inequalities are likely to be fulfilled, while the first two require n two be large enough relatively to p and q.

Main notations The semiparametric *Fisher information* matrix reads as follows,

$$\breve{D}^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \doteq D^2_{\boldsymbol{\theta}\boldsymbol{\theta}} - A_{\boldsymbol{\theta}\boldsymbol{\eta}} D^{-2}_{\boldsymbol{\eta}\boldsymbol{\eta}} A^\top_{\boldsymbol{\theta}\boldsymbol{\eta}}, \tag{2.4}$$

where

$$\begin{aligned} D^2_{\boldsymbol{\theta}\boldsymbol{\theta}} &\doteq \Psi(\Phi\boldsymbol{\eta}^*)\,\Psi(\Phi\boldsymbol{\eta}^*)^\top, \\ A_{\boldsymbol{\theta}\boldsymbol{\eta}} &\doteq \Psi(\Phi\boldsymbol{\eta}^*)\,\operatorname{diag}\bigl\{\Psi'(\Phi\boldsymbol{\eta}^*)^\top\boldsymbol{\theta}^*\bigr\}\Phi, \\ D^2_{\boldsymbol{\eta}\boldsymbol{\eta}} &\doteq \sigma^{-2} I_q + \Phi^\top \operatorname{diag}\bigl\{\Psi'(\Phi\boldsymbol{\eta}^*)^\top\boldsymbol{\theta}^*\bigr\}^2\Phi. \end{aligned}$$

Set

$$\diamondsuit(\mathtt{x}) \asymp \sigma\left(\mathtt{w}_{p,1} + \mathtt{v}_q + \mathtt{w}_{p,2}\mathtt{v}_q\sqrt{n^{-1}(p+q+\mathtt{x})}\right)(p+q+\mathtt{x})n^{-1/2} + \square. \tag{2.5}$$

Theorem 2.1. *Under the above assumptions, there exists random vector $\breve{\boldsymbol{\xi}} \in \mathbb{R}^p$ satisfying for each $\mathtt{x} \geq 0$,*

$$\mathbb{P}\left(\|\breve{\boldsymbol{\xi}}\| > \sqrt{p} + \sqrt{2\mathtt{x}}\right) \leq 2e^{-\mathtt{x}},$$

such that the following statements hold.

- *If $\diamondsuit^2(\mathtt{x})/p \to 0$, then $\widetilde{\boldsymbol{\theta}}$ is root-n consistent and $\mathbb{E}\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 \asymp p/n$.*
- *If $\diamondsuit(\mathtt{x}) \to 0$, then $\widetilde{\boldsymbol{\theta}}$ is root-n asymptotically normal and the Fisher expansion holds*

$$\bigl\|\breve{D}_{\boldsymbol{\theta}\boldsymbol{\theta}}\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\bigr) - \breve{\boldsymbol{\xi}}\bigr\| \leq \diamondsuit(\mathtt{x})$$

on a set with probability at least $1 - 6e^{-\mathtt{x}}$.

- *If $p\,\diamondsuit^2(\mathtt{x}) \to 0$, then the Wilks expansion holds*

$$\bigl|L_\Phi(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\eta}}) - L_\Phi(\boldsymbol{\theta}^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}) - \|\breve{\boldsymbol{\xi}}\|^2/2\bigr| \leq \sqrt{p}\,\diamondsuit(\mathtt{x})$$

on a set with probability at least $1 - 6e^{-\mathtt{x}}$.

Discussion Note, that the estimator $\widetilde{\boldsymbol{\theta}}$ depends on the matrix Φ through the linear subspace $\{\Phi\boldsymbol{\eta} : \boldsymbol{\eta} \in I\!R^q\}$. It is reasonable to expect that the condition (iv) is actually a condition for the span of columns of Φ. Indeed, if we have another orthonormal matrix $\widetilde{\Phi}$, that generates the same subspace, then there is S such that

$$\widetilde{\Phi} = \Phi S, \qquad S^\top S = \boldsymbol{I}.$$

Introduce

$$V(\Phi) \stackrel{\text{def}}{=} n \max_{i=1,\ldots,n} \|\Phi_i\|^2,$$

where $\Phi_1, \ldots, \Phi_n$ are the rows of the matrix Φ. Then, the rows of matrix $\widetilde{\Phi}$ are $\Phi_i S$ and we have

$$V(\widetilde{\Phi}) = n \max_{i=1,\ldots,n} \|\Phi_i S\|^2 = V(\Phi), \tag{2.6}$$

since S is orthogonal. Thus, the value $V(\Phi)$ remains the same for all orthogonal matrices Φ representing the same subspace. Now let us analyse the value v_q from the condition (iv). Since Φ is orthonormal, we have $\operatorname{tr}(\Phi^\top\Phi) = q$. On the other hand,

$$n \max_{i\le n} \|\Phi_i\|^2 \ge \sum_{i=1}^n \|\Phi_i\|^2 = \operatorname{tr}(\Phi\Phi^\top),$$

implying

$$\mathsf{v}_q \ge \sqrt{q}. \tag{2.7}$$

Below we show two examples with $\mathsf{v}_q \asymp \sqrt{q}$.

Example 2.1. A possible representation of a "smooth" design $\boldsymbol{X}^*$ can be given by

$$X_i^* = g(i/n), \qquad g : [0,1] \to I\!R, \qquad i = 1, \ldots, n, \tag{2.8}$$

with smoothness of the function $g(t)$ corresponding to that of $\boldsymbol{X}^*$. Suppose we are given an orthonormal basis $\{h_m(t)\}_{m=1}^\infty$ with $\langle h_i, h_j\rangle_{L^2[0,1]} = \delta_{ij}$. Consider the corresponding sieve approximation of the function g:

$$g(t) \approx \sum_{m=1}^q \eta_m h_m(t).$$

This corresponds to Φ in (1.2) defined as follows,

$$\boldsymbol{\phi}_m = \sqrt{\frac{1}{n}} \left(h_m\left(\frac{1}{n}\right), \ldots, h_m(1) \right), \qquad \Phi = \big(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_q\big). \tag{2.9}$$

These vectors satisfy condition (iv), but fail to be orthogonal. Nevertheless, due to orthogonality of the basis $(h_m(t),\, m = 1, \dots, q)$ under mild conditions we can orthogonalize Φ without much loss in $V(\Phi)$.

Lemma 2.2. *Suppose, that each h_m is twice differentiable and*

$$\|h_m^{(s)}\|_\infty \le A_s m^s, \qquad s = 0, 1, 2.$$

If $A_0^2 q/n + (A_0 A_2 + A_1^2) q^3/(6n^2) \le \frac{1}{2}$, then the matrix Φ in (2.9) can be transformed into $\widetilde{\Phi}$, such that $\widetilde{\Phi}^\top \widetilde{\Phi} = \boldsymbol{I}_q$ and $V(\widetilde{\Phi}) \le 2V(\Phi)$.

Example 2.2. Another example is given by Haar basis, which might be relevant to image processing, when the function $g(t)$ in (2.8) is piecewise smooth. Suppose that $n = 2^k$ and consider the sequence of partitions $\mathcal{S}^{(0)}, \mathcal{S}^{(1)}, \dots, \mathcal{S}^{(k)}$ of the set of indices $\mathcal{N} = \{1, 2, \dots, n\}$. First, we have $\mathcal{S}^{(0)} = \{A_1^{(0)} = \mathcal{N}\}$ — the whole set. Further, $\mathcal{S}^{(l+1)}$ is made by splitting each $A_j^{(l)} \in \mathcal{S}^{(l)}$ into two equal parts $A_{2j-1}^{(l+1)}$ and $A_{2j}^{(l+1)}$. Denoting $\mathcal{S}^{(l)} = \{A_1^{(l)}, \dots, A_{2^l}^{(l)}\}$, we have

$$\begin{aligned}
&A_1^{(1)} = \{1, \dots, n/2\}, && A_1^{(2)} = \{1, \dots, n/4\}, && \dots \\
&A_2^{(1)} = \{n/2+1, \dots, n\}, && A_2^{(2)} = \{n/4+1, \dots, n/2\}, && \dots \\
& && A_3^{(2)} = \{n/2+1, \dots, 3n/4\}, && \\
& && A_4^{(2)} = \{3n/4+1, \dots, n\}, &&
\end{aligned}$$

Continuing the sequence we get for $l \le k-1$ and $1 \le r \le 2^l$ that $A_r^{(l)} = \{(r-1)n/2^l + 1, \dots, rn/2^l\}$. Note, that $\# A_r^{(l)} = n 2^{-l}$. Now, denoting by $\boldsymbol{\phi}(i)$ the ith component of the vector $\boldsymbol{\phi}$, introduce

$$\boldsymbol{\phi}_1(i) = \frac{1}{\sqrt{n}}, \qquad \boldsymbol{\phi}_{2^{l-1}+r}(i) = \sqrt{\frac{2^{l-1}}{n}} \begin{cases} 1, & i \in A_{2r-1}^{(l)}, \\ -1, & i \in A_{2r}^{(l)}, \\ 0, & \text{otherwise} \end{cases}$$

where, obviously, each $j = 2, \dots, n$, is uniquely representable as $j = 2^{l-1} + r$ with $1 \le l \le k$ and $1 \le r \le 2^{l-1}$. By the construction, $(\boldsymbol{\phi}_j)_{j=1}^n$ is an orthonormal basis in $\mathbb{R}^n$. Moreover, it's easy to see, that for $q = 2^l$ the span of vectors $\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_q$ is

$$\{\boldsymbol{X} \in \mathbb{R}^n : X_i \text{ is constant over each } A \in \mathcal{S}^{(l)}\},$$

and that the projector $\Phi\Phi^\top$ on that subspace, when applied to $\boldsymbol{X} \in \mathbb{R}^n$, averages values X_i over $i \in A$ for each $A \in \mathcal{S}^{(l)}$. By (2.6) we can consider any other orthonormal $\widetilde{\Phi}$,

generating the same subspace. One of them is

$$\widetilde{\Phi} = \sqrt{\frac{2^l}{n}} \begin{pmatrix} \overbrace{1\,1\,\ldots\,1}^{n/2^l} & \overbrace{0\,0\,\ldots 0}^{n/2^l} & \ldots & \overbrace{0\,0\,\ldots 0}^{n/2^l} \\ 0\,0\,\ldots\,0 & 1\,1\,\ldots\,1 & \ldots & 0\,0\,\ldots\,0 \\ & & \ldots & \\ 0\,0\,\ldots 0 & 0\,0\,\ldots\,0 & \ldots & 1\,1\,\ldots\,1 \end{pmatrix},$$

for which condition (v) is satisfied with $\mathsf{v}_q = \sqrt{2^l} = \sqrt{q}$, which is exactly the lower bound (2.7). Note, that when n is not a power of 2, one can make the same sequence of partition dividing at each step $A_r^{(l-1)}$ into $A_{2r-1}^{(l)}$ and $A_{2r}^{(l)}$ such that $\bigl|\#A_j^{(l)} - \#A_{j+1}^{(l)}\bigr| \le 1$ for each $l \le \log_2 n$, $j < 2^l$.

2.2 Application to composite function estimation

Consider a regression model of the form

$$Y_i = f\bigl(g(i/n)\bigr) + \varepsilon_i, \qquad Z_i = g(i/n) + \nu_i, \qquad i = 1, \ldots, n, \tag{2.10}$$

where f and g are two real valued functions. Here the problem is to estimate the function f from the observations $(Y_1, Z_1), \ldots, (Y_n, Z_n)$. This problem is related to the problem of composite function estimation which recently got much attention in the literature (see, e.g. Juditsky et al. (2009) and references therein). However our main interest here lies in estimating the function f and not the whole composite function $f(g(\cdot))$. Let us apply finite-dimensional sieve approximations to f and g:

$$f(x) \approx f_p(x) = \sum_{m=1}^{p} \theta_m \psi_m(x), \qquad g(x) \approx g_q(x) = \sum_{m=1}^{q} \eta_m \phi_m(x).$$

Thus, our problem is reduced to a parametric model (2.1) with parameter $(\boldsymbol{\theta}, \boldsymbol{\eta}) \in I\!R^{p+q}$ and the estimator $\widetilde{f}_{p,q}$ of f is given by $\widetilde{\boldsymbol{\theta}}$ from (2.2) corresponding to Φ given in (2.9):

$$\widetilde{f}_{p,q}(x) = \Psi(x)^\top \widetilde{\boldsymbol{\theta}} = \sum_{m=1}^{p} \widetilde{\theta}_m \psi_m(x).$$

The performance of the estimator depends on the values p, q and on the accuracy of corresponding sieve approximation. One can describe the dependence of the bias of $\widetilde{f}_{p,q}$ on the values of p and q through the so-called *Sobolev elipsoids*

$$S_\psi(\beta, Q) = \left\{ f(x) = \sum_{m=1}^{\infty} \theta_m \psi_m(x) \ : \ \boldsymbol{\theta} \in \ell^2, \ \sum_{m=1}^{\infty} m^{2\beta} \theta_m^2 \le Q^2 \right\}.$$

We refer to Tsybakov (2009) for the proof of the following result.

Lemma 2.3. *Let* $f \in S_\psi(\beta, Q)$, *then*

$$\|f - f_p\|_2 \le Qp^{-\beta}. \tag{2.11}$$

Moreover, if $\beta > s + 1/2$ *and each* $\psi_m(x)$, *is* $s \ge 0$ *times continuously differentiable with* $\|\psi_m^{(s)}\|_\infty \le Am^s$, $m = 1, 2 \dots$ *for some* $A > 0$, *then*

$$\|f_p^{(s)}\|_\infty \le \frac{AQ}{\sqrt{2(\beta - s) - 1}}, \qquad \|f^{(s)} - f_p^{(s)}\|_\infty \le \frac{AQ}{\sqrt{2(\beta - s) - 1}} p^{-(\beta - s) + \frac{1}{2}}. \tag{2.12}$$

In order to apply Theorem 2.1, we need to check conditions (i) - (vi). Here we introduce a list of conditions for the nonparametric problem, which ensure those of the reduced parametric problem.

(a) The bases (ψ_m) and (ϕ_m) satisfy, for some $A > 0$ and each $s = 0, 1, 2$,

$$\|\psi_m^{(s)}\|_\infty \le Am^s, \qquad \|\phi_m^{(s)}\|_\infty \le Am^s;$$

(b) $f \in S_\psi(\beta, Q)$ and $g \in S_\phi(\beta_X, Q)$;

(c) $\Psi(\boldsymbol{X}^*)\Psi(\boldsymbol{X}^*)^\top \ge (\mathtt{f}n)\boldsymbol{I}_p$;

(d) for some $\mathtt{C}_0$ and each interval $I \subset I\!R$ it holds

$$\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i^* \in I\} \le \mathtt{C}_0 \max\left(\mu(I), \frac{1}{n}\right),$$

where $\mu(I)$ is the length of the interval.

Discussion Let us remark on the assumption (c). One may argue that this condition is not exactly nonparametric, as it depends on the dimension p. In that case we suggest to consider another condition, that seems more interpretable. The following lemma states, that if the the points X_i^* are dense enough on the interval $[0, 1]$, then (c) is satisfied.

Lemma 2.4. *Let* (ψ_m) *satisfy (a). Suppose, that with some constant* $\mathtt{c}_0 > 0$ *and some small enough* $\delta > 0$ *it holds*

$$\frac{1}{n}\sum_{i=1}^n \mathbb{1}\big\{X_i^* \in I\big\} \ge \mathtt{c}_0\big(\mu(I) - \delta p^{-2}\big)$$

for each interval $I \subset [0, 1]$. *Then we have,*

$$n^{-1}\Psi(\boldsymbol{X}^*)\Psi(\boldsymbol{X}^*)^\top \ge \mathtt{c}_0\big(1 - \alpha(\delta)\big)\boldsymbol{I}_p,$$

with $\alpha(\delta) \to 0$ *as* $\delta \to 0$.

Let us check other conditions of Theorem 2.1. Condition (v) follows from the following lemma.

Lemma 2.5. *Let (a) and (d) hold, and (b) be satisfied with* $\beta, \beta_X > 3/2$. *Then,*

$$\|\boldsymbol{f}^* - \Psi(\boldsymbol{X}^*)^\top \boldsymbol{\theta}^*\| \le \mathtt{C}_0^{1/2} Q \sqrt{n} p^{-\beta} + \frac{\mathtt{C}_0^{1/2} AQ}{\sqrt{\beta - 3/2}} p^{-\beta+1},$$

and

$$\|\Phi \boldsymbol{\eta}^* - \boldsymbol{X}^*\| \le Q\sqrt{n} q^{-\beta_X} + \frac{AQ}{\sqrt{\beta_X - 3/2}} q^{-\beta_X+1}.$$

Additionally, if $p, q \le \sqrt{n}$, *then it holds*

$$\|\boldsymbol{f}^* - \Psi(\boldsymbol{X}^*)^\top \boldsymbol{\theta}^*\| \le Q' \sqrt{n} p^{-\beta}, \qquad \|\Phi \boldsymbol{\eta}^* - \boldsymbol{X}^*\| \le Q' \sqrt{n} q^{-\beta_X}, \tag{2.13}$$

with $Q' = \mathtt{C}_0^{1/2} Q \left(1 + \frac{A}{\sqrt{\min\{\beta, \beta_X\} - 3/2}}\right)$ *not depending on* p, q *and* n.

Condition (i) follows from inequality (2.12), and the rest of the conditions can be checked straightforward. Manipulating with the error term $\diamondsuit(\mathtt{x})$, given in (2.5), we can show the following result.

Proposition 2.6. *Let* $\beta > 2.5$ *and* $\beta_X > 3/2$. *Then, by choosing*

$$p \stackrel{\text{def}}{=} n^{\frac{1}{2\beta+1}}, \qquad q \stackrel{\text{def}}{=} p^{\beta/\beta_X} = n^{\frac{\beta}{2\beta\beta_X + \beta_X}},$$

we get the standard non-parametric rate of convergence

$$\|f - \widetilde{f}_{p,q}\|_2 \lesssim_{I\!P} n^{-\frac{\beta}{2\beta+1}}.$$

This rate is in fact optimal over the class $S_\psi(\beta, Q)$.

A stronger assumption, that f is Hölder smooth on $[0,1]$, i.e. $f \in \Sigma(\beta, L)$ (see, e.g. Tsybakov (2009)) for some $\beta > 0$ and $\psi_m(x) = \cos(2\pi m x)$, yields $f \in S(\beta, L/\pi^\beta)$ and the corresponding inequality (2.11). However, one can get a better bound in the sup-norm

$$\|f^{(s)} - f_p^{(s)}\|_\infty \le L' p^{-(\beta - s)} \ln p,$$

for some L' not depending on p and $\beta > s$ (compared to $\beta > s + 1/2$ in (2.12)). Thus, such an assumption helps to "relax" the conditions $\beta > 2.5$ and $\beta_X > 3/2$ in the Proposition 2.6. Assume that

(a') both basis are trigonometric: $\psi_m(x) = \phi_m(x) = \cos(2\pi m x)$;

(b') $f \in \Sigma(\beta, L)$ and $g \in \Sigma(\beta_X, L)$.

and the conditions (c), (d) remain the same. Then we have the following proposition.

Proposition 2.7. *Assume (a'), (b'), (c) and (d). Let $\beta > 2$ and $\beta_X \geq 3/2$. Then, by choosing*

$$p \stackrel{\text{def}}{=} n^{\frac{1}{2\beta+1}}, \qquad q \stackrel{\text{def}}{=} p^{\beta/\beta_X} = n^{\frac{\beta}{2\beta\beta_X+\beta_X}},$$

we get the standard non-parametric rate of convergence

$$\|f - \widetilde{f}_{p,q}\|_2 \lesssim_{I\!P} n^{-\frac{\beta}{2\beta+1}}.$$

This rate is in fact optimal over the class $\Sigma(\beta, L)$.

3 Simulation results

Consider first a semiparametric regression model (2.10) with

$$f(x) = \sum_{k=1}^{p} \theta_m^* \cos(\pi m x)$$

and

$$g(x) = \sum_{k=1}^{\infty} \frac{\cos(\pi k x)}{k^{2\beta}} = (-1)^{\beta-1}(2\pi)^{2\beta} \frac{B_{2\beta}(x/2)}{2(2\beta)!}, \quad x \in [0,1]$$

for some natural $\beta, p > 0$, where $B_m(x)$ is a Bernoulli polynomial of order m. Note that $g \in S_{\cos}(2\beta + 1/2, Q)$ for some $Q > 0$. The aim is to estimate the function f, i.e., the vector $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_p^*)$ based on the sample $(Y_1, Z_1), \ldots, (Y_n, Z_n)$. We assume that the errors ε_i, ν_i, $i = 1, \ldots, n$, in (2.10) are normal i.i.d. with zero mean and variance 0.1. We consider the following sieve approximation for g

$$g(x) \approx g_q(x) = \sum_{m=1}^{q} \eta_m^* \cos(\pi m x)$$

with $\eta_m^* = m^{-2\beta}$. In order to compute the corresponding ML estimates for $\boldsymbol{\theta}^*$ and $\boldsymbol{\eta}^* = (\eta_1^*, \ldots, \eta_q^*)$ we use the following strategy. First we compute a preliminary plug-in estimate $\boldsymbol{\eta}^{(0)}$ for $\boldsymbol{\eta}^*$ by applying the standard least-squares approach to the linear model

$$Z_i = g_q(i/n) + \nu_i, \quad i = 1, \ldots, n.$$

A preliminary estimate $\boldsymbol{\theta}^{(0)}$ for $\boldsymbol{\theta}^*$ is then obtained from the model

$$Y_i = f\big(g_q^{(0)}(i/n)\big) + \varepsilon_i, \quad i = 1, \ldots n,$$

with $g_q^{(0)}(x) = \sum_{m=1}^{q} \eta_m^{(0)} \phi_m(x)$. Then we maximize the ML function

$$L_\Phi(\boldsymbol{\theta}, \boldsymbol{\eta}) = -\|\boldsymbol{Y} - \Psi(\Phi\boldsymbol{\eta})^\top \boldsymbol{\theta}\|^2 - \|\boldsymbol{Z} - \Phi\boldsymbol{\eta}\|^2$$

iteratively over $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ starting at $(\boldsymbol{\theta}^{(0)}, \boldsymbol{\eta}^{(0)})$. To minimize $L_\Phi(\boldsymbol{\theta}, \boldsymbol{\eta})$ over $\boldsymbol{\eta}$, we use the Levenberg-Marquardt algorithm (`minpack.lm` library in `R`). Such alternation maximization procedure usually converges rather quickly. In Figure 3.1 we show the boxplots of the weighted norm $(n/p)\|\boldsymbol{\theta}^* - \boldsymbol{\theta}^{(10)}\|^2$ based on 1000 independent samples each of length n from an EIV model with parameters $\boldsymbol{\theta}^* = (1, \dots, 1)$, $\beta = 2$, $q = 5$ as function of n (left) and p (right). A stable behaviour can be observed for increasing values of n and p supporting the results of Theorem 2.1. Next we turn to a nonparametric model

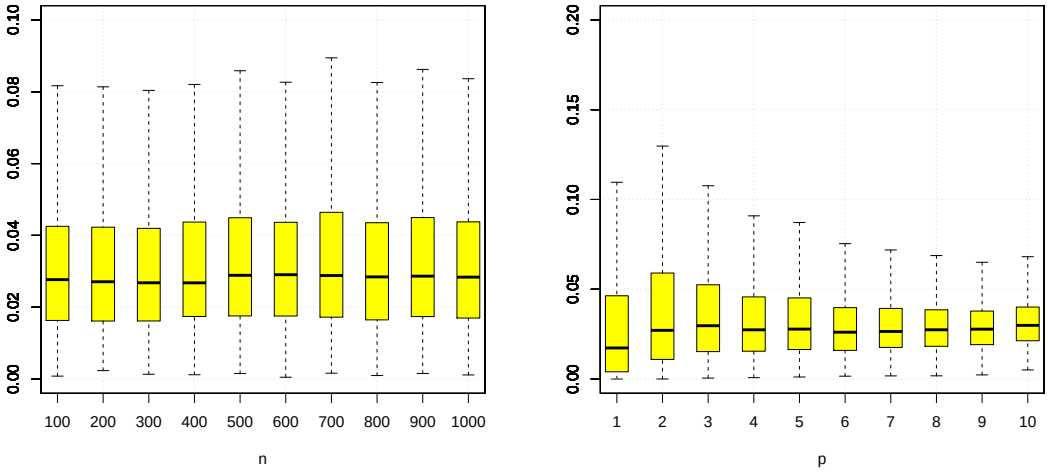


Figure 3.1: Left: the boxplots of the loss $(n/p)\|\boldsymbol{\theta}^* - \boldsymbol{\theta}^{(10)}\|^2$ as a function of the sample size n. Right: the boxplots of the loss $(n/p)\|\boldsymbol{\theta}^* - \boldsymbol{\theta}^{(10)}\|^2$ as a function of the dimension p.

with

$$f(x) = \sum_{k=1}^{\infty} \frac{\sin(\pi k x)}{k^{2\alpha+1}} = (-1)^{\alpha-1}(2\pi)^{2\alpha+1} \frac{B_{2\alpha+1}(x/2)}{2(2\alpha+1)!}, \quad x \in [0,1],$$

where $\alpha = 3$. In Figure 3.2 we show the empirical error $\|\boldsymbol{Y} - \Psi(\Phi\boldsymbol{\eta}^{(10)})^\top \boldsymbol{\theta}^{(10)}\|^2/n$ as function of p and q for $n = 100$. Clear minima can be observed in each plot and their locations are in accordance with the results of Proposition 2.6. As also can be seen, the dependence of estimation error on the dimension q of the nuisance parameter $\boldsymbol{\eta}$ is much weaker then on p (at least in the case of large p and q). Figure 3.3 graphically shows 500 realisations of the estimate $\widetilde{f}_{p,q}$ for $n = 100$ (left) and $n = 500$ (right) and suitably chosen p, q.

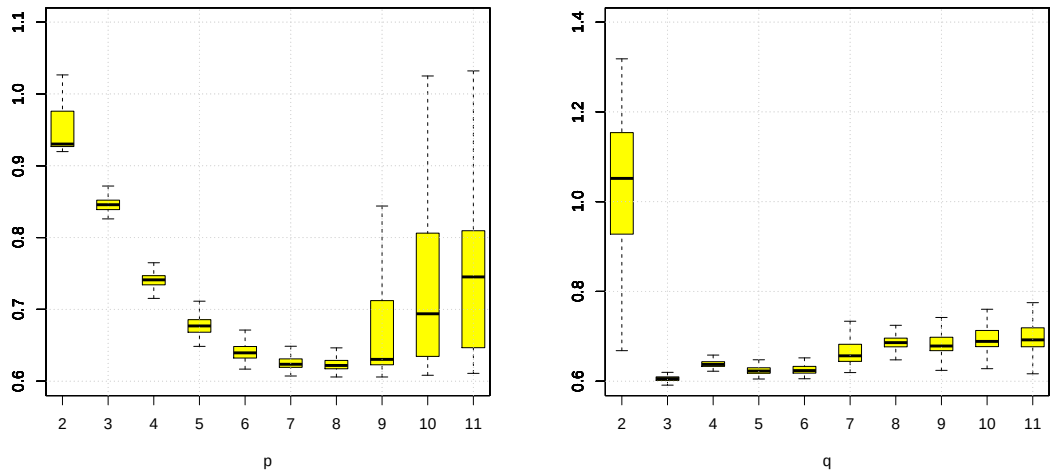


Figure 3.2: Left: the boxplots of the empirical loss $(1/n)\|\boldsymbol{f} - \Psi(\Phi\boldsymbol{\eta}^{(10)})^\top\boldsymbol{\theta}^{(10)}\|^2$ as a function of the dimension p. Right: the boxplots of the loss $(1/n)\|\boldsymbol{f} - \Psi(\Phi\boldsymbol{\eta}^{(10)})^\top\boldsymbol{\theta}^{(10)}\|^2$ as a function of the dimension q.

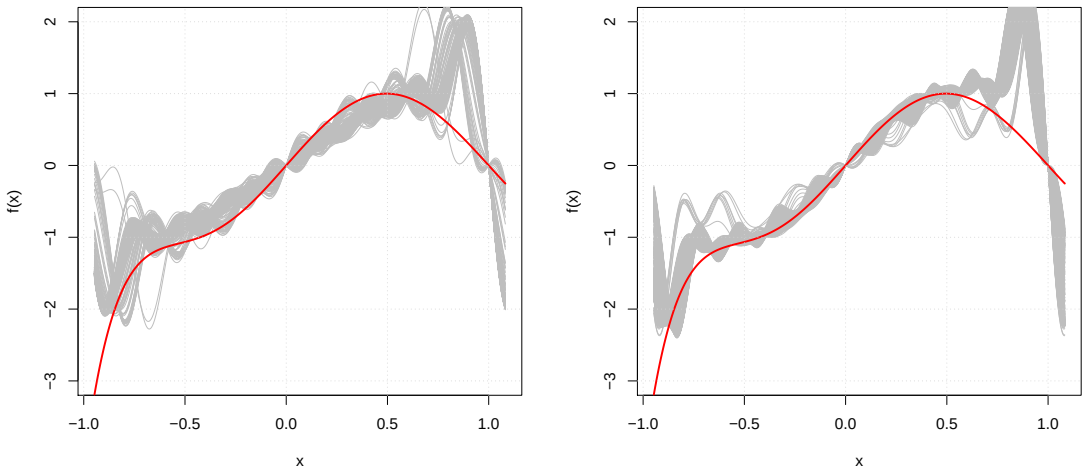


Figure 3.3: Function $f(x)$ (red) together with 500 estimates based on 500 independent samples each of the length $n = 100$ (left) and $n = 500$ (right) with properly chosen p and q.

4 Proofs

This section is organized as follows. First, we apply general finite sample theory imposed in Spokoiny (2012) and Andresen and Spokoiny (2013) to parametric problem (2.1),

which will lead us to the proof of Theorem 2.1. Then, we present proofs for the rest of statements, in the order in which they appear above.

4.1 Main objects

We start with describing main objects that will be helpful through the whole paper. The *stochastic term* $\zeta_{\Phi}(\boldsymbol{\theta},\boldsymbol{\eta})$ is easy to describe:

$$\zeta_{\Phi}(\boldsymbol{\theta},\boldsymbol{\eta}) = L_{\Phi}(\boldsymbol{\theta},\boldsymbol{\eta}) - I\!\!E L_{\Phi}(\boldsymbol{\theta},\boldsymbol{\eta}) = \boldsymbol{\varepsilon}^{\top}\Psi(\Phi\boldsymbol{\eta})^{\top}\boldsymbol{\theta} + \sigma^{-1}\boldsymbol{\nu}^{\top}\Phi\boldsymbol{\eta}.$$

Below we denote by $\Psi^{(s)}(\boldsymbol{X})$ the matrix with the entries $\psi_m^{(s)}(X_i)$ which is obtained by entry-wise differentiation of $\Psi(\boldsymbol{X})$. For a vector $\boldsymbol{a} = (a_i) \in I\!R^n$, the notation $\operatorname{diag}(\boldsymbol{a})$ means a $n \times n$ diagonal matrix with the diagonal entries a_i. Then the *score* $\nabla\zeta_{\Phi}(\boldsymbol{\theta},\boldsymbol{\eta})$ can be written in the block-wise form as

$$\nabla\zeta_{\Phi}(\boldsymbol{\theta},\boldsymbol{\eta}) = \begin{pmatrix} \nabla_{\boldsymbol{\theta}}\zeta_{\Phi}(\boldsymbol{\theta},\boldsymbol{\eta}) \\ \nabla_{\boldsymbol{\eta}}\zeta_{\Phi}(\boldsymbol{\theta},\boldsymbol{\eta}) \end{pmatrix} = \begin{pmatrix} \Psi(\Phi\boldsymbol{\eta})\boldsymbol{\varepsilon} \\ \sigma^{-1}\Phi^{\top}\boldsymbol{\nu} + \Phi^{\top}\operatorname{diag}\bigl\{\Psi'(\Phi\boldsymbol{\eta})^{\top}\boldsymbol{\theta}\bigr\}\boldsymbol{\varepsilon} \end{pmatrix}.$$

Here $\operatorname{diag}\{\Psi'(\boldsymbol{X})^{\top}\boldsymbol{\theta}\}$ means the diagonal matrix with the diagonal entries $\sum_m \theta_m\psi_m'(X_i)$. The *Hessian of the stochastic component* reads as

$$\nabla^2\zeta_{\Phi}(\boldsymbol{\theta},\boldsymbol{\eta}) = \begin{pmatrix} 0 & \Psi'(\Phi\boldsymbol{\eta})\operatorname{diag}(\boldsymbol{\varepsilon})\,\Phi \\ \Phi^{\top}\operatorname{diag}(\boldsymbol{\varepsilon})\,\Psi'(\Phi\boldsymbol{\eta})^{\top} & \Phi^{\top}\operatorname{diag}\bigl\{\Psi''(\Phi\boldsymbol{\eta})^{\top}\boldsymbol{\theta}\bigr\}\operatorname{diag}(\boldsymbol{\varepsilon})\,\Phi \end{pmatrix}$$

The lower right block of this matrix is diagonal with the diagonal entries $\sum_m \theta_m\psi_m''(X_i)\varepsilon_i$ for $i = 1,\ldots,n$ (not taking into account left and right multiplying by Φ).

Further, the expected log-likelihood reads (up to a constant term) as

$$-2I\!\!E L_{\Phi}(\boldsymbol{\theta},\boldsymbol{\eta}) = \|\boldsymbol{f}^* - \Psi(\Phi\boldsymbol{\eta})^{\top}\boldsymbol{\theta}\|^2 + \sigma^{-2}\|\Phi\boldsymbol{\eta} - \boldsymbol{X}^*\|^2.$$

The related matrix $D^2(\boldsymbol{\theta},\boldsymbol{\eta})$ is

$$D^2(\boldsymbol{\upsilon}) = -\nabla^2 I\!\!E L_{\Phi}(\boldsymbol{\theta},\boldsymbol{\eta}) = \begin{pmatrix} D^2_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\upsilon}) & A_{\boldsymbol{\theta}\boldsymbol{\eta}}(\boldsymbol{\upsilon}) \\ A^{\top}_{\boldsymbol{\theta}\boldsymbol{\eta}}(\boldsymbol{\upsilon}) & D^2_{\boldsymbol{\eta}\boldsymbol{\eta}}(\boldsymbol{\upsilon}) \end{pmatrix},$$

where

$$\begin{aligned} D^2_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\upsilon}) &= \Psi(\Phi\boldsymbol{\eta})\Psi(\Phi\boldsymbol{\eta})^{\top}, \\ A_{\boldsymbol{\theta}\boldsymbol{\eta}}(\boldsymbol{\upsilon}) &= \Psi(\Phi\boldsymbol{\eta})\operatorname{diag}\bigl\{\Psi'(\Phi\boldsymbol{\eta})^{\top}\boldsymbol{\theta}\bigr\}\Phi + \Psi'(\Phi\boldsymbol{\eta})\operatorname{diag}\bigl\{\boldsymbol{f}^* - \Psi(\Phi\boldsymbol{\eta})^{\top}\boldsymbol{\theta}\bigr\}\Phi, \\ D^2_{\boldsymbol{\eta}\boldsymbol{\eta}}(\boldsymbol{\upsilon}) &= \sigma^{-2}I_q + \Phi^{\top}\operatorname{diag}\bigl\{\Psi'(\Phi\boldsymbol{\eta})^{\top}\boldsymbol{\theta}\bigr\}^2\Phi + \Phi\operatorname{diag}\bigl\{\Psi''(\Phi\boldsymbol{\eta})^{\top}\boldsymbol{\theta}\bigr\}\operatorname{diag}\bigl\{\boldsymbol{f}^* - \Psi(\Phi\boldsymbol{\eta})^{\top}\boldsymbol{\theta}\bigr\}\Phi, \end{aligned}$$

and define also

$$D^2_{\boldsymbol{\theta\theta}} = \Psi(\Phi\boldsymbol{\eta}^*)\Psi(\Phi\boldsymbol{\eta}^*)^\top,$$
$$A_{\boldsymbol{\theta\eta}} = \Psi(\Phi\boldsymbol{\eta}^*)\,\mathrm{diag}\bigl\{\Psi'(\Phi\boldsymbol{\eta}^*)^\top\boldsymbol{\theta}^*\bigr\}\Phi,$$
$$D^2_{\boldsymbol{\eta\eta}} = \sigma^{-2}\boldsymbol{I}_q + \Phi^\top\,\mathrm{diag}\bigl\{\Psi'(\Phi\boldsymbol{\eta}^*)^\top\boldsymbol{\theta}^*\bigr\}^2\Phi,$$

which are the blocks of the matrix

$$D^2 = \begin{pmatrix} D^2_{\boldsymbol{\theta\theta}} & A_{\boldsymbol{\theta\eta}} \\ A^\top_{\boldsymbol{\theta\eta}} & D^2_{\boldsymbol{\eta\eta}} \end{pmatrix}.$$

The matrix D^2 is defined in a that way to be close to $D^2(\boldsymbol{\upsilon}^*)$, but is more convenient to work with.

Finally, we need to measure distance between $\boldsymbol{\upsilon}$'s. Motivated by the view of D^2 define

$$H_0^2 = \begin{pmatrix} D^2_{\boldsymbol{\theta\theta}} & 0 \\ 0 & \sigma^{-1}\boldsymbol{I}_q \end{pmatrix},$$

and the *local vicinity* for parameter $(\boldsymbol{\theta},\boldsymbol{\eta})$ among projection $(\boldsymbol{\theta}^*, \boldsymbol{\eta}^* = \Phi^\top\boldsymbol{X}^*)$

$$\Upsilon_\circ(\mathtt{r}) = \{\boldsymbol{\upsilon} = (\boldsymbol{\theta},\boldsymbol{\eta}) : \|H_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2 = \|D_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 + \sigma^{-2}\|\boldsymbol{\eta} - \boldsymbol{\eta}^*\|^2 \le \mathtt{r}^2\}.$$

4.2 Auxiliary lemmas

The first lemma describes the variability of the matrix $\Psi(\boldsymbol{X})$.

Lemma 4.1. *For any vector $\boldsymbol{\gamma} \in \mathbb{R}^p$ with $\|\boldsymbol{\gamma}\| \le 1$, it holds with $\mathtt{w}_{p,s}$ from (2.3)*

$$\bigl\|\Psi^{(s)}(\boldsymbol{X})^\top D^{-1}_{\boldsymbol{\theta\theta}}\boldsymbol{\gamma}\bigr\|_\infty \le \frac{\mathtt{w}_{p,s}}{\sqrt{n}}, \qquad s = 0,1,2$$

Proof. By Cauchy-Schwartz inequality

$$|\Psi^{(s)}(X_i)^\top D^{-1}_{\boldsymbol{\theta\theta}}\boldsymbol{\gamma}|^2 \le \|D^{-1}_{\boldsymbol{\theta\theta}}\boldsymbol{\gamma}\|^2 \times \sum_{i=1}^p \psi^{(s)}_m(X_i) \le \mathtt{f}^{-1}\sum_{i=1}^p \mu^2_{m,s}.$$

□

In addition, we need to bound the value $\Psi^{(s)}(\boldsymbol{X})^\top\boldsymbol{\theta}$ on the neighbourhood of $\boldsymbol{\theta}^*$.

Lemma 4.2. *For $s = 0,1,2$ and any $\boldsymbol{\theta} \in \mathbb{R}^p$ satisfying $\|D_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le \mathtt{r}$ and each $X \in \mathbb{R}$ it holds*

$$\bigl|\Psi^{(s)}(X)^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\bigr| \le \frac{\mathtt{w}_{p,s}\,\mathtt{r}}{\sqrt{n}},$$
$$\bigl|\Psi^{(s)}(X)^\top\boldsymbol{\theta}\bigr| \le \mathtt{C}_{f,s} + \frac{\mathtt{w}_{p,s}\,\mathtt{r}}{\sqrt{n}}.$$

Proof. Using previous lemma with $\boldsymbol{\gamma} = \mathtt{r}^{-1} D_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$,

$$\bigl|\Psi^{(s)}(X)^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\bigr| = \bigl|\Psi^{(s)}(X)^\top D_{\boldsymbol{\theta\theta}}^{-1} \times D_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\bigr| \le \frac{\mathtt{r}\mathtt{w}_{p,s}}{\sqrt{n}},$$

and the second follows by $\bigl|\Psi^{(s)}(X)^\top\boldsymbol{\theta}\bigr| \le \bigl|\Psi^{(s)}(X)^\top\boldsymbol{\theta}^*\bigr| + \bigl|\Psi^{(s)}(X)^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\bigr|$. □

In particular, if $\mathtt{r} \le \sqrt{n}/\mathtt{w}_{p,s}$, then $|\Psi^{(s)}(X)^\top\boldsymbol{\theta}| \le \mathtt{C}_{f,s} + 1$.

Lemma 4.3. *If* $\sigma^{-1}\|\boldsymbol{X} - \boldsymbol{X}^*\| \le \mathtt{r}$, *then it holds for* $s = 0, 1, 2$

$$\bigl\|\bigl\{\Psi^{(s)}(\boldsymbol{X}) - \Psi^{(s)}(\boldsymbol{X}^*)\bigr\}^\top D_{\boldsymbol{\theta\theta}}^{-1}\boldsymbol{\gamma}\bigr\| \le \frac{\sigma\,\mathtt{w}_{p,s+1}\,\mathtt{r}}{\sqrt{n}},$$

$$\bigl\|\bigl\{\Psi^{(s)}(\boldsymbol{X}) - \Psi^{(s)}(\boldsymbol{X}^*)\bigr\}^\top\boldsymbol{\theta}^*\bigr\| \le \sigma\,\mathtt{C}_{f,s+1}\,\mathtt{r}.$$

Proof. We have with some $\boldsymbol{X}'$ on the line connecting $\boldsymbol{X}$, $\boldsymbol{X}^*$

$$\{\Psi^{(s)}(\boldsymbol{X}) - \Psi^{(s)}(\boldsymbol{X}^*)\}^\top D_{\boldsymbol{\theta\theta}}^{-1}\boldsymbol{\gamma} = (\boldsymbol{X} - \boldsymbol{X}^*)^\top \operatorname{diag}\bigl\{\Psi^{(s+1)}(\boldsymbol{X}')^\top D_{\boldsymbol{\theta\theta}}^{-1}\boldsymbol{\gamma}\bigr\},$$

and the first inequality by Lemma 4.1.

By analogy, the second inequality follows from $\bigl|\Psi^{(s+1)}(\boldsymbol{X}')^\top\boldsymbol{\theta}^*\bigr| \le \mathtt{C}_{f,s+1}$ and

$$\{\Psi^{(s)}(\boldsymbol{X}) - \Psi^{(s)}(\boldsymbol{X}^*)\}^\top\boldsymbol{\theta}^* = (\boldsymbol{X} - \boldsymbol{X}^*)^\top \operatorname{diag}\bigl\{\Psi(\boldsymbol{X}')^\top\boldsymbol{\theta}^*\bigr\}.$$

□

The next lemma checks identifiability condition $(\mathcal{I})$ of Andresen and Spokoiny (2013).

Lemma 4.4. *Define*

$$\rho = \sqrt{\frac{\mathtt{C}_{f,1}^2}{\sigma^{-2} + \mathtt{C}_{f,1}^2}}.$$

Then,

$$\bigl\|D_{\boldsymbol{\theta\theta}}^{-1} A_{\boldsymbol{\theta\eta}} D_{\boldsymbol{\eta\eta}}^{-1}\bigr\| \le \rho, \qquad D^2 \ge (1-\rho)H_0^2, \qquad D^2 \le (1-\rho)^{-1}H_0^2.$$

Proof. Define, $\widetilde{D}_{\boldsymbol{\eta\eta}}^2 = D_{\boldsymbol{\eta\eta}}^2 - \sigma^{-2}\boldsymbol{I}_q$. Then,

$$\widetilde{D}^2 = \begin{pmatrix} D_{\boldsymbol{\theta\theta}}^2 & A_{\boldsymbol{\theta\eta}} \\ A_{\boldsymbol{\theta\eta}}^\top & \widetilde{D}_{\boldsymbol{\eta\eta}}^2 \end{pmatrix} = \begin{pmatrix} \Psi(\Phi\boldsymbol{\eta}^*)\Psi(\Phi\boldsymbol{\eta}^*)^\top & \Psi(\Phi\boldsymbol{\eta}^*)\operatorname{diag}\bigl\{\Psi'(\Phi\boldsymbol{\eta}^*)^\top\boldsymbol{\theta}^*\bigr\}\Phi \\ \Phi^\top \operatorname{diag}\bigl\{\Psi'(\Phi\boldsymbol{\eta}^*)^\top\boldsymbol{\theta}^*\bigr\}\Psi(\Phi\boldsymbol{\eta}^*) & \Phi^\top \operatorname{diag}\bigl\{\Psi'(\Phi\boldsymbol{\eta}^*)^\top\boldsymbol{\theta}^*\bigr\}^2\Phi \end{pmatrix}$$

$$= \begin{pmatrix} \Psi(\Phi\boldsymbol{\eta}^*) \\ \Phi^\top \operatorname{diag}\bigl\{\Psi'(\Phi\boldsymbol{\eta}^*)^\top\boldsymbol{\theta}^*\bigr\} \end{pmatrix} \begin{pmatrix} \Psi(\Phi\boldsymbol{\eta}^*) \\ \Phi^\top \operatorname{diag}\bigl\{\Psi'(\Phi\boldsymbol{\eta}^*)^\top\boldsymbol{\theta}^*\bigr\} \end{pmatrix}^\top,$$

meaning that $\widetilde{D}^2$ is nonnegative and $\|D_{\boldsymbol{\theta\theta}}^{-1} A_{\boldsymbol{\theta\eta}} \widetilde{D}_{\boldsymbol{\eta\eta}}^{-1}\| \le 1$. Thus,

$$\|D_{\boldsymbol{\theta\theta}}^{-1} A_{\boldsymbol{\theta\eta}} D_{\boldsymbol{\eta\eta}}^{-1}\| \le \|\widetilde{D}_{\boldsymbol{\eta\eta}} D_{\boldsymbol{\eta\eta}}^{-1}\| \le \|\operatorname{diag}\{\Psi'(\Phi\boldsymbol{\eta}^*)^\top \boldsymbol{\theta}^*\}^2 (\sigma^{-2} + \operatorname{diag}\{\Psi'(\Phi\boldsymbol{\eta}^*)^\top \boldsymbol{\theta}^*\}^2)^{-1}\|^{1/2},$$

which is less than ρ since the function $x/(\sigma^{-2}+x)$ is non decreasing and $|\Psi'(X)^\top \boldsymbol{\theta}^*| \le \mathtt{C}_{f,1}$ by the condition (i). So, by this inequality we have

$$D^2 \ge (1-\rho) \begin{pmatrix} D_{\boldsymbol{\theta\theta}}^2 & 0 \\ 0 & D_{\boldsymbol{\eta\eta}}^2 \end{pmatrix} \ge (1-\rho) H_0^2.$$

Further, using $D_{\boldsymbol{\eta\eta}}^2 \le (1+\sigma^2 \mathtt{C}_{f,1}^2) H_\sigma^2 = (1-\rho^2)^{-1} H_\sigma^2$ we get

$$D^2 \le (1+\rho) \begin{pmatrix} D_{\boldsymbol{\theta\theta}}^2 & 0 \\ 0 & D_{\boldsymbol{\eta\eta}}^2 \end{pmatrix} \le (1+\rho)(1-\rho^2)^{-1} H_0^2.$$

□

Lemma 4.5. *The vector* $\boldsymbol{b} = D^{-1} \nabla L_\Phi(\boldsymbol{\upsilon}^*)$ *satisfies* $\|\boldsymbol{b}\| \le \|\boldsymbol{f}^* - \Psi(\Phi\boldsymbol{\eta}^*)^\top \boldsymbol{\theta}^*\|$.

Proof. The explicit expression for the vector is as follows

$$\begin{aligned} \nabla I\!E L_\Phi(\boldsymbol{\upsilon}^*) &= \begin{pmatrix} \Psi(\Phi\boldsymbol{\eta}^*)\big(\boldsymbol{f}^* - \Psi(\Phi\boldsymbol{\eta}^*)^\top \boldsymbol{\theta}^*\big) \\ \Phi^\top \operatorname{diag}\big\{\Psi'(\Phi\boldsymbol{\eta}^*)^\top \boldsymbol{\theta}^*\big\}\big(\boldsymbol{f}^* - \Psi(\Phi\boldsymbol{\eta}^*)^\top \boldsymbol{\theta}^*\big) \end{pmatrix} \\ &= \begin{pmatrix} \Psi(\Phi\boldsymbol{\eta}^*) \\ \Phi^\top \operatorname{diag}\big\{\Psi'(\Phi\boldsymbol{\eta}^*)^\top \boldsymbol{\theta}^*\big\} \end{pmatrix} \big(\boldsymbol{f}^* - \Psi(\Phi\boldsymbol{\eta}^*)^\top \boldsymbol{\theta}^*\big). \end{aligned}$$

Define $A = \Psi(\Phi\boldsymbol{\eta}^*)$, $B = \Phi^\top \operatorname{diag}\big\{\Psi'(\Phi\boldsymbol{\eta}^*)^\top \boldsymbol{\theta}^*\big\}$. Then,

$$D^2 = \begin{pmatrix} AA^\top & AB^\top \\ AB^\top & \sigma^{-2} I_p + BB^\top \end{pmatrix} \ge \begin{pmatrix} A \\ B \end{pmatrix} \begin{pmatrix} A \\ B \end{pmatrix}^\top.$$

Hence,

$$\|D^{-1} \nabla I\!E L_\Phi(\boldsymbol{\upsilon}^*)\| \le \left\| D^{-1} \begin{pmatrix} A \\ B \end{pmatrix} \right\| \times \|\boldsymbol{f}^* - \Psi(\Phi\boldsymbol{\eta}^*)^\top \boldsymbol{\theta}^*\| \le \|\boldsymbol{f}^* - \Psi(\Phi\boldsymbol{\eta}^*)^\top \boldsymbol{\theta}^*\|.$$

□

4.3 Local perturbation of $D^2(\boldsymbol{\upsilon})$

What follows relates to condition $(\mathcal{L}_0)$ of Andresen and Spokoiny (2013). Our purpose is to bound $\|H_0^{-1}(D^2(\boldsymbol{\upsilon}) - D^2) H_0^{-1}\|$ for $\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})$. The result will depend on the smoothness of design basis Φ. It holds,

$$V(\Phi) = \sup_{\boldsymbol{\gamma} \in I\!R^q,\, \|\boldsymbol{\gamma}\|=1} \|\Phi\boldsymbol{\gamma}\|_\infty, \qquad \mathtt{v}_q \ge n^{1/2} V(\Phi),$$

and we only use $\mathtt{v}_q$ through inequality $\|\Phi\boldsymbol{\gamma}\|_\infty \le \frac{\mathtt{v}_q}{\sqrt{n}}\|\boldsymbol{\gamma}\|$ for each $\boldsymbol{\gamma} \in I\!R^q$.

Lemma 4.6. *Let* $\mathtt{r}$ *satisfies* $\mathtt{r} \le \sqrt{n}/\mathtt{w}_{p,1}$. *Then* $\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})$ *yields*

$$\begin{aligned}\|H_0^{-1}(D^2(\boldsymbol{\upsilon}) - D^2)H_0\| \;&\le\; \delta(\mathtt{r}) \\ &\stackrel{\text{def}}{=} \mathtt{C}^{**}\sigma\frac{\mathtt{w}_{p,1} + \mathtt{v}_q + \frac{\mathtt{w}_{p,2}\mathtt{v}_q\mathtt{r}}{\sqrt{n}}}{\sqrt{n}}\{\mathtt{r} + \square\},\end{aligned}$$

where

$$\begin{aligned}\square &= \|\boldsymbol{f}^* - \Psi(\Phi\boldsymbol{\eta}^*)^\top\boldsymbol{\theta}^*\|, \\ \mathtt{C}^* &= (\sigma \vee 1) \times \bigl(\mathtt{C}_{f,1} \vee \mathtt{C}_{f,2} + 1\bigr), \\ \mathtt{C}^{**} &= 8\mathtt{C}^* \vee 3(\mathtt{C}^*)^2.\end{aligned}$$

Proof. For $\boldsymbol{\alpha} = (\boldsymbol{\beta}, \boldsymbol{\gamma}) \in I\!R^{p+q}$ with $\|\boldsymbol{\alpha}\| \le 1$ we have

$$\begin{aligned}\bigl|\boldsymbol{\alpha}^\top H_0^{-1}(D^2(\boldsymbol{\upsilon}) - D^2)H_0^{-1}\boldsymbol{\alpha}\bigr| \le\; &\bigl|\boldsymbol{\beta}^\top D_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(D^2_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\upsilon}) - D^2_{\boldsymbol{\theta}\boldsymbol{\theta}})D_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\boldsymbol{\beta}\bigr| \\ &+2\sigma\bigl|\boldsymbol{\beta}^\top D_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(A_{\boldsymbol{\theta}\boldsymbol{\eta}}(\boldsymbol{\upsilon}) - A_{\boldsymbol{\theta}\boldsymbol{\eta}})\boldsymbol{\gamma}\bigr| + \sigma^2\bigl|\boldsymbol{\gamma}^\top(D^2_{\boldsymbol{\eta}\boldsymbol{\eta}}(\boldsymbol{\upsilon}) - D^2_{\boldsymbol{\eta}\boldsymbol{\eta}})\boldsymbol{\gamma}\bigr|.\end{aligned}$$

We will proceed with each term separately. First,

$$\begin{aligned}&\boldsymbol{\beta}^\top D_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\bigl\{D^2_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\upsilon}) - D^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\bigr\}D_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\boldsymbol{\beta} \\ &\quad= \boldsymbol{\beta}^\top D_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\bigl\{\Psi(\Phi\boldsymbol{\eta})\,\Psi(\Phi\boldsymbol{\eta})^\top - \Psi(\Phi\boldsymbol{\eta}^*)\,\Psi(\Phi\boldsymbol{\eta}^*)^\top\bigr\}D_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\boldsymbol{\beta} \\ &\quad= 2\boldsymbol{\gamma}^\top D_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\bigl\{\Psi(\Phi\boldsymbol{\eta}) - \Psi(\Phi\boldsymbol{\eta}^*)\bigr\}\,\Psi(\Phi\boldsymbol{\eta}^*)^\top D_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\boldsymbol{\gamma} + \bigl\|\bigl\{\Psi(\Phi\boldsymbol{\eta}) - \Psi(\Phi\boldsymbol{\eta}^*)\bigr\}^\top D_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\boldsymbol{\gamma}\bigr\|^2 \\ &\quad\le \frac{2\sigma\mathtt{w}_{p,1}\mathtt{r}}{\sqrt{n}} + \left(\frac{\sigma\mathtt{w}_{p,1}\mathtt{r}}{\sqrt{n}}\right)^2 \le \frac{3\sigma\mathtt{w}_{p,1}\mathtt{r}}{\sqrt{n}}.\end{aligned}$$

Second, we use decomposition

$$\begin{aligned}A_{\boldsymbol{\theta}\boldsymbol{\eta}}(\boldsymbol{\upsilon}) - A_{\boldsymbol{\theta}\boldsymbol{\eta}} = \;&\Psi(\Phi\boldsymbol{\eta}^*)\,\mathrm{diag}\{(\Psi'(\Phi\boldsymbol{\eta}) - \Psi'(\Phi\boldsymbol{\eta}^*))^\top\boldsymbol{\theta}^*\}\Phi \\ &+\Psi(\Phi\boldsymbol{\eta}^*)\,\mathrm{diag}\{\Psi'(\Phi\boldsymbol{\eta}^*)^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\}\Phi \\ &+(\Psi(\Phi\boldsymbol{\eta}) - \Psi(\Phi\boldsymbol{\eta}^*))\,\mathrm{diag}\{\Psi'(\Phi\boldsymbol{\eta})^\top\boldsymbol{\theta}\}\Phi \\ &+\Psi'(\Phi\boldsymbol{\eta})\,\mathrm{diag}\{\boldsymbol{f}^* - \Psi(\Phi\boldsymbol{\eta})^\top\boldsymbol{\theta}\}\Phi.\end{aligned}$$

Define the following vectors,

$$\begin{aligned}
&\boldsymbol{a} = (\Psi'(\Phi\boldsymbol{\eta}) - \Psi'(\Phi\boldsymbol{\eta}^*))^\top\boldsymbol{\theta}^*, && \|\boldsymbol{a}\| \leq \mathtt{C}_{f,2}\mathtt{r},\\
&\boldsymbol{b} = \Psi(\Phi\boldsymbol{\eta}^*)^\top D_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\boldsymbol{\beta}, && \|\boldsymbol{b}\|_\infty \leq \mathtt{w}_{p,0}/\sqrt{n}, \qquad \|\boldsymbol{b}\| = 1,\\
&\boldsymbol{c} = \Psi'(\Phi\boldsymbol{\eta}^*)^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*), && \|\boldsymbol{c}\|_\infty \leq \mathtt{w}_{p,1}\mathtt{r}/\sqrt{n},\\
&\boldsymbol{d} = (\Psi(\Phi\boldsymbol{\eta}) - \Psi(\Phi\boldsymbol{\eta}^*))^\top D_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\boldsymbol{\beta}, && \|\boldsymbol{d}\| \leq \sigma\mathtt{w}_{p,1}\mathtt{r}/\sqrt{n},\\
&\boldsymbol{e} = \Psi'(\Phi\boldsymbol{\eta})^\top\boldsymbol{\theta}, && \|\boldsymbol{e}\|_\infty \leq \mathtt{C}_{f,1},\\
&\boldsymbol{g} = \Phi\boldsymbol{\gamma} && \|\boldsymbol{g}\| \leq 1,\\
&\boldsymbol{h} = \Psi'(\Phi\boldsymbol{\eta})^\top D_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\boldsymbol{\beta}, && \|\boldsymbol{h}\|_\infty \leq \mathtt{w}_{p,1}/\sqrt{n},\\
&\boldsymbol{k} = \boldsymbol{f}^* - \Psi(\Phi\boldsymbol{\eta})^\top\boldsymbol{\theta}, && \|\boldsymbol{k}\| \leq \mathtt{C}^*\mathtt{r} + \square,
\end{aligned}$$

where the inequalities follow from Lemmas 4.1, 4.2 and 4.3. Then,

$$\begin{aligned}
\left|\boldsymbol{\beta}^\top D_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(A_{\boldsymbol{\theta}\boldsymbol{\eta}}(\boldsymbol{\upsilon}) - A_{\boldsymbol{\theta}\boldsymbol{\eta}})\boldsymbol{\gamma}\right| &= \left|\boldsymbol{b}^\top \operatorname{diag}(\boldsymbol{a})\boldsymbol{g} + \boldsymbol{b}^\top \operatorname{diag}(\boldsymbol{c})\boldsymbol{g} + \boldsymbol{d}^\top \operatorname{diag}(\boldsymbol{e})\boldsymbol{g} + \boldsymbol{h}^\top \operatorname{diag}(\boldsymbol{k})\boldsymbol{g}\right|\\
&= \left|\sum_{i=1}^n b_i a_i g_i + b_i c_i g_i + d_i e_i g_i + h_i k_i g_i\right|\\
&\leq \|\boldsymbol{b}\|_\infty\|\boldsymbol{a}\|\|\boldsymbol{g}\| + \|\boldsymbol{b}\|\|\boldsymbol{c}\|_\infty\|\boldsymbol{g}\| + \|\boldsymbol{d}\|\|\boldsymbol{e}\|_\infty\|\boldsymbol{g}\| + \|\boldsymbol{h}\|_\infty\|\boldsymbol{k}\|\|\boldsymbol{g}\|\\
&\leq \left\{\mathtt{C}^*\mathtt{w}_{p,0}\mathtt{r} + \mathtt{w}_{p,1}\mathtt{r} + \mathtt{C}^*\mathtt{w}_{p,1}\mathtt{r} + \mathtt{w}_{p,1}(\mathtt{C}^*\mathtt{r} + \square)\right\}/\sqrt{n}\\
&\leq \frac{3\mathtt{C}^*\mathtt{w}_{p,1}}{\sqrt{n}}\{\mathtt{r} + \square\},
\end{aligned}$$

taking into account that $\mathtt{w}_{p,0} \leq \mathtt{w}_{p,1}$.

Third, we write,

$$\begin{aligned}
D_{\boldsymbol{\eta}\boldsymbol{\eta}}^2(\boldsymbol{\upsilon}) - D_{\boldsymbol{\eta}\boldsymbol{\eta}}^2 = \Phi^\top \operatorname{diag}\left\{\left(\Psi'(\Phi\boldsymbol{\eta})^\top\boldsymbol{\theta}\right)^2 - \left(\Psi'(\Phi\boldsymbol{\eta}^*)^\top\boldsymbol{\theta}^*\right)^2\right\}\Phi\\
+\Phi^\top \operatorname{diag}\{\Psi''(\Phi\boldsymbol{\eta})^\top\boldsymbol{\theta}\} \operatorname{diag}\{\boldsymbol{f}^* - \Psi(\Phi\boldsymbol{\eta})^\top\boldsymbol{\theta}\}\Phi.
\end{aligned}$$

Define,

$$\begin{aligned}
&\boldsymbol{a} = \Psi'(\Phi\boldsymbol{\eta})^\top\boldsymbol{\theta} + \Psi'(\Phi\boldsymbol{\eta}^*)^\top\boldsymbol{\theta}^*, && \|\boldsymbol{a}\|_\infty \leq 2\mathtt{C}^*,\\
&\boldsymbol{b} = \left(\Psi'(\Phi\boldsymbol{\eta}) - \Psi'(\Phi\boldsymbol{\eta}^*)\right)^\top\boldsymbol{\theta}^*, && \|\boldsymbol{b}\| \leq \mathtt{C}^*\mathtt{r},\\
&\boldsymbol{c} = \Psi'(\Phi\boldsymbol{\eta})^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*), && \|\boldsymbol{c}\|_\infty \leq \mathtt{w}_{p,1}\mathtt{r}/\sqrt{n},\\
&\boldsymbol{d} = \Psi''(\Phi\boldsymbol{\eta})^\top\boldsymbol{\theta}, && \|\boldsymbol{d}\|_\infty \leq \mathtt{C}^* + \mathtt{w}_{p,2}\mathtt{r}/\sqrt{n},\\
&\boldsymbol{e} = \boldsymbol{f}^* - \Psi(\Phi\boldsymbol{\eta})^\top\boldsymbol{\theta}, && \|\boldsymbol{e}\| \leq \mathtt{C}^*\mathtt{r} + \square,\\
&\boldsymbol{g} = \Phi\boldsymbol{\gamma}, && \|\boldsymbol{g}\|_\infty \leq \mathtt{v}_q/\sqrt{n}, \qquad \|\boldsymbol{g}\| = 1.
\end{aligned}$$

Then,

$$\begin{aligned}
\left|\boldsymbol{\gamma}^\top (D^2_{\boldsymbol{\eta}\boldsymbol{\eta}}(\boldsymbol{\upsilon}) - D^2_{\boldsymbol{\eta}\boldsymbol{\eta}})\boldsymbol{\gamma}\right| &= \left|\boldsymbol{g}^\top \operatorname{diag}\{\boldsymbol{a}\} \operatorname{diag}\{\boldsymbol{b}+\boldsymbol{c}\}\boldsymbol{g} + \boldsymbol{g}^\top \operatorname{diag}\{\boldsymbol{d}\} \operatorname{diag}\{\boldsymbol{e}\}\boldsymbol{g}\right| \\
&= \left|\sum_{i=1}^n (g_i^2 a_i b_i + g_i^2 a_i c_i + g_i^2 d_i e_i)\right| \\
&\le \|\boldsymbol{a}\|_\infty \|\boldsymbol{g}\|_\infty \|\boldsymbol{g}\| \|\boldsymbol{b}\| + \|\boldsymbol{a}\|_\infty \|\boldsymbol{c}\|_\infty \|\boldsymbol{g}\|^2 + \|\boldsymbol{d}\|_\infty \|\boldsymbol{g}\|_\infty \|\boldsymbol{g}\| \|\boldsymbol{e}\| \\
&\le \left\{4(\mathtt{C}^*)^2 \mathtt{v}_q + 2\mathtt{C}^* \mathtt{w}_{p,1} + \mathtt{C}^* \mathtt{w}_{p,2} \mathtt{v}_q / \sqrt{n}\right\} \frac{\mathtt{r} + \square}{n^{1/2}}
\end{aligned}$$

Putting all together gives the required bound. □

4.4 Check of exponential moments of derivatives

The following two lemmas check $(\mathcal{ED}_1)$ and $(\mathcal{ED}_2)$ of Andresen and Spokoiny (2013), respectively.

Lemma 4.7. *For $\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})$ it holds for each $\boldsymbol{\gamma} \in I\!R^{p+q}$*

$$\log I\!E \exp\left(\boldsymbol{\gamma}^\top H_0^{-1} \nabla \zeta_\Phi(\boldsymbol{\upsilon})\right) \le \frac{\nu_1(\mathtt{r})^2 \|\boldsymbol{\gamma}\|^2}{2}$$

with $\nu_1(\mathtt{r}) = \sqrt{2}\nu_\varepsilon \left[1 + \sigma \mathtt{C}_{f,1} + \sigma \mathtt{w}_{p,1} \mathtt{r}/\sqrt{n}\right]$.

Proof. We have for $\boldsymbol{\gamma} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$, $\boldsymbol{\alpha} \in I\!R^p$ and $\boldsymbol{\beta} \in I\!R^q$

$$\begin{aligned}
\log I\!E \exp\left(\boldsymbol{\gamma}^\top H_0^{-1} \nabla \zeta_\Phi(\boldsymbol{\upsilon})\right) &= \log I\!E \exp\Big(\boldsymbol{\alpha}^\top D^{-1}_{\boldsymbol{\theta}\boldsymbol{\theta}} \Psi(\Phi\boldsymbol{\eta})\boldsymbol{\varepsilon} \\
&\qquad + \boldsymbol{\beta}^\top \Phi^\top \boldsymbol{\nu} + \sigma \boldsymbol{\beta}^\top \Phi^\top \operatorname{diag}\big\{\Psi'(\Phi\boldsymbol{\eta})^\top \boldsymbol{\theta}\big\}\boldsymbol{\varepsilon}\Big) \\
&\le \frac{\nu_\varepsilon^2 \|\boldsymbol{u}\|^2}{2} + \frac{\nu_\varepsilon^2 \|\boldsymbol{\beta}\|^2}{2},
\end{aligned}$$

with $\boldsymbol{u} = \boldsymbol{\alpha}^\top D^{-1}_{\boldsymbol{\theta}\boldsymbol{\theta}} \Psi(\Phi\boldsymbol{\eta}) + \sigma \boldsymbol{\beta}^\top \Phi^\top \operatorname{diag}\big\{\Psi'(\Phi\boldsymbol{\eta})^\top \boldsymbol{\theta}\big\}$. By Lemma 4.3 we get

$$\begin{aligned}
\|\boldsymbol{\alpha}^\top D^{-1}_{\boldsymbol{\theta}\boldsymbol{\theta}} \Psi(\Phi\boldsymbol{\eta})\| &\le \|\boldsymbol{\alpha}^\top D^{-1}_{\boldsymbol{\theta}\boldsymbol{\theta}} \Psi(\Phi\boldsymbol{\eta}^*)\| + \|\boldsymbol{\alpha}^\top D^{-1}_{\boldsymbol{\theta}\boldsymbol{\theta}} \{\Psi(\Phi\boldsymbol{\eta}) - \Psi(\Phi\boldsymbol{\eta}^*)\}\| \\
&\le (1 + \sigma \mathtt{w}_{p,1} \mathtt{r}/\sqrt{n}) \|\boldsymbol{\alpha}\|
\end{aligned}$$

Then, we use Lemma 4.2 to get $\|\boldsymbol{\beta}^\top \Phi^\top \operatorname{diag}\big(\Psi'(\Phi\boldsymbol{\eta})^\top \boldsymbol{\theta}\big)\| \le (\mathtt{C}_{f,1} + \mathtt{w}_{p,1}\mathtt{r}/\sqrt{n}) \|\boldsymbol{\beta}\|$. Summing up we get the required statement. □

Lemma 4.8. *Suppose $\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})$. Then it holds for each $\boldsymbol{\gamma}_{1,2} \in I\!R^{p+q}$*

$$\log I\!E \exp\left(\boldsymbol{\gamma}_1^\top H_0^{-1} \nabla^2 \zeta_\Phi(\boldsymbol{\upsilon}) H_0^{-1} \boldsymbol{\gamma}_2\right) \le \frac{\nu_2(\mathtt{r})^2 \|\boldsymbol{\gamma}_1\|^2 \|\boldsymbol{\gamma}_2\|^2}{2}$$

with $\nu_2(\mathtt{r})^2 = \sqrt{3}\nu_\varepsilon^2 \sigma \left[\mathtt{w}_{p,1} + \sigma(\mathtt{C}_{f,2} + \mathtt{w}_{p,2}\mathtt{r}/\sqrt{n})\mathtt{v}_q\right]/\sqrt{n}$.

Proof. Let $\boldsymbol{\gamma}_i = (\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i)$, $\boldsymbol{\alpha}_i \in I\!R^p$ and $\boldsymbol{\beta}_i \in I\!R^q$, $i = 1, 2$. Define then $\boldsymbol{b}_i = \Phi\boldsymbol{\beta}_i$. By orthonormality of columns of Φ we have $\|\boldsymbol{b}_i\| = \|\boldsymbol{\beta}_i\|$ and by condition (v) we have $\|\boldsymbol{b}_i\|_\infty \le \frac{\mathtt{v}_q}{\sqrt{n}}\|\boldsymbol{\beta}_i\|$.

So, by generalized Hölder's inequality we have

$$\begin{aligned}
&\log I\!E \exp\Big(\boldsymbol{\gamma}_1^\top H_0^{-1}\nabla^2\zeta_\Phi(\boldsymbol{\upsilon})H_0^{-1}\boldsymbol{\gamma}_2\Big) \\
&\quad\le \log I\!E \exp\Big(\sigma\boldsymbol{\alpha}_1^\top D_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\Psi'(\Phi\boldsymbol{\eta})\operatorname{diag}(\boldsymbol{\varepsilon})\,\boldsymbol{b}_2 + \sigma\boldsymbol{\alpha}_2^\top D_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\Psi'(\Phi\boldsymbol{\eta})\operatorname{diag}(\boldsymbol{\varepsilon})\,\boldsymbol{b}_1 \\
&\qquad\qquad +\sigma^2\boldsymbol{b}_1^\top \operatorname{diag}\big\{\Psi''(\Phi\boldsymbol{\eta})^\top\boldsymbol{\theta}\big\}\operatorname{diag}(\boldsymbol{\varepsilon})\,\boldsymbol{b}_2\Big) \\
&\quad\le \frac{1}{3}\log I\!E \exp\Big(3\,\sigma\boldsymbol{\alpha}_1^\top D_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\Psi'(\Phi\boldsymbol{\eta})\operatorname{diag}(\boldsymbol{\varepsilon})\,\boldsymbol{b}_2\Big) \\
&\qquad +\frac{1}{3}\log I\!E \exp\Big(3\,\sigma\boldsymbol{\alpha}_2^\top D_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\Psi'(\Phi\boldsymbol{\eta})\operatorname{diag}(\boldsymbol{\varepsilon})\,\boldsymbol{b}_1\Big) \\
&\qquad +\frac{1}{3}\log I\!E \exp\Big(3\,\sigma^2\boldsymbol{b}_1^\top \operatorname{diag}\big\{\Psi''(\Phi\boldsymbol{\eta})^\top\boldsymbol{\theta}\big\}\operatorname{diag}(\boldsymbol{\varepsilon})\,\boldsymbol{b}_2\Big).
\end{aligned}$$

Let's deal with each term separately. For $\boldsymbol{\alpha} \in I\!R^p, \boldsymbol{b} = \Phi\boldsymbol{\beta} \in I\!R^n$ we have

$$\boldsymbol{\alpha}^\top D_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\Psi'(\boldsymbol{X})\operatorname{diag}(\boldsymbol{\varepsilon})\boldsymbol{b} = \boldsymbol{u}^\top \operatorname{diag}(\boldsymbol{\varepsilon})\boldsymbol{b} = \sum_{i=1}^n u_i b_i \varepsilon_i,$$

where $\boldsymbol{u}^\top = \boldsymbol{\alpha}^\top D_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\Psi'(\Phi\boldsymbol{\eta})$ by Lemma 4.1 satisfy $\|\boldsymbol{u}\|_\infty \le \mathtt{w}_{p,1}\|\boldsymbol{\alpha}\|/\sqrt{n}$. Thus,

$$\begin{aligned}
\log I\!E \exp\Big(3\,\sigma\boldsymbol{\alpha}^\top D_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\Psi'(\Phi\boldsymbol{\eta})\operatorname{diag}(\boldsymbol{\varepsilon})\boldsymbol{b}\Big) &= \sum_{i=1}^n \log I\!E \exp(3\,\sigma u_i b_i \varepsilon_i) \\
&\le \frac{9\nu_\varepsilon^2\sigma^2\|\boldsymbol{u}\|_\infty^2\|\boldsymbol{b}\|^2}{2} \\
&\le \frac{9\nu_\varepsilon^2\sigma^2\mathtt{w}_{p,1}^2\|\boldsymbol{\alpha}\|^2\|\boldsymbol{\beta}\|^2/n}{2}.
\end{aligned}$$

Further,

$$\mathbf{b}_1^\top \operatorname{diag}\big(\Psi''(\Phi\boldsymbol{\eta})^\top\boldsymbol{\theta}\big)\operatorname{diag}(\boldsymbol{\varepsilon})\mathbf{b}_2 = \sum_{i=1}^n b_{1,i}k_i\varepsilon_i$$

with $\boldsymbol{k} = \operatorname{diag}\big\{\Psi''(\Phi\boldsymbol{\eta})^\top\boldsymbol{\theta}\big\}\mathbf{b}_2$, which by Lemma 4.2 and condition (v) satisfy $\|\boldsymbol{k}\|_\infty \le (\mathtt{C}_{f,2} + \mathtt{w}_{p,2}\mathtt{r}/\sqrt{n})\frac{\mathtt{v}_q}{\sqrt{n}}\|\boldsymbol{\beta}_2\|$. This brings us to

$$\log I\!E \exp\Big(3\sigma^2\mathbf{b}^\top \operatorname{diag}\big(\Psi''(\Phi\boldsymbol{\eta})^\top\boldsymbol{\theta}\big)\operatorname{diag}(\boldsymbol{\varepsilon})\mathbf{b}\Big) \le \frac{9\sigma^4(\mathtt{C}_{f,2} + \mathtt{w}_{p,2}\mathtt{r}/\sqrt{n})^2\frac{\mathtt{C}_\Phi^2}{n}\|\boldsymbol{\beta}_1\|^2\|\boldsymbol{\beta}_2\|^2}{2}.$$

Bringing those inequalities together we get

$$\begin{aligned}\log I\!E \exp\left(\boldsymbol{\gamma}_1^\top H_0^{-1}\nabla^2\zeta_{\Phi}(\boldsymbol{\upsilon})H_0^{-1}\boldsymbol{\gamma}_2\right) &\le \frac{3\nu_\varepsilon^2\sigma^2}{2}\Big[\frac{\mathtt{w}_{p,1}^2}{n}\left(\|\boldsymbol{\alpha}_1\|^2\|\boldsymbol{\beta}_2\|^2+\|\boldsymbol{\alpha}_1\|^2\|\boldsymbol{\beta}_2\|^2\right)\\ &\quad +\sigma^2(\mathtt{C}_{f,2}+\mathtt{w}_{p,2}\mathtt{r}/\sqrt{n})^2\frac{\mathtt{C}_{\Phi}^2}{n}\|\boldsymbol{\beta}_1\|^2\|\boldsymbol{\beta}_2\|^2\Big]\\ &\le \frac{3\nu_\varepsilon^2\sigma^2\left[\mathtt{w}_{p,1}^2+\sigma^2(\mathtt{C}_{f,2}+\mathtt{w}_{p,2}\mathtt{r}/\sqrt{n})^2\mathtt{v}_q^2\right]\|\boldsymbol{\gamma}_1\|^2\|\boldsymbol{\gamma}_2\|^2}{2n}.\end{aligned}$$

□

4.5 Large deviation bound

In this section we state large deviation inequality for the full estimator $(\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\eta}})$. First, we check concentration of the plug-in estimator.

Lemma 4.9. *It holds with probability at least* $1-4e^{-\mathtt{x}}$,

$$\widetilde{\boldsymbol{\upsilon}}^{(\mathsf{pl})}=\left(\widetilde{\boldsymbol{\theta}}^{(\mathsf{pl})},\widetilde{\boldsymbol{\eta}}^{(\mathsf{pl})}\right)\in\Upsilon_\circ(\mathtt{r}_1),$$

where $\mathtt{r}_1=4(1+\sigma\mathtt{C}_{f,1})\nu_\varepsilon\sqrt{p+q+\mathtt{x}}$.

Proof. First of all, we have with probability $1-2e^{-\mathtt{x}}$,

$$\sigma^{-1}\|\widetilde{\boldsymbol{\eta}}^{(\mathsf{pl})}-\boldsymbol{\eta}^*\|\le\nu_\varepsilon(\sqrt{q}+\sqrt{2\mathtt{x}}).$$

Then, using Lemma 4.3 with $\frac{\sigma\nu_\varepsilon\mathtt{w}_{p,1}(\sqrt{q}+\sqrt{2\mathtt{x}})}{\sqrt{n}}\le\frac{1}{6}$ by condition (vii),

$$\widetilde{\Psi}\widetilde{\Psi}^\top\ge D^2_{\boldsymbol{\theta}\boldsymbol{\theta}}/2,\qquad \left\|(\widetilde{\Psi}-\Psi(\Phi\boldsymbol{\eta}^*))^\top\boldsymbol{\theta}^*\right\|\le\sigma\mathtt{C}_{f,1}\nu_\varepsilon(\sqrt{q}+\sqrt{2\mathtt{x}}).$$

So,

$$\widetilde{\Psi}^\top(\widetilde{\boldsymbol{\theta}}^{(\mathsf{pl})}-\boldsymbol{\theta}^*)=\widetilde{\Psi}(\widetilde{\Psi}\widetilde{\Psi}^\top)^{-1}\widetilde{\Psi}\big(\boldsymbol{Y}-\Psi(\Phi\boldsymbol{\eta}^*)^\top\boldsymbol{\theta}^*+(\Psi(\Phi\boldsymbol{\eta}^*)-\widetilde{\Psi})^\top\boldsymbol{\theta}^*\big),$$

and, since $\left\|(\widetilde{\Psi}\widetilde{\Psi}^\top)^{-1/2}\widetilde{\Psi}\right\|=1$, we have

$$\left\|\widetilde{\Psi}^\top(\widetilde{\boldsymbol{\theta}}^{(\mathsf{pl})}-\boldsymbol{\theta}^*)\right\|\le\left\|(\widetilde{\Psi}\widetilde{\Psi}^\top)^{-1/2}\widetilde{\Psi}\boldsymbol{\varepsilon}\right\|+\sigma\mathtt{C}_{f,1}\nu_\varepsilon(\sqrt{q}+\sqrt{2\mathtt{x}}),$$

where conditioned on $\boldsymbol{Z}$, we have $\left\|(\widetilde{\Psi}\widetilde{\Psi}^\top)^{-1/2}\widetilde{\Psi}\boldsymbol{\varepsilon}\right\|\le\nu_\varepsilon(\sqrt{p}+\sqrt{2\mathtt{x}})$ with probability at least $1-2e^{-\mathtt{x}}$. □

Further we find a stationary point of log-likelihood as a maxima over the set

$$\Upsilon_\circ(\mathtt{R}_0),\qquad \mathtt{R}_0\stackrel{\mathrm{def}}{=}\frac{\sqrt{n}}{2\sigma\mathtt{w}_{p,1}}.$$

Lemma 4.10. *It holds with probability at least* $1 - e^{-\mathtt{x}}$,

$$\sup_{\upsilon \in \Upsilon_\circ(\mathtt{R}_0) \setminus \Upsilon_\circ(\mathtt{r}_2)} L_\Phi(\upsilon) - L_\Phi(\upsilon^*) \le 0, \tag{4.1}$$

where

$$\begin{aligned} \mathtt{r}_2 &= \mathtt{C}_0\{\sqrt{p+q+\mathtt{x}} \vee \square\}, \\ \mathtt{C}_0 &= 96\nu_\varepsilon\big[2(\sigma\mathtt{C}_{f,1} \vee 1) + 1\big](\sigma\mathtt{C}_{f,1} + 3/2). \end{aligned}$$

Proposition 4.11. *Let*

$$2\sigma\mathtt{C}_0\mathtt{w}_{p,1}\{\sqrt{p+q+\mathtt{x}} \vee \square\} \le \sqrt{n}.$$

Then, we have with probability at least $1 - 5e^{-\mathtt{x}}$

$$(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\eta}}),\ (\boldsymbol{\theta}^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}) \in \Upsilon_\circ(\mathtt{r}_0),$$

where $\mathtt{r}_0 = \mathtt{f}^{-1}\mathtt{F}(\mathtt{r}_1 + \mathtt{r}_2) + \mathtt{r}_1$.

Proof. The given inequality means $\mathtt{R}_0 > \mathtt{r}_2$, thus the maximum of $L_\Phi(\upsilon)$ over $\Upsilon_\circ(\mathtt{R}_0)$ lies inside $\Upsilon_\circ(\mathtt{r}_2)$ with probability at least $1 - e^{-\mathtt{x}}$ and, therefore, is a stationary point. So, we have such a point at most $\mathtt{r}_1 + \mathtt{r}_2$ far from the plug-in estimator. Thus, the closest stationary point to plug-in satisfies w.h.p.

$$\|H_0(\widetilde{\upsilon} - \upsilon^*)\| \le \mathtt{f}^{-1}\mathtt{F}(\mathtt{r}_1 + \mathtt{r}_2) + \mathtt{r}_1.$$

Obviously, the same arguments work for $(\boldsymbol{\theta}^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*})$. □

Remark 4.1. The inclusion $\widetilde{\upsilon}, \widetilde{\upsilon}_{\boldsymbol{\theta}^*} \in \Upsilon_\circ(\mathtt{r}_0)$ implies

$$\widehat{\upsilon}_{\boldsymbol{\theta}^*} = \Big(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*} - D^{-2}_{\boldsymbol{\eta}\boldsymbol{\eta}} A^\top_{\boldsymbol{\theta}\boldsymbol{\eta}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\Big) \in \Upsilon_\circ(3\mathtt{r}_0),$$

since

$$\begin{aligned} \|H_0(\widehat{\upsilon}_{\boldsymbol{\theta}^*} - \upsilon^*)\| &\le \mathtt{r}_0 + \sigma^{-1}\|\widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*} - \boldsymbol{\eta}^*\| + \|H_\sigma D^{-2}_{\boldsymbol{\eta}\boldsymbol{\eta}} A^\top_{\boldsymbol{\theta}\boldsymbol{\eta}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \\ &\le 2\mathtt{r}_0 + \|H_\sigma D^{-1}_{\boldsymbol{\eta}\boldsymbol{\eta}}\| \times \|D^{-1}_{\boldsymbol{\eta}\boldsymbol{\eta}} A^\top_{\boldsymbol{\theta}\boldsymbol{\eta}} D^{-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}\| \times \|D_{\boldsymbol{\theta}\boldsymbol{\theta}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \\ &\le 3\mathtt{r}_0, \end{aligned}$$

where $\|D^{-1}_{\boldsymbol{\eta}\boldsymbol{\eta}} A^\top_{\boldsymbol{\theta}\boldsymbol{\eta}} D^{-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}\| \le \rho < 1$ from Lemma $(\mathcal{I})$. This strange object we need to show Wilks expansion.

To proof Lemma 4.10 we are going to straightforwardly use Theorem 2.1 of Andresen and Spokoiny (2013), which requires two conditions: the first one controls the stochastic part of $L_\Phi(\upsilon) - L_\Phi(\upsilon^*)$ and the second one controls it's expectation. Bounding the stochastic part requires check of the exponential moments of the score $\nabla\zeta_\Phi(\upsilon)$, which is done by Lemma 4.7. Now we need to check that $-I\!\!E\big[L_\Phi(\upsilon) - L_\Phi(\upsilon^*)\big]$ grows quadratically.

Lemma 4.12. *Let* $A, \mathtt{b} > 0$ *satisfy*

$$\big[4(\sigma \mathtt{C}_{f,1} \vee 1) + 2\big]\sqrt{\mathtt{b} + 2A^{-2}} \le 1. \tag{4.2}$$

Then, $\upsilon \in \Upsilon_\circ(\mathtt{R}_0)$ *with* $\mathtt{r} = \|H_0(\upsilon - \upsilon^*)\| \ge A\square$ *implies*

$$-2I\!\!E\big[L_\Phi(\upsilon) - L_\Phi(\upsilon^*)\big] \ge \mathtt{b}\mathtt{r}^2.$$

Proof of Lemma 4.10. First,

$$\log I\!\!E \exp\big\{\gamma^\top H_0^{-1}\nabla\zeta_\Phi(\upsilon)\big\} \le \frac{\nu_1^2(\mathtt{R}_0)\|\gamma\|^2}{2}, \qquad \upsilon \in \Upsilon_\circ(\mathtt{R}_0),$$

where $\nu_1^2(\mathtt{R}_0)$ is given in Lemma 4.7. Since $\mathtt{R}_0 \le \mathtt{w}_{p,1}^{-1}\sqrt{n}/2$, we can take

$$\nu_1(\mathtt{R}_0) \le \nu_1 \stackrel{\text{def}}{=} \sqrt{2}\nu_\varepsilon\big(3/2 + \sigma\mathtt{C}_{f,1}\big).$$

Further, setting $A^2 = 2\mathtt{b}^{-1} = 16\big[2(\sigma\mathtt{C}_{f,1} \vee 1) + 1\big]^2$ in Lemma 4.12, we have by the theorem mentioned above, that (4.1) holds, when

$$\mathtt{r}_2 \ge \frac{6\nu_1}{\mathtt{b}}\sqrt{2(p+q) + \mathtt{x}},$$

which finishes the proof. □

Proof of Lemma 4.12. We need to show that for $\upsilon = (\theta, \eta)$ with $\|H_0(\upsilon - \upsilon^*)\| = \mathtt{r}$ and $A\square \le \mathtt{r} \le \mathtt{R}_0$ it holds,

$$\|f^* - \Psi(\Phi\eta)^\top\theta\|^2 + \sigma^{-2}\|\eta - \eta^*\|^2 - \|f^* - \Psi(\Phi\eta^*)^\top\theta^*\|^2 \ge \mathtt{b}\mathtt{r}^2. \tag{4.3}$$

We have,

$$f^* - \Psi(\Phi\eta)^\top\theta = \Psi(\Phi\eta^*)^\top(\theta - \theta^*) + (\Psi(\Phi\eta) - \Psi(\Phi\eta^*))^\top(\theta - \theta^*)$$
$$+(\Psi(\Phi\eta) - \Psi(\Phi\eta^*))^\top\theta^*.$$

Denote,

$$\begin{aligned}
&\boldsymbol{x}_{\boldsymbol{\theta}} = \Psi(\Phi\boldsymbol{\eta}^*)^\top(\boldsymbol{\theta}-\boldsymbol{\theta}^*), && \mathtt{r}_{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \|\boldsymbol{x}_\theta\|,\\
&\boldsymbol{x}_{\boldsymbol{\eta}} = \sigma^{-2}(\boldsymbol{\eta}-\boldsymbol{\eta}^*), && \mathtt{r}_{\boldsymbol{\eta}} \stackrel{\text{def}}{=} \|\boldsymbol{x}_{\boldsymbol{\eta}}\|,\\
&\boldsymbol{y}_1 = (\Psi(\Phi\boldsymbol{\eta}) - \Psi(\Phi\boldsymbol{\eta}^*))^\top(\boldsymbol{\theta}-\boldsymbol{\theta}^*), && \|\boldsymbol{y}_1\| \le \sigma n^{-1/2}\mathtt{w}_{p,1}\mathtt{r}_{\boldsymbol{\theta}}\mathtt{r}_{\boldsymbol{\eta}},\\
&\boldsymbol{y}_2 = (\Psi(\Phi\boldsymbol{\eta}) - \Psi(\Phi\boldsymbol{\eta}^*))^\top\boldsymbol{\theta}^*, && \|\boldsymbol{y}_2\| \le \sigma\mathtt{C}_{f,1}\mathtt{r}_{\boldsymbol{\eta}},\\
&\boldsymbol{y}_3 = \boldsymbol{f}^* - \Psi(\Phi\boldsymbol{\eta}^*)^\top\boldsymbol{\theta}^*, && \|\boldsymbol{y}_3\| = \square.
\end{aligned}$$

Then (4.3) rewrites as

$$\|\boldsymbol{x}_{\boldsymbol{\theta}} + \boldsymbol{y}_1 + \boldsymbol{y}_2 + \boldsymbol{y}_3\|^2 + \mathtt{r}_{\boldsymbol{\eta}}^2 - \square^2 \ge \mathtt{b}\left(\mathtt{r}_{\boldsymbol{\theta}}^2 + \mathtt{r}_{\boldsymbol{\eta}}^2\right).$$

If $\mathtt{r}_{\boldsymbol{\eta}}^2 - \square^2 \ge \mathtt{b}\left(\mathtt{r}_{\boldsymbol{\theta}}^2 + \mathtt{r}_{\boldsymbol{\eta}}^2\right)$ then (4.3) is obviously satisfied. Otherwise, defining $\widehat{\mathtt{b}} \stackrel{\text{def}}{=} \mathtt{b} + A^{-2}$, we have

$$\mathtt{r}_{\boldsymbol{\eta}} \le \sqrt{2\widehat{\mathtt{b}}}\,\mathtt{r}_{\boldsymbol{\theta}}, \qquad \square \le \sqrt{2}A^{-1}\mathtt{r}_{\boldsymbol{\theta}}.$$

Further, $\sigma^{-1}\|\boldsymbol{\eta}-\boldsymbol{\eta}^*\| \le (2\sigma\mathtt{w}_{p,1})^{-1}\sqrt{n}$ and therefore, $\|\boldsymbol{y}_1\| \le \mathtt{r}_{\boldsymbol{\theta}}/2$. So,

$$\mathtt{r}_{\boldsymbol{\theta}} - \|\boldsymbol{y}_1\| - \|\boldsymbol{y}_2\| - \|\boldsymbol{y}_3\| \ge \left(1 - \frac{1}{2} - \sigma\mathtt{C}_{f,1}\sqrt{2\widehat{\mathtt{b}}} - \sqrt{2}A^{-1}\right)\mathtt{r}_{\boldsymbol{\theta}}.$$

We need to show that the last expression to square is at least $\mathtt{b}\mathtt{r}_{\boldsymbol{\theta}}^2 + \square^2 \le \{\mathtt{b} + 2A^{-2}\}\mathtt{r}_{\boldsymbol{\theta}}^2$ and the proof is finished by checking that (4.2) ensures

$$\frac{1}{2} - \sigma\mathtt{C}_{f,1}\sqrt{2\widehat{\mathtt{b}}} - \sqrt{2}A^{-1} \ge \sqrt{\widehat{\mathtt{b}} + A^{-2}}.$$

□

4.6 Local linear approximation of the score

We start with the linear approximation of gradient of the likelihood. Define for $\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})$ the following process

$$\breve{\chi}(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} \breve{D}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\breve{\nabla}_{\boldsymbol{\theta}} L_\Phi(\boldsymbol{\upsilon}) - \breve{\boldsymbol{\xi}} + \breve{D}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*),$$

where operator $\breve{\nabla}_{\boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}} - A_{\boldsymbol{\theta}\boldsymbol{\eta}}D_{\boldsymbol{\eta}\boldsymbol{\eta}}^{-2}\nabla_{\boldsymbol{\eta}}$, $\breve{\boldsymbol{\xi}} = \breve{D}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\breve{\nabla}\zeta_\Phi(\boldsymbol{\upsilon}^*)$ and $\breve{D}_{\boldsymbol{\theta}\boldsymbol{\theta}}$ defined in (2.4).

In what follows we will also exploit function

$$z_{p+q}(\mathtt{x}) = \sqrt{2(p+q) + \mathtt{x}},$$

which is involved in the concentration inequalities for suprema of empirical processes.

Lemma 4.13. *On a set* $C(\mathtt{x},\mathtt{r}) \subset \Omega$ *of probability at least* $1 - e^{-\mathtt{x}}$ *it holds*

$$\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \bigl\| \breve{\chi}(\boldsymbol{\upsilon}) \bigr\| \leq \diamondsuit(\mathtt{x},\mathtt{r}) + \square,$$

where

$$\diamondsuit(\mathtt{x},\mathtt{r}) \, = \, 2(1-\rho)^{-1/2} \bigl\{ 6\nu_2(\mathtt{r}) z_{p+q}(\mathtt{x}) + \delta(\mathtt{r}) \bigr\} \mathtt{r}$$

and $\square = \| \boldsymbol{f}^* - \Psi(\Phi\boldsymbol{\eta}^*)^\top \boldsymbol{\theta}^* \|$ *from the condition (vi).*

Proof. Define the process, corresponding to the linear approximation of $\nabla L_\Phi(\boldsymbol{\upsilon})$ among the biased point $\boldsymbol{\upsilon}^*$,

$$\chi(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} D^{-1} \{ \nabla L_\Phi(\boldsymbol{\upsilon}) - \nabla L_\Phi(\boldsymbol{\upsilon}^*) + D^2(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) \}.$$

We can split this vector into two terms, the expectation and the stochastic part. First, using $I\!\!E \nabla L_\Phi = \nabla I\!\!E L_\Phi$ we get

$$I\!\!E \chi(\boldsymbol{\upsilon}) = D^{-1} \{ \nabla I\!\!E L_\Phi(\boldsymbol{\upsilon}) - \nabla I\!\!E L_\Phi(\boldsymbol{\upsilon}^*) + D^2(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) \},$$

which using $(\mathcal{L}_0)$ and $(\mathcal{I})$ can be easily bounded, see details in Andresen and Spokoiny (2013),

$$\| I\!\!E \chi(\boldsymbol{\upsilon}) \| \leq 2(1-\rho)^{-1/2} \mathtt{r} \delta(\mathtt{r}).$$

Using Theorem 4.1 of Andresen and Spokoiny (2013) along with Lemma 4.8, the value $\bigl\| H_0^{-1} \{ \nabla \zeta_\Phi(\boldsymbol{\upsilon}) - \nabla \zeta_\Phi(\boldsymbol{\upsilon}^*) \} \bigr\|$ is bounded by $6\nu_2(\mathtt{r}) \sqrt{2(p+q) + \mathtt{x}}$ uniformly over $\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})$ with porbability at least $1 - e^{-\mathtt{x}}$. To sum up, we get on a set $C(\mathtt{x},\mathtt{r}) \subset \Omega$

$$\begin{aligned} \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \bigl\| \chi(\boldsymbol{\upsilon}) \bigr\| \, &\leq \, \diamondsuit(\mathtt{x},\mathtt{r}) \\ &\stackrel{\text{def}}{=} 2(1-\rho)^{-1/2} \bigl\{ 3\nu_2(\mathtt{r}) z_{p+q}(\mathtt{x}) + \delta(\mathtt{r}) \bigr\} \mathtt{r}, \end{aligned}$$

where $I\!\!P(C(\mathtt{x},\mathtt{r})) \geq 1 - e^{-\mathtt{x}}$. Define,

$$\boldsymbol{b} = D^{-1} \nabla L_\Phi(\boldsymbol{\upsilon}^*), \qquad \| \boldsymbol{b} \| \leq \| \boldsymbol{f}^* - \Psi(\Phi\boldsymbol{\eta}^*)^\top \boldsymbol{\theta}^* \|,$$

with the inequality following from Lemma 4.5. Then, for $\breve{\Pi}_{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \bigl(I_p \quad -A_{\boldsymbol{\theta}\boldsymbol{\eta}} D_{\boldsymbol{\eta}\boldsymbol{\eta}}^{-2} \bigr)$ we have $\breve{\nabla}_{\boldsymbol{\theta}} = \breve{\Pi}_{\boldsymbol{\theta}} \nabla$ and the proof is finished by

$$\breve{\chi}(\boldsymbol{\upsilon}) = \breve{D}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1} \breve{\Pi}_{\boldsymbol{\theta}} D \, \bigl\{ \chi(\boldsymbol{\upsilon}) + \boldsymbol{b} \bigr\},$$

where we have $\| \breve{D}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1} \breve{\Pi}_{\boldsymbol{\theta}} D \| = 1$, since the matrix is orthonormal. $\square$

4.7 Semiparametric Fisher and square-rootWilks

The goal of the Fisher expansion is approximation $\breve{D}_{\boldsymbol{\theta\theta}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \approx \breve{\boldsymbol{\xi}}$, which can be done as follows. Once we have large deviation bound $(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\eta}}) \in \Upsilon_\circ(\mathtt{r}_0)$ with high probability, we can apply Lemma 4.13 with $\boldsymbol{\upsilon} = \widetilde{\boldsymbol{\upsilon}}$, which using $\nabla L_\Phi(\widetilde{\boldsymbol{\upsilon}}) = 0$ gives

$$\|\breve{D}_{\boldsymbol{\theta\theta}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \breve{\boldsymbol{\xi}}\| \le \diamondsuit(\mathtt{x}, \mathtt{r}_0) + \square,$$

on the set $\{\widetilde{\boldsymbol{\upsilon}} \in \Upsilon_\circ(\mathtt{r}_0)\} \cap C(\mathtt{x}, \mathtt{r}_0)$, which is of probability at least $1 - 4e^{-\mathtt{x}}$.

Remark 4.2. We don't require $\mathtt{r}_0$ to be exactly as in Proposition 4.11, the global concentration can be held by any other means and we only require $\Upsilon_\circ(\mathtt{r}_0)$ to contain the estimator $(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\eta}})$ with high probability.

Semiparametric square-root Wilks approximates the distribution of $\{2L_\Phi(\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*})\}^{1/2}$, by the the distribution of $\|\breve{\boldsymbol{\xi}}\|$. This might be shown via quadratic approximation of the likelihood. The Lemma 4.13 helps to ensure it.

Lemma 4.14. *Define,*

$$\begin{aligned}\alpha(\boldsymbol{\theta}, \boldsymbol{\eta}) \, = \, & L_\Phi(\boldsymbol{\theta}, \boldsymbol{\eta}) - L_\Phi\Bigl(\boldsymbol{\theta}^*, \boldsymbol{\eta} - D^{-2}_{\boldsymbol{\eta\eta}} A^\top_{\boldsymbol{\theta\eta}} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)\Bigr) \\ & - \bigl\{\breve{D}_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\bigr\}^\top \breve{\boldsymbol{\xi}} + \bigl\|\breve{D}_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\bigr\|^2/2.\end{aligned}$$

Then if $(\boldsymbol{\theta}, \boldsymbol{\eta}) \in \Upsilon_\circ(\mathtt{r})$, *it holds (pointwise)*

$$\frac{|\alpha(\boldsymbol{\theta}, \boldsymbol{\eta})|}{\bigl\|\breve{D}_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\bigr\|} \le \sup_{(\boldsymbol{\theta}, \boldsymbol{\eta}) \in \Upsilon_\circ(\mathtt{r})} \|\chi(\boldsymbol{\theta}, \boldsymbol{\eta})\|.$$

Proof. For a given pair $(\boldsymbol{\theta}, \boldsymbol{\eta})$ define $\breve{\boldsymbol{\eta}} = \boldsymbol{\eta} - D^{-2}_{\boldsymbol{\eta\eta}} A_{\boldsymbol{\theta\eta}} \boldsymbol{\theta}$ and $\breve{L}_\Phi(\boldsymbol{\theta}, \boldsymbol{\eta}) = L_\Phi(\boldsymbol{\theta}, \breve{\boldsymbol{\eta}})$, such that $\nabla_{\boldsymbol{\theta}} \breve{L}_\Phi(\boldsymbol{\theta}, \boldsymbol{\eta}) = \breve{\nabla}_{\boldsymbol{\theta}} L_\Phi(\boldsymbol{\theta}, \breve{\boldsymbol{\eta}})$. Define the approximating process

$$\breve{\alpha}(\boldsymbol{\theta}, \boldsymbol{\eta}) \stackrel{\text{def}}{=} \breve{L}_\Phi(\boldsymbol{\theta}, \boldsymbol{\eta}) - \breve{L}_\Phi(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - \{\breve{D}_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\}^\top \breve{\boldsymbol{\xi}} + \|\breve{D}_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2.$$

Note, that $\breve{\alpha}(\boldsymbol{\theta}^*, \boldsymbol{\eta}) \equiv 0$ for all $\boldsymbol{\eta}$. Thus, by mean value theorem $\bigl|\breve{\alpha}(\boldsymbol{\theta}, \boldsymbol{\eta})\bigr| \le \|\breve{D}_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \times \sup_{\boldsymbol{\theta}^\circ \in [\boldsymbol{\theta}, \boldsymbol{\theta}^*]} \bigl\|\breve{D}^{-1}_{\boldsymbol{\theta\theta}} \nabla_{\boldsymbol{\theta}} \breve{\alpha}(\boldsymbol{\theta}^\circ, \boldsymbol{\eta})\bigr\|$. Calculating the derivative gives

$$\breve{D}^{-1}_{\boldsymbol{\theta\theta}} \nabla_{\boldsymbol{\theta}} \breve{\alpha}(\boldsymbol{\theta}, \boldsymbol{\eta}) \, = \, \breve{D}^{-1}_{\boldsymbol{\theta\theta}} \breve{\nabla}_{\boldsymbol{\theta}} L_\Phi(\boldsymbol{\theta}, \breve{\boldsymbol{\eta}}) - \breve{\boldsymbol{\xi}} + \breve{D}^2_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) = \chi(\boldsymbol{\theta}, \breve{\boldsymbol{\eta}}),$$

which gives the desired result after noticing, that $\alpha(\boldsymbol{\theta}, \breve{\boldsymbol{\eta}}) = \breve{\alpha}(\boldsymbol{\theta}, \boldsymbol{\eta})$. □

As we mentioned in Remark 4.1, it holds $(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\eta}}), \Bigl(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*} - D^{-2}_{\boldsymbol{\eta\eta}} A_{\boldsymbol{\theta\eta}} (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\Bigr) \in \Upsilon_\circ(3\mathtt{r}_0)$ with probability $\ge 1 - 3e^{-\mathtt{x}}$. So, define

$$\diamondsuit(\mathtt{x}) \stackrel{\text{def}}{=} \diamondsuit(\mathtt{x}, 3\mathtt{r}_0) + \square.$$

Then by Lemmas 4.13, 4.14 we have on the set $C(\mathtt{x},3\mathtt{r}_0)\cap\{\widetilde{\boldsymbol{\upsilon}},\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}\in\Upsilon_\circ(\mathtt{r}_0)\}$ the following chain of inequalities, with $\widetilde{\boldsymbol{u}}=\breve{D}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)$

$$\begin{aligned}
L_\Phi(\widetilde{\boldsymbol{\upsilon}})-L_\Phi(\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}) &\le L_\Phi(\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\eta}})-L_\Phi\Big(\boldsymbol{\theta}^*,\widetilde{\boldsymbol{\eta}}-D_{\boldsymbol{\eta}\boldsymbol{\eta}}^{-2}A_{\boldsymbol{\theta}\boldsymbol{\eta}}^\top(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)\Big)\\
&\le \widetilde{\boldsymbol{u}}^\top\breve{\boldsymbol{\xi}}-\|\widetilde{\boldsymbol{u}}\|^2\Big/2+\diamondsuit^*(\mathtt{x})\|\widetilde{\boldsymbol{u}}\|\\
&\le \|\widetilde{\boldsymbol{u}}\|^2/2+\|\widetilde{\boldsymbol{u}}\|\big\{\|\widetilde{\boldsymbol{u}}-\breve{\boldsymbol{\xi}}\|+\diamondsuit^*(\mathtt{x})\big\}\\
&\le \|\widetilde{\boldsymbol{u}}\|^2/2+2\|\widetilde{\boldsymbol{u}}\|\diamondsuit(\mathtt{x}),
\end{aligned}$$

and on the other hand,

$$\begin{aligned}
L_\Phi(\widetilde{\boldsymbol{\upsilon}})-L_\Phi(\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*}) &\ge L_\Phi\Big(\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}-D_{\boldsymbol{\eta}\boldsymbol{\eta}}^{-2}A_{\boldsymbol{\theta}\boldsymbol{\eta}}(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)\Big)-L_\Phi(\boldsymbol{\theta}^*,\widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*})\\
&\ge \|\widetilde{\boldsymbol{u}}\|^2/2-2\|\widetilde{\boldsymbol{u}}\|\diamondsuit(\mathtt{x}),
\end{aligned}$$

which together gives

$$\Big|\{2L_\Phi(\widetilde{\boldsymbol{\upsilon}},\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*})\}^{1/2}-\|\widetilde{\boldsymbol{u}}\|\Big| \le \frac{\big|2L_\Phi(\widetilde{\boldsymbol{\upsilon}},\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*})-\|\widetilde{\boldsymbol{u}}\|^2\big|}{\|\widetilde{\boldsymbol{u}}\|}\le 4\diamondsuit(\mathtt{x}),$$

where we have used that on the same subset of Ω it holds $\|\widetilde{\boldsymbol{u}}-\breve{\boldsymbol{\xi}}\|\le\diamondsuit^*(\mathtt{x})$ from the Fisher expansion. Using this inequality one last time gives

$$\Big|\{2L_\Phi(\widetilde{\boldsymbol{\upsilon}},\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*})\}^{1/2}-\|\breve{\boldsymbol{\xi}}\|\Big|\le 5\diamondsuit(\mathtt{x}).$$

4.8 Proof of Theorem 2.1

From Proposition 4.11 we have $\mathtt{r}_0\asymp\sqrt{p+q+\mathtt{x}}\vee\square$. From Lemma 4.6 and Lemma 4.8 we have,

$$\begin{aligned}
\diamondsuit(\mathtt{x}) &\asymp \frac{\sigma(\mathtt{w}_{p,1}+\mathtt{v}_q)\mathtt{r}_0^2}{\sqrt{n}}+\frac{\sigma\mathtt{w}_{p,2}\mathtt{v}_q\mathtt{r}_0^3}{n}+\square,\\
&\lesssim \frac{\sigma(\mathtt{w}_{p,1}+\mathtt{v}_q)(p+q+\mathtt{x})}{\sqrt{n}}+\frac{\sigma\mathtt{w}_{p,2}\mathtt{v}_q(p+q+\mathtt{x})^{3/2}}{n}\\
&\quad+\square+\frac{\sigma(\mathtt{w}_{p,1}+\mathtt{v}_q)\square^2}{\sqrt{n}}+\frac{\sigma\mathtt{w}_{p,2}\mathtt{v}_q\square^3}{n}\\
&\lesssim \sigma\Big(\mathtt{w}_{p,1}+\mathtt{v}_q+\mathtt{w}_{p,2}\mathtt{v}_q\sqrt{n^{-1}(p+q+\mathtt{x})}\Big)(p+q+\mathtt{x})n^{-1/2}+\square,
\end{aligned}$$

by use of the inequalities given by condition (vii). It remains to check, that

$$\log I\!E\exp\big(\boldsymbol{\gamma}^\top\breve{\boldsymbol{\xi}}\big)\le\frac{\nu_\varepsilon^2\|\boldsymbol{\gamma}\|^2}{2},\qquad \boldsymbol{\gamma}\in I\!R^p.$$

Note, that D^2 corresponds to $\mathrm{Var}\big(\nabla\zeta_{\Phi}(\boldsymbol{\upsilon}^*)\big)$, when the errors ε_i, ν_i are standard normal. In general, one can easily check, that for $\boldsymbol{\xi} = D^{-1}\nabla\zeta_{\Phi}(\boldsymbol{\upsilon}^*)$ and for each $\boldsymbol{\gamma} \in I\!R^{p+q}$ it holds

$$\log I\!E \exp\big(\boldsymbol{\gamma}^\top \boldsymbol{\xi}\big) \le \frac{\nu_\varepsilon^2 \|\boldsymbol{\gamma}\|^2}{2}.$$

Now, the moments for the $\breve{\boldsymbol{\xi}}$ follow by applying this inequality to $\breve{D}^{-1}_{\boldsymbol{\theta}\boldsymbol{\theta}} \breve{I\!I}^\top_{\boldsymbol{\theta}} D\boldsymbol{\gamma}$, since $\breve{D}^{-1}_{\boldsymbol{\theta}\boldsymbol{\theta}} \breve{I\!I}^\top_{\boldsymbol{\theta}} D$ is an orthonormal matrix.

4.9 Proof of Lemma 2.2

Define $N = \Phi^\top\Phi$ and $\widetilde{\Phi} = \Phi N^{-1/2}$. Since N is symmetric, $\widetilde{\Phi}^\top\widetilde{\Phi} = N^{-1/2}\Phi\Phi^\top N^{-1/2} = \boldsymbol{I}$. Moreover, applying trapezoidal rule for integrating $\gamma(t) = h_m(t)h_k(t)$, we get

$$\begin{aligned}
|N_{mk} - \delta_{mk}| = |\boldsymbol{\phi}_m^\top\boldsymbol{\phi}_k - \delta_{mk}| &= \left|\frac{1}{n}\sum_{i=1}^{n} h_m(i/n)h_k(i/n) - \int_0^1 h_m(t)h_k(t)dt\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^{n} \gamma(i/n) - \int_0^1 \gamma(t)dt\right| \\
&= \frac{|\gamma(1)-\gamma(0)|}{2n} + \left|\frac{1}{n}\sum_{i=1}^{n} \frac{\gamma\big((i-1)/n\big)+\gamma(i/n)}{2} - \int_0^1 \gamma(t)dt\right| \\
&\le \frac{\|\gamma\|_\infty}{n} + \frac{\|\gamma''\|_\infty}{12n^2}.
\end{aligned}$$

Using the inequalities $\|\gamma\|_\infty \le A_0^2$ and

$$\|\gamma''\|_\infty = \|h_m h_k'' + h_m'' h_k + 2h_m' h_k'\|_\infty \le A_0A_2m^2 + 2A_1^2mk + A_0A_2k^2 \le 2(A_0A_2 + A_1^2)q^2,$$

we get that the absolute value of each element of the matrix $N - \boldsymbol{I}$ is less or equal to $\delta = \frac{A_0^2}{n} + \frac{(A_0A_2+A_1^2)q^2}{6n^2}$. Thus,

$$\|N-\boldsymbol{I}\| \le \|N-\boldsymbol{I}\|_{\mathrm{Frob}} = \|\mathrm{vec}(N-\boldsymbol{I})\|_2 \le \sqrt{q^2}\|\mathrm{vec}(N-\boldsymbol{I})\|_\infty \le q\delta \le \frac{1}{2},$$

with the last inequality following from the conditions of the lemma. Thus, $\|N^{-1}\| \le 2$ and

$$C(\widetilde{\Phi}) = n\max_{i\le n}\|N^{-1/2}\Phi_i\|^2 \le n\|N^{-1}\|\max_{i\le n}\|\Phi_i\|^2 \le 2C(\Phi).$$

4.10 Proof of Lemma 2.4

For $h \in C^1[0,1]$ and $N \in \mathbb{N}$ we have,

$$\int_0^1 h(x)dx \le \frac{1}{N}\sum_{j=1}^{N} \min_{x\in\left[\frac{j-1}{N},\frac{j}{N}\right)} h(x) + \frac{\|h'\|_\infty}{N}.$$

Take $N < \delta^{-1}p^2$. By condition of the lemma, each interval $\left[\frac{j-1}{N}, \frac{j}{N}\right)$ contains at least $\mathtt{c}_0 n \left(\frac{1}{N} - \delta p^{-2}\right)$ points. So,

$$\frac{1}{N} \min_{x \in \left[\frac{j-1}{N}, \frac{j}{N}\right)} h(x) \le \frac{(1 - N\delta p^{-2})^{-1}}{\mathtt{c}_0 n} \sum_{X_i^* \in \left[\frac{j-1}{N}, \frac{j}{N}\right)} h(X_i^*),$$

which, by summing over $j = 1, \ldots, N$, gives

$$\int_0^1 h(x)dx \le \frac{\left(\mathtt{c}_0(1 - N\delta p^{-2})\right)^{-1}}{n} \sum_{i=1}^n h(X_i^*) + \frac{\|h'\|_\infty}{N}.$$

Fix arbitrary unit vector $\boldsymbol{\gamma} \in I\!R^p$ and set $h(x) = (\Psi(x)^\top \boldsymbol{\gamma})^2$. By (a) we have

$$\int_0^1 h(x)dx = 1, \qquad \|h'\|_\infty \le 2A^2 p^2,$$

and taking $N(\delta) = \lfloor (1-\delta)\delta^{-1}p^2 \rfloor$

$$\frac{1}{n}\sum_{i=1}^n h(X_i^*) \ge \mathtt{c}_0\big(1 - \alpha(\delta)\big), \qquad \alpha(\delta) = \left(1 - \frac{2A^2p^2}{N(\delta)}\right)(1 - N(\delta)\delta p^{-2}).$$

Since $N(\delta) \ge \delta^{-1}p^2 - p^2 - 1$,

$$\alpha(\delta) \le (\delta + \delta p^{-2}) \left(1 - \frac{2A^2\delta}{1-\delta}\right).$$

It is left to notice, that $\frac{1}{n}\sum_{i=1}^n h(X_i^*) = \boldsymbol{\gamma}^\top \big[n^{-1}\Psi(\boldsymbol{X}^*)\Psi(\boldsymbol{X}^*)^\top\big]\boldsymbol{\gamma}$.

4.11 Proof of Lemma 2.5

For any $h \in C^1[0,1]$ it holds

$$\|h\|_2^2 \ge \frac{1}{n}\sum_{j=1}^n \max_{x \in \left[\frac{j-1}{n}, \frac{j}{n}\right]} h(x)^2 - \frac{2\|h\|_\infty \|h'\|_\infty}{n}.$$

The amount of points X_i^* lying in $\left[\frac{j-1}{n}, \frac{j}{n}\right]$ is bounded by $\mathtt{C}_0$, so it holds

$$\max_{x \in \left[\frac{j-1}{n}, \frac{j}{n}\right]} h(x)^2 \ge \mathtt{C}_0^{-1} \sum_{X_i^* \in \left[\frac{j-1}{n}, \frac{j}{n}\right]} h(X_i^*)^2,$$

which brings us to

$$\|h\|_2^2 \ge -\frac{2\|h\|_\infty \|h'\|_\infty}{n} + \frac{1}{\mathtt{C}_0 n}\sum_{i \le n} h(X_i^*)^2.$$

Applying this to $h = f - f_p$, using $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ and Lemma 2.3, gives

$$\|\boldsymbol{f}^* - \Psi(\boldsymbol{X}^*)^\top \boldsymbol{\theta}^*\| \le \mathtt{C}_0^{1/2}\sqrt{n}\|f - f_p\|_2 + (2\mathtt{C}_0)^{1/2}\sqrt{\|f - f_q\|_\infty \|f' - f_q'\|_\infty}$$
$$\le \mathtt{C}_0^{1/2} Q\sqrt{n} p^{-\beta} + \frac{(2\mathtt{C}_0)^{1/2} AQ}{\sqrt{2\beta - 3}} p^{-\beta+1}.$$

Further, since $\|\Phi\boldsymbol{\eta}^* - \boldsymbol{X}^*\|^2 = \sum_{i \le n}(g_q(i/n) - g(i/n))^2.$, we use arguments above and Lemma 2.3 to show

$$\|\Phi\boldsymbol{\eta}^* - \boldsymbol{X}^*\| \le \sqrt{n}\|g_q - g\|_2 + \sqrt{2\|g_q - g\|_\infty \|(g_q - g)'\|_\infty}$$
$$\le Q\sqrt{n} q^{-\beta_X} + \frac{\sqrt{2}AQ}{\sqrt{2\beta_X - 3}} q^{-\beta_X + 1}.$$

4.12 Proof of Proposition 2.6

First we check the conditions of Theorem 2.1.

(i). Due to (a) and Lemma 2.3, we have for $\beta > 2.5$ and $s = 0, 1, 2$

$$\|f_p^{(s)}\|_\infty \le \mathtt{C}_{f,s} = \frac{AQ}{\sqrt{2(\beta - s) - 1}}.$$

(ii). We need to transfer condition (c) from the matrix $\mathbb{F} = \Psi(\boldsymbol{X}^*)\Psi(\boldsymbol{X}^*)^\top$ to the matrix $D^2_{\boldsymbol{\theta}\boldsymbol{\theta}} = \Psi(\Phi\boldsymbol{\eta}^*)\Psi(\Phi\boldsymbol{\eta}^*)^\top$. By Lemma 4.1 it holds,

$$\big\|\mathbb{F}^{-1/2}\big[\Psi(\boldsymbol{X}^*) - \Psi(\Phi\boldsymbol{\eta}^*)\big]\big\|_{\mathrm{op}} \le n^{-1/2}\mathtt{w}_{p,1}\|\boldsymbol{X}^* - \Phi\boldsymbol{\eta}^*\|$$
$$\le \mathtt{C} p^{3/2} q^{-\beta_X},$$

where the last inequality follows from $\mathtt{w}_{p,1} = \mathtt{f}^{-1/2} A p^{3/2}$. So, for a unit vector $\boldsymbol{\gamma} \in I\!R^p$, since $\|\boldsymbol{\gamma}^\top \mathbb{F}^{-1/2}\Psi(\boldsymbol{X}^*)\| = 1$, it holds

$$\big|\boldsymbol{\gamma}^\top \mathbb{F}^{-1/2}(D^2_{\boldsymbol{\theta}\boldsymbol{\theta}} - \mathbb{F})\mathbb{F}^{-1/2}\boldsymbol{\gamma}\big| \le \delta\,(2 + \delta)\,,$$

where $\delta = \mathtt{C} p^{3/2} q^{-\beta_X}$. We have then,

$$D^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \ge (1 - \delta(2 + \delta))\mathbb{F} \ge \mathtt{f}(1 - \delta(2 + \delta))(n I_p),$$

which ensures condition (ii), since $p^{3/2} q^{-\beta_X} = n^{-(\beta - 3/2)/(2\beta + 1)}$ is small.

(iii). By (a) we can take $\mu_{m,s} = A m^s$.

(iv). Using Lemma 2.2, we get this condition satisfied with $\mathtt{C}_{\widetilde{\Phi}} = 2A\sqrt{q}$. Indeed

$$\mathtt{C}_\Phi = n\|\Phi_i\|^2 = \sum_{m=1}^{q} \phi_m^2(i/m) \le A^2 q.$$

(v). We have by (2.13)

$$\begin{aligned}\square &= \|\boldsymbol{f}^* - \Psi(\Phi\boldsymbol{\eta}^*)^\top\boldsymbol{\theta}^*\| \le \mathtt{C}_{f,1}\|\Phi\boldsymbol{\eta}^* - \boldsymbol{X}^*\| + \|\boldsymbol{f}^* - \Psi(\boldsymbol{X}^*)^\top\boldsymbol{\theta}^*\| \\ &\le \mathtt{C}\sqrt{n}(p^{-\beta} + q^{-\beta_X}),\end{aligned}$$

and $\square \le \sqrt{n}$ for large enough p and q.

(vii). Taking $\mathtt{w}_{p,s} = Ap^{s+\frac{1}{2}}$ the conditions are satisfied, since $p^{3/2}\sqrt{(p+q)/n}$ is small.

Using $n^{-1}\breve{D}^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \ge (1-\rho)\mathtt{f}\boldsymbol{I}_p$, which is given by By Lemma 4.4 and condition (ii), Theorem 2.1 provides us with

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \lesssim_{I\!P} \sqrt{\frac{p}{n}} + \Delta_{p,q},$$

where

$$\Delta_{p,q} = \frac{\sigma(p^{3/2}+q^{1/2})(p+q)}{n} + \sigma(p^{3/2}+q^{1/2})(p^{-\beta}+q^{-\beta_X})^2 + q^{-\beta_X} + p^{-\beta_X}.$$

Since $\widetilde{f}_{p,q}(x) = \Psi(x)^\top\widetilde{\boldsymbol{\theta}}$ and $f_p(x) = \Psi(x)^\top\boldsymbol{\theta}^*$, we have $\|\widetilde{f}_{p,q} - f_p\|_2 = \|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|$. Further, using $\|f - f_p\|_2 \lesssim p^{-\beta}$, given by Sobolev condition, we arrive at

$$\begin{aligned}\big\|f - \widetilde{f}_{p,q}\big\|_2 \lesssim_{I\!P} &\sqrt{\frac{p}{n}} + p^{-\beta} + q^{-\beta_X} \\ &+ \left\{\frac{\sigma(p^{3/2}+q^{1/2})(p+q)}{n} + \sigma(p^{3/2}+q^{1/2})(p^{-\beta}+q^{-\beta_X})^2\right\}.\end{aligned}$$

The choice of p and q provides

$$\sqrt{\frac{p}{n}} = p^{-\beta} = q^{-\beta_X} = n^{-\frac{\beta}{2\beta+1}}.$$

Next, we deal with the second term from the figure braces,

$$(p^{3/2}+q^{1/2})(p^{-\beta}+q^{-\beta})^2 \asymp n^{\frac{\frac{3}{2}\vee\frac{\beta}{2\beta_X}-2\beta}{2\beta+1}} \le n^{\frac{-\beta}{2\beta+1}},$$

where the inequality holds since $\beta \ge 2 > 3/2$ and $2\beta_X \ge 3 > 1$. Finally, we let us deal with the term

$$\begin{aligned}\frac{(p^{3/2}+q^{1/2})(p+q)}{n} &\asymp \frac{p^{5/2}+q^{3/2}+p^{3/2}q}{n} \\ &= n^{\left(\frac{5}{2}\vee\frac{3\beta}{2\beta_X}\vee\left(\frac{3}{2}+\frac{\beta}{\beta_X}\right)-2\beta-1\right)/(2\beta+1)} \le r_n(\beta) = n^{-\beta/(2\beta+1)},\end{aligned}$$

with the inequality equivalent to the union of the following inequalities

$$\begin{cases} \dfrac{5}{2} - 2\beta - 1 \le -\beta, \\ \dfrac{3\beta}{2\beta_X} - 2\beta - 1 \le -\beta, \\ \dfrac{3}{2} + \dfrac{\beta}{\beta_X} - 2\beta - 1 \le -\beta \end{cases} \Leftrightarrow \begin{cases} \dfrac{3}{2} \le \beta, \\ \dfrac{3}{2}\beta \le \beta\beta_X + \beta_X, \\ \dfrac{1}{2}\beta_X + \beta \le \beta\beta_X \end{cases} \Leftrightarrow \begin{cases} \beta \ge \dfrac{3}{2}, \\ (\beta+1)(\beta_X - 3/2) \ge -\dfrac{3}{2}, \\ (\beta - 1/2)(\beta_X - 1) \ge \dfrac{1}{2}, \end{cases}$$

where all of the last one are true, since $\beta \ge 2$ and $\beta_X \ge 3/2$.

References

Andresen, A. and Spokoiny, V. (2013). Critical dimension in profile semiparametric estimation. Manuscript. arXiv:1303.4640.

Butucea, C. and Taupin, M.-L. (2008). New M-estimators in semi-parametric regression with errors in variables. *Ann. Inst. Henri Poincaré, Probab. Stat.*, 44(3):393–421.

Carroll, R., Ruppert, D., and Stefanski, L. (1995). Measurement error in nonlinear models. number 63 in monographs on statistics and applied probability.

Cheng, C.-L. and Van Ness, J. W. (1999). *Statistical regression with measurement error*. John Wiley & Sons.

Fan, J., Truong, Y. K., and Wang, Y. (1991). *Nonparametric function estimation involving errors-in-variables*. Springer.

Fuller, W. A. (2009). *Measurement error models*, volume 305. John Wiley & Sons.

Juditsky, A. B., Lepski, O. V., and Tsybakov, A. B. (2009). Nonparametric estimation of composite functions. *The Annals of Statistics*, pages 1360–1404.

Schneeweiß, H. and Mittag, H.-J. (1986). *Lineare Modelle mit fehlerbehafteten Daten*. Physica-Verlag Heidelberg-Wien.

Spokoiny, V. (2012). Parametric estimation. Finite sample theory. *Ann. Statist.*, 40(6):2877–2909. arXiv:1111.3029.

Tsybakov, A. (2009). *Introduction to Nonparametric estimation*. Springer New York.

Wansbeek, T. J. and Meijer, E. (2000). *Measurement error and latent variables in econometrics*, volume 37. Elsevier Amsterdam.