Adaptive Estimation of Quadratic Functionals in Nonparametric Instrumental Variable Models

CHRISTOPH BREUNIG†    XIAOHONG CHEN‡

First version: August 2018; Revised version: January 2021

This paper considers adaptive estimation of quadratic functionals in the nonparametric instrumental variables (NPIV) models. Minimax estimation of a quadratic functional of a NPIV is an important problem in optimal estimation of a nonlinear functional of an ill-posed inverse regression with an unknown operator using one random sample. We first show that a leave-one-out, sieve NPIV estimator of the quadratic functional proposed by Breunig and Chen [2020] attains a convergence rate that coincides with the lower bound previously derived by Chen and Christensen [2018]. The minimax rate is achieved by the optimal choice of a key tuning parameter (sieve dimension) that depends on unknown NPIV model features. We next propose a data driven choice of the tuning parameter based on Lepski’s method. The adaptive estimator attains 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.

Keywords: nonparametric instrumental variables, ill-posed inverse, 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 Cristina Butucea, Tim Christensen, Richard Nickl and Sasha Tsybakov for discussions. We acknowledge helpful comments from participants at econometrics workshops at NYU/Stern, NUS, Chicago/Booth, and 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, 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 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 observational data. In economics and other social sciences, as well as in medical research, it is very difficult to estimate causal effects using observational data sets alone. When treatment assignment is not randomized, it is generally impossible to discern between the causal effect of treatments and spurious correlations that are induced by unobserved factors. 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, the nonparametric instrumental variables regressions (NPIV) 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 a 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 measurable function, \( X \) is a \( d \)-dimensional vector of 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 \( T \) is compact. This makes the estimation of \( h_0 \) an ill-posed inverse problem with an unknown operator \( T \). See, for example, Newey and Powell [2003], Carrasco et al. [2007], Blundell et al. [2007] and Horowitz [2011].

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

$$f(h_0) := \int h_0^2(x)\mu(x)dx$$

(1.2)

for a known non-negative weighting function $\mu$, which is assumed to be uniformly bounded from below on its support. Although Chen and Pouzo [2015] and Chen and Christensen [2018] already studied plug-in sieve NPIV estimation and inference on nonlinear functionals of $h_0$, there is no results on the minimax rate optimal estimation of any nonlinear functionals of $h_0$ yet. Since a quadratic functional is a leading example of a nonlinear functional and is also widely used in goodness-of-fit testing on $h_0$, this paper focuses on optimal estimation of the quadratic functional $f(h_0)$ in the NPIV model.

In this paper, we first analyze a simple leave-one-out, sieve NPIV estimator $\hat{f}_J$ for the quadratic functional $f(h_0)$ recently proposed by Breunig and Chen [2020]. We establish an upper bound on the convergence rate for this estimator, and show that it coincides with the lower bound previously derived by Chen and Christensen [2018]. Thus, the leave-one-out, sieve NPIV estimator $\hat{f}_J$ is minimax rate-optimal for $f(h_0)$.

Depending on the smoothing properties of the conditional expectation operator $T$, we distinguish between the mildly and severely ill-posed cases. For the severely ill-posed case, the minimax optimal convergence rate for estimating $f(h_0)$ is of $(\log n)^{-a}$ order, where $a > 0$ depends on the smoothness of the NPIV function $h_0$ and the degree of ill-posedness. For the mildly ill-posed case, the optimal convergence rate for estimating $f(h_0)$ exhibits the so-called elbow phenomena: the rate is of order $n^{-1/2}$ for the regular mildly ill-posed case; the rate is of order $n^{-b}$, where $0 < b < 1/2$ depends on the smoothness of $h_0$, the dimension of $X$ and the degree of ill-posedness. As already noticed in Chen and Christensen [2018], the minimax convergence rate for the severely ill-posed case is so slow that even a plug-in sieve NPIV estimator $f(\hat{h})$ of $f(h_0)$ achieves it, where $\hat{h}$ is the sieve NPIV estimator of $h_0$. However, the plug-in estimator $f(\hat{h})$ fails to achieve the minimax rate for the mildly ill-posed case. Nevertheless, we show that the simple leave-one-out sieve NPIV estimator $\hat{f}_J$ is minimax rate-optimal regardless the degree of ill-posedness.

The minimax optimal rate is achieved by the optimal choice of a key tuning parameter (sieve dimension) $J$ that depends on the unknown smoothness of the NPIV function $h_0$ and the unknown degree of ill-posedness. We next propose a data driven choice of the tuning parameter (sieve dimension) $J$ based on Lepski’s principle. The adaptive, leave-one-out sieve NPIV estimator $\hat{f}_J$ 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 an unknown density or a regression function from a random sample, Efromovich and Low [1996] shown that the extra $\sqrt{\log n}$ factor is the necessary price to pay for adaptation to the unknown smoothness of the function.

Previously for the nonparametric estimation of the NPIV function $h_0$ in the 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 $h_0$ in root-mean squared error metric using a combined model selection and Lepski method. These papers obtain adaptive rate of convergence up to a factor of $\sqrt{\log(n)}$ (of the minimax optimal rate) in both severely ill-posed and mildly ill-posed cases. Chen and Christensen [2015] propose adaptive estimation of $h_0$ in $L^\infty$ norm using Lepski method, and show that their data-driven procedure attains the minimax optimal rate in sup-norm and thus fully adaptive. Our data-driven choice of the sieve dimension is closest to that of Chen and Christensen [2015], which might explain why we also obtain minimax optimal adaptivity for quadratic functionals of a NPIV function $h_0$.

Minimax rate optimal estimation of a quadratic functional in density and regression settings has a long history in statistics. The elbow phenomenon was perhaps first discovered by Bickel and Ritov [1988] in estimation of the integrated square of a density. An analogous result was established by Donoho and Nussbaum [1990] in estimation of quadratic functionals in Gaussian sequence models. Efromovich and Low [1996], Laurent [1996] and Laurent and Massart [2000] establish that Pinsker type estimators of quadratic functionals can achieve the minimax optimal rates of convergence. For estimation of a integrated squared density, Giné and Nickl [2008] showed that a leave-one-out estimator attains the rate optimal and provided an adaptive estimator based on Lepski’s method. Collier et al. [2017] consider minimax rate-optimal estimation of a quadratic functional under sparsity constraints. Minimax rate optimal estimation of a quadratic functional has also been analyzed in density deconvolutions and inverse regressions in Gaussian sequence models. See, for example, Butucea [2007], Butucea and Meziani [2011] and Kroll [2019]. As far as we know, there is no published work on minimax rate-optimal adaptive estimation of a quadratic functional in NPIV models yet.

The rest of the paper is organized as follows. Section 2 presents the leave-one-out sieve NPIV estimator of the quadratic functional $f(h_0)$. It also derives the minimax
optimal convergence rates for our estimator of \( f(h_0) \). Section 3 first presents a simple data-driven procedure of choosing tuning parameters based on Lepski’s method. It then establishes the convergence rates of our adaptive estimator of the quadratic functional. Section 4 briefly concludes. All proofs can be found in the Appendices A–C.

2. Minimax Optimal Quadratic Functional Estimation

This section consists of three parts. Subsection 2.1 first recalls a lower bound result in Chen and Christensen [2018] for estimation of a quadratic functional of the NPIV function \( h_0 \). 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 rate coincides with the lower bound and hence is optimal.

2.1. The Lower Bound

We first introduce notation. For any random vector \( V \) with support \( \mathcal{V} \), we let \( L^2(\mathcal{V}) = \{ \phi : \mathcal{V} \mapsto \mathbb{R}, \| \phi \|_{L^2(\mathcal{V})} < \infty \} \) with the norm \( \| \phi \|_{L^2(\mathcal{V})} = \sqrt{\mathbb{E}[\phi^2(V)]} \). 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 \( b_n \lesssim a_n \).

Denote \( L^2_\mu = \{ \phi : [0,1]^d \mapsto \mathbb{R}, \| \phi \|_{\mu} < \infty \} \) with the norm \( \| \phi \|_{\mu} = \sqrt{\int \phi^2(x) \mu(x) \, dx} \). Let \( \{\tilde{\psi}\}_{j \geq 1} \) be an orthonormal basis and \( \tilde{\psi}_{j,k,G} \) be a tensor-product CDV wavelet basis in \( L^2_\mu \). For simplicity we assume that the structural function \( h_0 \) belongs to the Sobolev ellipsoid

\[
\mathcal{H}_2(p,L) = \left\{ \phi \in L^2_\mu : \sum_{j=1}^{\infty} j^{2p/d} \langle \phi, \tilde{\psi}_{j} \rangle^2_{\mu} \leq L \right\}, \quad \text{for } 0 < p, L < \infty.
\]

Finally we let \( T : L^2(X) \mapsto L^2(W) \) denote the conditional expectation operator given by \( (Th)(w) = \mathbb{E}[h(X)|W = w] \).

**Condition LB.** (i) \( \mathbb{E}[U^2|W = w] \geq \sigma^2 > 0 \) uniformly for \( w \in W \); (ii) there is some decreasing sequence \( \left( \nu_j \right)_{j \geq 1} \) such that \( \|Th\|^2_{L^2(W)} \lesssim \sum_{j,G,k} \nu(2^j)^2 \langle h, \tilde{\psi}_{j,k,G} \rangle^2_{\mu} \) for all \( h \in \mathcal{H}_2(p,L) \); and (iii) \( T[h - h_0] = 0 \) for any \( h \in L^2_\mu \) implies that \( f(h) = f(h_0) \).

Condition LB (ii) specifies the smoothing properties of the conditional expectation operator relative to \( \tilde{\psi}_{j,k,G} \). This assumption is commonly imposed in the related
literature, see also Chen and Reiß [2011] for an overview. As in the related literature, we distinguish between a mildly ill-posed case where $\nu(t) = t^{-\zeta}$ and a severely ill-posed case where $\nu(t) = \exp(-\frac{1}{2}t^\zeta)$ for some $\zeta > 0$. Condition LB (iii) ensures identification of the nonlinear functional $f(h_0)$. The next result was established by Chen and Christensen [2018, Theorem C.1] and hence, its proof is omitted.

**Lemma 2.1.** Let Condition LB be satisfied. Then, we have

$$\liminf_{n \to \infty} \inf_{\tilde{f}} \sup_{h \in \mathcal{H}_2(p,L)} \mathbb{P}_h \left( |\tilde{f} - f(h)| > cr_n \right) \geq c > 0,$$

for some constant $c > 0$, where $\inf_{\tilde{f}}$ is the infimum over all estimators that depend on the sample size $n$ and

1. Mildly ill-posed case:

$$r_n = \begin{cases} 
  n^{-4p/(4p+\zeta)+d}, & \text{if } p \leq \zeta + d/4, \\
  n^{-1/2}, & \text{if } p > \zeta + d/4. 
\end{cases} \quad (2.1)$$

2. Severely ill-posed case:

$$r_n = (\log n)^{-2p/\zeta}. \quad (2.2)$$

According to Lemma 2.1, there is the so-called elbow phenomena in the mildly ill-posed case 2.1: the regular case with a parametric rate of $n^{-1/2}$ when $p > \zeta + d/4$; and the irregular case with a nonparametric rate when $p \leq \zeta + d/4$.

### 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 $\hat{h}$ of $h_0$ can be written in matrix form as

$$\hat{h}(\cdot) = \psi^J(\cdot)^t [\Psi'B(B'B)^{-1}B']^{-1}B'' \Psi'[B(B'B)^{-1}B']^{-1}B'' \Psi'B(B'B)^{-1}B'Y$$

where $Y = (Y_1, \ldots, Y_n)^t$ and

$$\psi^J(x) = (\psi_1(x), \ldots, \psi_J(x))^t \quad \Psi = (\psi^J(X_1), \ldots, \psi^J(X_n))^t$$

$$b^K(w) = (b_1(w), \ldots, b_K(w))^t \quad B = (b^K(W_1), \ldots, b^K(W_n))^t$$
where \( \{ \psi_1, \ldots, \psi_J \} \) and \( \{ b_1, \ldots, b_K \} \) are collections of basis functions of dimension \( J \) and \( K \) used for approximating \( h_0 \) and the instrument space, respectively. Based on many simulation results in Blundell et al. [2007] and the minimax sup-norm rate results in Chen and Christensen [2018], the crucial regularization parameter is the dimension \( J \) of the sieve space used to approximate unknown function \( h_0 \). In this paper, we keep the relationship of \( J \) and \( K \) fixed, i.e., the function \( K(\cdot) \) does not depend on the sample size and satisfies \( K(J) \geq J \) for all \( J \). As pointed out by Chen and Christensen [2018], although one could estimate \( f(h_0) \) by the plug-in sieve NPIV estimator \( \hat{f}(\tilde{h}) \), it fails to achieve the lower bound in Lemma 2.1.

Recently Breunig and Chen [2020] propose a leave-one-out, sieve NPIV estimator for the quadratic functional \( f(h_0) \) as follows:

\[
\hat{f}_J = \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} Y_i Y_{i'} b^K(W_i) \hat{A}' G_\mu A b^K(W_{i'})
\]

where \( G_\mu = \int \psi^J(x) \psi^J(x)' \mu(x) dx \), and

\[
\hat{A} = n[\Psi'B(B'B)^{-1}B'\Psi]^\top - \Psi'B(B'B)^{-1}
\]

is an estimator of \( A = [S'G_b^{-1}S]^{-1}S'G_b^{-1} \), where

\[
S = \mathbb{E}[b^K(W_i) \psi^J(X_i)^\top], \quad G_b = \mathbb{E}[b^K(W_i)b^K(W_i)^\top].
\]

We will show that this simple leave-one-out estimator \( \hat{f}_J \) can achieve the lower bound for estimating \( f(h_0) \).

### 2.3. Rate of Convergence

We first define the sieve measure of ill-posedness which, roughly speaking, measures how much the conditional expectation operator \( T : L^2(X) \mapsto L^2(W) \) with \( (Th)(w) = \mathbb{E}[h(X)|W = w] \) smoothes out \( h \). Following Blundell et al. [2007] the sieve \( L^2 \) 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 \( \Psi_J = \text{clsp} \{ \psi_1, \ldots, \psi_J \} \subset L^2(X) \) denotes the sieve spaces for the endogenous variables. We call a NPIV model (1.1)
(i) mildly ill-posed if \( \tau_j \sim j^{\zeta/d} \) for some \( \zeta > 0 \); and

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

The next assumption captures conditions that we require to obtain our rate of convergence of the estimator.

**Assumption 1.** (i) \( h_0 \in \mathcal{H}_2(p, L) \) and \( \cup J \Psi_J \) is dense in \( (\mathcal{H}_2(p, L), \| \cdot \|_\mu) \); (ii) Condition LB (iii) is satisfied; and (iii) \( \sup_{w \in \Psi} \mathbb{E}[U^2 | W = w] \leq \sigma^2 < \infty \) and \( \mathbb{E}[U^2] < \infty \).

We introduce the projections \( \Pi_J h(x) = \psi^J(x) G_\mu^{-1} \langle \psi^J, h \rangle_\mu \). Let \( \zeta_{\psi,J} = \sup_x \| G_\mu^{-1/2} \psi^J(x) \| \) and \( \zeta_{b,K} = \sup_w \| G_\mu^{-1/2} b^K(w) \| \). Denote \( \zeta_J = \max(\zeta_{\psi,J}, \zeta_{b,K}) \) for \( K = K(J) \).

**Assumption 2.** (i) \( \tau_j \zeta_J^2 \sqrt{(\log J)/n} = O(1) \); (ii) \( \zeta_J \sqrt{\log J} \| h_0 - \Pi_J h_0 \|_\mu = O(1) \).

**Assumption 3.** (i) \( \sup_{b \in \Psi, J} \| \Pi_J T - T \| \| h \|_{L^2(W)} / \| h \|_\mu = o(\tau_J^{-1}) \); (ii) There exists constant \( C > 0 \) such that for all \( j \geq 1 \): \( \tau_j \| T(h_0 - \Pi_J h_0) \|_{L^2(W)} \leq C \| h_0 - \Pi_J h_0 \|_\mu \).

In the following, we introduce the vector \( \tau^J = (\tau_1, \ldots, \tau_J) \) where \( \tau_j > 0 \) for all \( 1 \leq j \leq J \) and \( \tau_j \leq \tau_{j'} \) for all \( 1 \leq j < j' \leq J \). For a \( r \times c \) matrix \( A \) with \( r \leq c \) and full row rank \( r \) we let \( A_\dagger \) denote its left pseudoinverse, namely \( (A^t A)^{-1} A^t \) where \( ^t \) denotes transpose and \( - \) denotes generalized inverse. Below, \( \| \cdot \| \) denotes the vector \( \ell_2 \) norm when applied to vectors and the operator norm induced by the vector \( \ell_2 \) norm when applied to matrices.

**Assumption 4.** \( \| \text{diag}(\tau^J)^{-1} (G_\mu^{-1/2} S G_\mu^{-1/2})^\top_i \| \leq D \) for some constant \( D > 0 \).

**Discussion of Assumptions:** Assumption 1 (i) imposes a regularity condition on the structural function \( h_0 \). Assumption 2 imposes bounds on the growth of the basis functions which are known for commonly used bases. For instance, \( \zeta_b = O(\sqrt{K}) \) and \( \zeta_{\psi} = O(\sqrt{J}) \) for (tensor-product) polynomial spline, wavelet and cosine bases, and \( \zeta_b = O(K) \) and \( \zeta_{\psi} = O(J) \) for (tensor-product) orthogonal polynomial bases. Assumption 3 (i) is a mild condition on the approximation properties of the basis used for the instrument space. In fact, \( \| (\Pi_K T - T) h \|_{L^2(W)} = 0 \) for all \( h \in \Psi_J \) when the basis functions for \( \text{clsp}\{b_1, \ldots, b_K\} \) and \( \Psi_J \) 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 Efroymovich 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.

The next result provides the rate of convergence of our estimator $\hat{f}_J$.

**Theorem 2.1.** Let Assumptions 1–4 be satisfied and assume $J \sim K$. Then, we have

$$\left| \hat{f}_J - f(h_0) \right|^2 = O_p \left( \frac{\tau^2_J J}{n^2} + \frac{1}{n} \left( \sum_{j=1}^J \tau^2_j <h_0, \tilde{\psi}_j>^2_n + \tau^2_J \sum_{j> J} <h_0, \tilde{\psi}_j>^2_n \right) + J^{-4p/d} \right). \quad (2.3)$$

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

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

2. **Severely ill-posed case:** choosing

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

implies

$$\left| \hat{f}_J - f(h_0) \right| = O_p \left( (\log n)^{-2p/\zeta} \right). \quad (2.5)$$

Theorem 2.1 provides an extension of Breunig and Chen [2020, Theorem F.1] under Assumption 4. Theorem 2.1 also provides concrete convergence rates of $\hat{f}_J$ once $J$ is chosen optimally in either the mildly ill-posed case or the severely ill-posed case. 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 $\zeta$, either the first or the second variance term is dominating which then leads to the so-called elbow phenomenon. In particular, the simple estimator $\hat{f}_J$ of Breunig and Chen [2020] achieves the lower bound rate in Lemma 2.1 and is thus, minimax rate optimal.

### 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 Lepski’s method; see Lepski [1990], Lepski and
Spokoiny [1997] and Lepski et al. [1997] for detailed descriptions of the method.

The oracle choice of the dimension parameter is given by

$$J_0 = \min \left\{ J \in \mathbb{N} : J^{-2p/d} \leq C_0 V(J) \right\}$$

where $V(J) = \tau J^{2 \sqrt{J \log n}} n^{-3}$ (3.1)

for some constant $C_0 > 0$. The dimension parameter $J_0$ characterizes the optimal choice of the dimension parameter given a-priori knowledge of smoothness of $h_0$ and the degree of ill-posedness of our estimation problem. We thus denote the $\hat{f}_{J_0}$ as the oracle estimator.

Our data driven choice $\hat{J}$ of sieve dimension is defined as follows

$$\hat{J} = \min \left\{ J \in \hat{I} : |\hat{f}_J - \hat{f}_{J'}| \leq c_0 (\hat{V}(J) + \hat{V}(J')) \right\}$$

for some constant $c_0 > 0$, some index set $\hat{I}$ to be introduced below, and where

$$\hat{V}(J) = \hat{\tau} J^{2 \sqrt{J \log n}} n^{-3}$$

and an estimator of the sieve measure of ill-posedness $\tau_J$ is given by

$$\hat{\tau}_J := \left[ s_{\min}((B'B/n)^{-1/2}(B'\Psi/n)G_{\mu}^{-1/2}) \right]^{-1}$$

where $s_{\min}(\cdot)$ is the minimal singular value.

Since our quadratic functional estimator is the simple leave-one-out sieve NPIV estimator, for simplicity we use the same index set as that in Chen and Christensen [2015] for their adaptive sieve NPIV estimation of $h_0$:

$$\hat{I} = \{ J : J_{\text{min}} \leq J \leq \hat{J}_{\text{max}} \},$$

where $J_{\text{min}} = \lfloor \log \log n \rfloor$ and

$$\hat{J}_{\text{max}} = \min \left\{ J > J_{\text{min}} : \hat{\tau}_J [\zeta(J)]^2 \sqrt{\ell(J)(\log n)}/n \geq 1 \right\}$$

where $\ell(J) = 0.1 \log \log J$, and $\zeta(J) = \sqrt{J}$ for spline, wavelet, or trigonometric sieve basis, and $\zeta(J) = J$ for orthogonal polynomial basis.

Since the definition of the feasible index set $\hat{I}$ relies on an upper bound $\hat{J}_{\text{max}}$, we follow Chen and Christensen [2015] to use a dimension parameter $\overline{J}$ to control the complexity of $\hat{I}$, which is allowed to grow slowly with the sample size $n$. previously, Chen and Christensen [2015, Theorem 3.2] establish that $\hat{J}_{\text{max}} \leq \overline{J}$ with probability
approaching one, where $\mathcal{T}$ satisfies the rate restrictions imposed in the next assumption. We also make use of the notation $\zeta = \zeta_{\mathcal{T}}$.

**Assumption 5.** $\tau_{\mathcal{T}} \zeta^2 (\log n)^2 / \sqrt{n} = o(1)$ and $\mathcal{T}^{2+\varepsilon} = O(n)$ for some $\varepsilon > 0$.

Assumption 5 is a uniform version of the rate conditions imposed in Assumption 2 (i). The next result establishes an upper bound for the adaptive estimator $\hat{f}_{\mathcal{T}}$.

**Theorem 3.1.** Let Assumptions 1–5 be satisfied. Then, we have

$$|\hat{f}_{\mathcal{T}} - f(h_0)| \leq 3 c_0 \tau_{h_0}^2 \sqrt{J_0 \log n / n} + |\hat{f}_{h_0} - f(h_0)|$$

with probability approaching one.

The proof of Theorem 3.1 is based on a Bernstein-type inequality for canonical U-statistics. Also note that Theorem 2.1 provides an upper bound for $|\hat{f}_{h_0} - f(h_0)|$.

The following result illustrates the general upper bound for the mildly and severely ill-posed case.

**Corollary 3.1.** Let Assumptions 1–5 be satisfied. Then, we have in the

1. **Mildly ill-posed case:**

$$|\hat{f}_{\mathcal{T}} - f(h_0)| = \begin{cases} O_p \left(\left(\sqrt{\log n / n}\right)^{4p/(4(p + \zeta + d))}\right), & \text{if } p \leq \zeta + d/4, \\ O_p \left(n^{-1/2}\right), & \text{if } p > \zeta + d/4. \end{cases} \quad (3.3)$$

2. **Severely ill-posed case:**

$$|\hat{f}_{\mathcal{T}} - f(h_0)| = O_p \left((\log n)^{-2p/\zeta}\right). \quad (3.4)$$

Corollary 3.1 shows that our data-driven choice of tuning parameter 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 extra $\sqrt{\log n}$ factor for the irregular mildly ill-posed case. We note that when $\zeta = 0$ 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 findings of Efroymovich and Low [1996] who showed that one must pay a factor of $\sqrt{\log n}$ penalty in adaptive estimation of a quadratic functional of unknown density and regression functions when $p \leq d/4$. 

11
4. Conclusion

In this paper we first show that the simple leave-one-out, sieve NPIV estimator of the quadratic functional proposed in Breunig and Chen [2020] is minimax rate optimal. We then propose an adaptive leave-one-out sieve NPIV estimator of the quadratic functional based on Lepski’s method. We show that the adaptive estimator achieves the minimax optimal rate for the severely ill-posed case as well as 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.

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 quadratic functional of a NPIV function will perform well in finite samples. Implementation of our data driven method still relies on choice of a calibration constant. To improve finite sample performance over the original Lepski method, Spokoiny and Vial [2009] offered a propagation approach, and Spokoiny and Willrich [2019] proposed a bootstrap calibration of Lepski’s method. We believe that the bootstrap calibration will be helpful in the NPIV setting as well.

A leading nonlinear ill-posed inverse problems with unknown operator is the nonparametric quantile instrumental variables (NPQIV) model:

$$Y = h_0(X) + U, \quad \mathbb{E}[\mathbb{1}\{U \leq 0\}|W] = \tau \in (0, 1).$$

Recently Dunker [2020] considers adaptive estimation of the NPQIV function $h_0$ in $L^2$ norm under an a stronger assumption that $U$ and $W$ are fully independent, without establishing minimax rate optimality. Previously Chen and Pouzo [2015] presented the estimation and inference on possibly nonlinear functionals of a NPQIV function $h_0$. It will be interesting to study the minimax rate adaptive estimation of a quadratic functional in a NPQIV model.
A. Proofs of Results in Section 2

We define \( s_J = s_{\min}(G_b^{-1/2}SG_{\mu}^{-1/2}) \) which satisfies

\[
s_J^{-1} = \sup_{h \in \Psi_J, \mu \neq 0} \frac{\|h\|_{\mu}^2}{\|\Pi_K T h\|_{L^2(W)}} \geq \tau_J \tag{A.1}
\]

for all \( K = K(J) \geq J > 0 \). Indeed, we note that

\[
s_J = \inf_{h \in \Psi_J} \frac{\|\Pi_K T h\|_{L^2(W)}}{\|h\|_{\mu}} \geq \inf_{h \in \Psi_J} \frac{\|\Pi_K T - T\|_{L^2(W)}}{\|h\|_{\mu}} = (1 - o(1))\tau_J^{-1} \tag{A.2}
\]

by Assumption 3.

In the following we let \( Q_J h(x) = \psi^J(x)'[S'G_b^{-1}S]^{-1}S'G_b^{-1} E[b^{K(J)}(W)h(X)] \). For a \( r \times c \) matrix \( A \) with \( r \leq c \) and full row rank \( r \) we let \( A^\dagger \) denote its left pseudoinverse, namely \((A^\prime A)^{-1}A^\prime \) where \( ^\prime \) denotes transpose and \( - \) denotes generalized inverse. Thus, we can write

\[
Q_J h(x) = \psi^J(x)'(G_b^{-1/2}S)\tilde{G}_b^{-1/2} E[b^{K(J)}(W)h(X)]
\]

\[
= \psi^J(x)'G_{\mu}^{-1/2}(G_b^{-1/2}SG_{\mu}^{-1/2})\tilde{G}_b^{-1/2} E[b^{K(J)}(W)h(X)].
\]

**Proof of Theorem 2.1.** Proof of (2.3). Under Assumptions 1–3, we may apply Breunig and Chen [2020, Theorem F.1] which yields

\[
\|f_J - f(h_0)\| = O_p \left(n^{-1/2} \sqrt{\tau} + n^{-1/2} \left\| \langle Q_J h_0, \psi^J \rangle_{\mu}(G_b^{-1/2}S)\tilde{G}_b^{-1/2} \right\| + J^{-2p/4} \right).
\]

In the remainder of this proof we bound the second summand on the right hand side.

Let us define \( \tilde{\psi}^J = G_\mu^{-1/2} \psi^J \) and \( \tilde{b}^{K} = G_\mu^{-1/2}b^{K} \). From the definition of \( Q_J h(x) = \psi^J(x)'(G_b^{-1/2}S)\tilde{G}_b^{-1/2} E[\tilde{b}^{K(J)}(W)h(X)] \) we infer \( \langle Q_J h_0, \tilde{\psi}^J \rangle_{\mu} = (G_b^{-1/2}SG_{\mu}^{-1/2})\tilde{G}_b^{-1/2} E[\tilde{b}^{K(J)}(W)h(X)] \). We thus calculate

\[
\left\| \langle Q_J h_0, \psi^J \rangle_{\mu}(G_b^{-1/2}S)\tilde{G}_b^{-1/2} \right\| = \left\| \langle Q_J h_0, \tilde{\psi}^J \rangle_{\mu}(G_b^{-1/2}SG_{\mu}^{-1/2})\tilde{G}_b^{-1/2} \right\|
\]

\[
\leq \left\| \text{diag}(\tau^J)(G_b^{-1/2}SG_{\mu}^{-1/2})\tilde{G}_b^{-1/2} E[\tilde{b}^{K(J)}(W)h_0(X)] \right\|
\]

\[
\times \left\| \text{diag}(\tau^J)^{-1}(G_b^{-1/2}SG_{\mu}^{-1/2})\tilde{G}_b^{-1/2} \right\|
\]

where \( \left\| \text{diag}(\tau^J)^{-1}(G_b^{-1/2}SG_{\mu}^{-1/2})\tilde{G}_b^{-1/2} \right\| \leq D \) due to Assumption 4. Consequently, the
definition of the projection $\Pi_J h(x) = \psi^J(x)'G^{-1}_\mu \langle \psi^J, h \rangle_\mu$ yields

$$
\| \langle Q_J h_0, \psi^J \rangle_\mu (G^{-1/2}_b S_i^-) \| \leq D \| \text{diag}(\tau^J) (G^{-1/2}_b S G^{-1/2}_\mu)_i^- \mathbb{E} [\tilde{b}^{K_J}(W) h_0(X)] \| \\
\leq D \| \text{diag}(\tau^J) (\tilde{\psi}^J, h_0)_\mu \| \\
+ D \| \text{diag}(\tau^J) (G^{-1/2}_b S G^{-1/2}_\mu)_i^- \mathbb{E} [\tilde{b}^{K_J}(W)(h_0(X) - \Pi_J h_0(X)))] \| \\
\leq D \sqrt{\sum_{j=1}^{J} \tau^2_j \langle h_0, \tilde{\psi}_j \rangle^2_\mu} \\
+ D^2 \tau^2_J \| \mathbb{E} [\tilde{b}^{K_J}(W)(h_0(X) - \Pi_J h_0(X))] \|
$$

where we used again Assumption 4 in the last inequality. Consider the second term on the right hand side. We obtain

$$
\tau^2_J \| \mathbb{E} [\tilde{b}^{K_J}(W)(h_0(X) - \Pi_J h_0(X))] \| = \tau^2_J \| \Pi_K T(h_0 - \Pi_J h_0) \|_{L^2(W)} \\
\leq \tau^2_J \| T(h_0 - \Pi_J h_0) \|_{L^2(W)} \\
= O(\tau_J \| h_0 - \Pi_J h_0 \|_\mu) \\
= O \left( \tau_J \sqrt{\sum_{j>\tilde{j}} \langle h_0, \tilde{\psi}_j \rangle^2_\mu} \right)
$$

by making use of Assumption 3 (ii).

Proof of (2.4). The choice of $J \sim n^{2d/(4(p+\zeta)+d)}$ implies

$$
n^{-2} \tau^4_J \sim n^{-2} J^{1+4\zeta/d} \sim n^{-8p/(4(p+\zeta)+d)}
$$

and for the bias term

$$
J^{-4p/d} \sim n^{-8p/(4(p+\zeta)+d)}.
$$

We now distinguish between the two regularity cases of the result. First, consider the case $p \leq \zeta + d/4$, where the function $j \mapsto j^{2(\zeta-p)/d+1/2}$ is increasing and consequently,
we observe

\[ n^{-1} \sum_{j=1}^{J} \langle h_0, \tilde{\psi}_j \rangle^2 \sim n^{-1} \sum_{j=1}^{J} \langle h_0, \tilde{\psi}_j \rangle^2 J^{2p/d - 1/2} j^{2(\zeta - p)/d + 1/2} \]

\[ \lesssim J^{2(\zeta - p)/d + 1/2} n^{-1} \sum_{j=1}^{J} \langle h_0, \tilde{\psi}_j \rangle^2 J^{2p/d - 1/2} \]

\[ \lesssim J^{2(\zeta - p)/d + 1/2} n^{-1} \sim n^{-8p/(4(p+\zeta)+d)}. \]

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

\[ n^{-1} \sum_{j > J} \langle h_0, \tilde{\psi}_j \rangle^2 \lesssim n^{-1} J^{2(\zeta - p)/d} \sum_{j > J} \langle h_0, \tilde{\psi}_j \rangle^2 j^{2p/d} \]

\[ \lesssim n^{-1} J^{2(\zeta - p)/d} \lesssim n^{-8p/(4(p+\zeta)+d)}. \]

Finally, it remains to consider the case \( p > \zeta + d/4 \). In this case, we have that

\[ \sum_{j=1}^{J} \langle h_0, \tilde{\psi}_j \rangle^2 \sim \sum_{j=1}^{J} \langle h_0, \tilde{\psi}_j \rangle^2 j^{2p/d} \]

\[ = O(1) \]

and consequently, the second variance term satisfies \( n^{-1} \| (QJh_0, \psi^J) \mu (G_b^{-1/2}S) \|^2 = O(n^{-1}) \) which is the dominating rate and thus, completes the proof of the result.

Proof of (2.5). The choice of

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

implies

\[ n^{-2} J^{-2} \sim n^{-2} J \exp(2J^{\zeta/d}) \]

\[ \sim \left( \log n - \frac{4p + d}{2\zeta} \log \log n \right)^{d/\zeta} (\log n)^{-(4p+d)/\zeta} \]

\[ \sim (\log n)^{d/\zeta} (\log n)^{-(4p+d)/\zeta} \]

\[ \sim (\log n)^{-4p/\zeta} \]

15
and $J^{-4p/d} \sim (\log n)^{-4p/\zeta}$. Moreover, since the function $j \mapsto j^{-2p/d} \exp(j^{\zeta/d})$ is increasing we obtain

$$\frac{1}{n} \sum_{j=1}^{J} \langle h_{0}, \tilde{\psi}_{j} \rangle_{\mu}^{2} \tau_{j}^{2} \sim \frac{1}{n} \sum_{j=1}^{J} \langle h_{0}, \tilde{\psi}_{j} \rangle_{\mu}^{2} J^{2p/d} j^{-2p/d} \exp(j^{\zeta/d})$$

$$\lesssim J^{-2p/d} \exp(J^{\zeta/d}) \frac{1}{n} \sum_{j=1}^{J} \langle h_{0}, \tilde{\psi}_{j} \rangle_{\mu}^{2} J^{2p/d}$$

$$\sim (\log n)^{-2p/\zeta} (\log n)^{-(2p+d)/\zeta}$$

and finally

$$\frac{1}{n} \sum_{j>J} \langle h_{0}, \tilde{\psi}_{j} \rangle_{\mu}^{2} \sim \frac{1}{n} \exp(J^{\zeta/d}) \sum_{j>J} \langle h_{0}, \tilde{\psi}_{j} \rangle_{\mu}^{2} J^{2p/d}$$

$$\lesssim n^{-1} J^{-2p/d} \exp(J^{\zeta/d})$$

$$\lesssim (\log n)^{-4p/\zeta},$$

which shows the result.

\[ \square \]

**B. Proofs of Results in Section 3**

Below, we make use of the notation $J_{0} = \{ J \in \mathbb{N} : J^{-2p/d} \leq C_{0} V(J) \}$ and

$$\hat{J} = \left\{ J \in \hat{I} : |\hat{f}_{J} - \hat{f}_{J'}| \leq c_{0}(\hat{V}(J) + \hat{V}(J')) \text{ for all } J' \in \hat{I} \text{ with } J' \geq J \right\}$$

and recall the definition $\hat{K} = \{ J : J_{\min} \leq J \leq \hat{J}_{\max} \}$. Below, we abbreviate “with probability approaching one” to “wpa1”. We make use of the notation $\underline{J} = \min \{ J > J_{\min} : 4\tau J\zeta^{2}(J) \sqrt{\ell(J)(\log n)/n} \geq 1 \}$. We introduce the set $\mathcal{E}_{n}^{*} = \{ J_{0} \in \hat{J} \} \cap \{ |\hat{\tau}_{j}^{-1} - s_{j}| \leq \eta s_{j} \text{ for all } J \leq J \leq J' \}$ for some $\eta \in (0, 1 - (2/3)^{1/4})$. By Lemma C.3 and C.5 it holds $\mathbb{P}(\mathcal{E}_{n}^{*}) = 1 + o(1)$.

**Proof of Theorem 3.1.** Due to Theorem 3.2 of Chen and Christensen [2015] we may assume that $\underline{J} \leq \hat{J}_{\max} \leq J$ on $\mathcal{E}_{n}^{*}$. The definition $\hat{J} = \min_{J \in \hat{J}} J$ implies $\hat{J} \leq J_{0}$.
on the set $\mathcal{E}_n^*$ and hence, we obtain

$$|\hat{f}_j - f(h_0)| \1_{\mathcal{E}_n^*} \leq |\hat{f}_j - J_0| \1_{\mathcal{E}_n^*} + |\hat{f}_j - f(h_0)| \1_{\mathcal{E}_n^*} \leq c_0 \left( \hat{V}(\hat{J}) + \hat{V}(J_0) \right) \1_{\mathcal{E}_n^*} + |\hat{f}_j - f(h_0)|.$$  

On the set $\mathcal{E}_n^*$, we have $|\hat{r}_j^1 - s_j| \leq \eta s_j$ which implies $\hat{r}_j^2 \leq s_j^2 (1 - \eta)^{-2}$ and thus, by the definition of $\hat{V}(\cdot)$ in (3.2) we have

$$|\hat{f}_j - f(h_0)| \1_{\mathcal{E}_n^*} \leq c_0 (1 - \eta)^{-2} (s_j^2 \hat{J} + s_j^2 J_0) \1_{\mathcal{E}_n^*} \sqrt{\log n/n} + |\hat{f}_j - f(h_0)|.$$  

Using inequality (A.2) we have $s_j^2 \leq (1 - \eta)^{-2} \tau_j^2$ for $J \in \{J_0, \hat{J}\}$ and $n$ sufficiently large, using that $\hat{J} \geq J_{\text{min}} = [\log \log n]$. Consequently, from the definition of $V(\cdot)$ in (3.1) we infer:

$$|\hat{f}_j - f(h_0)| \1_{\mathcal{E}_n^*} \leq c_0 (1 - \eta)^{-2} \left( \tau_j^2 \hat{J} + \tau_j^2 J_0 \right) \1_{\mathcal{E}_n^*} \sqrt{\log n/n} + |\hat{f}_j - f(h_0)| \leq c_0 (1 - \eta)^{-4} V(\hat{J}) + V(J_0) \1_{\mathcal{E}_n^*} + |\hat{f}_j - f(h_0)|$$

for $n$ sufficiently large, where the last inequality is due to $V(\hat{J}) \1_{\mathcal{E}_n^*} \leq V(J_0)$ since $\hat{J} \leq J_0$ on $\mathcal{E}_n^*$. Since $\eta \in (0, 1 - (2/3)^{1/4})$ we obtain $2(1 - \eta)^{-4} \leq 3$. By Lemma C.3 and C.5 it holds $\mathbb{P}(\mathcal{E}_n^*) = 1 + o(1)$ which completes the proof. \hfill \Box

**Proof of Corollary 3.1.** Proof of (3.3). The definition of the oracle choice in (3.1) implies $J_0 \sim (n/\sqrt{\log n})^{2d/(4(p+\zeta)+d)}$ in the mildly ill-posed case. Thus, we obtain

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

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

$$n^{-1} \sum_{j=1}^{J_0} \tau_j^2 \langle h_0, \psi_j \rangle^2 \leq J_0^{2(p-\zeta)/d+1/2} n^{-1} \lesssim (\sqrt{\log n/n})^{8p/(4(p+\zeta)+d)}.$$
Moreover, using $h_0 \in H_2(p, L)$, i.e., $\sum_{j \geq 1} \langle h_0, \tilde{\psi}_j \rangle^2 \mu j^{2p/d} \leq L$, we obtain
\[ n^{-1} \tau_{j_0}^2 \sum_{j > j_0} \langle h_0, \tilde{\psi}_j \rangle^2 \mu \lesssim (\sqrt{\log n/n})^{8p/(4(p+\zeta)+d)}. \]

Finally, it remains to consider the case $p > \zeta + d/4$, where as in the proof of Theorem 2.1 we have
\[ \sum_{j=1}^{j_0} \tau_j^2 \langle h_0, \tilde{\psi}_j \rangle^2 \mu = O(1), \]
implying $n^{-1} \| Q_{j_0} h_0, \psi^{j_0} \mu (G_b^{-1/2} S)^r_l \|^2 = O(n^{-1})$ which is the dominating rate and thus, completes the proof of the result.

Proof of (3.4). In the severely ill-posed case, the definition of the oracle choice in (3.1) implies
\[ J_0 \sim \left( \log(n/\sqrt{\log n}) - \frac{4p + d}{2\zeta} \log \log(n/\sqrt{\log n}) \right)^{d/\zeta} \]
and ensures that the variance and squared bias term are of the same order. Specifically, we obtain
\[ n^{-2} (\log n) \tau_{j_0}^4 J_0 \sim n^{-2} (\log n) J_0 \exp(2J_0^{\zeta/d}) \]
\[ \sim \left( \log(n/\sqrt{\log n}) \right)^{d/\zeta} \left( \log(n/\sqrt{\log n}) \right)^{-(4p+d)/\zeta} \]
\[ \sim \left( \log n - \frac{1}{2} \log \log n \right)^{-4p/\zeta} \]
\[ \sim (\log n)^{-4p/\zeta} \]
and $J_0^{-4p/d} \sim (\log n)^{-4p/\zeta}$. Moreover, since the function $j \mapsto j^{-2p/d} \exp(j^{\zeta/d})$ is increasing we obtain
\[ n^{-1} \sum_{j=1}^{j_0} \tau_j^2 \langle h_0, \tilde{\psi}_j \rangle^2 \mu \lesssim (\log n)^{-4p/\zeta}. \]
and finally
\[ n^{-1} \tau_{j_0}^2 \sum_{j > j_0} \langle h_0, \tilde{\psi}_j \rangle^2 \mu \lesssim (\log n)^{-4p/\zeta}, \]
which completes the proof. □

C. Supplementary Lemmas

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

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

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

$$R_1(Z_i, Z_{i'}) = U_i \mathbb{1}_{M_i} b^{K}(W_i)' A' G \mu A b^{K}(W_{i'}) U_{i'} \mathbb{1}_{M_{i'}}$$

$$- \mathbb{E}[U_i \mathbb{1}_{M_i} b^{K}(W_i)' A' G \mu A \mathbb{E}[b^{K}(W_{i'}) U_{i'} \mathbb{1}_{M_{i'}}]] \quad (C.1)$$

where $M_i = \{|U_i| \leq M_n\}$ with $M_n = \bar{\xi}^{-1} \sqrt{n}/(\log J)$. 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

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

where the kernel $R_2$ is given by

$$R_2(Z_i, Z_{i'}) = U_i \mathbb{1}_{M_i} b^{K}(W_i)' A' G \mu A b^{K}(W_{i'}) U_{i'} \mathbb{1}_{M_{i'}}$$

$$- \mathbb{E}[U_i \mathbb{1}_{M_i} b^{K}(W_i)' A' G \mu A \mathbb{E}[b^{K}(W_{i'}) U_{i'} \mathbb{1}_{M_{i'}}]].$$

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

$$\mathbb{P}_h \left( \left| \sum_{1 \leq i < i' \leq n} R(Z_i, Z_{i'}) \right| \geq C \left( \Lambda_1 \sqrt{u} + \Lambda_2 u + \Lambda_3 u^{3/2} + \Lambda_4 u^2 \right) \right) \leq 6 \exp(-u)$$
where
\[
\Lambda_1^2 = \frac{n(n-1)}{2} \mathbb{E}[R^2(Z_1, Z_2)],
\]
\[
\Lambda_2 = n \sup_{\|\nu\|_{L^2(Z)} \leq 1} \mathbb{E}[R(Z_1, Z_2)\nu(Z_1)\kappa(Z_2)],
\]
\[
\Lambda_3 = \sqrt{n} \sup_z |\mathbb{E}[R^2(Z_1, z)]|,
\]
\[
\Lambda_4 = \sup_{z_1, z_2} |R(z_1, z_2)|.
\]

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 (iii) be satisfied. Given kernel \(R = R_1\), it holds:
\[
\Lambda_1^2 \leq \frac{\sigma^4 n(n-1)}{2} J s_J^{-4}
\] (C.2)
\[
\Lambda_2 \leq 2\sigma^2 n s_J^{-2}
\] (C.3)
\[
\Lambda_3 \leq \sigma^2 \sqrt{n} M_n \zeta_{b,K} s_J^{-2}
\] (C.4)
\[
\Lambda_4 \leq M_n^2 \zeta_{b,K}^2 s_J^{-2}.
\] (C.5)

**Proof.** The result is due to Lemma F.1 and Lemma G.2 of Breunig and Chen [2020]. \(\square\)

**Lemma C.3.** Let Assumption 5 be satisfied. Then, for \(\eta \in (0, 1)\) it holds \(|\hat{\tau}_J^{-1} - s_J| \leq \eta s_J\) for all \(\underline{J} \leq J \leq \overline{J}\) wpa1.

**Proof.** The result is due to Lemma E.16 of Chen and Christensen [2015]. \(\square\)

**Lemma C.4.** Let Assumptions 1–5 be satisfied and let \(\underline{J} \leq \hat{J}_{\text{max}} \leq \overline{J}\) hold wpa1. Then, uniformly for all \(J_{\text{min}} \leq J \leq \overline{J}\) it holds
\[
|\hat{f}_J - f(Q_J h_0)| \leq 2\sigma^2 s_J^{-2} \sqrt{J \log n} \frac{\sqrt{\log n}}{n - 1} =: c_n(J)
\] (C.6) wpa1.

**Proof.** First, observe that by making use of Assumption 5, i.e., \(\tau J^{-2 \sqrt{\log n} / n} = o(1)\), it holds \(c_n(J) = o(1)\) uniformly in \(\underline{J} \leq J \leq \overline{J}\). Following the proof of Breunig
and Chen [2020, Theorem F.1] we have the decomposition

\[
\hat{f}_J - f(Q_J h_0) = U_{n,1}(J) + U_{n,2}(J) + \frac{2}{n(n-1)} \sum_{i<i'} U_i U_{i'} b^{K(J)}(W_i)' (A' G_{\mu} A - \hat{A}' G_{\mu} \hat{A}) b^{K(J)}(W_{i'}). \\
=: d_{n,3}(J)
\]

We obtain

\[
P\left(\max_{J \leq J} \left\{ c_n(J)^{-1} | \hat{f}_J - f(Q_J h_0) | \right\} > 1 \right) \leq P\left(\max_{J \leq J} \left\{ c_n(J)^{-1} | U_{n,1}(J) | \right\} > 1/3 \right) + P\left(\max_{J \leq J} \left\{ c_n(J)^{-1} | U_{n,2}(J) | \right\} > 1/3 \right) + P\left(\max_{J \leq J} \left\{ c_n(J)^{-1} | U_{n,3}(J) | \right\} > 1/3 \right) \leq: T_1 + T_2 + T_3.
\]

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

\[
\Lambda_1 \sqrt{u} + \Lambda_2 u + \Lambda_3 u^{3/2} + \Lambda_4 u^2 \\
\leq \Lambda_1 \sqrt{2 \log J} + 2 \Lambda_2 \log J + \Lambda_3 (2 \log J)^{3/2} + 4 \Lambda_4 (\log J)^2 \\
\leq \sigma^2 \nu s_J^{-2} \sqrt{J \log n} + 4 \sigma^2 \nu s_J^{-2} \log n + \sigma^2 \nu s_J^{-2} \sqrt{2 \log n} + 4 \nu s_J^{-2} \\
\frac{2}{\sqrt{n}} c_n(J)
\]

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

\[
\Lambda_1 \sqrt{u} + \Lambda_2 u + \Lambda_3 u^{3/2} + \Lambda_4 u^2 \leq 2 \sigma^2 \nu s_J^{-2} \sqrt{J \log n} = \frac{n(n-1)}{2} c_n(J)
\]

21
by the definition of $c_n(J)$. Consequently, Lemma C.1 with $u = 2 \log J$ yields

$$T_1 \leq \sum_{J \leq \bar{J}} \mathbb{P} \left( \left| \sum_{i < i'} R_i(Z_i, Z_{i'}) \right| \geq \frac{n(n-1)}{2} c_n(J) \right)
\leq 6 \exp \left( (\log \bar{J}) - 2(\log \bar{J}) \right)
\leq 6 \frac{J}{\bar{J}}
= o(1)$$

where the last equation is due to the lower bound $\bar{J} \geq J_{\min} = \lfloor \log \log n \rfloor$.

Consider $T_2$. We first observe for all $J \leq \bar{J}$ that

$$\left| \frac{2}{n(n-1)} \sum_{i < i'} U_i 1_{M_i} b^{K(J)}(W_i)' A' G_{\mu} b^{K(J)}(W_{i'}) U_{i'} 1_{M_{i'}} \right|
\leq \frac{2}{n(n-1)} \sum_{i < i'} |U_i 1_{M_i} | |U_{i'} 1_{M_{i'}} | \| \zeta_2^2 \| \| G^{1/2}_\mu A G^{1/2}_b \| ^2.$$  

For all $J \geq 1$ such that $s_J = s_{\min}(G^{-1/2}_b S G^{-1/2}_\mu) > 0$ it holds

$$\left\| G^{1/2}_\mu A G^{1/2}_b \right\| = \left\| G^{1/2}_\mu [S' G^{-1}_b S]^{-1} S' G^{-1/2}_b \right\|
= \left\| G^{-1/2}_\mu S' G^{-1/2}_b S G^{-1/2}_\mu \right\|
= \left\| \left( G^{-1/2}_b S G^{-1/2}_\mu \right)^{-1} \right\|
= s_J^{-1}.$$  

Consequently, using that $M_n = \zeta^{-1} \sqrt{n}/(\log \bar{J})$ we obtain by Markov’s inequality and the definition of $c_n(J)$ that

$$T_2 \leq 6 \mathbb{E} |U 1_{\{|U| > M_n\}}| \mathbb{E} |U 1_{\{|U| > M_n\}}| \max_{J \leq \bar{J}} \frac{\zeta_2^2}{s_J^2 c_n(J)}
\leq 6 M_n^{-6} \left( \mathbb{E}[U^4] \right)^2 \max_{J \leq \bar{J}} \frac{\zeta_2^2}{s_J^2 c_n(J)}
= O \left( n^{-2} (\log \bar{J})^6 \zeta^8 \bar{J}^4 \sqrt{\log n} \right)
= o(1)$$

where the last equation is due to the rate condition imposed in Assumption 5.
Consider $T_3$. We make use of the inequality
\[
\sum_{i \neq i'} U_i U_{i'} b^{K(J)}(W_i) \left( A'G_\mu A - \hat{A}'G_\mu \hat{A} \right) b^{K(J)}(W_{i'}) \\
\leq \left\| \sum_{i} U_i \tilde{b}^{K(J)}(W_i) \right\|^2 \left\| G_b^{1/2} (A'G_\mu A - \hat{A}'G_\mu \hat{A}) G_b^{1/2} \right\| \\
= \sum_{i \neq i'} U_i U_{i'} \tilde{b}^{K(J)}(W_i) \tilde{b}^{K(J)}(W_{i'}) \left\| G_b^{1/2} (A'G_\mu A - \hat{A}'G_\mu \hat{A}) G_b^{1/2} \right\| \\
+ \frac{1}{n} \sum_i \left\| U_i \tilde{b}^{K(J)}(W_i) \right\|^2 \left\| G_b^{1/2} (A'G_\mu A - \hat{A}'G_\mu \hat{A}) G_b^{1/2} \right\|.
\]
Note that $n^{-1} \sum_i \left\| U_i \tilde{b}^{K(J)}(W_i) \right\|^2 \leq K(J) + o_p(1)$ uniformly in $J$. Further, the definition $c_n(J) = 2\sigma^2 s_J^{-2} \sqrt{J \log n/(n-1)}$ implies
\[
T_3 = \mathbb{P} \left( \max_{J \leq \bar{J}} \frac{\sigma^2}{n \sqrt{J \log n}} \sum_{i < i'} U_i U_{i'} b^{K(J)}(W_i) \left( A'G_\mu A - \hat{A}'G_\mu \hat{A} \right) b^{K(J)}(W_{i'}) > \frac{1}{3\sigma^2} \right) \\
\leq \mathbb{P} \left( \max_{J \leq \bar{J}} \frac{\sigma^2}{n \sqrt{J \log n}} \sum_{i \neq i'} U_i U_{i'} \tilde{b}^{K(J)}(W_i) \tilde{b}^{K(J)}(W_{i'}) > \frac{1}{6\sigma^2} \right) \\
+ \mathbb{P} \left( \max_{J \leq \bar{J}} \left( \| G_b^{1/2} (A'G_\mu A - \hat{A}'G_\mu \hat{A}) G_b^{1/2} \| \right) > 1 \right) \\
+ \mathbb{P} \left( \max_{J \leq \bar{J}} \left( K(J) \| G_b^{1/2} (A'G_\mu A - \hat{A}'G_\mu \hat{A}) G_b^{1/2} \| \right) > \frac{1}{6\sigma^2} \right) + o(1) \\
=: T_{31} + T_{32} + T_{33} + o(1).
\]
Now, $T_{31} = o(1)$ directly by following the first step in the proof of Breunig and Chen [2020, Theorem 3.3]. Finally, $T_{32} = o(1)$ and $T_{33} = o(1)$ by following the proof of Chen and Christensen [2015, Lemma E.16].

\[\underline{\text{Lemma C.5.}}\]
Let Assumptions 1–3 be satisfied and let $\underline{J} \leq \hat{J}_{\max} \leq \bar{J}$ hold wpa1. Then, we have $\mathbb{P}(J_0 \in \hat{J}) = 1 + o(1)$.

Proof. Let $\mathcal{E}_n$ denote the event upon which $\underline{J} \leq \hat{J}_{\max} \leq \bar{J}$ and observe that $\mathbb{P}(\mathcal{E}_n^c) = o(1)$ by hypothesis. The triangular inequality yields
\[
| \hat{J}_h - \hat{f}_J | \leq | \hat{f}_h - f(Q_j h_0) | + | \hat{f}_J - f(Q_j h_0) | + | f(Q_j h_0) - f(h_0) | + | f(Q_j h_0) - f(h_0) |.
\]
By Lemma C.4, uniformly for all $J_{\min} \leq J \leq \bar{J}$ it holds
\[
| \hat{f}_J - f(Q_j h_0) | \leq 2\sigma^2 s_J^{-2} \sqrt{J \log n} \frac{n-1}{n} \]
on some set $\mathcal{E}_{n,1} \subseteq \mathcal{E}_n$ where $\mathbb{P}(\mathcal{E}_{n,1}^c) = o(1)$. By the definition of $\mathcal{J}_0$ we infer for all $J \in \mathcal{J}_0$ that

$$|f(Q_Jh_0) - f(h_0)| \leq C_0 \tau_J^2 \sqrt{J \log n/n} \leq C_0 s_J^2 \sqrt{J \log n/n}$$

where the last inequality is due to (A.1). Hence, we conclude

$$|\hat{f}_{J_0} - \hat{f}_J| \leq (C_0 + 2\sigma^2) \left( s_J^{-2} \sqrt{J_0 \log n/n} + s_J^{-2} \sqrt{J \log n/n} \right).$$

Due to Lemma C.3 it holds $s_J^{-2} \leq (1 + \eta)^2 \tau_J^2$ for some $\eta \in (0, 1)$, uniformly for all $J_0 \leq J \leq J$, 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

$$|\hat{f}_{J_0} - \hat{f}_J| \leq (C_0 + 2\sigma^2)(1 + \eta)^2 \left( \tau_{J_0}^2 \sqrt{J_0 \log n/n} + \tau_J^2 \sqrt{J \log n/n} \right)$$

$$= (C_0 + 2\sigma^2)(1 + \eta)^2 \left( \hat{V}(J_0) + \hat{V}(J) \right)$$

uniformly for all $J_0 < J \leq J$. We conclude that $J_0 \in \hat{J}$ on $\mathcal{E}_{n,1} \cap \mathcal{E}_{n,2}$ and $\mathbb{P}(\mathcal{E}_{n,1} \cap \mathcal{E}_{n,2}) = 1 - o(1)$.

\section*{References}

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

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

C. Breunig and X. Chen. Adaptive, rate-optimal testing in instrumental variables models. \textit{arXiv preprint arXiv:2006.09587}, 2020.

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

C. Butucea. Goodness-of-fit testing and quadratic functional estimation from indirect observations. \textit{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.

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, adaptivity and inference in nonparametric instrumental variables estimation. *arXiv preprint arXiv:1508.03365v1*, 2015.

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 quasi likelihood ratio inference 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.

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.

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

F. Dunker. Nonparametric instrumental variable regression and quantile regression with full independence. *arXiv preprint arXiv:1511.03977v2*, 2020.

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

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

E. Giné and R. Nickl. A simple adaptive estimator of the integrated square of a density. *Bernoulli*, 14(1):47–61, 2008.
J. L. Horowitz. Applied nonparametric instrumental variables estimation. *Econometrica*, 79(2):347–394, 2011.

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. Efficient estimation of integral functionals of a density. *The Annals of Statistics*, 24(2):659–681, 1996.

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.

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.