Simple Adaptive Estimation of Quadratic Functionals in Nonparametric IV Models*

Christoph Breunig† Xiaohong Chen‡

First version: August 2018; Revised version: January 2022

This paper considers adaptive, minimax estimation of a quadratic functional in a nonparametric instrumental variables (NPIV) model, which is an important problem in optimal estimation of a nonlinear functional of an ill-posed inverse regression with an unknown operator. We first show that a leave-one-out, sieve NPIV estimator of the quadratic functional can attain a convergence rate that coincides with the lower bound previously derived in Chen and Christensen [2018]. The minimax rate is achieved by the optimal choice of the sieve dimension (a key tuning parameter) that depends on the smoothness of the NPIV function and the degree of ill-posedness, both are unknown in practice. We next propose a Lepski-type data-driven choice of the key sieve dimension adaptive to the unknown NPIV model features. The adaptive estimator of the quadratic functional is shown to attain the minimax optimal rate in the severely ill-posed case and in the regular mildly ill-posed case, but up to a multiplicative $\sqrt{\log n}$ factor in the irregular mildly ill-posed case.

Keywords: nonparametric instrumental variables, ill-posed inverse problem with an unknown operator, quadratic functional, minimax estimation, leave-one-out, adaptation, Lepski's method.

---

*We are grateful to Volodia Spokoiny for his friendship and insightful comments, and admire his creativity and high standard on research. We thank Enno Mammen and an anonymous referee for helpful comments, and Cristina Butucea, Tim Christensen, Richard Nickl and Sasha Tsybakov for initial discussions. Early versions have been presented at the conference "Celebrating Whitney Newey's Contributions to Econometrics" at MIT in May 2019, the North American Summer Meeting of the Econometric Society in Seattle in June 2019, and the conference "Foundations of Modern Statistics" in occasion of Volodia Spokoiny's 60th birthday at WIAS in November 2019.

†Department of Economics, Emory University, Rich Memorial Building, Atlanta, GA 30322, USA. Email: christoph.breunig@emory.edu

‡Cowles Foundation for Research in Economics, Yale University, Box 208281, New Haven, CT 06520, USA. E-mail: xiaohong.chen@yale.edu

1. Introduction

Long before the recent popularity of instrumental variables in modern machine learning causal inference, reinforcement learning and biostatistics, the instrumental variables technique has been widely used in economics. For instance, instrumental variables regressions are frequently used to account for omitted variables, mis-measured regressors, endogeneity in simultaneous equations and other complex situations in economic observational data. In economics and other social sciences, as well as in medical research, it is very difficult to estimate causal effects when treatment assignment is not randomized. Instrumental variables are commonly used to provide exogenous variation that is associated with the treatment status, but not with the outcome variable (beyond its direct effect on the treatments).

To avoid mis-specification of parametric functional forms, nonparametric instrumental variables (NPIV) regressions have gained popularity in econometrics and modern causal inference in statistics and machine learning. The simplest NPIV model assumes that a random sample $\{(Y_i, X_i, W_i)\}_{i=1}^n$ is drawn from an unknown joint distribution of (Y, X, W) satisfying

$$Y = h_0(X) + U, \quad \mathbb{E}[U|W] = 0, \tag{1.1}$$

where h_0 is an unknown continuous function, X is a d-dimensional vector of continuous endogenous regressors in the sense that $\mathbb{E}[U|X] \neq 0$, W is a vector of conditioning variables (instrumental variables) such that $\mathbb{E}[U|W] = 0$. The structural function h_0 can be identified as a solution to an integral equation of first kind with an unknown operator:

$$\mathbb{E}[Y|W = w] = (Th_0)(w) := \int h_0(x) f_{X|W}(x|w) dx,$$

where the conditional density $f_{X|W}$ (and hence the conditional expectation operator T) is unknown. Under mild conditions, the conditional density $f_{X|W}$ is continuous and the operator T smoothes out "low regular" (or wiggly) parts of h_0. This makes the nonparametric estimation (recovery) of h_0 a difficult ill-posed inverse problem with an unknown smoothing operator T. See, for example, Newey and Powell [2003], Hall and Horowitz [2005], Carrasco, Florens, and Renault [2007], Blundell, Chen, and Kristensen [2007], Chen and Reiß [2011] and Darolles et al. [2011]. For a given smoothness of h_0, the difficulty of recovering h_0 depends on the smoothing property of the conditional expectation operator T. The literature distinguishes between the mildly and severely ill-posed regimes, and the optimal convergence rates for nonparametrically estimating h_0 are different in the two regimes.

This paper considers adaptive, minimax rate-optimal estimation of a quadratic functional of h_0 in the NPIV model (1.1):

$$f(h_0) := \int h_0^2(x)\mu(x)dx \tag{1.2}$$

for a known positive, continuous weighting function μ, which is assumed to be uniformly bounded below from zero and from above on some subset of of the support of X. Let $\widehat{h}$ be a sieve NPIV estimator of the NPIV function h_0 (see e.g., Blundell et al. [2007]). Chen and Pouzo [2015] and Chen and Christensen [2018] considered inference on a slightly more general nonlinear functional $g(h_0)$ using plug-in sieve NPIV estimator $g(\widehat{h})$. However, there is no result on any adaptive, minimax rate-optimal estimation of any nonlinear functional $g(h_0)$ of the NPIV function h_0 yet. Since a quadratic functional is a leading example of a smooth nonlinear functional in h_0, Chen and Christensen [2018, Theorem C.1] established the minimax lower bound for estimating a quadratic functional $f(h_0)$ in a NPIV model. They also point out that a plug-in sieve NPIV estimator $f(\widehat{h})$ of the quadratic functional $f(h_0)$ can achieve the lower bound in the severely ill-posed regime, but fails to achieve the lower bound in the mildly ill-posed regime. Moreover, none of the existing work considers adaptive minimax rate-optimal estimation of the quadratic functional $f(h_0)$ in a NPIV model.

In this paper, we first propose a simple leave-one-out sieve NPIV estimator $\widehat{f}_J$ for the quadratic functional $f(h_0)$, and establish an upper bound on its convergence rate. By choosing the sieve dimension J optimally to balance the squared bias and the variance parts, we show that the resulting convergence rate of $\widehat{f}_J - f(h_0)$ coincides with the lower bound of Chen and Christensen [2018, Theorem C.1]. In this sense the estimator $\widehat{f}_J$ is minimax rate-optimal for $f(h_0)$ regardless whether the NPIV model is severely ill-posed or mildly ill-posed. In particular, for the severely ill-posed case, the optimal convergence rate is of the order $(\log n)^{-\alpha}$, where $\alpha > 0$ depends on the smoothness of the NPIV function h_0 and the degree of severe ill-posedness. For the mildly ill-posed case, the optimal convergence rate of $\widehat{f}_J - f(h_0)$ exhibits the so-called *elbow phenomena*: the rate is of the parametric order $n^{-1/2}$ for the regular mildly ill-posed case, and is of the order $n^{-\beta}$ for the irregular mildly ill-posed case, where $\beta \in (0, 1/2)$ depends on the smoothness of h_0, the dimension of X and the degree of mild ill-posedness.

The minimax optimal estimation rate of $\widehat{f}_J - f(h_0)$ is achieved by the optimal choice of the sieve dimension J (a key tuning parameter) that depends on the unknown smoothness of h_0 and the unknown degree of ill-posedness. We next propose a data driven choice $\widehat{J}$ of the sieve dimension based on a modified Lepski method.[1] The modification is needed to account for the estimation of the unknown degree of ill-posedness. The adaptive, leave-one-out sieve NPIV estimator $\widehat{f}_{\widehat{J}}$ of $f(h_0)$ is shown to attain the minimax optimal rate in the severely ill-posed case and in the regular mildly ill-posed case, but up to a multiplicative $\sqrt{\log n}$ in the irregular mildly ill-posed case. We note that even for adaptive estimation of a quadratic functional of a direct regression in a Gaussian white noise model, Efromovich and Low [1996] already shown that the extra $\sqrt{\log n}$ factor is the necessary price to pay for adaptation to the unknown smoothness of the regression function.

Previously for the nonparametric estimation of h_0 in the NPIV model (1.1), Horowitz [2014] considers adaptive estimation of h_0 in L^2 norm using a model selection procedure. Breunig and Johannes [2016] consider adaptive estimation of a linear functional of the NPIV function h_0 in a root-mean squared error metric using a combined model selection and Lepski method. These papers obtain adaptive rate of convergence up to a multiplicative factor of $\sqrt{\log(n)}$ (of the minimax optimal rate) in both severely ill-posed and mildly ill-posed cases. Chen, Christensen, and Kankanala [2021] propose adaptive estimation of h_0 in L^∞ norm using a modified Lepski method and tight random matrix inequalities to account for the estimated measure of ill-posedness. They show that their data-driven procedure attains the minimax optimal rate in L^∞ norm and is fully adaptive to the unknown smoothness of h_0 in both severely and mildly ill-posed regimes. Our data-driven choice of the sieve dimension is closest to that of Chen et al. [2021], which might explain why we also obtain minimax optimal adaptivity for the quadratic functional $f(h_0)$ in both severely and mildly ill-posed regimes.

While Horowitz [2014], Breunig and Johannes [2016] and Chen et al. [2021] use plug-in sieve NPIV estimators in their adaptive estimation of a linear functional of h_0, we use a leave-one-out sieve NPIV estimator $\widehat{f}_J$ for the quadratic functional $f(h_0) = \int h_0^2(x)\mu(x)dx$. Recently Breunig and Chen [2021] propose a test statistic that is based on a standardized leave-one-out estimator of a quadratic distance for a null hypothesis of $\mathbb{E}[(h_0(X) - h^R(X))^2\mu(X)] = 0$ in a NPIV model (for some parametric, semiparametric or shape restricted h^R). They construct an adaptive minimax test using a random exponential scan procedure. We use the unstandardized

[1] See Lepski [1990], Lepski and Spokoiny [1997] and Lepski, Mammen, and Spokoiny [1997] for detailed descriptions of the original Lepski principle.

leave-one-out estimator $\widehat{f}_J$ in our modified Lepski procedure for adaptive minimax estimation of $f(h_0)$ in a NPIV model. It is well-known that adaptive minimax testing and adaptive minimax estimation are related but different (see, e.g., Giné and Nickl [2021]). In particular, while both papers apply a tight Bernstein-type inequality for U-statistics (Houdré and Reynaud-Bouret [2003]) in the proofs, the adaptive optimal rates are different. For instance, the adaptive minimax L^2 separation rate of testing in Breunig and Chen [2021] is always slower than $n^{-1/2}$, while our adaptive minimax estimation for $f(h_0)$ can achieve the parametric rate of $n^{-1/2}$ for regular mildly ill-posed NPIV models.

Minimax rate-optimal estimation of a quadratic functional in density and direct regression (in Gaussian white noise) settings has a long history in statistics. See, for example, Bickel and Ritov [1988], Donoho and Nussbaum [1990], Fan [1991], Efromovich and Low [1996], Laurent and Massart [2000], Cai and Low [2006], Giné and Nickl [2008], Collier, Comminges, and Tsybakov [2017] and the references therein. To the best of our knowledge, there are not many published papers on minimax estimation of a quadratic functional in difficult inverse problems. See Butucea [2007], Butucea and Meziani [2011], Chesneau [2011] and Kroll [2019] for deconvolutions and inverse regressions in Gaussian sequence models. Moreover, Chesneau [2011] seems the only published work on adaptive estimation of a quadratic functional in a special deconvolution (with a known operator). Our paper is the first to propose a simple estimator that is adaptive minimax rate-optimal for a quadratic functional in a NPIV model, and also contributes to inverse problems with unknown operators.

The rest of the paper is organized as follows. Section 2 presents the leave-one-out sieve NPIV estimator of the quadratic functional $f(h_0)$, and derives its optimal convergence rates. Section 3 first presents a simple data-driven procedure of choosing the sieve dimension using a modified Lepski method. It then establishes the optimal convergence rates of our adaptive estimator of the quadratic functional. Section 4 provides a brief conclusion and discusses several extensions. All proofs can be found in the Appendices A–C.

2. Minimax Optimal Quadratic Functional Estimation

This section consists of three parts. The first subsection introduces model preliminaries and notation. Subsection 2.2 introduces a simple leave-one-out, sieve NPIV estimator of the quadratic functional $f(h_0)$. Subsection 2.3 establishes the convergence rate of the proposed estimator, and shows that the convergence rate coincides with the lower bound and hence is optimal.

2.1. Preliminaries and Notation

We first introduce notation that is used throughout the paper. For any random vector V with support $\mathcal{V}$, we let $L^2(V) = \{\phi : \mathcal{V} \to \mathbb{R}, \|\phi\|_{L^2(V)} < \infty\}$ with the norm $\|\phi\|_{L^2(V)} = \sqrt{\mathbb{E}[\phi^2(V)]}$. If $\{a_n\}$ and $\{b_n\}$ are sequences of positive numbers, we use the notation $a_n \lesssim b_n$ if $\limsup_{n\to\infty} a_n/b_n < \infty$ and $a_n \sim b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

We consider a known positive, continuous weighting function μ, which is assumed to be uniformly bounded below from zero and from above on some subset of $\mathcal{X}$, denoted by X_μ. Denote $L^2_\mu = \{h : \mathcal{X}_\mu \to \mathbb{R}, \|h\|_\mu < \infty\}$ with the norm $\|h\|_\mu = \sqrt{\int h^2(x)\mu(x)dx}$. We consider basis functions $\{\psi_j\}_{j\geq 1}$ to approximate the NPIV function h_0. Its orthonormalized analog with respect to $\|\cdot\|_\mu$ is denoted by $\{\widetilde{\psi}_j\}_{j\geq 1}$. We assume that the structural function h_0 belongs to the Sobolev ellipsoid

$$\mathcal{H}_2(p, L) = \Big\{h \in L^2_\mu : \sum_{j=1}^{\infty} j^{2p/d}\langle h, \widetilde{\psi}_j\rangle^2_\mu \leq L\Big\}, \quad \text{for } \; d/2 < p < \infty, \quad 0 < L < \infty \; .$$

Let $T : L^2(X) \mapsto L^2(W)$ denote the conditional expectation operator given by $(Th)(w) = \mathbb{E}[h(X)|W = w]$. Finally let $\{\psi_1, ..., \psi_J\}$ and $\{b_1, ..., b_K\}$ be collections of sieve basis functions of dimension J and K for approximating functions in $L^2(X)$ and $L^2(W)$, respectively. We define the *sieve measure of ill-posedness* which, roughly speaking, measures how much the conditional expectation operator T smoothes out h. Following Blundell et al. [2007] the sieve L^2_μ measure of ill-posedness is

$$\tau_J = \sup_{h\in\Psi_J, h\neq 0} \frac{\|h\|_\mu}{\|Th\|_{L^2(W)}} = \sup_{h\in\Psi_J, h\neq 0} \frac{\sqrt{f(h)}}{\|Th\|_{L^2(W)}},$$

where $\Psi_J = \text{clsp}\{\psi_1, ..., \psi_J\} \subset L^2(X)$ denotes the sieve spaces for the endogenous variables. We call a NPIV model (1.1)

(i) mildly ill-posed if $\tau_j \sim j^{a/d}$ for some $a > 0$; and

(ii) severely ill-posed if $\tau_j \sim \exp(\frac{1}{2} j^{a/d})$ for some $a > 0$.

2.2. A Leave-one-out, Sieve NPIV Estimator

Let $\{(Y_i, X_i, W_i)\}_{i=1}^n$ denote a random sample from the NPIV model (1.1). The sieve NPIV (or series 2SLS) estimator $\widehat{h}$ of h_0 can be written in matrix form as follows (see, e.g., Chen and Christensen [2018])

$$\widehat{h}(\cdot) = \psi^J(\cdot)'[\Psi' P_B \Psi]^- \Psi' P_B \mathbf{Y} = \psi^J(\cdot)' \widehat{A} B' \mathbf{Y}/n$$

where $P_B = B(B'B)^- B'$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)'$,

$$\begin{aligned} \psi^J(x) &= (\psi_1(x), \ldots, \psi_J(x))' \qquad & \Psi &= (\psi^J(X_1), \ldots, \psi^J(X_n))' \\ b^K(w) &= (b_1(w), \ldots, b_K(w))' \qquad & B &= (b^K(W_1), \ldots, b^K(W_n))' \end{aligned}$$

and $\widehat{A} = n[\Psi' P_B \Psi]^- \Psi' B(B'B)^-$ is an estimator of $A = [S' G_b^{-1} S]^{-1} S' G_b^{-1}$, with $S = \mathbb{E}[b^K(W_i)\psi^J(X_i)']$ and $G_b = \mathbb{E}[b^K(W_i) b^K(W_i)']$.

As pointed out by Chen and Christensen [2018], although one could estimate $f(h_0)$ by the plug-in sieve NPIV estimator $f(\widehat{h})$, it fails to achieve the minimax lower bound. We propose a leave-one-out sieve NPIV estimator for the quadratic functional $f(h_0)$ as follows:

$$\widehat{f}_J = \frac{2}{n(n-1)} \sum_{1 \le i < i' \le n} Y_i b^K(W_i)' \widehat{A}' G_\mu \widehat{A}\, b^K(W_{i'}) Y_{i'}$$

where $G_\mu = \int \psi^J(x) \psi^J(x)' \mu(x) dx$. We will show that this simple leave-one-out estimator $\widehat{f}_J$ can achieve the lower bound for estimating $f(h_0)$.

Based on many simulation results in Blundell et al. [2007] and Chen and Christensen [2018], the crucial *regularization* parameter in sieve NPIV estimation of h_0 is the dimension J of the sieve space used to approximate unknown function h_0. In this paper, we simply let $K(J) = c_K J$ for some constant $c_K \ge 1$. Further, we let $\zeta_{\psi,J} = \sup_x \|G_\mu^{-1/2} \psi^J(x)\|$ and $\zeta_{b,K} = \sup_w \|G_b^{-1/2} b^K(w)\|$. For instance, $\zeta_{\psi,J} = O(\sqrt{J})$ and $\zeta_{b,K} = O(\sqrt{K})$ for (tensor-product) polynomial spline, wavelet and cosine bases. Denote $\zeta_J = \max(\zeta_{\psi,J}, \zeta_{b,K})$ for $K = K(J)$. In the rest of the paper we restrict sieve bases to the ones such that $\zeta_J = O(\sqrt{J})$.

2.3. Rate of Convergence

We first introduce assumptions that are used to derive our rate of convergence of the estimator $\widehat{f}_J$. We denote the sieve Least Squares (LS) projection of h onto $\Psi_J = \text{clsp}\{\psi_1, ..., \psi_J\}$ as $\Pi_J h(x) = \psi^J(x)' G_\mu^{-1} \langle \psi^J, h \rangle_\mu$. For $h_0 \in \mathcal{H}_2(p, L)$ we have $\|h_0 - \Pi_J h_0\|_\mu \le L J^{-p/d}$ which is used throughout this paper. This implies that $\sqrt{J(\log J)} \|h_0 - \Pi_J h_0\|_\mu = o(1)$ as J goes to infinity (since $p > d/2$).

Assumption 1. *(i) $T[h - h_0] = 0$ for any $h \in L^2_\mu$ implies that $f(h) = f(h_0)$; (ii) $\sup_{w \in \mathcal{W}} \mathbb{E}[Y^2|W = w] \leq \overline{\sigma}^2_Y < \infty$ and $\mathbb{E}[Y^4] < \infty$; (iii) the densities of X and W are Lebesgue continuous and uniformly bounded below from zero and from above on the closed rectangular supports $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{W} \subset \mathbb{R}^{d_w}$, respectively.*

Assumption 2. $\tau_J J \sqrt{(\log J)/n} = O(1)$.

Below we let $\Pi_K g(w) = b^K(w)' G_b^{-1} \mathbb{E}[b^K(W) g(W)]$ denote the sieve LS projection of $g \in L^2(W)$ onto $B_K = \text{clsp}\{b_1, ..., b_K\}$.

Assumption 3. *(i) $\sup_{h \in \Psi_J} \tau_J \|(\Pi_K T - T)h\|_{L^2(W)} / \|h\|_\mu \leq v_J$ where $v_J < 1$ for all J and $v_J \to 0$ as $J \to \infty$. (ii) there exists a constant $C > 0$ such that $\tau_J \|T(h_0 - \Pi_J h_0)\|_{L^2(W)} \leq C \|h_0 - \Pi_J h_0\|_\mu$.*

For a $r \times c$ matrix M with $r \leq c$ and full row rank r we let M_l^- denote its left pseudoinverse, namely $(M'M)^- M'$ where $'$ denotes transpose and $^-$ denotes generalized inverse. Below, $\|\cdot\|$ respectively denotes the vector ℓ_2 norm when applied to a vector and the operator norm $\|A\| := \sup_{x : \|x\|=1} \|Ax\|$ when applied to a matrix A. Let $(s_1, \ldots, s_J)$ denote the singular values, in non-increasing order, of $G_b^{-1/2} S G_\mu^{-1/2}$. In particular $s_J = s_{\min}(G_b^{-1/2} S G_\mu^{-1/2})$.

Assumption 4. $\left\| diag(s_1, \ldots, s_J) \left(G_b^{-1/2} S G_\mu^{-1/2}\right)_l^- \right\| \leq D$ *for some constant $D > 0$.*

Discussion of Assumptions: Assumption 1(i) ensures identification of the nonlinear functional $f(h_0)$. Assumption 2 restricts the growth of the sieve dimension J. Assumption 3(i) is a mild condition on the approximation properties of the basis used for the instrument space and is first imposed in Chen et al. [2021]. In fact, $\|(\Pi_K T - T)h\|_{L^2(W)} = 0$ for all $h \in \Psi_J$ when the basis functions for B_K (with $K \geq J$) and Ψ_J form either a Riesz basis or an eigenfunction basis for the conditional expectation operator. Assumption 3(ii) is the usual L^2 "stability condition" imposed in the NPIV literature (cf. Assumption 6 in Blundell et al. [2007]). Note that Assumption 3(ii) is also automatically satisfied by Riesz bases. Assumption 4 is a modification of the sieve measure of ill-posedness and was used by Efromovich and Koltchinskii [2001]. Assumption 4 is also related to the extended link condition in Breunig and Johannes [2016] to establish optimal upper bounds in the context of minimax optimal estimation of linear functionals in NPIV models. Finally we note that by definition, s_J satisfies

$$s_J = \inf_{h \in \Psi_J, h \neq 0} \frac{\|\Pi_K T h\|_{L^2(W)}}{\|h\|_\mu} \leq \tau_J^{-1} \tag{2.1}$$

for all $K = K(J) \geq J > 0$. Assumption 3(i) further implies that

$$s_J \geq \inf_{h \in \Psi_J, h \neq 0} \frac{\|Th\|_{L^2(W)}}{\|h\|_\mu} - \sup_{h \in \Psi_J, h \neq 0} \frac{\|(\Pi_K T - T)h\|_{L^2(W)}}{\|h\|_\mu} = c_\tau \tau_J^{-1}, \tag{2.2}$$

for some constant $c_\tau > 0$. We shall maintain Assumption 3(i) and use the equivalence of s_J and τ_J^{-1} in the paper.

The next result provides an upper bound on the rate of convergence for the estimator $\widehat{f}_J$.

Theorem 2.1. *Let Assumptions 1–3 hold. Then:*

$$\widehat{f}_J - f(h_0) = O_p\left(\frac{\tau_J^2 \sqrt{J}}{n} + \frac{\left\| \langle h_0, \psi^J \rangle_\mu' (G_b^{-1/2} S)_l^- \right\| + \tau_J \|h_0 - \Pi_J h_0\|_\mu}{\sqrt{n}} - \|h_0 - \Pi_J h_0\|_\mu^2 \right). \tag{2.3}$$

If in addition $h_0 \in \mathcal{H}_2(p, L)$ and Assumption 4 holds, then:

1. *Mildly ill-posed case: choosing $J \sim n^{2d/(4(p+a)+d)}$ implies*

$$\widehat{f}_J - f(h_0) = \begin{cases} O_p\left(n^{-4p/(4(p+a)+d)}\right), & \text{if } p \leq a + d/4, \\ O_p\left(n^{-1/2}\right), & \text{if } p > a + d/4. \end{cases} \tag{2.4}$$

2. *Severely ill-posed case: choosing*

$$J \sim \left(\log n - \frac{4p+d}{2a} \log \log n \right)^{d/a}$$

implies

$$\widehat{f}_J - f(h_0) = O_p\left((\log n)^{-2p/a} \right). \tag{2.5}$$

Theorem 2.1 presents an upper bound on the convergence rates of $\widehat{f}_J$ to $f(h_0)$. When the sieve dimension J is chosen optimally, the convergence rate (2.4) coincides with the minimax lower bound in Chen and Christensen [2018, Theorem C.1] for the mildly ill-posed case, while the convergence rate (2.5) coincides with the minimax lower bound in Chen and Christensen [2018, Theorem C.1] for the severely ill-posed case. Moreover, within the mildly ill-posed case, depending on the smoothness of h_0 relatively to the dimension of X and the degree of mildly ill-posedness a, either the first or the second variance term in (2.3) dominates, which leads to the so-called elbow phenomenon: the regular case with a parametric rate of $n^{-1/2}$ when $p > a + d/4$; and the irregular case with a nonparametric rate when $p \leq a+d/4$. In particular, Theorem 2.1 shows that the simple leave-one-out estimator $\widehat{f}_J$ is minimax rate optimal provided that the sieve dimension J is chosen optimally.

Chen and Christensen [2018, Theorem C.1] actually established lower bound for estimating a quadratic functional of a derivative of h_0 in a NPIV model as well. Using Fourier, spline and wavelet bases, we can easily show that our simple leave-one-out, sieve NPIV estimator of the quadratic functional of a derivative of h_0 also achieve the lower bound, and hence is minimax rate-optimal. We do not present such a result here since it is a very minor extension of Theorem 2.1.

3. Rate Adaptive Estimation

The minimax rate of convergence depends on the optimal choice of sieve dimension J, which depends on the unknown smoothness p of the true NPIV function h_0 and the unknown degree of ill-posedness. In this section we propose a data-driven choice of the sieve dimension J based on a modified Lepski method; see Lepski [1990], Lepski and Spokoiny [1997] and Lepski et al. [1997] for early development of this popular method.

In this section we follow Chen et al. [2021] and let Ψ_J be a tensor-product Cohen-Daubechies-Vial (CDV) wavelet (see, e.g., chapter 4.3.5 of Giné and Nickl [2021]) or dyadic B-spline sieve (see, e.g., Appendix A.1 of Chen et al. [2021]) for $\mathcal{H}_2(p, L)$. Let $\mathcal{T}$ denote the set of possible sieve dimensions J. For example for (order r) B-splines, $\mathcal{T} = \{J = (2^l + r - 1)^d : l \in \mathbb{N} \cup \{0\}\}$. Since $\widehat{f}_J$ is based on a sieve NPIV estimator, we can simply use a random index set $\widehat{\mathcal{I}}$ that is proposed in Chen et al. [2021] for their sup-norm rate adaptive sieve NPIV estimation of h_0:

$$\widehat{\mathcal{I}} = \{J \in \mathcal{T} : 0.1(\log \widehat{J}_{\max})^2 \leq J \leq \widehat{J}_{\max}\},$$

where

$$\widehat{J}_{\max} = \min\left\{J \in \mathcal{T} : \widehat{s}_J^{-1} J\sqrt{\log J} \leq 10\sqrt{n} < \widehat{s}_{J^+}^{-1} J^+ \sqrt{\log J^+}\right\}, \tag{3.1}$$

$\widehat{s}_J$ is the smallest singular value of $(B'B/n)^{-1/2}(B'\Psi/n)G_\mu^{-1/2}$, and $J^+ = \min\{j \in \mathcal{T} : j > J\}$.

We define our data driven choice $\widehat{J}$ of "optimal" sieve dimension for estimating

$f(h_0)$ as follows:

$$\widehat{J} = \min\left\{J \in \widehat{\mathcal{I}} : |\widehat{f}_J - \widehat{f}_{J'}| \le c_0(\widehat{V}(J) + \widehat{V}(J')) \text{ for all } J' \in \widehat{\mathcal{I}} \text{ with } J' > J\right\} \quad (3.2)$$

for some constant $c_0 > 0$ and

$$\widehat{V}(J) = \frac{\sqrt{J(\log n)}}{n\,\widehat{s}_J^2} \vee \frac{1}{\sqrt{n}}, \quad (3.3)$$

where $a \vee b := \max\{a, b\}$. The random index set $\widehat{\mathcal{I}}$ is used to compute our data driven choice (3.2) since the unknown measure of ill-posedness τ_J is estimated by $\widehat{s}_J^{-1}$.

We introduce a non-random index set $\mathcal{I} = \{J \in \mathcal{T} : J \le \overline{J}\}$, where $\overline{J} = \sup\left\{J \in \mathcal{T} : \tau_J J\sqrt{(\log J)/n} \le \bar{c}\right\}$ for some sufficiently large constant $\bar{c} > 0$. Let $\mathcal{B} = \{h \in L^2_\mu : \|h\|_\infty \le L\}$ and $\overline{p} > \underline{p} \ge 3d/4$. The following assumption strengthens some conditions imposed in the previous section.

Assumption 5. *(i)* $\sup_{h\in\mathcal{H}_2(p,L)\cap\mathcal{B}} \|h - \Pi_J h\|_\mu \le cJ^{-p/d}$ *for some finite constant* $c > 0$ *for all* $p \in [\underline{p}, \overline{p}]$*, with* Ψ_J *being CDV wavelet or dyadic B-spline basis; (ii)* $\sup_{w\in\mathcal{W}} \mathbb{E}[Y^4|W = w] \le \overline{\sigma}_Y^4 < \infty$*; (iii) Assumptions 3(ii) and 4 hold for all* $J \in \mathcal{I}$*.*

The next result establishes an upper bound for the adaptive estimator $\widehat{f}_{\widehat{J}}$.

Theorem 3.1. *Let Assumptions 1(i)(iii), 3(i), and 5 hold. Then, we have in the*

1. *mildly ill-posed case:*

$$\sup_{p\in[\underline{p},\overline{p}]} \sup_{h_0\in\mathcal{H}_2(p,L)\cap\mathcal{B}} \mathbb{P}_{h_0}\left(\left|\widehat{f}_{\widehat{J}} - f(h_0)\right| > C_1 r_n\right) = o(1) \quad (3.4)$$

   for some constant $C_1 > 0$ *and where*

$$r_n = \begin{cases} \left(\sqrt{\log n}/n\right)^{4p/(4(p+a)+d)}, & \text{if } p \le a + d/4, \\ n^{-1/2}, & \text{if } p > a + d/4. \end{cases}$$

2. *severely ill-posed case:*

$$\sup_{p\in[\underline{p},\overline{p}]} \sup_{h_0\in\mathcal{H}_2(p,L)\cap\mathcal{B}} \mathbb{P}_{h_0}\left(\left|\widehat{f}_{\widehat{J}} - f(h_0)\right| > C_2(\log n)^{-2p/a}\right) = o(1) \quad (3.5)$$

   for some constant $C_2 > 0$.

Theorem 3.1 shows that our data-driven choice of the key sieve dimension can lead to fully adaptive rate-optimal estimation of $f(h_0)$ for both the severely ill-posed case and the regular mildly ill-posed case, while it has to pay a price of an extra $\sqrt{\log n}$ factor for the irregular mildly ill-posed case (i.e., when $p \leq a + d/4$). We note that when $a = 0$ in the mildly ill-posed case, the NPIV model (1.1) becomes the regression model with $X = W$. Thus our result is in agreement with the theory in Efromovich and Low [1996], which showed that one must pay a factor of $\sqrt{\log n}$ penalty in adaptive estimation of a quadratic functional in a Gaussian white noise model when $p \leq d/4$.

In adaptive estimation of a nonparametric regression function $\mathbb{E}[Y|X = \cdot] = h(\cdot)$, it is known that Lepski method has the tendency of choosing small sieve dimension, and hence may not perform well in empirical work. We wish to point out that due to the ill-posedness of the NPIV model (1.1), the optimal sieve dimension for estimating $f(h_0)$ is smaller than the optimal sieve dimension for estimating $f(\mathbb{E}[Y|X = \cdot])$. Therefore, we suspect that our simple adaptive estimator of a quadratic functional of a NPIV function will perform well in finite samples.

4. Conclusion and Extensions

In this paper we first show that a simple leave-one-out sieve NPIV estimator of the quadratic functional $f(h_0)$ is minimax rate optimal. We then propose an adaptive leave-one-out sieve NPIV estimator of the $f(h_0)$ based on a modified Lepski method to account for the unknown degree of ill-posedness. We show that the adaptive estimator achieves the minimax optimal rate for the severely ill-posed case and for the regular mildly ill-posed case, while a multiplicative $\sqrt{\log n}$ term is the price to pay for the irregular mildly ill-posed NPIV problem.

Like all existing work using Lepski method, implementation of our data-driven choice relies on a calibration constant. To improve finite sample performance over the original Lepski method, Spokoiny and Vial [2009] suggest a propagation approach, Chernozhukov, Chetverikov, and Kato [2014] and Spokoiny and Willrich [2019] propose bootstrap calibrations in kernel density estimation and in linear regressions with Gaussian errors respectively. Chen et al. [2021] propose a bootstrap implementation of a modified Lepski method in their minimax adaptive sup-norm estimation in a NPIV model, and show its good performance in finite samples. Their bootstrap implementation can be easily extended to calibrate the constant in our adaptive estimation of the quadratic functional in a NPIV model. We leave this to future refinement.

Our results can be extended in several directions. First, we can relax the Sobolev ball assumption imposed on h_0 in the NPIV model. We can let the NPIV function

h_0 belong to a bump algebra space. The result by Collier et al. [2017] on minimax estimation of a quadratic functional under sparsity constraints can be useful for this extension. Second, we focus on adaptive estimation of a quadratic functional of the NPIV function h_0 in this paper. There are works on minimax-rate estimation and adaptive estimation for more general smooth nonlinear functionals of densities and of nonparametric regressions; see, e.g., Birgé and Massart [1995], Liu, Mukherjee, Robins, and Tchetgen [2021] and the references therein. We can combine our approach here with those in the literature for extensions to other smooth nonlinear functionals of the NPIV function h_0. Such an extension will allow for adaptive minimax estimation of nonlinear policy functionals in economics and modern causal inference.

A. Proofs of Results in Section 2

Recall the 2SLS projection of h onto Ψ_J is given by:

$$Q_J h(x) = \psi^J(x)'[S'G_b^{-1}S]^{-1}S'G_b^{-1}\,\mathbb{E}[b^{K(J)}(W)h(X)] = \psi^J(x)'A\,\mathbb{E}[b^{K(J)}(W)h(X)].$$

For a $r \times c$ matrix M with $r \leq c$ and full row rank r we let M_l^- denote its left pseudoinverse, namely $(M'M)^-M'$. Let $\widetilde{\psi}^J = G_\mu^{-1/2}\psi^J$ and $\widetilde{b}^K = G_b^{-1/2}b^K$. Thus, we have $AG_b^{1/2} = (G_b^{-1/2}S)_l^-$ and

$$G_\mu^{1/2}AG_b^{1/2} = (G_b^{-1/2}SG_\mu^{-1/2})_l^-.$$

In particular, we can write

$$\begin{aligned} Q_J h(x) &= \psi^J(x)'(G_b^{-1/2}S)_l^-\,\mathbb{E}[\widetilde{b}^{K(J)}(W)h(X)] \\ &= \widetilde{\psi}^J(x)'(G_b^{-1/2}SG_\mu^{-1/2})_l^-\,\mathbb{E}[\widetilde{b}^{K(J)}(W)h(X)]. \end{aligned}$$

The minimal or maximal eigenvalue of a quadratic matrix M is denoted by $\lambda_{\min}(M)$ or $\lambda_{\max}(M)$. Recall that

$$\widehat{f}_J = \frac{1}{n(n-1)}\sum_{i\neq i'} Y_iY_{i'}b^K(W_i)'\widehat{A}'G_\mu\widehat{A}\,b^K(W_{i'}).$$

Proof of Theorem 2.1. Proof of Result (2.3). Note that

$$f(Q_J h_0) = \int \Big(\psi^J(x)'(G_b^{-1/2}S)_l^- \,\mathbb{E}[\widetilde{b}^K(W)h_0(X)]\Big)^2 \mu(x)dx$$
$$= \big\|G_\mu^{1/2}(G_b^{-1/2}S)_l^- \,\mathbb{E}[\widetilde{b}^K(W)h_0(X)]\big\|^2 = \|\,\mathbb{E}[V^J]\|^2$$

using the notation $V_i^J = Y_i G_\mu^{1/2} A b^K(W_i)$. Thus, the definition of the estimator $\widehat{f}_J$ implies

$$\widehat{f}_J - f(Q_J h_0) = \frac{1}{n(n-1)} \sum_{j=1}^J \sum_{i\neq i'} \big(V_{ij}V_{i'j} - \mathbb{E}[V_{1j}]^2\big) \tag{A.1}$$
$$+ \frac{1}{n(n-1)} \sum_{i\neq i'} Y_i Y_{i'} b^K(W_i)'\Big(A'G_\mu A - \widehat{A}'G_\mu \widehat{A}\Big) b^K(W_{i'}), \tag{A.2}$$

where we bound both summands on the right hand side separately in the following. Consider the summand in (A.1), we observe

$$\mathbb{E}\Big|\sum_{j=1}^J \sum_{i\neq i'} \big(V_{ij}V_{i'j} - \mathbb{E}[V_{1j}]^2\big)\Big|^2$$
$$= 2n(n-1)(n-2) \underbrace{\sum_{j,j'=1}^J \mathbb{E}\Big[\big(V_{1j}V_{2j} - \mathbb{E}[V_{1j}]^2\big)\big(V_{3j'}V_{2j'} - \mathbb{E}[V_{1j'}]^2\big)\Big]}_{I}$$
$$+ n(n-1) \underbrace{\sum_{j,j'=1}^J \mathbb{E}\Big[\big(V_{1j}V_{2j} - \mathbb{E}[V_{1j}]^2\big)\big(V_{1j'}V_{2j'} - \mathbb{E}[V_{1j'}]^2\big)\Big]}_{II}.$$

By Assumption 1(ii) it holds $\sup_{w\in\mathcal{W}} \mathbb{E}[Y^2|W=w] \leq \overline{\sigma}_Y^2$, which together with Breunig and Chen [2021, Lemma E.7] implies $\lambda_{\max}\big(\mathbb{V}ar(Y\widetilde{b}^K(W)\big) \leq \overline{\sigma}_Y^2$. To bound the summand I we observe that

$$I = \sum_{j,j'=1}^J \mathbb{E}[V_{1j}]\,\mathbb{E}[V_{1j'}]\,\mathbb{C}ov(V_{1j}, V_{1j'}) = \mathbb{E}[V_1^J]'\,\mathbb{C}ov(V_1^J, V_1^J)\,\mathbb{E}[V_1^J]$$
$$\leq \lambda_{\max}\big(\mathbb{V}ar(Y\widetilde{b}^K(W))\big)\big\|(G_b^{-1/2}SG_\mu^{-1/2})_l^-\,\mathbb{E}[V_1^J]\big\|^2$$
$$= \overline{\sigma}_Y^2\Big\|\langle Q_J h_0, \psi^J\rangle_\mu'(G_b^{-1/2}S)_l^-\Big\|^2$$

by using the notation $V_i^J = Y_i(G_b^{-1/2} S G_\mu^{-1/2})_l^- \widetilde{b}^K(W_i)$. Consider II. We observe

$$
\begin{aligned}
II &= n(n-1)\sum_{j,j'=1}^{J}\mathbb{E}[V_{1j}V_{1j'}]^2 - n(n-1)\Big(\sum_{j=1}^{J}\mathbb{E}[V_{1j}]^2\Big)^2 \\
&\leq n(n-1)\sum_{j,j'=1}^{J}\mathbb{E}[V_{1j}V_{1j'}]^2 \leq 2\overline{\sigma}_Y^2 n(n-1)s_J^{-4}J
\end{aligned}
$$

where the last inequality stems from Breunig and Chen [2021, Lemma E.1] together with $\sup_{w\in\mathcal{W}}\mathbb{E}[Y^2|W=w] \leq \overline{\sigma}_Y^2$. Consequently, we obtain

$$
\mathbb{E}\Big|\frac{1}{n(n-1)}\sum_{j=1}^{J}\sum_{i\neq i'}\big(V_{ij}V_{i'j} - \mathbb{E}[V_{1j}]^2\big)\Big|^2 \leq 4\overline{\sigma}_Y^4\left(\frac{1}{n}\big\|\langle Q_J h_0, \psi^J\rangle_\mu'(G_b^{-1/2}S)_l^-\big\|^2 + \frac{J}{n^2 s_J^4}\right). \tag{A.3}
$$

The second summand in (A.2) can be bounded following the same proof as that of Breunig and Chen [2021, Lemma E.4] (replacing their $(Y_i - h_0(X_i))$ with our Y_i and our Assumption 1(ii)), which yields

$$
\widehat{f}_J - f(Q_J h_0) = O_p\left(\frac{1}{\sqrt{n}}\big\|\langle Q_J h_0, \psi^J\rangle_\mu'(G_b^{-1/2}S)_l^-\big\| + \frac{\sqrt{J}}{n s_J^2}\right).
$$

Next, by the definition of Q_J we have: $\langle Q_J h_0, \widetilde{\psi}^J\rangle_\mu = (G_b^{-1/2} S G_\mu^{-1/2})_l^-\,\mathbb{E}[\widetilde{b}^{K(J)}(W)h_0(X)]$. Thus, we have

$$
\begin{aligned}
\big\|\langle Q_J h_0, \psi^J\rangle_\mu'(G_b^{-1/2}S)_l^-\big\| \leq{}& \big\|\langle h_0, \psi^J\rangle_\mu'(G_b^{-1/2}S)_l^-\big\| \\
&+ s_J^{-2}\big\|\,\mathbb{E}[\widetilde{b}^{K(J)}(W)(h_0(X) - \Pi_J h_0(X))]\big\|,
\end{aligned}
$$

By inequality (2.2) and Assumption 3(ii), we have

$$
\begin{aligned}
s_J^{-2}\big\|\,\mathbb{E}[\widetilde{b}^{K(J)}(W)(h_0(X) - \Pi_J h_0(X))]\big\| &= O\Big(\tau_J^2\|\Pi_K T(h_0 - \Pi_J h_0)\|_{L^2(W)}\Big) \\
&= O\big(\tau_J\|h_0 - \Pi_J h_0\|_\mu\big).
\end{aligned}
$$

It remains to evaluate

$$
f(Q_J h_0) - f(h_0) = \|Q_J h_0\|_\mu^2 - \Big[\|\Pi_J h_0\|_\mu^2 + \|h_0 - \Pi_J h_0\|_\mu^2\Big].
$$

Consider the first summand on the right hand side. There exist unitary matrices M_1,

M_2 with $\dot{b}^K := M_1\widetilde{b}^K$ and $\dot{\psi}^J := M_2\widetilde{\psi}^J$ such that $\mathbb{E}[\dot{b}^{K(J)}(W)\dot{\psi}^J(X)']$ has an upper $J \times J$ matrix $diag(s_1, \ldots, s_J)$ and is zero otherwise. We thus derive

$$\begin{aligned}\|Q_J h_0\|_\mu^2 &= \left\|(G_b^{-1/2} S G_\mu^{-1/2})_l^- \,\mathbb{E}[\widetilde{b}^{K(J)}(W)h_0(X)]\right\|^2 \\ &= \sum_{j=1}^J s_j^{-2}\,\mathbb{E}[\dot{b}_j(W)h_0(X)]^2 = \sum_{j=1}^J \langle h_0, \dot{\psi}_j\rangle_\mu^2 = \|\Pi_J h_0\|_\mu^2,\end{aligned}$$

and hence $f(Q_J h_0) - f(h_0) = -\|h_0 - \Pi_J h_0\|_\mu^2$. This completes the proof of Result (2.3).

For the proofs of Results (2.4) and (2.5), we note that $h_0 \in \mathcal{H}_2(p, L)$ implies

$$\|h_0 - \Pi_J h_0\|_\mu \le L J^{-p/d}.$$

Moreover, by inequality (2.2) and Assumption 4 we have:

$$\left\|\langle h_0, \psi^J\rangle_\mu'(G_b^{-1/2}S)_l^-\right\| = \left\|\langle h_0, \widetilde{\psi}^J\rangle_\mu'(G_b^{-1/2}S'G_\mu^{-1/2})_l^-\right\| \le D c_\tau^{-1}\sqrt{\sum_{j=1}^J \tau_j^2 \langle h_0, \widetilde{\psi}_j\rangle_\mu^2}.$$

These bounds are used below to derive the concrete rates of convergence in the mildly and severely ill-posed regimes.

Proof of Result (2.4) for the mildly ill-posed case. The choice of $J \sim n^{2d/(4(p+a)+d)}$ implies

$$n^{-2}\tau_J^4 J \sim n^{-2}J^{1+4a/d} \sim n^{-8p/(4(p+a)+d)}$$

and for the bias term $J^{-4p/d} \sim n^{-8p/(4(p+a)+d)}$. We now distinguish between the two regularity cases of the result. First, consider the case $p \le a + d/4$, where the mapping $j \mapsto j^{2(a-p)/d+1/2}$ is increasing and consequently, we observe

$$\begin{aligned}n^{-1}\sum_{j=1}^J \langle h_0, \widetilde{\psi}_j\rangle_\mu^2\,\tau_j^2 &\sim n^{-1}\sum_{j=1}^J \langle h_0, \widetilde{\psi}_j\rangle_\mu^2\, j^{2p/d-1/2} j^{2(a-p)/d+1/2} \\ &\lesssim n^{-1}J^{2(a-p)/d+1/2} \sim n^{-8p/(4(p+a)+d)}.\end{aligned}$$

Moreover, we obtain

$$n^{-1}\tau_J^2 J^{-2p/d} \sim n^{-1}J^{2(a-p)/d} \lesssim n^{-8p/(4(p+a)+d)}.$$

Finally, it remains to consider the case $p > a + d/4$. In this case, we have that

$$\sum_{j=1}^{J} \langle h_0, \widetilde{\psi}_j \rangle_\mu^2 \, \tau_J^2 \lesssim \sum_{j=1}^{J} \langle h_0, \widetilde{\psi}_j \rangle_\mu^2 j^{2p/d} = O(1)$$

and consequently, the second variance term satisfies $n^{-1} \big\| \langle Q_J h_0, \psi^J \rangle_\mu (G_b^{-1/2} S)_l^- \big\|^2 = O(n^{-1})$ which is the dominating rate and thus, completes the proof of the result.

Proof of Result (2.5) for the severely ill-posed case. The choice of

$$J \sim \left(\log n - \frac{4p + d}{2a} \log \log n \right)^{d/a}$$

implies

$$n^{-2} \tau_J^4 J \sim n^{-2} J \exp(2 J^{a/d}) \sim \left(\log n - \frac{4p + d}{2a} \log \log n \right)^{d/a} (\log n)^{-(4p+d)/a} \sim (\log n)^{-4p/a}.$$

We further analyze for the bias part

$$J^{-4p/d} \sim \left(\log n - \frac{4p + d}{2a} \log \log n \right)^{-4p/a} \sim (\log n)^{-4p/a}.$$

Moreover, since the mapping $j \mapsto j^{-2p/d} \exp(j^{a/d})$ is increasing we obtain

$$\begin{aligned} n^{-1} \sum_{j=1}^{J} \langle h_0, \widetilde{\psi}_j \rangle_\mu^2 \, \tau_j^2 &\sim n^{-1} \sum_{j=1}^{J} \langle h_0, \widetilde{\psi}_j \rangle_\mu^2 j^{2p/d} j^{-2p/d} \exp(j^{a/d}) \\ &\lesssim n^{-1} \exp(J^{a/d}) J^{-2p/d} \sim (\log n)^{-2p/a} (\log n)^{-(2p+d)/a} \lesssim (\log n)^{-4p/a} \end{aligned}$$

and finally

$$n^{-1} \tau_J^2 J^{-2p/d} \sim n^{-1} \exp(J^{a/d}) J^{-2p/d} \lesssim (\log n)^{-4p/a},$$

which shows the result. $\square$

B. Proofs of Results in Section 3

We denote $\mathcal{H} = \bigcup_{p\in[\underline{p},\overline{p}]} \mathcal{H}_2(p, L) \cap \mathcal{B}$ and recall that $\mathcal{B} = \mathcal{B}(L) = \{h : \|h\|_\infty < L\}$. Below, we make use of the notation

$$\widehat{\mathcal{J}} = \Big\{J \in \widehat{\mathcal{I}} : |\widehat{f}_J - \widehat{f}_{J'}| \leq c_0(\widehat{V}(J) + \widehat{V}(J')) \text{ for all } J' \in \widehat{\mathcal{I}} \text{ with } J' > J\Big\}$$

and recall the definition $\widehat{\mathcal{I}} = \{J \in \mathcal{T} : 0.1(\log \widehat{J}_{\max})^2 \leq J \leq \widehat{J}_{\max}\}$. We denote

$$\overline{J}(c) = \sup\{J \in \mathcal{T} : \tau_J J\sqrt{(\log J)/n} \leq c\}$$

for some constant $c > 0$. The oracle choice of the dimension parameter is given by

$$J_0 = J_0(p, c_0) = \sup\Big\{J \in \mathcal{T} : V(J) \leq c_0 J^{-2p/d}\Big\}, V(J) = \tau_J^2 \frac{\sqrt{J(\log n)}}{n} \vee \frac{1}{\sqrt{n}} \quad \text{(B.1)}$$

for some constant $c_0 > 0$. We introduce the set

$$\mathcal{E}_n^* = \{J_0 \in \widehat{\mathcal{J}}\} \cap \{|\widehat{s}_J - s_J| \leq \eta s_J \text{ for all } J \in \mathcal{I}\}$$

for some $\eta \in (0, 1)$.

Proof of Theorem 3.1. Proof of Result (3.4) for the mildly ill-posed case. Due to Chen et al. [2021, Lemma B.5] we have $\overline{J}(c_1) \leq \widehat{J}_{\max} \leq \overline{J}(c_2)$ for some constants $c_1, c_2 > 0$ on $\mathcal{E}_n^*$. The definition $\widehat{J} = \min_{J\in\widehat{\mathcal{J}}} J$ implies $\widehat{J} \leq J_0$ on the set $\mathcal{E}_n^*$ and hence, we obtain

$$\begin{aligned}\big|\widehat{f}_{\widehat{J}} - f(h_0)\big| \mathbb{1}_{\mathcal{E}_n^*} &\leq \big|\widehat{f}_{\widehat{J}} - \widehat{f}_{J_0}\big| \mathbb{1}_{\mathcal{E}_n^*} + |\widehat{f}_{J_0} - f(h_0)| \mathbb{1}_{\mathcal{E}_n^*}\\ &\leq c_0\Big(\widehat{V}(\widehat{J}) + \widehat{V}(J_0)\Big) \mathbb{1}_{\mathcal{E}_n^*} + |\widehat{f}_{J_0} - f(h_0)|.\end{aligned}$$

On the set $\mathcal{E}_n^*$, we have $|\widehat{s}_J - s_J| \leq \eta s_J$, for some $\eta \in (0, 1)$, which implies $\widehat{s}_J^{-2} \leq s_J^{-2}(1-\eta)^{-2}$ and thus, by the definition of $\widehat{V}(\cdot)$ in (3.3) we have

$$\big|\widehat{f}_{\widehat{J}} - f(h_0)\big| \mathbb{1}_{\mathcal{E}_n^*} \leq c_0(1-\eta)^{-2}\left(\Big(s_{\widehat{J}}^{-2}\sqrt{\widehat{J}} + s_{J_0}^{-2}\sqrt{J_0}\Big) \mathbb{1}_{\mathcal{E}_n^*} \frac{\sqrt{\log n}}{n}\right) \vee \frac{1}{\sqrt{n}} + |\widehat{f}_{J_0} - f(h_0)|.$$

Using inequality (2.2) together with Assumption 3(i) yields $s_J^{-2} \leq c_\tau \tau_J^2$ for all J, see inequality (2.2). Consequently, from the definition of $V(\cdot)$ in (B.1) we infer:

$$
\begin{aligned}
\big|\widehat{f}_{\widehat{J}} - f(h_0)\big|\,\mathbb{1}_{\mathcal{E}_n^*} &\leq \frac{c_0\, c_\tau}{(1-\eta)^2}\left(\Big(\tau_{\widehat{J}}^2\sqrt{\widehat{J}} + \tau_{J_0}^2\sqrt{J_0}\Big)\,\mathbb{1}_{\mathcal{E}_n^*}\frac{\sqrt{\log n}}{n}\right) \vee \frac{1}{\sqrt{n}} + |\widehat{f}_{J_0} - f(h_0)| \\
&\leq \frac{c_0\, c_\tau}{(1-\eta)^2}\Big(V(\widehat{J}) + V(J_0)\Big)\,\mathbb{1}_{\mathcal{E}_n^*} + |\widehat{f}_{J_0} - f(h_0)| \\
&\leq \frac{2c_0\, c_\tau}{(1-\eta)^2} V(J_0) + |\widehat{f}_{J_0} - f(h_0)|
\end{aligned}
$$

for n sufficiently large, where the last inequality is due to $V(\widehat{J})\,\mathbb{1}_{\mathcal{E}_n^*} \leq V(J_0)$ since $\widehat{J} \leq J_0$ on $\mathcal{E}_n^*$. By Lemmas C.3 and C.7 it holds $\mathbb{P}(\mathcal{E}_n^*) = 1 + o(1)$.

The definition of the oracle choice in (B.1) implies $J_0 \sim (n/\sqrt{\log n})^{2d/(4(p+a)+d)}$ in the mildly ill-posed case. Thus, we obtain

$$
n^{-2}(\log n)\tau_{J_0}^4 J_0 \sim n^{-2}(\log n) J_0^{1+4a/d} \sim (\sqrt{\log n}/n)^{8p/(4(p+a)+d)}
$$

which coincides with the rate for the bias term. We now distinguish between the two cases in the mildly ill-posed case. First, consider the case $p \leq a + d/4$. In this case, the mapping $j \mapsto j^{2(a-p)/d+1/2}$ is increasing in j and consequently, we observe

$$
n^{-1}\sum_{j=1}^{J_0} \tau_j^2 \langle h_0, \widetilde{\psi}_j\rangle_\mu^2 \lesssim J_0^{2(a-p)/d+1/2} n^{-1} \lesssim (\sqrt{\log n}/n)^{8p/(4(p+a)+d)}.
$$

Moreover, using $h_0 \in \mathcal{H}_2(p, L)$, i.e., $\sum_{j\geq 1}\langle h_0, \widetilde{\psi}_j\rangle_\mu^2 j^{2p/d} \leq L$, we obtain

$$
n^{-1}\tau_{J_0}^2 \sum_{j>J_0}\langle h_0, \widetilde{\psi}_j\rangle_\mu^2 \lesssim (\sqrt{\log n}/n)^{8p/(4(p+a)+d)}.
$$

Finally, it remains to consider the case $p > a + d/4$, where as in the proof of Theorem 2.1 we have $\sum_{j=1}^{J_0} \tau_j^2 \langle h_0, \widetilde{\psi}_j\rangle_\mu^2 = O(1)$, implying $n^{-1}\big\|\langle Q_{J_0}h_0, \psi^{J_0}\rangle_\mu (G_b^{-1/2}S)_l^-\big\|^2 = O(n^{-1})$ which is the dominating rate and thus, completes the proof for the mildly ill-posed case.

Proof of Result (3.5) for the severely ill-posed case. We have

$$
\begin{aligned}
&\big|\widehat{f}_{\widehat{J}} - f(h_0)\big|\,\mathbb{1}_{\mathcal{E}_n^*} \leq \big|\widehat{f}_{\widehat{J}} - f(Q_{\widehat{J}}h_0)\big|\,\mathbb{1}_{\mathcal{E}_n^*} + \max_{\overline{J}(c_1)\leq J\leq \overline{J}(c_2)} |f(Q_J h_0 - h_0)|\,\mathbb{1}_{\mathcal{E}_n^*} \\
&\quad\leq 2\overline{\sigma}_Y^2\left(s_{\overline{J}(c_2)}^{-2}\frac{\sqrt{\overline{J}(c_2)\log \overline{J}(c_2)}}{n-1} + \frac{\|\langle Q_{\overline{J}(c_2)}h_0, \psi^{\overline{J}(c_2)}\rangle_\mu' (G_b^{-1/2}S)_l^-\|}{\sqrt{n}}\right) + \big(\overline{J}(c_1)\big)^{-2p/d}
\end{aligned}
$$

with probability approaching one by Lemma C.5. From Chen et al. [2021, Lemma B.2] it holds, in the severely ill-posed case, $J_0^+ = \inf\{J \in \mathcal{T} : J > J_0\} \geq \overline{J}(c_1)$ for all n sufficiently large and thus, by the definition of $\overline{J}(\cdot)$ we have

$$\big|\widehat{f}_{\widehat{J}} - f(h_0)\big|\, \mathbb{1}_{\mathcal{E}_n^*} \leq (2\overline{\sigma}_Y^2 + 1)\big(C\overline{J}(c_2)\big)^{-2p/d}$$

with probability approaching one, using that $\overline{J}(c_1) \geq C\overline{J}(c_2)$ for some constant $C > 0$. From the definition of $\overline{J}(\cdot)$ we have $(c \log n)^{d/a} \leq \overline{J}(c_2)$ for any $c \in (0, 1)$ and n sufficiently large. This implies

$$\big|\widehat{f}_{\widehat{J}} - f(h_0)\big|\, \mathbb{1}_{\mathcal{E}_n^*} = O_p\big((\log n)^{-2p/a}\big),$$

which completes the proof. □

C. Supplementary Lemmas

We first introduce additional notation. First we consider a U-statistic

$$\mathcal{U}_{n,1} = \frac{2}{n(n-1)} \sum_{i<i'} R_1(Z_i, Z_{i'})$$

where $Z_i = (Y_i, W_i)$ and the kernel R_1 is given by

$$\begin{aligned} R_1(Z_i, Z_{i'}) = {} & Y_i\, \mathbb{1}_{M_i}\, b^{K(J)}(W_i)' A' G_\mu A b^{K(J)}(W_{i'}) Y_{i'}\, \mathbb{1}_{M_{i'}} \\ & - \mathbb{E}[Y_i\, \mathbb{1}_{M_i}\, b^{K(J)}(W_i)]' A' G_\mu A\, \mathbb{E}[b^{K(J)}(W_{i'}) Y_{i'}\, \mathbb{1}_{M_{i'}}] \end{aligned} \tag{C.1}$$

where $M_i = \{|Y_i| \leq M_n\}$ with $M_n = J^{-1/4}\sqrt{n}/(\log \overline{J})$. Note that the kernel R_1 is a symmetric function such that $\mathbb{E}[R_1(Z_i, z)] = 0$ for all z. We also introduce the U-statistic

$$\mathcal{U}_{n,2} = \frac{2}{n(n-1)} \sum_{i<i'} R_2(Z_i, Z_{i'})$$

where the kernel R_2 is given by

$$\begin{aligned} R_2(Z_i, Z_{i'}) = {} & Y_i b^{K(J)}(W_i)' A' G_\mu A b^{K(J)}(W_{i'}) Y_{i'}\, \mathbb{1}_{M_i^c \cup M_{i'}^c} \\ & - \mathbb{E}\left[Y_i b^{K(J)}(W_i)' A' G_\mu A b^{K(J)}(W_{i'}) Y_{i'}\, \mathbb{1}_{M_i^c \cup M_{i'}^c}\right]. \end{aligned}$$

We make use of the following exponential inequality established by Houdré and

Reynaud-Bouret [2003].

Lemma C.1 (Houdré and Reynaud-Bouret [2003]). *Let U_n be a degenerate U-statistic of order 2 with kernel R based on a simple random sample $Z_1, \ldots, Z_n$. Then there exists a generic constant $C > 0$, such that*

$$\mathbb{P}\left(\Big| \sum_{1 \le i < i' \le n} R(Z_i, Z_{i'}) \Big| \ge C\Big(\Lambda_1 \sqrt{u} + \Lambda_2 u + \Lambda_3 u^{3/2} + \Lambda_4 u^2 \Big) \right) \le 6 \exp(-u)$$

where

$$\begin{aligned}
\Lambda_1^2 &= \frac{n(n-1)}{2}\, \mathbb{E}[R^2(Z_1, Z_2)], \\
\Lambda_2 &= n \sup_{\|\nu\|_{L^2(Z)} \le 1, \|\kappa\|_{L^2(Z)} \le 1} \mathbb{E}[R(Z_1, Z_2)\nu(Z_1)\kappa(Z_2)], \\
\Lambda_3 &= \sqrt{n \sup_z |\, \mathbb{E}[R^2(Z_1, z)]|}, \\
\Lambda_4 &= \sup_{z_1, z_2} |R(z_1, z_2)|.
\end{aligned}$$

The next result provides upper bounds for the estimates $\Lambda_1, \Lambda_2, \Lambda_3, \Lambda_4$ when the kernel R coincides with R_1 given in (C.1).

Lemma C.2. *Let Assumption 1(ii) be satisfied. Given kernel $R = R_1$, it holds:*

$$\Lambda_1^2 \le \frac{\overline{\sigma}_Y^4 n(n-1)}{2} J s_J^{-4} \tag{C.2}$$

$$\Lambda_2 \le 2\overline{\sigma}_Y^2\, n\, s_J^{-2} \tag{C.3}$$

$$\Lambda_3 \le \overline{\sigma}_Y^2 M_n \sqrt{n\, c_K J} s_J^{-2} \tag{C.4}$$

$$\Lambda_4 \le c_K M_n^2 J s_J^{-2}. \tag{C.5}$$

Proof. The result follows from the proofs of Lemma E.1 and Lemma F.2 of Breunig and Chen [2021] using Assumption 1(ii). □

Lemma C.3. *Let Assumption 1(iii) hold and Ψ_J be CDV wavelet or dyadic B-spline basis. Then, for any constant $\eta \in (0,1)$ we have*

$$\mathbb{P}\Big(\sup_{J \in \mathcal{I}} s_J^{-1} |\widehat{s}_J - s_J| \le \eta \Big) \to 1.$$

Proof. The result is due to the proof of Lemma C.7 of Chen et al. [2021]. □

Lemma C.4. *Let Assumptions 1(iii) and 3(i) hold and Ψ_J be CDV wavelet or dyadic B-spline basis. Then, for any constants $c_1, c_2 > 0$ we have*

$$\mathbb{P}\Big(\overline{J}(c_1) \leq \widehat{J}_{\max} \leq \overline{J}(c_2)\Big) \to 1.$$

Proof. The result is due to Chen et al. [2021, Lemma B.5]. □

Below, for simplicity of notation, we denote $\overline{J} := \overline{J}(c)$ for some constant $c > 0$ and we make use of the notation

$$c_n(J) := 2\overline{\sigma}_Y^2 s_J^{-2}\,\frac{\sqrt{J\log\overline{J}}}{n-1} + \|\langle Q_J h_0, \psi^J\rangle_\mu'(G_b^{-1/2}S)_l^-\|/\sqrt{n}$$

Lemma C.5. *Let Assumptions 1(i)(iii), 3(i), and 5 be satisfied. Then, we have*

$$\inf_{h_0\in\mathcal{H}} \mathbb{P}_{h_0}\Big(\big|\widehat{f}_J - f(Q_J h_0)\big| \leq c_n(J) \quad \forall J \in \mathcal{I}\Big) \to 1. \tag{C.6}$$

Proof. First, observe that by making use of Assumption 5 it holds $c_n(J) = o(1)$ uniformly in $J \in \mathcal{I}$. We make use of the decomposition

$$\begin{aligned}
&\widehat{f}_J - f(Q_J h_0)\\
&= \frac{2}{n(n-1)}\sum_{i<i'}\big(Y_i b^{K(J)}(W_i) - \mathbb{E}[Yb^{K(J)}(W)]\big)' A' G_\mu A\big(Y_{i'} b^{K(J)}(W_{i'}) - \mathbb{E}[Yb^{K(J)}(W)]\big)\\
&\quad + \frac{4}{n}\sum_i \mathbb{E}[Yb^{K(J)}(W)]' A' G_\mu A\big(Y_i b^{K(J)}(W_i) - \mathbb{E}[Yb^{K(J)}(W)]\big)\\
&\quad + \frac{2}{n(n-1)}\sum_{i<i'} Y_i Y_{i'} b^{K(J)}(W_i)'\big(A'G_\mu A - \widehat{A}'G_\mu\widehat{A}\big) b^{K(J)}(W_{i'}).
\end{aligned}$$

Using the U-statistic notation introduced at the beginning of this section we obtain

$$\begin{aligned}
&\mathbb{P}_{h_0}\left(\max_{J\in\mathcal{I}}\Big\{c_n(J)^{-1}\big|\widehat{f}_J - f(Q_J h_0)\big|\Big\} > 1\right) \leq \mathbb{P}\left(\max_{J\in\mathcal{I}}\Big\{c_n(J)^{-1}\big|\mathcal{U}_{n,1}(J)\big|\Big\} > \frac{1}{4}\right)\\
&\quad + \mathbb{P}_{h_0}\left(\max_{J\in\mathcal{I}}\Big\{c_n(J)^{-1}\big|\mathcal{U}_{n,2}(J)\big|\Big\} > \frac{1}{4}\right)\\
&\quad + \mathbb{P}_{h_0}\left(\max_{J\in\mathcal{I}}\Big\{c_n(J)^{-1}\Big|\frac{4}{n}\sum_i \mathbb{E}[Yb^{K(J)}(W)]'A'G_\mu A\big(Y_i b^{K(J)}(W_i) - \mathbb{E}[Yb^{K(J)}(W)]\big)\Big|\Big\} > \frac{1}{4}\right)\\
&\quad + \mathbb{P}_{h_0}\left(\max_{J\in\mathcal{I}}\Big\{c_n(J)^{-1}\Big|\frac{2}{n(n-1)}\sum_{i<i'} Y_i Y_{i'} b^{K(J)}(W_i)'\big(A'G_\mu A - \widehat{A}'G_\mu\widehat{A}\big) b^{K(J)}(W_{i'})\Big|\Big\} > \frac{1}{4}\right)\\
&\quad =: T_1 + T_2 + T_3 + T_4.
\end{aligned}$$

Consider T_1. We make use of U-statistics deviation results. To do so, consider $\Lambda_1, \dots, \Lambda_4$ as given in Lemma C.1. From Lemma C.2 we infer with $u = 2\log \overline{J}$ and $M_n = J^{-1/4}\sqrt{n}/(\log \overline{J})$ that for all $J \le \overline{J}$ we have

$$\begin{aligned}\Lambda_1\sqrt{u} + \Lambda_2 u + &\Lambda_3 u^{3/2} + \Lambda_4 u^2 \\ &\le \Lambda_1\sqrt{2\log \overline{J}} + 2\Lambda_2 \log \overline{J} + \Lambda_3(2\log \overline{J})^{3/2} + 4\Lambda_4(\log \overline{J})^2 \\ &\le \overline{\sigma}_Y^2 n s_J^{-2}\sqrt{J\log \overline{J}} + 4\overline{\sigma}_Y^2 n s_J^{-2}\log \overline{J} + \overline{\sigma}_Y^2 n s_J^{-2}\, J^{1/4}\sqrt{2\log \overline{J}} + 4n s_J^{-2}\sqrt{J}\end{aligned}$$

for n sufficiently large. Hence, we obtain for n sufficiently large

$$\Lambda_1\sqrt{u} + \Lambda_2 u + \Lambda_3 u^{3/2} + \Lambda_4 u^2 \le 2\overline{\sigma}_Y^2 n s_J^{-2}\sqrt{J\log \overline{J}} \le \frac{n(n-1)}{2} c_n(J)$$

by the definition of $c_n(J)$ and Lemma C.1 with $u = 2\log \overline{J}$ yields

$$T_1 \le \sum_{J\in\mathcal{I}} \mathbb{P}_{h_0}\left(\Big|\sum_{i<i'} R_1(Z_i, Z_{i'})\Big| \ge \frac{n(n-1)}{2} c_n(J)\right) \le 6\exp\big((\log \overline{J}) - 2(\log \overline{J})\big) \le \frac{6}{\overline{J}}$$

and thus, $T_1 = o(1)$ since $\overline{J}$ diverges. Consider T_2. Markov's inequality together with $\#\mathcal{I} \le \log_2(\overline{J}) \le 2\log(\overline{J})$ yield by following the derivation of the upper bound (A.3):

$$\begin{aligned}T_2 &\le \sum_{J\in\mathcal{I}} c_n(J)^{-1}\sqrt{\mathbb{E}_{h_0}|\mathcal{U}_{n,2}(J)|^2} \le 2\log(\overline{J}) \max_{J\in\mathcal{I}} c_n(J)^{-1}\sqrt{\mathbb{E}_{h_0}|\mathcal{U}_{n,2}(J)|^2} \\ &= O\Big(n^{-1/2}\log(\overline{J}) \max_{J\in\mathcal{I}} \frac{\sqrt{J}}{M_n^2 c_n(J) s_J^2}\Big) = O\Big(n^{-1/2}\log(\overline{J})^{3/2}\sqrt{\overline{J}}\Big) = o(1),\end{aligned}$$

using that $M_n^{-2} = \sqrt{J}(\log \overline{J})^2/n$ and $(c_n(J)s_J^2)^{-1} \le n/\sqrt{J\log \overline{J}}$. Lemma C.6 implies $T_3 = o(1)$. Consider T_4. We have that

$$\max_{J\in\mathcal{I}}\Big\{n^{-1}\Big(\log(J)\sum_{j=1}^{J} s_j^{-4}\Big)^{-1/2}\sum_{i<i'} Y_i Y_{i'} b^{K(J)}(W_i)'\big(A'G_\mu A - \widehat{A}'G_\mu\widehat{A}\big)b^{K(J)}(W_{i'})\Big|\Big\} = o_p(1)$$

by following Lemma E.5 of Breunig and Chen [2021] (with their $Y_i - h_0(X_i)$ and v_J replaced by our Y_i and $(\sum_{j=1}^{J} s_j^{-4})^{1/2}$ respectively, and, in particular, we do not need to impose a lower bound on $\mathbb{E}[Y^2|W]$ since our estimator is un-standardized.) The previous equation implies $T_4 = o(1)$. □

Lemma C.6. *Let Assumptions 1(i)(iii), 3(i), and 5 be satisfied. Then, there exists* *a constant $C > 0$ such that*

$$\inf_{h_0 \in \mathcal{H}} \mathbb{P}_{h_0}\left(\sup_{J \in \mathcal{I}}\left(c_n^{-1}(J)\Big|\frac{1}{n}\sum_i Y_i a_J(W_i) - \mathbb{E}[Y a_J(W)]\Big|\right) \leq C\right) \to 1,$$

where $a_J(w) = \langle Q_J h_0, \psi^J\rangle'_\mu (G_b^{-1/2} S)_l^- \widetilde{b}^{K(J)}(w)$.

Proof. The result follows by an application of Talagrand's inequality analogously to the proof of Chen et al. [2021, Lemma C.2], based on the following upper bounds:

$$\mathbb{E}\,|Y_i a_J(W_i) - \mathbb{E}[Y a_J(W)]|^2 \leq \mathbb{E}\,|Y a_J(W)|^2 \leq \sigma_Y^2 \|\langle Q_J h_0, \psi^J\rangle'_\mu (G_b^{-1/2} S)_l^- \|^2$$

by using Assumption 5(ii) and

$$|Y_i a_J(W_i) - \mathbb{E}[Y a_J(W)]|\, \mathbb{1}_{M_n} \leq \sqrt{K(J)} M_n \|\langle Q_J h_0, \psi^J\rangle'_\mu (G_b^{-1/2} S)_l^- \|,$$

where $M_n = \big\{ \max_j |Y_i \widetilde{b}_j(W_i) - \mathbb{E}[Y \widetilde{b}_j(W)]| \leq n^{1/6}\big\}$. □

Lemma C.7. *Let Assumptions 1(i)(iii), 3(i), and 5 be satisfied and consider the mildly ill-posed case, i.e., $\tau_j \sim j^{a/d}$. Then, we have $\inf_{h_0 \in \mathcal{H}} \mathbb{P}_{h_0}(J_0 \in \widehat{\mathcal{J}}) \to 1$.*

Proof. Let $\mathcal{E}_n$ denote the event upon which $\overline{J}(c_1) \leq \widehat{J}_{\max} \leq \overline{J}(c_2)$ for some constant $c_1, c_2 > 0$ and observe that $\mathbb{P}(\mathcal{E}_n^c) = o(1)$ by Lemma C.4. For all $J > J_0$, we make use of the upper bound

$$\big|\widehat{f}_{J_0} - \widehat{f}_J\big| \leq \big|\widehat{f}_{J_0} - f(Q_{J_0} h_0)\big| + \big|\widehat{f}_J - f(Q_J h_0)\big| + 2\big|f(Q_{J_0} h_0 - h_0)\big| + 2\big|f(Q_J h_0 - h_0)\big|.$$

By Lemma C.5, uniformly for all $J \in \mathcal{I}$ it holds

$$\big|\widehat{f}_J - f(Q_J h_0)\big| \leq 2\overline{\sigma}_Y^2 s_J^{-2} \frac{\sqrt{J \log \overline{J}}}{n-1} + \|\langle Q_J h_0, \psi^J\rangle'_\mu (G_b^{-1/2} S)_l^- \|/\sqrt{n}$$

on some set $\mathcal{E}_{n,1} \subseteq \mathcal{E}_n$ where $\mathbb{P}(\mathcal{E}_{n,1}^c) = o(1)$. For all $J > J_0$ we have

$$\begin{aligned}\Big|f(Q_J h_0 - h_0)\Big| &\leq C\|\Pi_J h_0 - h_0\|_\mu^2 \leq C L J^{-2p/d} \leq C L (J_0^+)^{-2p/d} \\ &\leq C_0 L \tau_{J_0^+}^2 \sqrt{J_0^+ \log n}/n,\end{aligned}$$

where in the last inequality we used the definition of $J_0^+ = \inf\{J \in \mathcal{J} : J > J_0\}$.

Hence, we conclude for all $J \geq J_0^+$ that

$$\left|\widehat{f}_{J_0} - \widehat{f}_J\right| \leq (C_0 + 2\overline{\sigma}_Y^2)\left(s_{J_0}^{-2}\sqrt{J_0 \log n}/n + s_J^{-2}\sqrt{J \log n}/n\right) \vee 1/\sqrt{n}.$$

Due to Lemma C.3 it holds $s_J^{-2} \leq (1+\eta)^2\, \widehat{s}_J^{-2}$ for some $\eta \in (0,1)$, uniformly for $J \in \mathcal{I} \cap \{J > J_0\}$, on some set $\mathcal{E}_{n,2}$ with $\mathbb{P}(\mathcal{E}_{n,2}^c) = o(1)$. Consequently, on $\mathcal{E}_{n,1} \cap \mathcal{E}_{n,2}$ it holds

$$\begin{aligned}\left|\widehat{f}_{J_0} - \widehat{f}_J\right| &\leq (C_0 + 2\overline{\sigma}_Y^2)(1+\eta)^2\left(\widehat{s}_{J_0}^{-2}\sqrt{J_0 \log n}/n + \widehat{s}_J^{-2}\sqrt{J \log n}/n\right) \vee 1/\sqrt{n} \\ &= (C_0 + 2\overline{\sigma}_Y^2)(1+\eta)^2\left(\widehat{V}(J_0) + \widehat{V}(J)\right)\end{aligned}$$

uniformly for $J \in \mathcal{I} \cap \{J > J_0\}$. We conclude that $J_0 \in \widehat{\mathcal{J}}$ on $\mathcal{E}_{n,1} \cap \mathcal{E}_{n,2}$ and $\mathbb{P}(\mathcal{E}_{n,1} \cap \mathcal{E}_{n,2}) \to 1$. □

References

P. J. Bickel and Y. Ritov. Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 381–393, 1988.

L. Birgé and P. Massart. Estimation of integral functionals of a density. *The Annals of Statistics*, pages 11–29, 1995.

R. Blundell, X. Chen, and D. Kristensen. Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669, 2007.

C. Breunig and X. Chen. Adaptive, rate-optimal hypothesis testing in nonparametric iv models. *arXiv preprint arXiv:2006.09587v2*, 2021.

C. Breunig and J. Johannes. Adaptive estimation of functionals in nonparametric instrumental regression. *Econometric Theory*, 32(3):612–654, 2016.

C. Butucea. Goodness-of-fit testing and quadratic functional estimation from indirect observations. *The Annals of Statistics*, 35(5):1907–1930, 2007.

C. Butucea and K. Meziani. Quadratic functional estimation in inverse problems. *Statistical Methodology*, 8(1):31–41, 2011.

T. T. Cai and M. G. Low. Optimal adaptive estimation of a quadratic functional. *The Annals of Statistics*, 34(5):2298–2325, 2006.

M. Carrasco, J.-P. Florens, and E. Renault. Linear inverse problems in structural econometrics: Estimation based on spectral decomposition and regularization. In J. Heckman and E. Leamer, editors, *Handbook of Econometrics*, volume 6B. North Holland, 2007.

X. Chen and T. M. Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, 9 (1):39–84, 2018.

X. Chen and D. Pouzo. Sieve wald and qlr inferences on semi/nonparametric conditional moment models. *Econometrica*, 83(3):1013–1079, 2015.

X. Chen and M. Reiß. On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory*, 27(03):497–521, 2011.

X. Chen, T. Christensen, and S. Kankanala. Adaptive estimation and uniform confidence bands for nonparametric iv. *arXiv preprint arXiv:2107.11869*, 2021.

V. Chernozhukov, D. Chetverikov, and K. Kato. Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 42(5):1787–1818, 2014.

C. Chesneau. On adaptive wavelet estimation of a quadratic functional from a deconvolution problem. *Annals of the Institute of Statistical Mathematics*, 63(2):405–429, 2011.

O. Collier, L. Comminges, and A. B. Tsybakov. Minimax estimation of linear and quadratic functionals on sparsity classes. *The Annals of Statistics*, 45(3):923–958, 2017.

S. Darolles, Y. Fan, J.-P. Florens, and E. Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.

D. L. Donoho and M. Nussbaum. Minimax quadratic estimation of a quadratic functional. *Journal of Complexity*, 6(3):290–323, 1990.

S. Efromovich and V. Koltchinskii. On inverse problems with unknown operators. *IEEE Transactions on Information Theory*, 47(7):2876–2894, 2001.

S. Efromovich and M. Low. On optimal adaptive estimation of a quadratic functional. *The Annals of Statistics*, 24(3):1106–1125, 1996.

J. Fan. On the estimation of quadratic functionals. *The Annals of Statistics*, pages 1273–1294, 1991.

E. Giné and R. Nickl. A simple adaptive estimator of the integrated square of a density. *Bernoulli*, 14(1):47–61, 2008.

E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press, 2021.

P. Hall and J. L. Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33:2904–2929, 2005.

J. L. Horowitz. Adaptive nonparametric instrumental variables estimation: Empirical choice of the regularization parameter. *Journal of Econometrics*, 180(2):158–173, 2014.

C. Houdré and P. Reynaud-Bouret. Exponential inequalities, with constants, for u-statistics of order two. In *Stochastic inequalities and applications*, pages 55–69. Springer, 2003.

M. Kroll. Rate optimal estimation of quadratic functionals in inverse problems with partially unknown operator and application to testing problems. *ESAIM: Probability and Statistics*, 23:524–551, 2019.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, pages 1302–1338, 2000.

O. V. Lepski. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability and its Applications*, 35:454–466, 1990.

O. V. Lepski and V. G. Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, pages 2512–2546, 1997.

O. V. Lepski, E. Mammen, and V. G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *The Annals of Statistics*, pages 929–947, 1997.

L. Liu, R. Mukherjee, J. M. Robins, and E. T. Tchetgen. Adaptive estimation of nonparametric functionals. *Journal of Machine Learning Research*, 22(99):1–66, 2021.

W. K. Newey and J. L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71:1565–1578, 2003.

V. Spokoiny and C. Vial. Parameter tuning in pointwise adaptation using a propagation approach. *The Annals of Statistics*, 37(5B):2783–2807, 2009.

V. Spokoiny and N. Willrich. Bootstrap tuning in gaussian ordered model selection. *The Annals of Statistics*, 47(3):1351–1380, 2019.