Adapting to Endogeneity in Nonparametric Models*

Zhen Xie

Department of Economics

Northwestern University

zhenxie@u.northwestern.edu

This version: November 12, 2026
Newest version here

Abstract

This paper proposes a new method for estimating nonparametric structural functions by adapting to the unknown magnitude of endogeneity. Instrumental variables can address endogeneity, but their use may come at a substantial cost in estimation accuracy: nonparametric instrumental variables (NPIV) estimators can converge at very slow rates. By contrast, nonparametric least squares (NP-OLS) can estimate the conditional mean function at a much faster rate, but this function generally differs from the structural function under endogeneity. We formalize this trade-off as a minimax estimation problem over classes of data generating processes indexed by the magnitude of endogeneity. We first show that the optimal rate of convergence ranges from the faster NP-OLS rate to the slower NPIV rate as endogeneity increases. We then construct a feasible shrinkage estimator combining NP-OLS and NPIV that nearly attains the optimal rate in each regime without prior knowledge of the magnitude of endogeneity. The key to this adaptation is a one-sided downward-adjusted estimate of the endogeneity bias component. Simulations demonstrate that our adaptive estimator exploits substantial accuracy gains under limited endogeneity while performing comparably to NPIV under severe endogeneity.

Keywords: Endogeneity, Nonparametric instrumental variables, Adaptive estimation, Minimax rate optimality, Ill-posed inverse problem.

JEL classification codes: C13, C14, C26.

*I am deeply grateful to Ivan Canay, Joel Horowitz, Federico Bugni, and Eric Auerbach for their guidance and encouragement throughout this project. I also thank Vivek Bhattacharya, Federico Crippa, Peter Hull, Sid Kankanala, Ben Krueger, Andres Santos, Bruno Strulovici, Amilcar Velez, and Daniel Wilhelm for helpful comments and suggestions. All errors are my own.

1 Introduction

Structural relationships are central to many economic questions because they characterize the mechanisms that determine economic outcomes. Nonparametric methods allow researchers to study these relationships without imposing a particular functional form.[1] In many applications, however, the regressors of interest are endogenous in the sense that they are correlated with unobserved determinants of the outcome. Instrumental variables can be used to address such endogeneity; in nonparametric settings, this leads to the nonparametric instrumental variables (NPIV) model (Newey and Powell, 2003; Hall and Horowitz, 2005; Blundell, Chen, and Kristensen, 2007; Darolles, Fan, Florens, and Renault, 2011).

Addressing endogeneity with instrumental variables can come at a substantial cost in estimation accuracy. In nonparametric models, this cost is particularly pronounced: NPIV estimation requires solving an ill-posed inverse problem, so the resulting estimators can converge at very slow rates (Hall and Horowitz, 2005; Blundell et al., 2007; Chen and Reiss, 2011). By contrast, nonparametric least squares (NP-OLS) can estimate the conditional mean function at a much faster rate, but generally incurs endogeneity bias when used to estimate the structural function. The researcher therefore faces a trade-off between estimation accuracy and robustness to endogeneity.

This paper develops an approach to estimating nonparametric structural functions that balances this trade-off. As endogeneity increases, the optimal estimation strategy shifts from exploiting the accuracy gains of NP-OLS toward avoiding endogeneity bias through NPIV. We formalize this trade-off as a minimax estimation problem, evaluating the worst-case mean squared error (MSE) over the class of data generating processes with a given magnitude of endogeneity. We measure this magnitude by the squared L^2 distance between the conditional mean and structural functions, which coincide under exogeneity. We allow this magnitude to vary with sample size to capture different endogeneity regimes in the asymptotic analysis. Within this framework, we first characterize the optimal benchmark for estimation at each magnitude of endogeneity, and then construct a feasible estimator that nearly attains this benchmark without knowing this magnitude in advance.

Our first result characterizes how the minimax optimal rate for estimating the structural function varies with the magnitude of endogeneity. The minimax optimal rate is the fastest achievable rate, among all estimators, at which the MSE converges to zero uniformly over the corresponding class of data generating processes. We show that it ranges from the faster NP-OLS rate under limited endogeneity (Stone, 1982) to the slower NPIV rate under severe endogeneity (Hall and Horowitz, 2005; Chen and Reiss, 2011), with an intermediate regime in which the rate is determined by the magnitude of endogeneity itself. Thus, the robustness-efficiency trade-off we highlight is a fundamental feature of the problem of estimating the structural function, rather than a consequence of comparing particular estimators. This result also clarifies that the slow NPIV rate is *not* an unavoidable consequence of addressing endogeneity; it becomes the fundamental limit only

[1] For surveys of nonparametric methods in econometrics, see Härdle and Linton (1994); Yatchew (1998); DiNardo and Tobias (2001); Chen (2007); Ichimura and Todd (2007); Matzkin (2007); Horowitz (2009); Matzkin (2013).

when endogeneity is sufficiently severe.

We next construct a family of shrinkage estimators that adjust NP-OLS using an NPIV-based estimate of endogeneity bias. Our estimator starts from NP-OLS and subtracts a fraction $\lambda \in [0, 1]$ of the estimated endogeneity bias. We obtain this bias estimate from a suitably centered difference between sieve NP-OLS and NPIV estimators. Sieve methods approximate unknown functions by finitely many basis functions, such as polynomials or splines (Newey, 1997; Newey and Powell, 2003; Chen, 2007; Blundell et al., 2007). The sieve dimension, or the number of basis functions, controls the flexibility of the nonparametric approximation. The shrinkage parameter λ governs the robustness-efficiency trade-off: increasing λ reduces endogeneity bias at the cost of higher sampling variance. An oracle jointly chooses these two tuning parameters to minimize MSE within this family. We show that the resulting oracle estimator attains the minimax optimal rate in each endogeneity regime. The oracle choice, however, depends on unknown features of the data generating process, including the magnitude of endogeneity, and is therefore infeasible in practice.

We then construct a feasible shrinkage estimator by developing a data-driven method for jointly selecting the shrinkage parameter and sieve dimension. The selection criterion is derived from an asymptotic upper bound on MSE that separates the relevant risk into an NP-OLS component and an additional component involving the unknown endogeneity bias. In particular, the criterion requires estimating the squared L^2-norm of the endogeneity bias projected onto the sieve space. The main challenge is that a naive plug-in estimate based on the difference between NP-OLS and NPIV estimators inherits the sampling variation amplified by the ill-posed inverse problem. When endogeneity is limited, the plug-in estimate can be spuriously large, leading the selection procedure to move too aggressively toward NPIV. To address this issue, we introduce a one-sided downward adjustment that subtracts an inflated variance penalty from the plug-in estimate and truncates the resulting estimate at zero. We show that this adjustment controls the relevant upward stochastic fluctuations at the NP-OLS scale, allowing the estimator to exploit the accuracy gains available under limited and moderate endogeneity.

Our second result shows that the feasible shrinkage estimator nearly attains the minimax optimal rate in each endogeneity regime. More precisely, it attains the optimal rate up to a logarithmic factor, which arises only under severe endogeneity when the NPIV problem is mildly ill-posed. Thus, our estimator adapts to the unknown magnitude of endogeneity: it achieves nearly the same optimal rate of convergence as if this magnitude were known in advance.[2] We also show that its MSE is within a logarithmic factor of the minimum MSE attainable within the shrinkage family.

Monte Carlo simulations illustrate the finite-sample MSE gains from adapting to endogeneity. Under exogeneity, the adaptive estimator nearly coincides with NP-OLS and reduces MSE by 92.8% relative to NPIV. As endogeneity increases, the estimator places greater weight on the NPIV-based endogeneity component and eventually performs comparably to NPIV. The simulations

[2]Historically, an estimator is called *adaptive* if, without knowledge of certain features of the data generating process, it achieves the asymptotic performance attainable if those features were known, see Beran (1974); Stone (1975); Bickel (1982); Manski (1984); Newey (1988).

also reveal a finite-sample cost of adaptation, particularly in the intermediate regime where the downward adjustment can lead to insufficient correction for endogeneity bias. Finally, we apply the adaptive estimator to household gasoline demand data and find that the procedure selects an interior shrinkage parameter, $\hat{\lambda} = 0.33$, corresponding to a partial NPIV-based correction to NP-OLS.

Related literature. This paper contributes to a large literature on nonparametric estimation in models with endogeneity, see Carrasco, Florens, and Renault (2007); Horowitz (2011, 2014b); Newey (2013); Chen and Qiu (2016) for reviews. Applications include consumer demand and Engel curves (Blundell et al., 2007; Blundell, Horowitz, and Parey, 2012), differentiated-products demand (Berry and Haile, 2014, 2024; Compiani, 2022), production functions (Gandhi, Navarro, and Rivers, 2020), and international trade (Adao, Costinot, and Donaldson, 2017; Adão, Arkolakis, and Ganapati, 2020). A central challenge in this literature is that NPIV estimation requires solving an ill-posed inverse problem, which can lead to convergence rates as slow as logarithmic. We show that the fundamental difficulty of estimating the structural function depends not only on the ill-posedness of the NPIV problem, but also on the magnitude of endogeneity. In particular, when endogeneity is limited or moderate, the structural function can be estimated at faster rates by exploiting information in the reduced-form conditional mean function.

This paper also contributes to the literature on robustness-efficiency trade-offs (Hodges and Lehmann, 1952; Bickel, 1984; Armstrong, Kline, and Sun, 2025). This literature studies how to exploit additional restrictions to improve efficiency while remaining robust to their violation. Here, the corresponding trade-off is governed by the magnitude of endogeneity. Existing work has focused on parametric settings in which the relevant estimators converge at a common parametric rate, so the trade-off is reflected in asymptotic risk. By contrast, in our setting, NPIV estimation involves an ill-posed inverse problem and can therefore converge substantially more slowly than NP-OLS. Consequently, the trade-off is reflected in the minimax optimal rate.

The idea of exploiting OLS information to improve the estimation of structural parameters has a long history in econometrics. In linear models, early work pursued this idea through estimators that use OLS information alongside IV to improve finite-sample performance (Theil, 1958; Nagar, 1959; Sawa, 1973a, 1973b; Feldstein, 1974). Hausman (1978) proposed a specification test for exogeneity, naturally leading to a pre-test estimator that uses OLS when exogeneity is not rejected and IV otherwise. Shrinkage and averaging methods were developed in Maasoumi (1978); Hansen (2017); Angrist, Hull, Pathak, and Walters (2017). We extend this line of work to a nonparametric setting where the efficiency loss from instrumental variables is amplified by ill-posed inverse problems. Our contribution to this literature is to establish adaptive minimax rate optimality for our proposed feasible shrinkage estimator.

Methodologically, our data-driven procedure lies within the literature on model and tuning parameter selection. One strand studies MSE-based selection and averaging under potential model

misspecification, where the bias from imposing potentially invalid restrictions must be estimated (Claeskens and Hjort, 2003; DiTraglia, 2016; Cheng, Liao, and Shi, 2019; de Chaisemartin and D'Haultfœuille, 2020).[3] Our feasible selection criterion is similarly derived from an asymptotic MSE expansion that includes an endogeneity bias term. In our setting, however, this component is estimated using NPIV, so its sampling variation can dominate the selection criterion when the underlying endogeneity bias is small. Adaptation therefore requires controlling its upper stochastic fluctuations at the faster NP-OLS scale. Another strand studies adaptive estimation in NPIV models through a data-driven selection of the regularization parameter (Horowitz, 2014a; Breunig and Johannes, 2016; Breunig and Chen, 2019; Chen, Christensen, and Kankanala, 2025). Most closely related is Horowitz (2014a), who considers estimation of a nonparametric structural function as in this paper, and provides a data-dependent choice of sieve dimension that adapts to unknown smoothness of the structural function and the unknown degree of ill-posedness. Our procedure adds the magnitude of endogeneity as an additional dimension of adaptation by jointly selecting the sieve dimension and shrinkage parameter.

Outline. Section 2 introduces the setting and illustrates the robustness-efficiency trade-off by comparing NP-OLS and NPIV estimators. Section 3 develops a framework for analyzing this trade-off and provides an overview of the main results. Section 4 characterizes the optimal benchmark for a given magnitude of endogeneity by deriving the corresponding minimax optimal rate. Section 5 develops a feasible shrinkage estimator that nearly attains this benchmark without prior knowledge of the magnitude of endogeneity. Section 6 evaluates the finite-sample performance of the adaptive estimator through Monte Carlo simulations. Section 7 applies the estimator to household gasoline demand data. Section 8 concludes. All proofs are contained in the Appendix.

2 Set-up and Motivation

2.1 Setting

We observe an i.i.d. sample $\{(Y_i, X_i, Z_i)\}_{i=1}^n$, where $Y_i \in \mathbb{R}$ denotes the observed outcome, $X_i \in \mathbb{R}^{d_x}$ is a vector of possibly endogenous regressors, and $Z_i \in \mathbb{R}^{d_z}$ is a vector of instruments with $d_z \geq d_x$. Throughout the theoretical analysis, both X_i and Z_i are continuously distributed. We consider the nonparametric structural model

$$Y_i = h_0(X_i) + U_i \tag{2.1}$$

where h_0 is the structural function of interest and U_i is the unobserved determinant of the outcome.

The structural interpretation of (2.1) is that changing X_i from x to x', holding U_i fixed, changes the outcome by $h_0(x') - h_0(x)$. The model (2.1) generalizes the linear structural model by allow-

[3] Related work, building on James and Stein (1961), considers shrinkage toward potentially biased estimators; see Green and Strawderman (1991); Kim and White (2001); Judge and Mittelhammer (2004); Mittelhammer and Judge (2005); Hansen (2016); Fessler and Kasy (2019).

ing marginal response $\nabla_x h_0(x)$ to vary with x. This flexibility is important for policy-relevant functionals that depend on the shape of the response function, as with average derivative effects, elasticities, and welfare measures (Newey and Stoker, 1993; Hausman and Newey, 1995).

We distinguish the structural function $h_0(x)$ from the conditional mean function $E[Y_i \mid X_i = x]$. The conditional mean function is a feature of the observed joint distribution of (Y_i, X_i), describing how average observed outcomes vary with X_i. Under model (2.1),

$$g_0(x) := E[Y_i \mid X_i = x] = h_0(x) + E[U_i \mid X_i = x].$$

Thus, these two functions coincide only under the exogeneity condition $E[U_i \mid X_i] = 0$. When this condition fails, X_i is endogenous and $g_0(x)$ does not admit a structural interpretation.[4] The classical simultaneous equations model in demand estimation provides the canonical example. Because equilibrium price is jointly determined by demand and supply shocks, a naive regression of quantity on price generally does not trace the structural demand curve (Working, 1927; Haavelmo, 1944).

Instrumental variables provide a way to identify the structural function when X_i is endogenous. The following assumption provides sufficient conditions for identification of the structural function.

Assumption 2.1 (Identification, Newey and Powell (2003))**.** *We assume that*

(i) Instrument validity. $E[U_i \mid Z_i] = 0$*;*

(ii) Completeness condition. Let $\mathcal{H}$ *denote the admissible class of structural functions containing* h_0*, the condition holds if for any* $h_1, h_2 \in \mathcal{H}$*, the relation* $E[(h_1 - h_2)(X_i) \mid Z_i] = 0$ *a.s. implies that* $h_1 = h_2$ *a.s.*

Under Assumption 2.1, the structural function h_0 can be identified by the following conditional moment restriction:

$$E[Y_i - h_0(X_i) \mid Z_i] = 0 \quad \text{a.s.} \tag{2.2}$$

Instrument validity ensures that the structural function h_0 satisfies the conditional moment restriction (2.2), while completeness ensures that it is the unique solution within $\mathcal{H}$. The completeness condition is the nonparametric analogue of the rank condition in linear IV models: it requires that no two distinct admissible structural functions have the same conditional expectation given Z_i.

Example 2.1 (Household gasoline demand)**.** Consider household gasoline demand, which we study in Section 7. Let Y_i denote annual household gasoline consumption and X_i the local gasoline price. The structural function $h_0(x)$ describes how gasoline consumption responds to changes in price, holding unobserved determinants of demand fixed. These determinants, captured by U_i, may include household preferences and commuting needs. Unobserved local demand shocks may also affect gasoline prices. For example, greater commuting needs across households in a local market

[4] In the linear model, endogeneity is usually expressed as $E[X_i U_i] \neq 0$ (Wooldridge, 2010, p. 50). In the nonparametric setting, the corresponding notion of endogeneity is $E[U_i \mid X_i = x] \neq 0$ for some $x \in \mathbb{R}^{d_x}$; see Blundell and Powell (2003).

may increase gasoline consumption and raise equilibrium prices. Such common demand shocks can induce correlation between X_i and U_i, making gasoline prices endogenous.

Instrumental variables can address this endogeneity by exploiting supply-side variation in gasoline prices. Following Blundell et al. (2012), we use the distance from a major oil platform to the household's state capital, denoted as Z_i, as an instrument for gasoline price. Distance from the source of supply may affect local gasoline prices through transportation costs. Instrument validity requires that unobserved determinants of gasoline demand have mean zero conditional on this distance. Under Assumption 2.1, this supply-side variation identifies the demand function. ■

Remark 2.1. The theoretical analysis in this paper focuses on continuously distributed regressors and instruments. When both have finite support, identification of the structural function requires that the conditional probability matrix of X_i given Z_i have full column rank, and hence that the instrument have at least as many support points as the regressor (Newey and Powell, 2003; Freyberger and Horowitz, 2015). Under these conditions, IV and least squares estimators attain $\sqrt{n}$ rates for their respective targets (Horowitz, 2011, Section 6.1). The ill-posed inverse problem discussed in the next subsection does not arise in this finite-dimensional setting. ■

Notation. Let W denote a generic random vector with support $\mathcal{W}$, and let $\{W_i\}_{i=1}^n$ denote the corresponding observations. Let $L^2(W)$ denote the Hilbert space of measurable functions $f : \mathcal{W} \to \mathbb{R}$ satisfying $E[f(W)^2] < \infty$. For $f, g \in L^2(W)$, define the population inner product and norm by $\langle f, g\rangle_W = E[f(W)g(W)], \|f\|_W^2 = \langle f, f\rangle_W$. Their empirical counterparts are $\langle f, g\rangle_{W,n} = \frac{1}{n}\sum_{i=1}^n f(W_i)g(W_i), \|f\|_{W,n}^2 = \langle f, f\rangle_{W,n}$. For nonnegative quantities a and b, we write $a \lesssim b$ if $a \leq Cb$ for some constant $C > 0$, $a \gtrsim b$ if $b \lesssim a$, and $a \asymp b$ if both inequalities hold. We also write $a \wedge b = \min\{a, b\}$ and $[a]_+ = \max\{a, 0\}$. For a matrix A, let A^- denote its Moore-Penrose inverse.

2.2 Ill-posedness of NPIV Estimation

Under Assumption 2.1, the structural function h_0 is identified by the moment condition (2.2). However, its estimation poses a substantial challenge in the nonparametric setting. To see why, define the conditional expectation operator $T : L^2(X) \to L^2(Z)$ by $(Th)(z) = E[h(X) \mid Z = z]$ and let $m(z) = E[Y \mid Z = z]$. The conditional moment restriction (2.2) can then be written as

$$(Th_0)(z) = E[h_0(X_i) \mid Z_i = z] = m(z).$$

Thus, recovering the structural function h_0 amounts to inverting the operator T.

The main difficulty in estimating h_0 is that this inversion can be unstable. The conditional expectation operator averages h_0 over the distribution of X given Z, thereby smoothing out variation in the structural function. As a result, two substantially different structural functions can have conditional expectations that are arbitrarily close. Small estimation errors in m can therefore translate into large estimation errors in recovering h_0. This instability, known as an ill-posed inverse

problem, typically requires regularization (Carrasco et al., 2007; Horowitz, 2014b) and can lead to slow convergence rates for NPIV estimators (Hall and Horowitz, 2005; Chen and Reiss, 2011).

The instability of inversion is inherent to the NPIV estimation problem and is not specific to any particular estimator. To illustrate its implications for estimation accuracy, we consider the sieve NPIV estimator. Sieve methods approximate an unknown function by a finite linear combination of basis functions (Chen, 2007). We use J basis functions of X to approximate the structural function and K basis functions of Z as instruments. Specifically, let

$$p^J(x) = (p_{J1}(x), ..., p_{JJ}(x))', \qquad q^K(z) = (q_{K1}(z), ..., q_{KK}(z))'$$

denote the corresponding vectors of basis functions. These basis functions span the sieve spaces $\mathcal{H}_J = \text{clsp}\{p_{J1}, \ldots, p_{JJ}\} \subset L^2(X)$ and $\mathcal{Q}_K = \text{clsp}\{q_{K1}, \ldots, q_{KK}\} \subset L^2(Z)$, respectively, where clsp denotes the closed linear span. Define the sample matrices

$$\mathbf{P}_J = \left(p^J(X_1), ..., p^J(X_n)\right)' \in \mathbb{R}^{n \times J}, \qquad \mathbf{Q}_K = \left(q^K(Z_1), ..., q^K(Z_n)\right)' \in \mathbb{R}^{n \times K}.$$

We take K to grow at the same asymptotic rate as J, with $J \leq K \leq C_K J$ for some finite constant C_K. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)'$. The sieve two-stage least squares (2SLS) estimator of h_0 is given by

$$\widetilde{h}_{\text{IV},J}(x) = p^J(x)' \left[\mathbf{P}_J'\mathbf{Q}_K \left(\mathbf{Q}_K'\mathbf{Q}_K\right)^- \mathbf{Q}_K'\mathbf{P}_J\right]^- \mathbf{P}_J'\mathbf{Q}_K \left(\mathbf{Q}_K'\mathbf{Q}_K\right)^- \mathbf{Q}_K'\mathbf{Y}. \tag{2.3}$$

Computationally, the sieve NPIV estimator is simply conventional 2SLS applied to the basis functions: the first stage projects $p^J(X)$ onto the linear span of $q^K(Z)$, and the second stage regresses Y_i on the fitted values from the first stage.

We now turn to the mean squared error (MSE) of the sieve 2SLS estimator. To quantify the effect of ill-posedness on estimation, we follow Blundell et al. (2007) and define the sieve measure of ill-posedness as

$$\tau_J := \sup_{h \in \mathcal{H}_J, h \neq 0} \frac{\|h\|_X}{\|Th\|_Z} \tag{2.4}$$

The quantity τ_J measures the largest amplification of errors induced by inversion of T over the sieve space $\mathcal{H}_J$. A larger τ_J therefore indicates greater instability when recovering h from Th. We distinguish two regimes based on the asymptotic behavior of τ_J: the *mildly ill-posed* case, where $\tau_J \asymp J^{\varsigma/d_x}$, and the *severely ill-posed* case, where $\tau_J \asymp \exp(\frac{1}{2}J^{\varsigma/d_x})$, for some $\varsigma > 0$.

Suppose that h_0 has smoothness index $s > 0$, with squared sieve approximation error of order J^{-2s/d_x}. Under standard regularity conditions, the MSE of the sieve 2SLS estimator satisfies (Blundell et al., 2007):

$$E\left\|\widetilde{h}_{\text{IV},J} - h_0\right\|_X^2 \lesssim \frac{\tau_J^2 J}{n} + J^{-2s/d_x}. \tag{2.5}$$

The upper bound in (2.5) reflects the familiar bias-variance trade-off in sieve estimation, with variance amplified by ill-posedness. The first term captures sampling variance: J/n is the usual

variance associated with estimating J sieve coefficients, while τ_J^2 captures the additional variance amplification due to ill-posedness. The second term captures the squared sieve approximation bias, which decreases as J increases. Thus, increasing the sieve dimension reduces sieve approximation bias but increases sampling variance, with the latter amplified by ill-posedness.

Balancing the variance and sieve approximation bias in (2.5) yields the following MSE convergence rates for the sieve 2SLS estimator:

$$r_{n,\mathrm{IV}} := \begin{cases} n^{-2s/(2s+2\varsigma+d_x)}, & \text{mildly ill-posed,} \\ (\log n)^{-2s/\varsigma}, & \text{severely ill-posed.} \end{cases} \tag{2.6}$$

Under the corresponding regularity conditions, the rate $r_{n,\mathrm{IV}}$ coincides with the minimax optimal MSE rate for the NPIV estimation problem (Hall and Horowitz, 2005; Chen and Reiss, 2011). Thus, the slow convergence rates are not specific to the sieve 2SLS estimator but reflect the fundamental difficulty of estimating the structural function in the worst case over the class of NPIV models.

The severely ill-posed case in (2.6) can arise even under familiar designs. For example, when a scalar endogenous regressor and its instrument are jointly Gaussian with non-zero correlation, the sieve measure of ill-posedness τ_J grows exponentially with the sieve dimension J. Consequently, the optimal sieve dimension grows only logarithmically with the sample size n, giving the corresponding logarithmic rate of convergence, see Blundell et al. (2007, Remark 4).

2.3 Robustness-Efficiency Trade-off

The ill-posedness of NPIV estimation has a familiar counterpart in linear models, where the use of instrumental variables comes at a cost in efficiency. Under exogeneity, IV is generally less efficient than OLS because it relies only on the variation in the regressors explained by the instruments (Nakamura and Nakamura, 1998). This efficiency loss motivates the specification test of exogeneity proposed by Hausman (1978), which compares OLS and IV estimates to assess whether the more efficient OLS estimator can be used. Blundell and Horowitz (2007) develop a corresponding test for nonparametric models. We therefore consider nonparametric least squares (NP-OLS) as an alternative to NPIV for estimating the structural function.

NP-OLS estimates the conditional mean function $g_0(x) = E[Y|X = x]$. Using the same sieve basis in Section 2.2, the sieve NP-OLS estimator with sieve dimension J is

$$\widehat{g}_J(x) = p^J(x)' \left(\mathbf{P}_J'\mathbf{P}_J\right)^- \mathbf{P}_J'\mathbf{Y}. \tag{2.7}$$

The estimator is obtained by regressing Y_i directly on the basis functions $p^J(X_i)$ without using instruments, thereby avoiding the ill-posed inverse problem associated with NPIV estimation. However, it generally suffers from endogeneity bias when used to estimate the structural function.

The trade-off between estimation accuracy and endogeneity bias is reflected in the MSE of

NP-OLS. Suppose that g_0 has the same smoothness index s as h_0 in Section 2.2. Under standard regularity conditions for sieve least squares, the MSE of NP-OLS as an estimator of h_0 satisfies

$$E\left\|\widehat{g}_J - h_0\right\|_X^2 \lesssim \frac{J}{n} + J^{-2s/d_x} + \left\|E[U \mid X]\right\|_X^2. \tag{2.8}$$

Compared with the MSE for NPIV in (2.5), NP-OLS avoids the variance amplification due to ill-posedness but incurs an additional endogeneity bias. When this bias is fixed and non-zero, NP-OLS is inconsistent for the structural function.

Under exogeneity $E[U \mid X] \equiv 0$, the endogeneity bias in (2.8) vanishes. Balancing the sampling variance and sieve approximation error in (2.8) yields the following MSE convergence rate for NP-OLS:

$$r_{n,\mathrm{OLS}} := n^{-2s/(2s+d_x)}. \tag{2.9}$$

Under the corresponding regularity conditions, this rate coincides with the minimax optimal MSE rate for estimating the conditional mean function (Stone, 1982). In particular, the NP-OLS rate (2.9) is faster than the NPIV rate defined in (2.6).

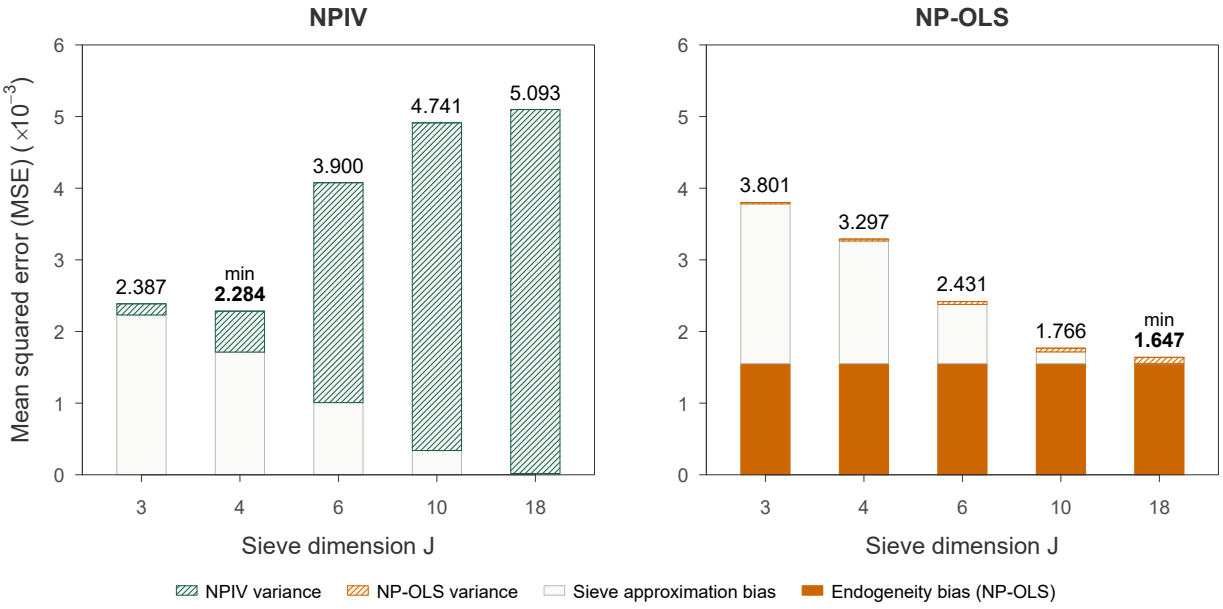


Figure 1: *Mean Squared Error Decomposition for NPIV and NP-OLS.*

Note. The figure reports the Monte Carlo mean squared errors (MSEs) of NP-OLS and sieve 2SLS (NPIV) across sieve dimensions J, using the simulation design in Section 6 with $\|E[U|X]\|_X^2 / \mathrm{Var}(U) = 0.10$. The stacked bars display the variance and sieve approximation bias components of MSE, together with the endogeneity bias for NP-OLS. The label "min" indicates the sieve dimension that minimizes MSE for each estimator.

Although NP-OLS is inconsistent for the structural function under fixed non-zero endogeneity, it can achieve a smaller finite-sample MSE than NPIV when endogeneity is limited. Figure 1 illustrates this possibility using the simulation design described in Section 6, with a small but non-zero magnitude of endogeneity. The figure compares the MSEs of NPIV and NP-OLS across sieve

dimensions, decomposing them into sampling variance, sieve approximation bias, and endogeneity bias for NP-OLS. As the sieve dimension increases, the sampling variance of NPIV grows much more rapidly than that of NP-OLS. At $J = 6$, the sampling variance of NPIV is approximately 76 times that of NP-OLS. At their respective MSE-optimal sieve dimensions, NP-OLS achieves an MSE approximately 28% lower than NPIV. Thus, when endogeneity is limited, the lower sampling variance of NP-OLS can outweigh its endogeneity bias, resulting in a smaller MSE than NPIV.

Taken together, the MSE bounds and Figure 1 illustrate the trade-off between robustness to endogeneity and estimation accuracy. NPIV eliminates endogeneity bias but has greater sampling variance due to ill-posedness, while NP-OLS avoids this variance amplification but suffers from the endogeneity bias. At a given sample size, NP-OLS may have a lower MSE than NPIV when endogeneity is limited, but a higher MSE as endogeneity increases. Thus, their relative estimation accuracy depends on the unknown magnitude of endogeneity bias, and neither estimator uniformly dominates the other.

3 Framework and Overview of Results

This section formalizes the robustness-efficiency trade-off in estimating the structural function as a minimax estimation problem indexed by the magnitude of endogeneity. We then provide an overview of our main results: the minimax optimal rate of convergence at a given magnitude of endogeneity and a feasible shrinkage estimator that nearly attains the corresponding rate without knowing this magnitude in advance.

3.1 Framework of Analysis

The comparison in Section 2.3 shows that the relative finite-sample mean squared errors of NP-OLS and NPIV estimators depend on the magnitude of endogeneity. When endogeneity is limited, NP-OLS may have a smaller MSE because its lower sampling variance can outweigh its endogeneity bias. However, conventional asymptotics that hold endogeneity fixed do not capture this possibility. For any fixed positive magnitude of endogeneity, NP-OLS has a nonvanishing endogeneity bias, while the MSE of NPIV converges to zero as the sample size increases. Thus, NPIV eventually outperforms NP-OLS regardless of how small the endogeneity bias is. To capture this trade-off in our asymptotic analysis, we allow the magnitude of endogeneity to vary with sample size.

Formally, we consider a sequence of statistical experiments indexed by the sample size n. In the nth experiment, the observations $\{\mathcal{D}_{n,i}\}_{i=1}^{n}$, where $\mathcal{D}_{n,i} = (Y_{n,i}, X_{n,i}, Z_{n,i})$, are i.i.d. with common law F_n. The structural function h_0 and the joint distribution of (X_i, Z_i) are held fixed across experiments, while the magnitude of endogeneity is allowed to vary with n. For notational convenience, we suppress the n-index whenever this dependence does not lead to confusion.

We measure the magnitude of endogeneity in the nth experiment by the squared $L^2(X)$-norm

of the endogeneity bias:

$$\rho_n := \left\| E_{F_n}[U \mid X] \right\|_X^2 \tag{3.1}$$

Equivalently, ρ_n measures the squared $L^2(X)$ distance between the conditional mean function and the structural function. For each n, $\rho_n = \rho(F_n)$ is a deterministic feature of the underlying population law F_n. The sequence $\{\rho_n\}_{n\geq 1}$ should not be interpreted as implying that the underlying population distribution changes as more observations are collected. Instead, it is an asymptotic device for comparing the performance of estimators across different endogeneity regimes.

The sequence ρ_n indexes the full spectrum of endogeneity regimes, with exact exogeneity as a boundary case. When $\rho_n \equiv 0$, we have $E_{F_n}[U|X] = 0$ for all n, so the regressor X is exogenous. When $\rho_n \to 0$, endogeneity may be present under every F_n but its magnitude vanishes asymptotically. Conversely, when ρ_n is bounded away from zero, the framework accommodates nonlocal endogeneity, including the case where the magnitude of endogeneity is held fixed as $n \to \infty$. Across all these regimes, the relevant asymptotic comparison is governed by the magnitude of ρ_n relative to the MSE convergence rates of NP-OLS and NPIV.

The reparameterization (3.1) is closely related to an extensive literature on local misspecification.[5] As in this literature, we allow the magnitude of endogeneity to vary with the sample size. However, we do *not* require endogeneity to vanish asymptotically: the unknown sequence ρ_n may converge to zero at an arbitrary rate or remain bounded away from zero, thereby accommodating both local and global endogeneity. In this respect, our formulation is closest to Cheng et al. (2019), who likewise allow for both local and global model misspecification.

More fundamentally, the comparison between NP-OLS and NPIV raises the question of how accurately the structural function can be estimated at a given magnitude of endogeneity. To answer this question, we formulate a minimax estimation problem that explicitly accounts for the magnitude of endogeneity. Specifically, we consider classes of data generating processes indexed by this magnitude and evaluate each estimator by its worst-case mean squared error (MSE) over the corresponding class. The minimax risk is defined as the infimum of this worst-case MSE over *all* estimators. We then characterize how its rate of convergence varies with the magnitude of endogeneity, thereby characterizing the optimal robustness-efficiency trade-off.

3.2 Overview of Results

Our first main result characterizes the minimax optimal rate for estimating the structural function at a given magnitude of endogeneity. Let $\mathcal{F}(\rho_n)$ denote the class of data generating processes with the magnitude of endogeneity ρ_n, formally defined in Section 4.1. We establish the following lower

[5] See, among many others, Kiefer and Skoog (1984); Newey (1985); Hjort and Claeskens (2003); Schorfheide (2005); Bugni, Canay, and Guggenberger (2012); Kitamura, Otsu, and Evdokimov (2013); Hansen (2014); DiTraglia (2016); Andrews, Gentzkow, and Shapiro (2017, 2020); Bugni and Ura (2019); Armstrong and Kolesár (2021); Bonhomme and Weidner (2022); Montiel Olea, Plagborg-Møller, Qian, and Wolf (2026).

bound on the minimax risk:

$$\inf_{\widehat{h}} \sup_{F_n \in \mathcal{F}(\rho_n)} E_{F_n} \left\| \widehat{h} - h_0 \right\|_X^2 \gtrsim \begin{cases} r_{n,\mathrm{OLS}}, & \rho_n \lesssim r_{n,\mathrm{OLS}} \\ \rho_n, & r_{n,\mathrm{OLS}} \lesssim \rho_n \lesssim r_{n,\mathrm{IV}} \\ r_{n,\mathrm{IV}}, & \rho_n \gtrsim r_{n,\mathrm{IV}} \end{cases} \tag{3.2}$$

where the infimum is taken over all estimators $\hat{h}$ measurable with respect to the sample, and $r_{n,\mathrm{OLS}}$ and $r_{n,\mathrm{IV}}$ denote the optimal mean squared error rates for NP-OLS and NPIV estimation, respectively, as defined in (2.9) and (2.6).

The lower bound shows that no estimator can attain a faster rate of convergence than that in (3.2) uniformly over the function class $\mathcal{F}(\rho_n)$. We show below that this lower bound can be attained by our feasible estimator up to a logarithmic factor. Together, these results characterize the minimax optimal rate of convergence, up to this factor, and establish how the fundamental limit on estimation accuracy varies with the magnitude of endogeneity.

The minimax optimal rate exhibits three regimes as the magnitude of endogeneity increases. Under limited endogeneity, it coincides with the faster NP-OLS rate. Under moderate endogeneity, it is determined by the magnitude of endogeneity itself. Under severe endogeneity, it coincides with the slower NPIV rate. Thus, the conventional slow NPIV rate becomes the fundamental limit only when endogeneity is sufficiently severe. When endogeneity is limited or moderate, the structural function can be estimated at faster rates by exploiting information in the conditional mean function.

Remark 3.1. Our optimality result is *asymptotic* and concerns minimax rates of convergence. As in classical nonparametric minimax theory, the optimal rate characterizes the asymptotic order of the minimax risk up to multiplicative constants (Tsybakov, 2009, Section 2.1). This notion differs from sharp asymptotic minimaxity, which additionally requires determining the sharp constant in the asymptotic minimax risk (Tsybakov, 1998). ■

To attain the optimal benchmark in (3.2), we construct a shrinkage estimator that adjusts NP-OLS using an NPIV-based estimate of endogeneity bias. Let $\Delta(x) := E[U \mid X = x] = g_0(x) - h_0(x)$ denote the endogeneity bias. For a shrinkage parameter $\lambda \in [0, 1]$ and sieve dimension J, we consider

$$\widehat{h}_{\lambda,J} = \widehat{g}_J - \lambda \widehat{\Delta}_J$$

where $\hat{g}_J$ is the sieve NP-OLS estimator defined in (2.7), and $\hat{\Delta}_J$ estimates the endogeneity bias using a suitably centered difference between NP-OLS and NPIV estimators, as formally defined in Section 4.2. The shrinkage parameter λ governs the robustness-efficiency trade-off: increasing λ moves the estimator from NP-OLS toward an NPIV-based estimator, reducing endogeneity bias at the cost of greater sampling variance. The sieve dimension J controls the flexibility of the nonparametric approximation.

We select the tuning parameters (λ, J) jointly by minimizing a feasible risk criterion derived from an asymptotic upper bound on MSE. Specifically, we construct

$$\widehat{R}_n(\lambda, J) := 2\widehat{V}_{g,J} - \left\|\hat{g}_J\right\|_{X,n}^2 + \lambda^2 \widehat{V}_{\Delta,J} + \left[(1-\lambda)^2 - 1\right] \left[\left\|\hat{\Delta}_J\right\|_{X,n}^2 - \hat{\kappa}_{n,J} \hat{V}_{\Delta,J} \right]_+ \tag{3.3}$$

where $\hat{V}_{g,J}$ and $\hat{V}_{\Delta,J}$ estimate the leading variance terms of $\hat{g}_J$ and $\hat{\Delta}_J$ respectively, and $\hat{\kappa}_{n,J}$ is an inflation factor defined in Section 5.1. In (3.3), the first two terms correspond to the NP-OLS component of the risk criterion, while the remaining terms account for the NPIV-based adjustment for endogeneity bias.

Minimizing the feasible criterion (3.3) yields the feasible shrinkage estimator

$$\widehat{h}_{\hat{\lambda},\hat{J}} = \widehat{g}_{\hat{J}} - \hat{\lambda}\widehat{\Delta}_{\hat{J}}, \qquad (\hat{\lambda}, \hat{J}) \in \underset{\lambda \in [0,1],\ J \in \mathcal{J}_n}{\arg\min}\ \widehat{R}_n(\lambda, J) \tag{3.4}$$

where $\mathcal{J}_n$ denotes the set of candidate sieve dimensions.

A distinctive feature of the criterion (3.3) is the one-sided downward adjustment to the plug-in estimate of squared endogeneity bias:

$$\left[\left\|\hat{\Delta}_J\right\|_{X,n}^2 - \hat{\kappa}_{n,J} \hat{V}_{\Delta,J} \right]_+ .$$

Because the unadjusted plug-in estimate $\|\widehat{\Delta}_J\|_{X,n}^2$ inherits the variance amplification of the ill-posed NPIV problem, it can substantially overstate the true bias even when endogeneity is small. By subtracting an inflated variance penalty and truncating the result at zero, the adjustment prevents NPIV sampling variation from dominating the selection criterion. In particular, this adjustment controls the relevant upward stochastic fluctuations at the faster NP-OLS variance scale, allowing the feasible estimator to attain the optimal rate under limited and moderate endogeneity.

Our second main result shows that the feasible estimator (3.4) nearly attains the minimax optimal rate without prior knowledge of the magnitude of endogeneity. In particular, we establish that

$$\sup_{F_n \in \mathcal{F}(\rho_n)} E_{F_n} \left\|\widehat{h}_{\hat{\lambda},\hat{J}} - h_0\right\|_X^2 \lesssim \begin{cases} r_{n,\mathrm{OLS}}, & \rho_n \lesssim r_{n,\mathrm{OLS}} \\ \rho_n, & r_{n,\mathrm{OLS}} \lesssim \rho_n \lesssim \tilde{r}_{n,\mathrm{IV}} \\ \tilde{r}_{n,\mathrm{IV}}, & \rho_n \gtrsim \tilde{r}_{n,\mathrm{IV}} \end{cases}$$

where $\tilde{r}_{n,\mathrm{IV}}$ is the adaptive NPIV rate, which differs from the optimal NPIV rate $r_{n,\mathrm{IV}}$ by a logarithmic factor under mild ill-posedness and has the same order under severe ill-posedness. Thus, our feasible estimator (3.4) automatically exploits the accuracy gains available under limited and moderate endogeneity while remaining robust to severe endogeneity.

4 Minimax Rate Optimality

This section characterizes the minimax optimal rate for estimating the structural function at a given magnitude of endogeneity. We first derive a minimax lower bound that exhibits three regimes, determined by the magnitude of endogeneity relative to the NP-OLS and NPIV rates. We then construct an oracle shrinkage estimator and show that its risk matches the lower bound up to constant factors in each regime.

4.1 Minimax Lower Bound

We now establish the minimax lower bound previewed in Section 3.2. We first introduce conditions on the smoothness of the structural functions and the degree of ill-posedness of the NPIV problem.

Assumption 4.1 (Smoothness condition, Breunig and Chen, 2024)**.** *Recall the sieve spaces* $\mathcal{H}_J = \text{clsp}\{p_{J1}, \ldots, p_{JJ}\} \subset L^2(X)$*. We assume that the sequence* $\{\mathcal{H}_J\}_{J\geq 1}$ *is nested, so that* $\mathcal{H}_J \subseteq \mathcal{H}_{J+1}$ *for every* $J \geq 1$*. Furthermore, let* $\{\psi_j\}_{j=1}^{\infty}$ *be an* $L^2(X)$*-orthonormal basis associated with this nested sequence, such that* $\mathcal{H}_J = \text{clsp}\{\psi_1, \ldots, \psi_J\}$ *for all* $J \geq 1$*. We assume that the structural function* h_0 *belongs to the Sobolev class*

$$\mathcal{H}(s, L_h) := \left\{ h \in L^2(X) : \sum_{j=1}^{\infty} j^{2s/d_x} \langle h, \psi_j \rangle_X^2 \leq L_h^2 \right\}, \tag{4.1}$$

where $s > 0$ *denotes the smoothness index. We also assume that* $g_0 \in \mathcal{H}(s, L_g)$*. In particular, the condition* (4.1) *implies a sieve approximation error of order* J^{-s/d_x}*.*

The degree of ill-posedness of the NPIV inverse problem is governed by the smoothing property of the conditional expectation operator T relative to the basis $\{\psi_j\}$ in (4.1). When $X|Z$ admits a conditional density, $(Th)(z) = E[h(X) \mid Z = z] = \int h(x) f_{X|Z}(x|z) \mathrm{d}x$, so T averages variation in h according to the conditional distribution of X given Z. We characterize this smoothing property through the exact link condition (Chen and Reiss, 2011, p. 509, Breunig and Chen, 2024, Equ.3.3): for a positive decreasing sequence $\{\nu_j\}_{j\geq 1}$ satisfying $\nu_j \to 0$,

$$c \sum_{j=1}^{\infty} \nu_j^2 \langle h, \psi_j \rangle_X^2 \leq \|Th\|_Z^2 \leq C \sum_{j=1}^{\infty} \nu_j^2 \langle h, \psi_j \rangle_X^2 \qquad \text{for all } h \in \mathcal{H}. \tag{4.2}$$

for some constants $C \geq c > 0$. In particular, taking $h = \psi_j$ gives $\|T\psi_j\|_Z \asymp \nu_j$, so ν_j measures the extent to which T smooths the j-th basis direction. The faster ν_j decreases to zero, the stronger the smoothing property of T. Consequently, it becomes harder to detect the corresponding features of h in the $L^2(X)$ metric from Th, leading to greater instability and a higher degree of ill-posedness in the inverse problem. The exact link condition (4.2) also implies that the sieve measure of ill-posedness (2.4) satisfies $\tau_J \asymp \nu_J^{-1}$. Accordingly, $\nu_j \asymp j^{-\varsigma/d_x}$ corresponds to the mildly ill-posed case, while $\nu_j \asymp \exp(-j^{\varsigma/d_x}/2)$ corresponds to the severely ill-posed case.

Assumption 4.2. *Recall that $U_i = Y_i - h_0(X_i)$. We assume that:*

(i) $\mathrm{Var}(U_i \mid X_i, Z_i) \geq \underline{\sigma}^2 > 0$ *a.s.;*

(ii) the densities of X_i and Z_i are uniformly bounded above and away from zero on their supports, which are Cartesian products of bounded intervals;

(iii) the right-hand inequality of (4.2) holds.

Assumption 4.2 follows Assumption 1 of Breunig and Chen (2024), except that condition (i) strengthens their conditional variance restriction by conditioning on (X_i, Z_i). Condition (iii) excludes models with continuously distributed exogenous regressors that serve as their own instruments. Chen and Christensen (2018, Section 3.3) consider such partially endogenous models and show that their rates of convergence results still hold under suitable regularity conditions. Extending our theoretical results to this setting is left for future work.

Theorem 4.1. *For each n and $\rho_n \geq 0$, let $\mathcal{F}(\rho_n)$ denote the class of data generating processes F_n satisfying Assumptions 2.1, 4.1 and 4.2, with the magnitude of endogeneity ρ_n as reparameterized in (3.1). Then for all sufficiently large n,*

$$\inf_{\widehat{h}} \sup_{F_n \in \mathcal{F}(\rho_n)} E_{F_n} \left\| \widehat{h} - h_0 \right\|_X^2 \gtrsim \begin{cases} r_{n,\mathrm{OLS}}, & \rho_n \lesssim r_{n,\mathrm{OLS}} \\ \rho_n, & r_{n,\mathrm{OLS}} \lesssim \rho_n \lesssim r_{n,\mathrm{IV}} \\ r_{n,\mathrm{IV}}, & \rho_n \gtrsim r_{n,\mathrm{IV}} \end{cases}$$

where the implied constant in $\gtrsim$ is common to all three regimes and independent of n and ρ_n.

We compare Theorem 4.1 with Chen and Reiss (2011, Theorem 1 and Corollary 1) who derive a minimax lower bound for estimating the structural function h_0 under $L^2(X)$ loss uniformly over the parameter space $\mathcal{F}$ which in our notation corresponds to

$$\mathcal{F} = \bigcup_{\rho \geq 0} \mathcal{F}(\rho).$$

Their result is therefore governed by the least favorable magnitude of endogeneity and has order $r_{n,\mathrm{IV}}$. By contrast, our Theorem 4.1 evaluates the minimax risk separately over each endogeneity-indexed class $\mathcal{F}(\rho_n)$. This finer characterization shows that the standard NPIV rate arises only when endogeneity is sufficiently severe, whereas faster rates are achievable when endogeneity is limited or moderate. Thus, while taking the union of $\mathcal{F}(\rho)$ over all $\rho \geq 0$ recovers the worst-case bound of Chen and Reiss (2011), indexing the parameter space by ρ_n reveals how the fundamental difficulty of estimating h_0 varies with the magnitude of endogeneity.

Remark 4.1. Unlike Chen and Reiss (2011), who consider a *fixed* parameter space, our parameter space $\mathcal{F}(\rho_n)$ in Theorem 4.1 varies with the sample size n. This distinction affects the formulation

of the minimax problem, but not the construction of the lower bound: in both settings, the least favorable alternatives may still vary with n. ■

Remark 4.2. Our minimax criterion in Theorem 4.1 is closely related to the classical minimax analysis of restricted parameter spaces.[6] For example, Bickel (1981) studies how minimax risk varies with the restriction on the parameter space. Likewise, in our setting, ρ_n indexes the departure from exogeneity, and we evaluate the minimax risk over the corresponding class $\mathcal{F}(\rho_n)$. However, our approach differs from Kempthorne (1988) which seeks a single estimator that performs well simultaneously under multiple risk criteria. Applied here, such an approach could trade off the maximal risk under exogeneity against maximal risk over a larger class allowing substantial endogeneity. By contrast, we characterize the minimax risk *separately* at each magnitude of endogeneity. ■

4.2 An Oracle Shrinkage Estimator

We now construct the shrinkage estimator introduced in Section 3.2 and establish its oracle rate optimality. We first describe how to construct the NPIV-based estimator of endogeneity bias. We then consider an oracle that jointly selects the shrinkage parameter and sieve dimension to minimize mean squared error (MSE) within this family. We show that the resulting oracle estimator attains the minimax optimal rate established in Section 4.1.

Recall that the endogeneity bias is $\Delta(x) := E[U \mid X = x] = g_0(x) - h_0(x)$ so that $h_0(x) = g_0(x) - \Delta(x)$. This decomposition suggests estimating the structural function by adjusting NP-OLS for endogeneity bias. For a shrinkage weight $\lambda \in [0, 1]$ and sieve dimension J, we consider

$$\widehat{h}_{\lambda,J} = \widehat{g}_J - \lambda \widehat{\Delta}_J \tag{4.3}$$

where $\hat{g}_J$ is the sieve NP-OLS estimator defined in (2.7), and $\hat{\Delta}_J$ is an NPIV-based estimator of Δ constructed below.

The shrinkage parameter λ governs the robustness-efficiency trade-off between NP-OLS and NPIV. When $\lambda = 0$, the estimator (4.3) coincides with NP-OLS which enjoys lower sampling variance but incurs endogeneity bias. As λ increases, the estimator (4.3) corrects a larger fraction of this bias but relies more heavily on the NPIV-based endogeneity component whose sampling variation is amplified by ill-posedness. Thus, λ balances the lower sampling variance of NP-OLS against the reduction in endogeneity bias from using NPIV information.

A natural estimator of Δ, motivated by Hausman (1978), is the difference between NP-OLS and sieve 2SLS estimators. However, these estimators generally have different population projection targets. To see this, let $\Pi_J : L^2(X) \to \mathcal{H}_J$ and $\Pi_K : L^2(Z) \to \mathcal{Q}_K$ denote least square projections

[6] See e.g. Marchand and Strawderman (2004) for a review. A related econometric literature studies the performance of estimators over expanded models indexed by a bound on the magnitude of misspecification relative to a baseline model; see Armstrong (2025, Sections 2.2 and 7.1) for a recent review and Kitamura et al. (2013); Bonhomme and Weidner (2022); Armstrong et al. (2025) for examples.

onto the sieve spaces $\mathcal{H}_J$ and $\mathcal{Q}_K$:

$$\Pi_J f := \arg\min_{h \in \mathcal{H}_J} \|h - f\|_X, \qquad \Pi_K m := \arg\min_{q \in \mathcal{Q}_K} \|q - m\|_Z$$

for $f \in L^2(X)$ and $m \in L^2(Z)$. The $L^2(X)$-projection of Δ onto $\mathcal{H}_J$ is $\Delta_J := \Pi_J \Delta = \Pi_J g_0 - \Pi_J h_0$. While the NP-OLS estimator targets $\Pi_J g_0$, the sieve 2SLS targets the population projection $\Xi_J h_0$ where $\Xi_J : L^2(X) \to \mathcal{H}_J$ is defined by

$$\Xi_J f := \arg\min_{h \in \mathcal{H}_J} \left\| \Pi_K T(h - f) \right\|_Z.$$

Unlike Π_J, the operator Ξ_J is generally not an orthogonal projection in $L^2(X)$. Its approximation error $\Xi_J h_0 - h_0$ need not be orthogonal to $\mathcal{H}_J$, yet our MSE analysis requires such orthogonality. To address this issue, we follow the two-stage projection approach of Horowitz (2014a): we first construct a common pilot estimator of Δ, then project it onto lower-dimensional sieve spaces $\mathcal{H}_J$. Because the second step is an orthogonal projection, the resulting estimator $\widehat{\Delta}_J$ targets Δ_J directly.

To implement this two-stage approach, let $J_{\Delta,\max}$ denote a deterministic pilot sieve dimension. In the first stage, we construct a pilot estimator of Δ by

$$\widetilde{\Delta}_{J_{\Delta,\max}} = \widehat{g}_{J_{\Delta,\max}} - \widetilde{h}_{\mathrm{IV},J_{\Delta,\max}} \tag{4.4}$$

where $\widetilde{h}_{\mathrm{IV},J}$ is the sieve 2SLS estimator defined in (2.3).

In the second stage, we project this common pilot estimator (4.4) onto lower-dimensional sieve spaces. For any $f \in L^2(X)$, define the empirical least square projection onto $\mathcal{H}_J$ by

$$\widehat{\Pi}_J f := \arg\min_{h \in \mathcal{H}_J} \|h - f\|_{X,n}.$$

We then estimate the endogeneity bias Δ by

$$\widehat{\Delta}_J = \widehat{\Pi}_{J \wedge J_{\Delta,\max}} \widetilde{\Delta}_{J_{\Delta,\max}} \tag{4.5}$$

The truncation at $J_{\Delta,\max}$ has an important implication for the interpretation of the shrinkage estimator. When $J < J_{\Delta,\max}$, the estimator (4.5) can be written as $\hat{h}_{\lambda,J} = (1-\lambda)\hat{g}_J + \lambda \hat{\Pi}_J \widetilde{h}_{\mathrm{IV},J_{\Delta,\max}}$ which is a weighted average of NP-OLS and the projected sieve 2SLS pilot. However, when $J \geq J_{\Delta,\max}$, the endogeneity component remains fixed at the pilot dimension, while NP-OLS uses a larger sieve. In this case, even with $\lambda = 1$, the estimator becomes $\hat{h}_{1,J} = \widetilde{h}_{\mathrm{IV},J_{\Delta,\max}} + (\hat{g}_J - \hat{g}_{J_{\Delta,\max}})$ which generally differs from the sieve 2SLS pilot $\widetilde{h}_{\mathrm{IV},J_{\Delta,\max}}$ because it contains the additional NP-OLS components beyond the pilot dimension.

Under the regularity conditions stated below, the estimator $\hat{\Delta}_J$ is asymptotically centered at $\Delta_{J \wedge J_{\Delta,\max}}$, whose approximation error is orthogonal to the corresponding sieve space in $L^2(X)$. In practice, we replace $J_{\Delta,\max}$ with its empirical counterpart $\hat{J}_{\Delta,\max}$ (see Appendix A.2) in both

stages of the construction (4.4)-(4.5).

We next impose regularity conditions to control the sampling variation and approximation errors of the NPIV-based estimator $\hat{\Delta}_J$. Following the notation in Chen and Christensen (2018), define the population Gram and cross-moment matrices by

$$G_X = G_{X,J} = E[p^J(X)p^J(X)'], \quad G_Z = G_{Z,K} = E[q^K(Z)q^K(Z)'], \quad S = S_{KJ} = E[q^K(Z)p^J(X)'].$$

We assume throughout that G_X, G_Z and $S'G_Z^{-1}S$ are nonsingular. Define

$$\sigma_{JK} = \inf_{h \in \mathcal{H}_J} \frac{\|\Pi_K T h\|_Z}{\|h\|_X} = \sigma_{\min}\left(G_Z^{-1/2} S G_X^{-1/2}\right)$$

where $\sigma_{\min}(\cdot)$ denotes the smallest singular value. We also define the normalized sieve basis envelope $\zeta_{X,J} = \sup_x \|G_X^{-1/2}p^J(x)\|$, $\zeta_{Z,K} = \sup_z \|G_Z^{-1/2}q^K(z)\|$, and let $\zeta_J = \max(\zeta_{X,J}, \zeta_{Z,K})$. We assume either $\zeta_J = O(\sqrt{J})$ or $\zeta_J = O(J)$.

Assumption 4.3. *Let $U_i = Y_i - h_0(X_i)$ and $\varepsilon_i = Y_i - g_0(X_i)$ where $E[\varepsilon_i \mid X_i] = 0$. We assume $\sup_z E[U_i^2 \mid Z_i = z] \le \bar{\sigma}_U^2 < \infty$, $E[U_i^4] < \infty$, and $\sup_x E[\varepsilon_i^2 \mid X_i = x] \le \bar{\sigma}_\varepsilon^2$.*

Assumption 4.4. *We assume that*

(i) The sieve dimensions satisfy

$$\frac{\sigma_{JK}^{-2}\zeta_J\sqrt{(K\log J)/n}}{\left(J^{-1}\sum_{j=1}^J \nu_j^{-2}\right)^{1/2}} = O(1), \qquad \frac{\sigma_{JK}^{-1}\zeta_J J^{-s/d_x}\sqrt{\log J}}{\left(J^{-1}\sum_{j=1}^J \nu_j^{-2}\right)^{1/2}} = O(1)$$

(ii) $\sigma_{JK}^{-1}\|\Pi_K T(\Pi_J h - h)\|_Z \le C_T\|\Pi_J h - h\|_X$ for a constant $C_T > 0$, uniformly over $h \in \mathcal{H}$;

(iii) Let $\mathcal{H}_{J,1} := \{h \in \mathcal{H}_J : \|h\|_X = 1\}$. The instrument sieve approximates $T\mathcal{H}_J$ sufficiently well in the sense that $\sup_{h\in\mathcal{H}_{J,1}} \tau_J\|(\Pi_K T - T)h\|_Z = o(1)$;

(iv) the left-hand inequality of (4.2) holds.

Assumption 4.4 is analogous to Assumptions 2-3 of Breunig and Chen (2024). Condition (i) restricts the growth of the sieve dimensions to control errors from sieve estimation of the conditional expectation operator T and sieve approximation. Condition (ii) is analogous to the L^2 stability condition commonly imposed in sieve NPIV estimation (Blundell et al., 2007, Assumption 6). Condition (iii) ensures that σ_{JK}^{-1} and τ_J are asymptotically equivalent, while condition (iv) implies $\tau_J \lesssim \nu_J^{-1}$, see Breunig and Chen (2024, p. 2038) for more details.

We now consider an oracle that jointly chooses the shrinkage weight λ and sieve dimension J to minimize MSE within the shrinkage family. The following theorem shows that, under this oracle choice, the resulting estimator attains the rate in the minimax lower bound in Theorem 4.1.

Theorem 4.2. *Let Assumptions 2.1 and 4.1 to 4.4 hold. Then we have that*

$$\sup_{F_n \in \mathcal{F}(\rho_n)} \inf_{\lambda \in [0,1], J} E_{F_n} \left\| \widehat{h}_{\lambda,J} - h_0 \right\|_X^2 \lesssim \begin{cases} r_{n,\mathrm{OLS}}, & \rho_n \lesssim r_{n,\mathrm{OLS}} \\ \rho_n, & r_{n,\mathrm{OLS}} \lesssim \rho_n \lesssim r_{n,\mathrm{IV}} \\ r_{n,\mathrm{IV}}, & \rho_n \gtrsim r_{n,\mathrm{IV}} \end{cases}$$

where $r_{n,\mathrm{OLS}}$ and $r_{n,\mathrm{IV}}$ denote the optimal mean squared error rates for NP-OLS and NPIV estimation, respectively, as defined in (2.9) and (2.6).

Theorem 4.2 shows that the minimax optimal rate established in Theorem 4.1 can be attained within the proposed shrinkage family under an oracle choice of (λ, J). However, the oracle choice depends on unknown features of the data generating process and is therefore infeasible. In Section 5.1, we develop a data-driven procedure for selecting (λ, J) and show that the resulting estimator nearly attains this rate.

5 Adaptive Estimation

This section constructs a feasible shrinkage estimator and establishes its adaptive minimax optimality. Section 5.1 introduces a data-driven criterion for jointly selecting the shrinkage parameter and sieve dimension. Section 5.2 shows that the resulting estimator attains the minimax rate across the endogeneity regimes, up to the cost of adaptation. Section 5.3 explains the one-sided downward adjustment of the estimated endogeneity component that is central to adaptation. Finally, Section 5.4 compares our procedure with a pre-test estimator and shows that the latter is not rate-optimal.

5.1 The Feasible Shrinkage Estimator

Section 4.2 shows that the shrinkage estimator (4.3) attains the minimax optimal rate in each endogeneity regime when the shrinkage parameter λ and sieve dimension J are chosen by an oracle to minimize mean squared error (MSE). This oracle choice is infeasible because it depends on unknown features of the data generating process. We now develop a feasible procedure that jointly selects (λ, J) by minimizing an empirical criterion derived from an asymptotic upper bound on MSE, as outlined in Section 3.2.

Recall that the shrinkage estimator is $\hat{h}_{\lambda,J} = \hat{g}_J - \lambda \hat{\Delta}_J$, where $\hat{g}_J$ is the sieve NP-OLS estimator and $\hat{\Delta}_J$ is an NPIV-based estimator of Δ. The structural function is $h_0 = g_0 - \Delta$. We can therefore decompose the estimation error as

$$\widehat{h}_{\lambda,J} - h_0 = (\widehat{g}_J - g_0) - \left(\lambda \widehat{\Delta}_J - \Delta\right).$$

Consequently, the MSE is bounded by

$$E\left\|\widehat{h}_{\lambda,J} - h_0\right\|_X^2 \leq 2E\left\|\hat{g}_J - g_0\right\|_X^2 + 2E\left\|\lambda\hat{\Delta}_J - \Delta\right\|_X^2 \tag{5.1}$$

The upper bound (5.1) separates the mean squared error into an NP-OLS component and an endogeneity bias component. We first derive an asymptotic risk criterion for each component and then construct its feasible counterpart.

To obtain the asymptotic risk criterion, we first establish stochastic expansions of the two estimators that hold uniformly over the set of candidate sieve dimensions $\mathcal{J}_n$. Recall that Π_J denotes the $L^2(X)$-orthogonal projection onto $\mathcal{H}_J$ and $\Delta_J = \Pi_J\Delta$. The following proposition provides the required uniform expansions. As in Section 4.2, we present the result using a deterministic pilot dimension $J_{\Delta,\max}$, which is replaced by $\hat{J}_{\Delta,\max}$ in practice.

Proposition 5.1 (Uniform Asymptotic Expansion)**.** *Let Assumptions 2.1 and 4.1 to 4.4 hold. We have the following decomposition*

$$\hat{g}_J - \Pi_J g_0 = \Psi_{g,J} + r_{g,J}, \qquad \textit{uniformly over } J \in \mathcal{J}_n$$

$$\hat{\Delta}_J - \Delta_J = \Psi_{\Delta,J} + r_{\Delta,J} \qquad \textit{uniformly over } J \in \mathcal{J}_n, J \leq J_{\Delta,\max},$$

where $\Psi_{g,J}$ and $\Psi_{\Delta,J}$ are mean-zero leading stochastic terms satisfying $E[\Psi_{g,J}] = E[\Psi_{\Delta,J}] = 0$. The remainder terms are uniformly negligible relative to the corresponding leading terms:

$$\max_{J\in\mathcal{J}_n} \frac{\|r_{g,J}\|_X}{\left(E\|\Psi_{g,J}\|_X^2\right)^{1/2}} = o_p(1), \qquad \max_{\substack{J\in\mathcal{J}_n \\ J\leq J_{\Delta,\max}}} \frac{\|r_{\Delta,J}\|_X}{\left(E\|\Psi_{\Delta,J}\|_X^2\right)^{1/2}} = o_p(1).$$

The expressions of the leading stochastic terms and proofs are given in Appendix A.3.

Following Horowitz (2014a), let $E_A[\cdot]$ denote the expectation of the leading term in the relevant asymptotic expansion. In particular, $E_A[\hat{g}_J - \Pi_J g_0] = 0$ and $E_A[\hat{\Delta}_J - \Delta_J] = 0$. We define the corresponding asymptotic variance components as

$$V_{g,J} := E\left\|\Psi_{g,J}\right\|_X^2, \qquad V_{\Delta,J} := E\left\|\Psi_{\Delta,J}\right\|_X^2.$$

These expansions yield a decomposition of each component in the MSE upper bound (5.1). We first consider the NP-OLS component. Since $\hat{g}_J \in \mathcal{H}_J$, and Π_J is the $L^2(X)$-orthogonal projection onto $\mathcal{H}_J$, the sieve approximation error $g_0 - \Pi_J g_0$ is orthogonal to $\mathcal{H}_J$. Hence,

$$E_A\left\|\hat{g}_J - g_0\right\|_X^2 = V_{g,J} + \left\|\Pi_J g_0 - g_0\right\|_X^2 = V_{g,J} - \left\|\Pi_J g_0\right\|_X^2 + \left\|g_0\right\|_X^2. \tag{5.2}$$

We next consider the endogeneity bias component in (5.1). For $J \leq J_{\Delta,\max}$, the estimator $\hat{\Delta}_J$

is asymptotically centered at $\Delta_J = \Pi_J \Delta$, and the sieve approximation error $\Delta - \Delta_J$ is orthogonal to $\mathcal{H}_J$. It follows that

$$E_A \left\| \lambda \hat{\Delta}_J - \Delta \right\|_X^2 = \lambda^2 V_{\Delta,J} + \left[(1-\lambda)^2 - 1 \right] \left\| \Delta_J \right\|_X^2 + \left\| \Delta \right\|_X^2 \tag{5.3}$$

Because the construction of $\hat{\Delta}_J$ is truncated at $J_{\Delta,\max}$, the same decomposition applies for larger sieve dimensions J replaced by $J \wedge J_{\Delta,\max}$.

Combining the two decompositions (5.2) and (5.3) gives

$$E_A \left\| \widehat{h}_{\lambda,J} - h_0 \right\|_X^2 \leq 2R_n(\lambda, J) + 2 \left\| g_0 \right\|_X^2 + 2 \left\| \Delta \right\|_X^2 \tag{5.4}$$

where

$$R_n(\lambda, J) = \underbrace{V_{g,J} - \left\| \Pi_J g_0 \right\|_X^2}_{\text{NP-OLS}} + \underbrace{\lambda^2 V_{\Delta, J \wedge J_{\Delta,\max}} + \left[(1-\lambda)^2 - 1 \right] \left\| \Delta_{J \wedge J_{\Delta,\max}} \right\|_X^2}_{\text{endogeneity}}. \tag{5.5}$$

The last two terms in (5.4) do not depend on (λ, J). Thus, minimizing $R_n(\lambda, J)$ is equivalent to minimizing the asymptotic MSE upper bound, motivating a feasible selection criterion based on its empirical counterpart.

We now construct a feasible counterpart to $R_n(\lambda, J)$ by replacing its unknown population quantities with sample analogues. Let $\hat{V}_{g,J}$ and $\hat{V}_{\Delta,J}$ denote the uniformly consistent estimators of $V_{g,J}$ and $V_{\Delta,J}$, respectively, whose constructions are given in Appendix A.4. Because $\hat{\Delta}_J$ is truncated at the pilot dimension $J_{\Delta,\max}$, its estimated variance $\widehat{V}_{\Delta,J}$ is evaluated at $J \wedge J_{\Delta,\max}$ throughout.

We first consider the NP-OLS component in (5.5). Since $E_A \|\hat{g}_J\|_X^2 = \|\Pi_J g_0\|_X^2 + V_{g,J}$, the population criterion satisfies that $V_{g,J} - \|\Pi_J g_0\|_X^2 = 2V_{g,J} - E_A \|\hat{g}_J\|_X^2$. This motivates the use of

$$2\widehat{V}_{g,J} - \left\| \hat{g}_J \right\|_{X,n}^2.$$

This criterion is equivalent, up to an additive constant independent of J, to the generalized Mallows criterion for sieve least squares, see Li (1987); Andrews (1991).

We next estimate the endogeneity component in (5.5), which depends on the unknown squared bias $\|\Delta_J\|_X^2$. A naive plug-in estimate $\|\hat{\Delta}_J\|_{X,n}^2$ can substantially overstate the true bias because $\hat{\Delta}_J$ inherits the sampling variation amplified by the ill-posedness of the NPIV problem. Since $(1-\lambda)^2 - 1 \leq 0$ in the population criterion (5.5), such overestimation can cause the criterion to place greater shrinkage weight on the NPIV component, even when endogeneity is small.

To address this issue, we use a one-sided downward-adjusted estimate of $\|\Delta_J\|_X^2$. For each $J \in \mathcal{J}_n$, define

$$\widehat{B}_{\Delta,J} = \left[\left\| \hat{\Delta}_J \right\|_{X,n}^2 - \hat{\kappa}_{n,J} \hat{V}_{\Delta,J} \right]_+ \tag{5.6}$$

where $[a]_+ = \max\{a, 0\}$. For any fixed $\eta > 0$, the inflation factor is

$$\widehat{\kappa}_{n,J} = \left(2 + \frac{\eta}{4}\right)\left[\log\left(e + \frac{3\widehat{V}_{\Delta,J}}{\widehat{V}_{g,J}}\right) + \log(2|\mathcal{J}_n|)\right]. \tag{5.7}$$

We set $\eta = 0.04$ in our implementation. The adjustment (5.6) subtracts an inflated variance penalty from $\|\hat{\Delta}_J\|_{X,n}^2$ and truncates the resulting estimate at zero, so the estimated squared endogeneity bias enters the selection criterion only when it exceeds the variance penalty. The inflation factor (5.7) accounts for the relative sampling variance of the endogeneity component and selection over the candidate sieve dimensions. As shown in Section 5.3, this adjustment controls upward stochastic fluctuations at the NP-OLS variance scale, which is essential for adaptation.

Combining the two components, we obtain the feasible counterpart of the population risk criterion $R_n(\lambda, J)$:

$$\widehat{R}_n(\lambda, J) := 2\widehat{V}_{g,J} - \left\|\hat{g}_J\right\|_{X,n}^2 + \lambda^2\widehat{V}_{\Delta,J} + \left[(1-\lambda)^2 - 1\right]\widehat{B}_{\Delta,J}. \tag{5.8}$$

We jointly select the shrinkage parameter and sieve dimension by minimizing this criterion:

$$(\hat{\lambda}, \hat{J}) \in \underset{\lambda\in[0,1],\ J\in\mathcal{J}_n}{\arg\min}\ \widehat{R}_n(\lambda, J). \tag{5.9}$$

The resulting feasible shrinkage estimator is

$$\widehat{h}_{\hat{\lambda},\hat{J}} = \widehat{g}_{\hat{J}} - \hat{\lambda}\widehat{\Delta}_{\hat{J}}. \tag{5.10}$$

Both tuning parameters are selected from the data, without prior knowledge of the magnitude of endogeneity. The next subsection establishes the adaptation and rate optimality of this estimator.

5.2 Adaptation Results

In this subsection, we establish that the feasible shrinkage estimator (5.10) is nearly minimax rate optimal and adapts to the unknown magnitude of endogeneity. We then discuss how our results relate to the existing literature and compare the mean squared error of the feasible estimator with that of its infeasible oracle counterpart.

To establish the adaptation result, we strengthen the moment conditions on the error terms.

Assumption 5.1 (Conditional Bernstein moments)**.** *Recall that $U_i = Y_i - h_0(X_i), \varepsilon_i = Y_i - g_0(X_i)$, and let $\bar{\sigma}_U^2$ and $\bar{\sigma}_\varepsilon^2$ be as in Assumption 4.3. Then there exist finite constants $C_U, C_\varepsilon > 0$, independent of n and F_n, such that for every integer $k \geq 2$*

$$E\big[\,|U_i|^k \mid X_i, Z_i\big] \leq \frac{k!}{2}\,\bar{\sigma}_U^2\, C_U^{k-2}\ \ a.s., \qquad E\big[\,|\varepsilon_i|^k \mid X_i\big] \leq \frac{k!}{2}\,\bar{\sigma}_\varepsilon^2\, C_\varepsilon^{k-2}\ \ a.s.$$

Assumption 5.1 imposes conditional Bernstein moment bounds on the error terms U_i and ε_i. These conditions allow the errors to be unbounded while restricting the growth of their conditional moments sufficiently to obtain Bernstein-type concentration inequality for the empirical sums, see Massart (2007, Proposition 2.9) and Boucheron, Lugosi, and Massart (2013, Theorem 2.10).

The following theorem establishes the rate of convergence of the feasible shrinkage estimator (5.10) across the three endogeneity regimes.

Theorem 5.1. *Let Assumptions 2.1, 4.1 to 4.4 and 5.1 hold. Then, for all sufficiently large n, uniformly over $F_n \in \mathcal{F}(\rho_n)$,*

$$\sup_{F_n \in \mathcal{F}(\rho_n)} E_{F_n} \left\| \widehat{h}_{\hat{\lambda},\hat{\jmath}} - h_0 \right\|_X^2 \lesssim \begin{cases} r_{n,\mathrm{OLS}}, & \rho_n \lesssim r_{n,\mathrm{OLS}} \\ \rho_n, & r_{n,\mathrm{OLS}} \lesssim \rho_n \lesssim \tilde{r}_{n,\mathrm{IV}} \\ \tilde{r}_{n,\mathrm{IV}}, & \rho_n \gtrsim \tilde{r}_{n,\mathrm{IV}} \end{cases}$$

where $\tilde{r}_{n,\mathrm{IV}}$ denotes the adaptive NPIV rate. Compared with the minimax NPIV rate $r_{n,\mathrm{IV}}$ in (2.6),

$$\tilde{r}_{n,\mathrm{IV}} := \begin{cases} \left(\frac{\log n}{n}\right)^{\frac{2s}{2s+2\varsigma+d_x}} = (\log n)^{\frac{2s}{2s+2\varsigma+d_x}} r_{n,\mathrm{IV}}, & \textit{mildly ill-posed,} \\ (\log n)^{-2s/\varsigma} = r_{n,\mathrm{IV}}, & \textit{severely ill-posed.} \end{cases}$$

Theorem 5.1 establishes the minimax rate optimality and adaptation properties of the feasible shrinkage estimator (5.10). In our setting, the minimax optimal rate in Theorem 4.1 depends on the magnitude of endogeneity, the smoothness of the structural function, and the degree of ill-posedness. Our feasible shrinkage estimator (5.10) nearly attains this rate *without* prior knowledge of any of these features, up to a logarithmic adaptation cost in the severe endogeneity regime.

In the severe endogeneity regime, our feasible estimator is rate-adaptive up to a power of $\log n$ under mild ill-posedness, but attains the optimal rate under severe ill-posedness.[7] The reason there is no adaptation cost under severe ill-posedness is that the NPIV variance grows exponentially with sieve dimension, so the logarithmic inflation term affects the optimal sieve dimension only at a lower order and leaves the rate unchanged. Under mild ill-posedness, by contrast, the variance grows polynomially, so the same inflation reduces the optimal sieve dimension and induces a logarithmic factor into the rate of convergence.

Remark 5.1. We compare the notion of adaptation in Theorem 5.1 with that of Armstrong et al. (2025). They study adaptation to an unknown magnitude of misspecification, evaluating feasible estimators by their worst-case relative risk against an oracle that knows a bound on the misspecification. In our setting, adaptation instead concerns whether a single feasible estimator can attain

[7] This aligns with the adaptive NPIV estimation results of Chen and Christensen (2015a). To the best of our knowledge, it remains an open question whether a fully data-driven sieve NPIV estimator can attain the minimax $L^2(X)$ NPIV rate without logarithmic loss under general regularity conditions in the mildly ill-posed case.

the minimax optimal rate of convergence over each class $\mathcal{F}(\rho_n)$ without knowing the magnitude of endogeneity. The distinction arises because the minimax optimal rate varies with the magnitude of endogeneity in our setting, while it remains the same across misspecification bounds in their parametric setting. ■

Remark 5.2. Cheng et al. (2019) propose a data-dependent average of a conservative generalized method of moments (GMM) estimator based only on moment conditions known to be valid, and an aggressive GMM estimator that also uses potentially misspecified moments. They show that their averaging estimator uniformly asymptotically dominates the conservative estimator over different degrees of misspecification. Our minimax framework provides a different perspective. We show that NPIV, the conservative estimator in our setting, is not rate-optimal under limited or moderate endogeneity. Whether our proposed feasible shrinkage estimator uniformly dominates NPIV remains an open question for future research. ■

We also compare the mean squared error (MSE) of the feasible shrinkage estimator with the oracle benchmark studied in Section 4.2. The oracle risk is the minimum MSE attainable within the shrinkage family if (λ, J) were chosen with knowledge of the data generating process. The following proposition shows that the MSE of the feasible estimator (5.10) is within a logarithmic factor of the oracle risk.

Proposition 5.2. *Under the conditions of Theorem 5.1, we have that*

$$E_{F_n}\left\|\widehat{h}_{\hat{\lambda},\hat{J}} - h_0\right\|_X^2 \lesssim (\log n) \inf_{\substack{\lambda\in[0,1]\\ J\in\mathcal{J}_n}} E_{F_n}\left\|\widehat{h}_{\lambda,J} - h_0\right\|_X^2$$

where the implicit constant is independent of n.

Proposition 5.2 shows that the data-driven selector $(\hat{\lambda}, \hat{J})$ mimics the infeasible oracle up to a logarithmic factor. This result complements Theorem 5.1: the feasible shrinkage estimator (5.10) not only nearly attains the minimax optimal rate across endogeneity regimes but also performs nearly as well as the best candidate within the shrinkage family.

Remark 5.3. Proposition 5.2 connects to a broad literature on model and tuning parameter selection, where a data-driven procedure is evaluated by comparing its risk with the infeasible oracle benchmark within a family of candidates. A central optimality criterion in this literature is asymptotic selection optimality, which requires the risk of the selected estimator to asymptotically match the oracle risk (Li, 1987; Andrews, 1991, Hansen, 2022, Section 28.13, Chetverikov, 2024, pp.5-7). Our result makes the same oracle comparison but allows a logarithmic multiplicative factor and extends it to the joint selection of the shrinkage parameter and sieve dimension. ■

Remark 5.4. Related work has established oracle optimality results for model averaging and adaptive NPIV estimation. Hansen (2007); Hansen and Racine (2012) establish asymptotic oracle optimality for their data-driven averaging weights over a class of candidate weights. While their

results concern averaging least squares regression estimators, our shrinkage parameter λ governs the robustness-efficiency trade-off induced by endogeneity. For the sieve dimension, Horowitz (2014a) constructs a data-driven sieve choice for NPIV whose asymptotic integrated mean squared error is within a logarithmic factor of the infeasible optimum. Proposition 5.2 establishes an oracle inequality for the joint selection of the shrinkage parameter and sieve dimension, accounting for the ill-posedness of estimating endogeneity bias. ■

5.3 One-Sided Downward Adjustment

The adaptation result in Section 5.2 hinges on the treatment of the endogeneity component $\|\Delta_J\|_X^2$ in the selection criterion. Since $\widehat{\Delta}_J$ contains the NPIV component, its sampling variation is amplified by ill-posedness and may dominate $\|\Delta_J\|_X^2$ under limited or moderate endogeneity. The adaptation proof must therefore prevent this NPIV-scale stochastic error from entering the selection criterion when the endogeneity bias is small. For this purpose, it suffices to control the upper fluctuations of $\hat{\Delta}_J - \Delta_J$. Specifically, the key step in our proof is to establish

$$E\left[\left\|\hat{\Delta}_J - \Delta_J\right\|_{X,n}^2 - \kappa_{n,J} V_{\Delta,J}\right]_+ \leq \frac{3V_{\Delta,J}}{|\mathcal{J}_n|(e + 3V_{\Delta,J}/V_{g,J})} \leq \frac{V_{g,J}}{|\mathcal{J}_n|}$$

where $\kappa_{n,J}$ denotes the population counterpart of the inflation factor $\hat{\kappa}_{n,J}$ in (5.7).

A Bernstein-type concentration argument yields the first inequality and shows that $\kappa_{n,J} V_{\Delta,J}$ is an upper function for $\|\hat{\Delta}_J - \Delta_J\|_{X,n}^2$ in the sense that its expected positive excess is small. The concentration level is calibrated so that this excess is at the NP-OLS variance scale. Consequently, the stochastic remainder from estimating Δ_J remains at the NP-OLS scale and therefore does not suffer from the ill-posedness when endogeneity is limited or moderate. This one-sided control is the key to the adaptation result. By contrast, unbiasedness alone is insufficient: centering an estimator at $\|\Delta_J\|_X^2$ does not control its positive fluctuations, which remain at the NPIV scale.

This need for one-sided control distinguishes our criterion from existing MSE-based model selection procedures under potential misspecification, where the unknown misspecification bias must also be estimated. In parametric settings, Cheng et al. (2019) use an uncorrected quadratic plug-in, Claeskens and Hjort (2003) construct an asymptotically unbiased estimate of the squared bias, and DiTraglia (2016) consider its positive-part modification. In our setting, the estimated bias component relies on NPIV whose sampling variation is amplified by ill-posedness. We therefore use a downward-adjusted estimate to control its upper stochastic fluctuations. This one-sided adjustment is essential for adaptation under limited and moderate endogeneity.

5.4 Comparison with a Pre-Test Estimator

A natural alternative to our shrinkage estimator is a pre-test estimator which first tests whether X_i is exogenous (Hausman, 1978; Blundell and Horowitz, 2007) and then selects between NP-OLS

and NPIV. Let $\phi_n \in \{0, 1\}$ denote the rejection indicator for testing

$$H_0 : E[U_i \mid X_i] = 0, \text{ a.s.} \qquad \text{versus} \qquad H_1 : P\left(E[U_i \mid X_i] \neq 0\right) > 0$$

with $\phi_n = 1$ corresponding to a rejection. The resulting pre-test estimator uses NP-OLS when exogeneity is not rejected and NPIV otherwise.

A pre-test estimator based on a fixed-level test cannot achieve adaptation over the limited and moderate endogeneity regimes. To illustrate this, consider exact exogeneity $E[U_i \mid X_i] \equiv 0$. If the test has asymptotic size $\alpha \in (0, 1)$, then $P_{F_n}(\phi_n = 1) \to \alpha$, so NPIV is selected with nonvanishing probability even though NP-OLS is correctly specified. Because NPIV converges at the slower rate, this non-vanishing selection probability keeps the risk of the pre-test estimator at the NPIV scale. By contrast, the minimax risk under exact exogeneity is at the faster NP-OLS scale. The pre-test estimator is therefore not rate-optimal.

The underlying issue of the pre-test estimator is that testing exogeneity and minimizing mean squared error are different objectives. A test of exogeneity asks whether endogeneity can be statistically distinguished from zero, while MSE-optimal estimation asks whether the endogeneity bias is large enough for the reduction in bias from using NPIV to outweigh its loss of precision. This distinction is essential in our setting because ill-posedness makes the variance of NPIV substantially larger than that of NP-OLS. Adaptation therefore requires the probability of selecting NPIV under limited or moderate endogeneity to vanish sufficiently fast for the resulting NPIV-scale variance to be asymptotically negligible. A fixed-level test does not provide this risk-based calibration, because it continues to select NPIV with nonvanishing probability even when the endogeneity bias is small.

The non-optimality of the pre-test estimator in our setting is related to the classical literature on the risk properties of pre-test estimators. Specifically, Sclove, Morris, and Radhakrishnan (1972) show that pre-test estimators can be dominated by alternative estimators, while Bock, Yancey, and Judge (1973) show that their MSE can substantially exceed that of the unrestricted estimator for some parameter values. See also Judge and Bock (1983, Sections 4.1.2 and 5.1.4) and Leamer (1983, Section 3.1) for surveys of these risk properties. In our setting, the pre-test estimator likewise has unfavorable risk properties, as it fails to attain the minimax optimal rate.

6 Monte Carlo Simulation

We investigate the finite-sample performance of the adaptive estimator through Monte Carlo simulations. The Monte Carlo designs vary the magnitude of endogeneity while holding fixed the structural function, the joint distribution of the regressor and instrument, and the error variance. We compare the mean squared errors (MSEs) of the adaptive estimator, NP-OLS, and NPIV, with their tuning parameters selected from the data. The simulations show that the adaptive estimator achieves substantial MSE gains relative to NPIV under limited endogeneity and performs com-

parably to NPIV when endogeneity is substantial. The results also reveal a finite-sample cost of adaptation at intermediate magnitudes of endogeneity.

6.1 Monte Carlo Design

Following Breunig and Chen (2024, Equation 5.2), we consider the structural function

$$h_0(x) = 0.1\left[1 - 2\Phi\left(\frac{x - 1/2}{0.1}\right)\right]$$

where $\Phi(\cdot)$ denotes the standard normal distribution function. The function is smooth and monotonically decreasing, with a relatively steep slope around $x = 1/2$. We generate observations from the structural model (2.1) where

$$\begin{pmatrix} X_i \\ Z_i \\ U_i \end{pmatrix} \sim N\left(\mathbf{0}_3, \begin{pmatrix} 1 & 0.5 & \sqrt{\rho_n} \\ 0.5 & 1 & 0 \\ \sqrt{\rho_n} & 0 & \sigma_U^2 \end{pmatrix}\right)$$

and $\sigma_U^2 = \mathrm{Var}(U_i)$. By construction, the instrument validity condition $E[U_i \mid Z_i] = 0$ holds, and $\|E[U_i \mid X_i]\|_X^2 = \rho_n$ corresponds to the magnitude of endogeneity defined in (3.1). We fix the error variance σ_U^2 such that the signal-to-noise ratio $\mathrm{Var}(h_0(X_i))/\sigma_U^2 = 0.5$ and vary the normalized magnitude of endogeneity ρ_n/σ_U^2.

We implement the estimators using quadratic B-spline basis functions for both the regressor and the instrument. Following Chen et al. (2025), we consider a nested sequence of sieve spaces indexed by resolution levels j. At each level, the basis for the regressor X has $2^j - 1$ interior knots placed at equally spaced empirical quantiles, yielding a sieve dimension of $J = 2^j + 2$. In our simulation, the candidate sieve dimensions for the regressor are $\mathcal{J} = \{3, 4, 6, 10, 18, 34, 66\}$. For the instrument, we use a finer quadratic B-spline basis, taking $K = 2J - 2$ for the adaptive estimator and $K = 8J - 14$ for the NPIV estimator.

We compare three feasible estimators, with their tuning parameters selected from the data. For the adaptive estimator $\hat{h}_{\hat{\lambda},J}$, we jointly select $(\hat{\lambda}, \hat{J})$ by minimizing the empirical risk criterion in (5.9). For the NP-OLS estimator $\hat{g}_J$ defined in (2.7), we select its sieve dimension by

$$\widehat{J}_{\mathrm{OLS}} \in \arg\min_{J \in \mathcal{J}} \quad 2\widehat{V}_{g,J} - \left\|\hat{g}_J\right\|_{X,n}^2.$$

For the NPIV estimator, following Horowitz (2014a), we construct a common sieve 2SLS pilot at the pilot sieve dimension $\hat{J}_{\mathrm{IV,max}}$, and project it onto lower-dimensional sieve spaces as described in Section 4.2. Denote the resulting estimators by

$$\widehat{h}_{\mathrm{IV},J} := \widehat{\Pi}_J \widetilde{h}_{\mathrm{IV},\hat{J}_{\mathrm{IV,max}}} \qquad J \le \hat{J}_{\mathrm{IV,max}}.$$

To select the sieve dimension, we adopt the empirical risk criterion of Horowitz (2014a, Equation 3.3). Let $\widehat{V}_{\mathrm{IV},J}$ denote the estimated variance of $\widehat{h}_{\mathrm{IV},J}$. We select

$$\widehat{J}_{\mathrm{IV}} \in \underset{J \leq \hat{J}_{\mathrm{IV,max}}}{\arg\min} \quad \frac{2}{3}\log(n)\widehat{V}_{\mathrm{IV},J} - \left\|\widehat{h}_{\mathrm{IV},J}\right\|_{X,n}^2.$$

We evaluate the finite-sample performance of the three estimators by comparing their mean squared errors as the normalized magnitude of endogeneity ρ_n/σ_U^2 varies. For each value of ρ_n/σ_U^2, we conduct 1000 Monte Carlo replications, each with $n = 3000$ observations. We compute MSE by averaging the integrated squared error $\|\widehat{h} - h_0\|_X^2$ across replications, where the norm is evaluated under the marginal distribution of X.

6.2 Simulation Results

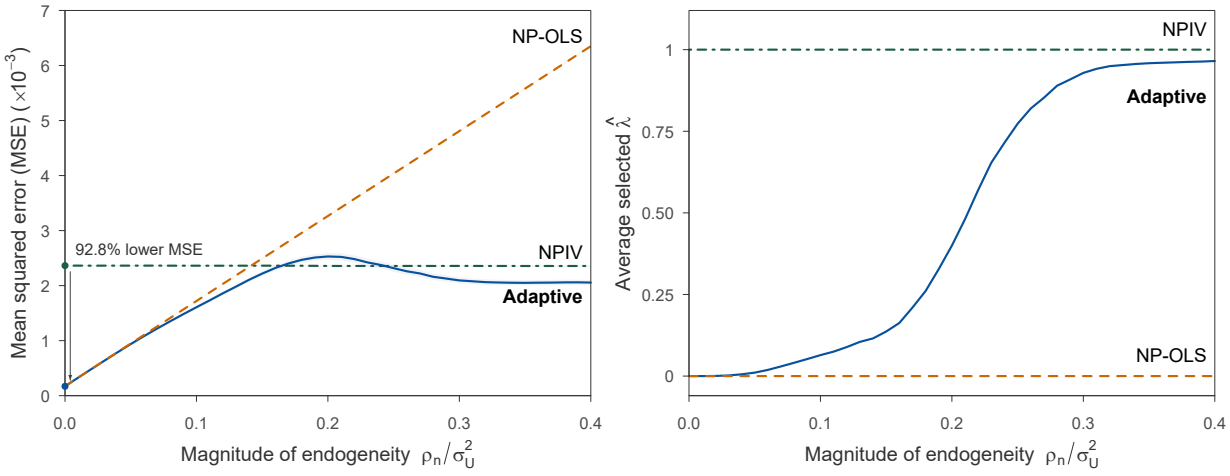


Figure 2: *Mean Squared Error versus Magnitude of Endogeneity.*

Note. The left panel reports the Monte Carlo mean squared error (MSE) of the adaptive estimator, NP-OLS, and NPIV as the magnitude of endogeneity ρ_n varies. The right panel reports the average selected shrinkage parameter $\hat{\lambda}$ for the adaptive estimator. The annotation indicates the percentage reduction in MSE of the adaptive estimator relative to NPIV under exogeneity.

Figure 2 illustrates how the MSEs of the three estimators vary with the magnitude of endogeneity. The MSE of NP-OLS increases approximately linearly with endogeneity, while that of NPIV remains relatively stable. Under exogeneity, the adaptive estimator coincides with NP-OLS and reduces MSE by 92.8% relative to NPIV. As endogeneity increases, the MSE of the adaptive estimator initially rises, exceeds that of NPIV at intermediate magnitudes, and subsequently declines as the estimator places greater weight on the NPIV-based endogeneity component. When endogeneity is substantial, the adaptive estimator performs comparably to NPIV, although the two estimators need not coincide in finite samples. As explained in Section 4.2, even when the shrinkage parameter is set to $\lambda = 1$, the NPIV-based endogeneity estimate $\hat{\Delta}_J$ is truncated at a pilot sieve

dimension, while the NP-OLS component may use a larger sieve dimension.

The MSE patterns in Figure 2 can be explained by the data-dependent choice of the shrinkage parameter. Under exogeneity, the proposed criterion with the one-sided downward adjustment selects $\hat{\lambda} = 0$ in nearly all Monte Carlo replications, so the adaptive estimator coincides with NP-OLS. By contrast, the naive plug-in criterion based on $\|\hat{\Delta}_J\|^2_{X,n}$ selects an average shrinkage parameter of 0.38, even though the true endogeneity bias is zero.

However, this downward adjustment comes at a finite-sample cost of adaptation. At intermediate magnitudes of endogeneity, the downward adjustment can be overly conservative, leading to insufficient correction for endogeneity bias and a higher MSE than NPIV. As endogeneity increases further, the endogeneity bias becomes easier to distinguish from sampling variation, allowing the selection criterion to choose larger shrinkage parameters. Consequently, the adaptive estimator places more weight on the endogeneity component, reducing its MSE and eventually performing comparably to NPIV.

Joint selection of the sieve dimension provides an additional source of MSE gains under limited endogeneity, particularly in this simulation design with a highly nonlinear structural function. Under exogeneity, the adaptive estimator coincides with NP-OLS, selecting an average sieve dimension of 15.9, compared with 3.1 for NPIV. By placing little weight on the NPIV-based endogeneity component, the adaptive estimator can use a richer sieve to reduce approximation bias without incurring the substantial variance cost associated with NPIV. Thus, in the nonparametric setting, the MSE gains from shrinking toward NP-OLS arise not only from lower sampling variance but also from the ability to use a larger sieve dimension to reduce approximation bias.

Overall, the simulations illustrate both the benefits and the finite-sample costs of adaptation. Under limited endogeneity, the adaptive estimator achieves substantial MSE gains relative to NPIV by exploiting the greater accuracy and flexibility of NP-OLS. At intermediate magnitudes of endogeneity, the one-sided downward adjustment can lead to insufficient correction for endogeneity bias, resulting in higher MSE than NPIV. As endogeneity increases further, the adaptive estimator places greater weight on the NPIV-based component and performs comparably to NPIV.

7 Empirical Demonstration: Gasoline Demand

We illustrate the proposed adaptive estimator in an application to household gasoline demand using data from Bento, Goulder, Jacobsen, and Von Haefen (2009). The data combine the 2001 National Household Travel Survey with regional gasoline prices from the American Chamber of Commerce Researchers Association (ACCRA). They match local market gasoline prices to households geographically and construct annual gasoline consumption from vehicle miles traveled and vehicle-specific fuel economy. We restrict the sample to households with children (strata 9-12 in their data) yielding 5619 households. In this sample, average annual gasoline consumption is 1121 gallons, average gasoline price is $1.44 per gallon, and average household income is approximately

$61,800. The 5th and 95th percentiles of gasoline prices are $1.27 and $1.62 per gallon, respectively.

We consider a nonparametric specification of household gasoline demand with a set of control variables:

$$\log Y_i = h_0(\log X_i) + C_i'\beta_0 + U_i$$

where Y_i denotes annual gasoline consumption, X_i is gasoline price, and C_i is a vector of controls consisting of log household income and quarter indicators. The structural function h_0 characterizes how gasoline consumption responds to changes in price, holding other determinants of demand fixed. In particular, the price elasticity of gasoline demand at price x is given by

$$\left.\frac{\partial \log Y_i}{\partial \log X_i}\right|_{X_i=x} = h_0'(\log x)$$

which may vary with the price level. Our specification extends the log-log linear regressions in Table 2 of Blundell et al. (2012) by allowing a flexible price response.

To address the potential endogeneity of gasoline prices, we follow Blundell et al. (2012) and use the distance from a major oil platform in the Gulf of Mexico to the household's state capital as an instrument, denoted by Z_i. As discussed in Example 2.1, this distance provides a potential source of supply-side variation in gasoline prices through transportation costs. We maintain the conditional moment restriction $E[U_i \mid Z_i, C_i] = 0$.

We estimate the structural demand function using the proposed adaptive estimator alongside NP-OLS and NPIV. Following the sieve construction in Section 6.1, we use quadratic B-spline basis functions for log gasoline price and the instrument, with knots placed at their respective empirical quantiles. We account for the control variables using the Frisch-Waugh-Lovell theorem, partialling out the controls from log gasoline consumption and the spline basis functions for both price and the instrument. For fixed sieve dimensions, this procedure yields the same spline coefficient estimates as including the controls directly in the corresponding sieve least squares and two-stage least squares regressions. We select the tuning parameters following the procedures in Section 6.1, computing the empirical norms and heteroskedasticity-robust variance estimates after partialling out the controls.

The left panel of Figure 3 presents the gasoline demand functions estimated by the adaptive estimator, NP-OLS, and NPIV. All three estimates suggest generally downward-sloping demand curves, although their shapes differ. At the sample median gasoline price, the adaptive estimator suggests less price-responsive demand, with an estimated price elasticity of -1.60, compared with -1.97 for NP-OLS and -1.99 for NPIV.

The right panel of Figure 3 shows that the feasible risk criterion of the adaptive estimator is minimized at $\hat{\lambda} = 0.33$ and $\hat{J} = 3$. This interior choice of shrinkage parameter yields a partial NPIV-based correction to NP-OLS, reflecting the trade-off between the reduction in estimated endogeneity bias and the additional sampling variance induced by the correction.

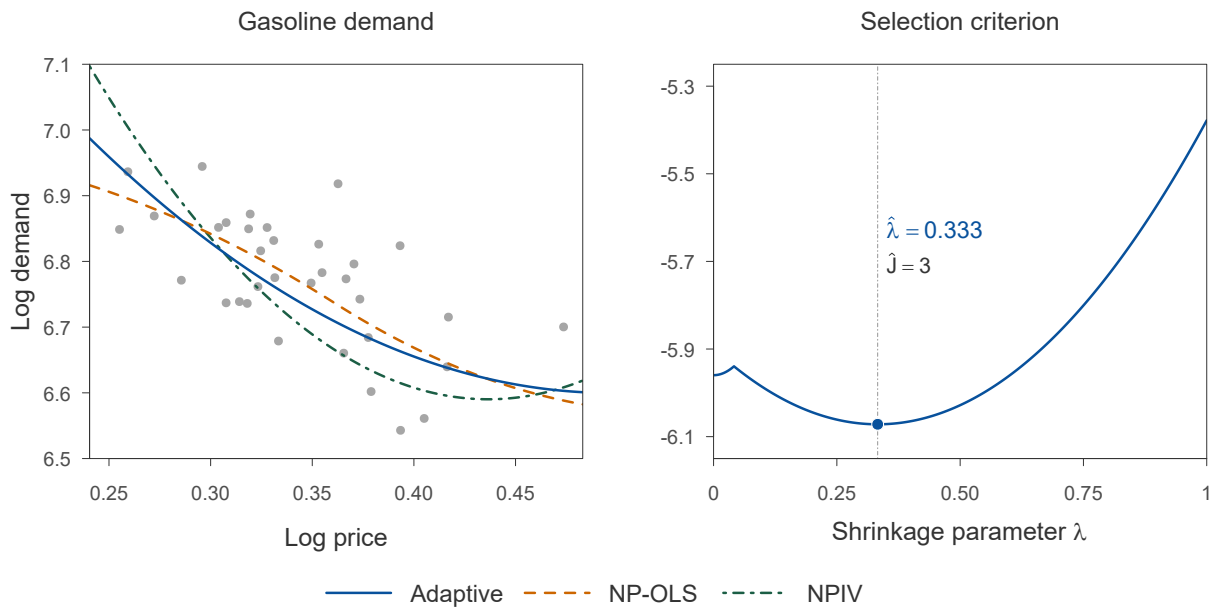


Figure 3: *Estimated Gasoline Demand and Shrinkage Parameter Selection.*

Note. The left panel compares the estimated gasoline demand functions from the adaptive estimator, NP-OLS, and NPIV. The gray points represent state-level averages of log gasoline price and consumption. The right panel plots the control-adjusted feasible risk criterion in (5.9), minimized over the sieve dimension J, as a function of the shrinkage parameter λ. The adaptive estimator selects $\hat{\lambda} = 0.33$ and $\hat{J} = 3$.

8 Conclusion

Endogeneity creates a trade-off between bias and variance in structural estimation. Least squares can estimate the conditional mean function at a faster rate but incurs endogeneity bias, while instrumental variables address this bias at the cost of greater sampling variance. To characterize this trade-off, we develop an asymptotic framework that allows the magnitude of endogeneity to vary with sample size, accommodating a continuum of endogeneity regimes. We formulate a minimax estimation problem over classes of data generating processes indexed by the magnitude of endogeneity. We show that the optimal rate of convergence ranges from the faster NP-OLS rate under limited endogeneity to the slower NPIV rate under severe endogeneity, with an intermediate regime determined by the magnitude of endogeneity itself. Thus, the trade-off induced by endogeneity is a fundamental feature of the problem of estimating nonparametric structural functions.

Our methodological contribution is a feasible shrinkage estimator that adapts to the unknown magnitude of endogeneity. The estimator starts from NP-OLS and subtracts an NPIV-based estimate of endogeneity bias, with the shrinkage parameter and sieve dimension jointly selected from the data. The key to adaptation is a one-sided downward adjustment of the estimated squared endogeneity bias that controls its upward stochastic fluctuations. We show that the resulting estimator nearly attains the corresponding minimax optimal rate in each endogeneity regime, without

prior knowledge of the magnitude of endogeneity.

More generally, our results suggest that an instrumental variables estimator need not yield the most accurate estimate of the structural function in terms of mean squared error. Although IV addresses endogeneity bias, it can induce substantially greater sampling variance than least squares, particularly in nonparametric models. The relevant question for estimation is therefore not simply whether the regressors are endogenous, but whether the reduction in endogeneity bias outweighs the additional sampling variance. This consideration motivates our adaptive estimator, which balances these two sources of estimation error. Overall, these results show that exploiting information from the reduced-form conditional mean function can improve the accuracy of nonparametric structural estimation, even in the presence of endogeneity.

References

Adão, R., C. Arkolakis, and S. Ganapati (2020): "From heterogeneous firms to heterogeneous trade elasticities: The aggregate implications of firm export decisions," Tech. rep., National Bureau of Economic Research.

Adao, R., A. Costinot, and D. Donaldson (2017): "Nonparametric counterfactual predictions in neoclassical models of international trade," *American Economic Review*, 107, 633–689.

Andrews, D. W. (1991): "Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors," *Journal of Econometrics*, 47, 359–377.

Andrews, I., M. Gentzkow, and J. M. Shapiro (2017): "Measuring the sensitivity of parameter estimates to estimation moments," *The Quarterly Journal of Economics*, 132, 1553–1592.

——— (2020): "On the informativeness of descriptive statistics for structural estimates," *Econometrica*, 88, 2231–2258.

Angrist, J. D., P. D. Hull, P. A. Pathak, and C. R. Walters (2017): "Leveraging lotteries for school value-added: Testing and estimation," *The Quarterly Journal of Economics*, 132, 871–919.

Armstrong, T. B. (2025): "Misspecification in Econometrics: A Selective Review.," .

Armstrong, T. B., P. Kline, and L. Sun (2025): "Adapting to misspecification," *Econometrica*, 93, 1981–2005.

Armstrong, T. B. and M. Kolesár (2021): "Sensitivity analysis using approximate moment condition models," *Quantitative Economics*, 12, 77–108.

Bento, A. M., L. H. Goulder, M. R. Jacobsen, and R. H. Von Haefen (2009): "Distributional and efficiency impacts of increased US gasoline taxes," *American Economic Review*, 99, 667–699.

Beran, R. (1974): "Asymptotically efficient adaptive rank estimates in location models," *The Annals of Statistics*, 63–74.

Berry, S. T. and P. A. Haile (2014): "Identification in differentiated products markets using market level data," *Econometrica*, 82, 1749–1797.

——— (2024): "Nonparametric identification of differentiated products demand using micro data," *Econo-*

metrica, 92, 1135–1162.

Bickel, P. J. (1981): "Minimax estimation of the mean of a normal distribution when the parameter space is restricted," *The Annals of Statistics*, 9, 1301–1309.

——— (1982): "On adaptive estimation," *The Annals of Statistics*, 647–671.

——— (1984): "Parametric robustness: small biases can be worthwhile," *The Annals of Statistics*, 12, 864–879.

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

Blundell, R. and J. L. Horowitz (2007): "A non-parametric test of exogeneity," *The Review of Economic Studies*, 74, 1035–1058.

Blundell, R., J. L. Horowitz, and M. Parey (2012): "Measuring the price responsiveness of gasoline demand: Economic shape restrictions and nonparametric demand estimation," *Quantitative Economics*, 3, 29–51.

Blundell, R. and J. L. Powell (2003): "Endogeneity in nonparametric and semiparametric regression models," *Econometric society monographs*, 36, 312–357.

Bock, M., T. Yancey, and G. Judge (1973): "The statistical consequences of preliminary test estimators in regression," *Journal of the American Statistical Association*, 68, 109–116.

Bonhomme, S. and M. Weidner (2022): "Minimizing sensitivity to model misspecification," *Quantitative Economics*, 13, 907–954.

Boucheron, S., G. Lugosi, and P. Massart (2013): *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford: Oxford University Press.

Breunig, C. and X. Chen (2019): "Simple adaptive estimation of quadratic functionals in nonparametric IV models," in *Foundations of Modern Statistics*, Springer, 459–482.

——— (2024): "Adaptive, Rate-Optimal Hypothesis Testing in Nonparametric IV Models," *Econometrica*, 92, 2027–2067.

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

Bugni, F. A., I. A. Canay, and P. Guggenberger (2012): "Distortions of asymptotic confidence size in locally misspecified moment inequality models," *Econometrica*, 80, 1741–1768.

Bugni, F. A. and T. Ura (2019): "Inference in dynamic discrete choice problems under local misspecification," *Quantitative Economics*, 10, 67–103.

Carrasco, M., J.-P. Florens, and E. Renault (2007): "Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization," *Handbook of econometrics*, 6, 5633–5751.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of econometrics*, 6, 5549–5632.

Chen, X., T. Christensen, and S. Kankanala (2025): "Adaptive estimation and uniform confidence bands for nonparametric structural functions and elasticities," *Review of Economic Studies*, 92, 162–

196.

Chen, X. and T. M. Christensen (2015a): "Optimal Sup-Norm Rates, Adaptivity and Inference in Nonparametric Instrumental Variables Estimation," *arXiv preprint arXiv:1508.03365*, version 1.

——— (2018): "Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression," *Quantitative Economics*, 9, 39–84.

Chen, X. and Y. J. J. Qiu (2016): "Methods for nonparametric and semiparametric regressions with endogeneity: A gentle guide," *Annual Review of Economics*, 8, 259–290.

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

Cheng, X., Z. Liao, and R. Shi (2019): "On uniform asymptotic risk of averaging GMM estimators," *Quantitative Economics*, 10, 931–979.

Chetverikov, D. (2024): "Tuning parameter selection in econometrics," *arXiv preprint arXiv:2405.03021*.

Claeskens, G. and N. L. Hjort (2003): "The focused information criterion," *Journal of the American Statistical Association*, 98, 900–916.

Compiani, G. (2022): "Market counterfactuals and the specification of multiproduct demand: A nonparametric approach," *Quantitative Economics*, 13, 545–591.

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

de Chaisemartin, C. and X. D'Haultfœuille (2020): "Empirical mse minimization to estimate a scalar parameter," *arXiv preprint arXiv:2006.14667*.

DiNardo, J. and J. L. Tobias (2001): "Nonparametric density and regression estimation," *Journal of Economic Perspectives*, 15, 11–28.

DiTraglia, F. J. (2016): "Using invalid instruments on purpose: Focused moment selection and averaging for GMM," *Journal of Econometrics*, 195, 187–208.

Feldstein, M. (1974): "Errors in variables: A consistent estimator with smaller MSE in finite samples," *Journal of the American Statistical Association*, 69, 990–996.

Fessler, P. and M. Kasy (2019): "How to use economic theory to improve estimators: Shrinking toward theoretical restrictions," *Review of Economics and Statistics*, 101, 681–698.

Freyberger, J. and J. L. Horowitz (2015): "Identification and shape restrictions in nonparametric instrumental variables estimation," *Journal of Econometrics*, 189, 41–53.

Gandhi, A., S. Navarro, and D. A. Rivers (2020): "On the identification of gross output production functions," *Journal of Political Economy*, 128, 2973–3016.

Green, E. J. and W. E. Strawderman (1991): "A James-Stein type estimator for combining unbiased and possibly biased estimators," *Journal of the American Statistical Association*, 86, 1001–1006.

Haavelmo, T. (1944): "The probability approach in econometrics," *Econometrica: Journal of the Econometric Society*, iii–115.

Hall, P. and J. L. Horowitz (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables," *The Annals of Statistics*, 33, 2904–2929.

Hansen, B. E. (2007): "Least squares model averaging," *Econometrica*, 75, 1175–1189.

——— (2014): "Model averaging, asymptotic risk, and regressor groups," *Quantitative Economics*, 5, 495–530.

——— (2016): "Efficient shrinkage in parametric models," *Journal of Econometrics*, 190, 115–132.

——— (2017): "Stein-like 2SLS estimator," *Econometric Reviews*, 36, 840–852.

——— (2022): *Econometrics*, Princeton University Press.

Hansen, B. E. and J. S. Racine (2012): "Jackknife model averaging," *Journal of Econometrics*, 167, 38–46.

Härdle, W. and O. Linton (1994): "Applied nonparametric methods," *Handbook of econometrics*, 4, 2295–2339.

Hausman, J. A. (1978): "Specification tests in econometrics," *Econometrica: Journal of the econometric society*, 1251–1271.

Hausman, J. A. and W. K. Newey (1995): "Nonparametric estimation of exact consumers surplus and deadweight loss," *Econometrica: Journal of the Econometric Society*, 1445–1476.

Hjort, N. L. and G. Claeskens (2003): "Frequentist model average estimators," *Journal of the American Statistical Association*, 98, 879–899.

Hodges, J. L., J. and E. L. Lehmann (1952): "The Use of Previous Experience in Reaching Statistical Decisions," *The Annals of Mathematical Statistics*, 23, 396–407.

Horowitz, J. L. (2009): *Semiparametric and nonparametric methods in econometrics*, vol. 12, Springer New York.

——— (2011): "Applied nonparametric instrumental variables estimation," *Econometrica*, 79, 347–394.

——— (2014a): "Adaptive nonparametric instrumental variables estimation: Empirical choice of the regularization parameter," *Journal of Econometrics*, 180, 158–173.

——— (2014b): "Ill-Posed Inverse Problems in Economics," *Annual Review of Economics*, 6, 21–51.

Ichimura, H. and P. E. Todd (2007): "Implementing nonparametric and semiparametric estimators," *Handbook of econometrics*, 6, 5369–5468.

James, W. and C. Stein (1961): "Estimation with Quadratic Loss," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press, vol. 1, 361–379.

Judge, G. G. and M. E. Bock (1983): "Biased estimation," *Handbook of econometrics*, 1, 599–649.

Judge, G. G. and R. C. Mittelhammer (2004): "A semiparametric basis for combining estimation problems under quadratic loss," *Journal of the American Statistical Association*, 99, 479–487.

Kempthorne, P. J. (1988): "Controlling risks under different loss functions: The compromise decision problem," *The Annals of Statistics*, 16, 1594–1608.

Kiefer, N. M. and G. R. Skoog (1984): "Local asymptotic specification error analysis," *Econometrica: Journal of the Econometric Society*, 873–885.

Kim, T.-H. and H. White (2001): "James-Stein-type estimators in large samples with application to the

least absolute deviations estimator," *Journal of the American Statistical Association*, 96, 697–705.

Kitamura, Y., T. Otsu, and K. Evdokimov (2013): "Robustness, infinitesimal neighborhoods, and moment restrictions," *Econometrica*, 81, 1185–1201.

Leamer, E. E. (1983): "Model choice and specification analysis," *Handbook of econometrics*, 1, 285–330.

Li, K.-C. (1987): "Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics*, 15, 958–975.

Maasoumi, E. (1978): "A modified Stein-like estimator for the reduced form coefficients of simultaneous equations," *Econometrica: Journal of the Econometric Society*, 695–703.

Manski, C. F. (1984): "Adaptive estimation of non–linear regression models," *Econometric reviews*, 3, 145–194.

Marchand, E. and W. E. Strawderman (2004): "Estimation in restricted parameter spaces: A review," *Lecture notes-monograph series*, 21–44.

Massart, P. (2007): *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*, Springer.

Matzkin, R. L. (2007): "Nonparametric identification," *Handbook of econometrics*, 6, 5307–5368.

——— (2013): "Nonparametric identification in structural economic models," *Annu. Rev. Econ.*, 5, 457–486.

Mittelhammer, R. C. and G. G. Judge (2005): "Combining estimators to improve structural model estimation and inference under quadratic loss," *Journal of econometrics*, 128, 1–29.

Montiel Olea, J. L., M. Plagborg-Møller, E. Qian, and C. K. Wolf (2026): "Double robustness of local projections and some unpleasant varithmetic," *Econometrica*, 94, 1313–1343.

Nagar, A. L. (1959): "The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations," *Econometrica: Journal of the Econometric Society*, 575–595.

Nakamura, A. and M. Nakamura (1998): "Model specification and endogeneity," *Journal of Econometrics*, 83, 213–237.

Newey, W. K. (1985): "Generalized method of moments specification testing," *Journal of econometrics*, 29, 229–256.

——— (1988): "Adaptive estimation of regression models via moment restrictions," *Journal of Econometrics*, 38, 301–339.

——— (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of econometrics*, 79, 147–168.

——— (2013): "Nonparametric instrumental variables estimation," *American Economic Review*, 103, 550–556.

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

Newey, W. K. and T. M. Stoker (1993): "Efficiency of weighted average derivative estimators and index models," *Econometrica: Journal of the Econometric Society*, 1199–1223.

Sawa, T. (1973a): "Almost unbiased estimator in simultaneous equations systems," *International Economic*

Review, 14, 97–106.

——— (1973b): "The mean square error of a combined estimator and numerical comparison with the TSLS estimator," *Journal of Econometrics*, 1, 115–132.

Schorfheide, F. (2005): "VAR forecasting under misspecification," *Journal of Econometrics*, 128, 99–136.

Sclove, S. L., C. Morris, and R. Radhakrishnan (1972): "Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution," *The Annals of Mathematical Statistics*, 1481–1490.

Stone, C. J. (1975): "Adaptive maximum likelihood estimators of a location parameter," *The annals of Statistics*, 267–284.

——— (1982): "Optimal global rates of convergence for nonparametric regression," *The annals of statistics*, 1040–1053.

Theil, H. (1958): *Economic Forecasts and Policy*, vol. 15 of *Contributions to Economic Analysis*, Amsterdam: North-Holland Publishing Company.

Tsybakov, A. B. (1998): "Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes," *The Annals of Statistics*, 26, 2420–2469.

——— (2009): *Introduction to Nonparametric Estimation*, Springer Series in Statistics, New York: Springer.

Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*, MIT Press, 2 ed.

Working, E. J. (1927): "What do statistical" demand curves" show?" *The Quarterly Journal of Economics*, 212–235.

Yatchew, A. (1998): "Nonparametric regression techniques in economics," *Journal of Economic Literature*, 36, 669–721.

Appendix A Proof of Main Results

Appendix A contains the proofs of the main theoretical results. In Section A.1, we establish the minimax lower bound. In Sections A.2 to A.5, we develop the auxiliary results needed for the analysis of the adaptive estimator, including the construction of the pilot sieve, uniform asymptotic expansions, consistency of the variance estimators, and concentration inequalities. Finally, in Section A.6, we use these results to establish the adaptive rate optimality of the proposed estimator.

A.1 Proof of Theorem 4.1

We establish the minimax lower bound using Assouad's hypercube method (Tsybakov, 2009, Section 2.7.2), as in Chen and Reiss (2011). The argument constructs a finite family of alternative indexed by binary vectors, reducing the estimation problem to a collection of binary hypothesis tests. By choosing alternatives that are well separated in $L^2(X)$ but induced distributions that are difficult to distinguish, we obtain a lower bound on the estimation risk.

The key challenge in our setting is to construct alternatives with distinct structural functions but the same magnitude of endogeneity ρ_n, while satisfying the conditional moment restriction (2.2) and the smoothness conditions in Assumption 4.1. We address this issue by constructing two families of alternatives. In the first construction, the structural and conditional mean functions vary together, holding the endogeneity bias fixed and yielding the lower bound of order $r_{n,\mathrm{OLS}}$. In the second construction, the conditional mean function remains fixed, while a common component is chosen to keep the magnitude of endogeneity unchanged as the structural function varies. The smoothing property of T then yields the lower bound $\min\{\rho_n, r_{n,\mathrm{IV}}\}$.

Fix an integer $J \geq 1$ and an orthonormal collection $\{v_0, v_{J,1}, \ldots, v_{J,J}\}$ in $L^2(X)$, to be constructed below. We consider a collection of structural functions

$$h_\theta = bv_0 + a\sum_{j=1}^{J} \theta_j v_{J,j}, \qquad \theta \in \{-1, 1\}^J.$$

where $a > 0$ and b are constants independent of θ. Suppose that each h_θ is the structural function under some sample distribution $F_{n,\theta} \in \mathcal{F}(\rho_n)$. We will construct these sample distributions in the following steps.

The construction contains 2^J alternatives, one for each possible combination of signs. For any θ, let $\theta^{(j)}$ denote the vector obtained by reversing its j-th sign while leaving all other coordinates unchanged. The corresponding structural functions satisfy

$$h_\theta - h_{\theta^{(j)}} = 2a\theta_j v_{J,j}, \qquad \left\|h_\theta - h_{\theta^{(j)}}\right\|_X^2 = 4a^2.$$

Now consider any estimator $\hat{h}$ based on n observations. For each $j = 1, ..., J$, define that

$$\widehat{\theta}_j = \begin{cases} 1, & \langle \hat{h}, v_{J,j} \rangle_X \geq 0, \\ -1, & \langle \hat{h}, v_{J,j} \rangle_X < 0. \end{cases}$$

Since v_0 is orthogonal to $v_{J,j}$, the true coefficient of h_θ in direction $v_{J,j}$ is $\langle h_\theta, v_{J,j} \rangle_X = a\theta_j$. Whenever $\widehat{\theta}_j \neq \theta_j$, the estimated coefficient has the wrong sign and therefore differs from the true coefficient by at least a. Bessel's inequality (Kreyszig, 1978, Theorem 3.4-6) therefore gives

$$\left\| \widehat{h} - h_\theta \right\|_X^2 \geq \sum_{j=1}^{J} \left| \left\langle \widehat{h} - h_\theta, v_{J,j} \right\rangle_X \right|^2 = \sum_{j=1}^{J} \left(\left\langle \widehat{h}, v_{J,j} \right\rangle_X - a\theta_j \right)^2 \geq a^2 \sum_{j=1}^{J} \mathbf{1}\{\widehat{\theta}_j \neq \theta_j\}.$$

Let KL denote the Kullback-Leibler divergence between probability measures. By Assouad's lemma (Tsybakov, 2009, Lemma 2.12 and Theorem 2.12(ii)) and Pinsker's inequality (Tsybakov, 2009, Lemma 2.5(i)),

$$\max_{\theta \in \{-1,1\}^J} E_{F_{n,\theta}^{\otimes n}} \left\| \widehat{h} - h_\theta \right\|_X^2 \geq \frac{Ja^2}{2} \left[1 - \sqrt{\frac{1}{2} \max_{\theta \in \{-1,1\}^J} \max_{1 \leq j \leq J} \mathrm{KL}\left(F_{n,\theta}^{\otimes n}, F_{n,\theta^{(j)}}^{\otimes n} \right)} \right]$$

where $F_{n,\theta}^{\otimes n}$ denote the joint distribution of n i.i.d. observations from $F_{n,\theta}$.

Consequently, if the sample distributions corresponding to neighboring alternatives satisfy the uniform Kullback-Leibler divergence bound

$$\max_{\theta \in \{-1,1\}^J} \max_{1 \leq j \leq J} \mathrm{KL}\left(F_{n,\theta}^{\otimes n}, F_{n,\theta^{(j)}}^{\otimes n} \right) \leq \frac{1}{2} \tag{A.1}$$

then

$$\inf_{\widehat{h}} \max_{\theta \in \{-1,1\}^J} E_{F_{n,\theta}^{\otimes n}} \left\| \widehat{h} - h_\theta \right\|_X^2 \geq \frac{Ja^2}{4}. \tag{A.2}$$

It remains to construct admissible alternatives for which the separation Ja^2 is large while the KL divergence between neighboring distributions satisfies (A.1).

Step 1. Construction of perturbation directions

To apply the bound (A.2), we construct orthonormal perturbation directions $\{v_{J,1}, \ldots, v_{J,J}\}$ such that their smoothness can be controlled while their conditional expectations given Z_i are small.

First, the upper link condition (4.2) implies that T is compact (Kreyszig, 1978, Theorems 8.1-4 and 8.1-5). Let $T^* : L^2(Z) \to L^2(X)$ denote the adjoint of T, with $(T^*q)(x) = E[q(Z_i) \mid X_i = x]$. The operator T^*T is compact, self-adjoint, and positive (Kreyszig, 1978, Sections 3.9 and 9.3); moreover, conditional expectation is a contraction, so $0 \leq T^*T \leq I$. By the spectral properties of compact operators (Kreyszig, 1978, Theorem 8.3-1 and 8.3-3), let $\mathcal{V}$ be the span of its eigenfunctions

with eigenvalues greater than $1/2$. Choose a fixed integer J_0 such that $C\nu_{J_0+1}^2 \leq 1/4$. Then it follows that

$$\dim(\mathcal{V}) \leq J_0, \qquad 0 \leq T^*T \leq \tfrac{1}{2}I \quad \text{on } \mathcal{V}^\perp.$$

To see why, suppose instead that $\dim(\mathcal{V}) > J_0$. Since the sieve space $\mathcal{H}_{J_0}$ has dimension J_0, by Kreyszig (1978, Theorem 2.6-10 and Lemma 3.3-5), there must exist a nonzero function $f \in \mathcal{V} \cap \mathcal{H}_{J_0}^\perp$, implying that

$$\frac{1}{2}\|f\|_X^2 < \|Tf\|_Z^2 \leq C \sum_{j=J_0+1}^{\infty} \nu_j^2 \langle f, \psi_j \rangle_X^2 \leq C\nu_{J_0+1}^2 \|f\|_X^2 \leq \frac{1}{4}\|f\|_X^2$$

leading to a contradiction.

We now construct the orthonormal functions. Since $\dim(\mathcal{V}) \leq J_0$ and $\dim(\mathcal{H}_{J_0+1}) = J_0 + 1$, there exists a unit vector $v_0 \in \mathcal{H}_{J_0+1} \cap \mathcal{V}^\perp$. By the orthonormality of $\{\psi_j\}_{j\geq 1}$,

$$\sum_{j=1}^{\infty} j^{2s/d_x} \langle v_0, \psi_j \rangle_X^2 = \sum_{j=1}^{J_0+1} j^{2s/d_x} \langle v_0, \psi_j \rangle_X^2 \leq (J_0+1)^{2s/d_x} \|v_0\|_X^2 = (J_0+1)^{2s/d_x}. \tag{A.3}$$

Next, for every $J \geq J_0 + 1$, the subspace $\text{clsp}\{\psi_{J+1}, \ldots, \psi_{3J}\} \cap \mathcal{V}^\perp$ has dimension at least $2J - J_0 \geq J$. We can therefore choose J orthonormal functions $v_{J,1}, \ldots, v_{J,J}$ from this subspace. They are also orthogonal to v_0, since $v_0 \in \mathcal{H}_{J_0+1}$ and $J \geq J_0 + 1$.

By construction, every $f \in \text{clsp}\{v_{J,1}, \ldots, v_{J,J}\}$ is a linear combination of $\psi_{J+1}, \ldots, \psi_{3J}$. The orthonormality of $\{\psi_j\}_{j\geq 1}$ and the monotonicity of j^{2s/d_x} therefore imply

$$\sum_{j=1}^{\infty} j^{2s/d_x} \left\langle f, \psi_j \right\rangle_X^2 \leq (3J)^{2s/d_x} \left\| f \right\|_X^2, \tag{A.4}$$

Similarly, since $\{\nu_j\}_{j\geq 1}$ is decreasing, the upper link condition (4.2) gives

$$\left\| Tf \right\|_Z^2 \leq C\nu_{J+1}^2 \left\| f \right\|_X^2. \tag{A.5}$$

Step 2. Construction of Gaussian submodels

We now construct a collection of data generating processes under which the structural functions h_θ satisfy the conditional model restrictions (2.2). We hold the joint distribution of (X_i, Z_i) fixed across alternatives and specify the conditional distribution of Y_i given (X_i, Z_i) so that the structural function h and the conditional mean function g can vary separately.

Recall from Step 1 that $0 \le T^*T \le \frac{1}{2}I$ on $\mathcal{V}^\perp$. Thus, the inverse operator

$$\mathcal{A} := (I - T^*T)^{-1} = \sum_{\ell=0}^{\infty} (T^*T)^\ell \tag{A.6}$$

is well defined on $\mathcal{V}^\perp$. It can be verified that $\|\mathcal{A}f\|_X \le 2\|f\|_X$.

For every $f \in \mathcal{V}^\perp$, define that $\mu_f(x,z) := (T\mathcal{A}f)(z) - (T^*T\mathcal{A}f)(x)$. The map $f \mapsto \mu_f$ is linear. By iterated expectations and the identity $(I - T^*T)\mathcal{A}f = f$,

$$\begin{aligned} E[\mu_f(X_i, Z_i) \mid X_i] &= (T^*T\mathcal{A}f)(X_i) - (T^*T\mathcal{A}f)(X_i) = 0, \\ E[\mu_f(X_i, Z_i) \mid Z_i] &= \big[T(I - T^*T)\mathcal{A}f\big](Z_i) = (Tf)(Z_i). \end{aligned} \tag{A.7}$$

Conditioning on X_i, using the definition of the adjoint operator and the Neumann series representation of $\mathcal{A}$ in (A.6), we obtain

$$\begin{aligned} E\Big[\mu_f(X_i, Z_i)^2\Big] &= \Big\|T\mathcal{A}f\Big\|_Z^2 - \Big\|T^*T\mathcal{A}f\Big\|_X^2 = \Big\langle f, T^*T\mathcal{A}f\Big\rangle_X \\ &= \sum_{\ell=0}^{\infty} \Big\langle f, (T^*T)^{\ell+1} f\Big\rangle_X \le 2\Big\langle f, T^*Tf\Big\rangle_X = 2\Big\|Tf\Big\|_Z^2. \end{aligned}$$

where the inequality follows because T^*T is positive and has operator norm at most $1/2$ on $\mathcal{V}^\perp$.

We now use this construction to obtain admissible DGPs in $\mathcal{F}(\rho_n)$. Since the minimax risk over $\mathcal{F}(\rho_n)$ is bounded below by the minimax risk over any subclass, it suffices to consider a Gaussian submodel (Chen and Reiss, 2011). For any $h \in \mathcal{H}(s, L_h)$ and $g \in \mathcal{H}(s, L_g)$ satisfying $h - g \in \mathcal{V}^\perp$ and $\|g - h\|_X^2 = \rho_n$, consider the Gaussian model

$$Y_i = g(X_i) + \mu_{h-g}(X_i, Z_i) + e_i, \qquad e_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2), \tag{A.8}$$

where e_i is independent of (X_i, Z_i). By (A.7) with $f = h - g$, the Gaussian model (A.8) satisfies that $E[Y_i \mid X_i] = g(X_i), E[U_i \mid Z_i] = 0$ and $E[U_i \mid X_i] = g(X_i) - h(X_i)$, where $U_i = Y_i - h(X_i)$. Thus, the Gaussian model (A.8) defines a DGP in $\mathcal{F}(\rho_n)$, provided the fixed joint distribution of (X_i, Z_i) satisfies the remaining assumptions.

Finally, the Gaussian construction provides an explicit expression for the KL divergence between alternatives. Let $F_{n,\theta}$ denote the distribution corresponding to (h_θ, g_θ), and define

$$M_\theta(x,z) := g_\theta(x) + \mu_{h_\theta - g_\theta}(x,z),$$

the conditional mean of Y_i given $(X_i, Z_i) = (x,z)$. Since the joint distribution of (X_i, Z_i) and the conditional variance σ^2 are common to all alternatives, the KL divergence between their sample

distributions is

$$\mathrm{KL}\left(F_{n,\theta}^{\otimes n},\, F_{n,\theta'}^{\otimes n}\right) \;=\; \frac{n}{2\sigma^2}\, E\left[\,\{M_\theta(X_i,Z_i) \;-\; M_{\theta'}(X_i,Z_i)\}^2\,\right]. \tag{A.9}$$

We can now verify the testing condition (A.1).

Step 3. The lower bound of order $r_{n,\mathrm{OLS}}$

We construct alternatives in which the structural function h_θ and the conditional mean function g_θ vary by the same perturbations, while the endogeneity bias remains fixed across alternatives.

We choose a fixed constant $\delta_* > 0$, independent of n and ρ_n, sufficiently small that

$$3^{2s/d_x}\delta_*^2 \le \frac{1}{4}\min\{L_h^2, L_g^2\}, \qquad \delta_*^2 \le \min\left\{\frac{1}{2}, \frac{\sigma^2}{4}\right\}.$$

We set $J = \lceil n^{d_x/(2s+d_x)} \rceil, a^2 = \delta_*^2 J^{-1-2s/d_x}$. For all sufficiently large n, $J \ge J_0 + 1$. For each $\theta \in \{-1,1\}^J$, define

$$h_\theta = \sqrt{\rho_n}\, v_0 + a\sum_{j=1}^{J}\theta_j v_{J,j}, \qquad g_\theta = a\sum_{j=1}^{J}\theta_j v_{J,j}.$$

By construction, $h_\theta - g_\theta = \sqrt{\rho_n}\, v_0 \in \mathcal{V}^\perp$ and $\|g_\theta - h_\theta\|_X^2 = \rho_n$.

We next verify that these alternatives satisfy the smoothness restrictions. Since v_0 and $v_{J,1}, \ldots, v_{J,J}$ are supported on disjoint sets of basis functions, (A.3) and (A.4) imply

$$\sum_{j=1}^{\infty} j^{2s/d_x}\langle h_\theta, \psi_j\rangle_X^2 \le \rho_n(J_0+1)^{2s/d_x} + (3J)^{2s/d_x} J a^2 \le \frac{L_h^2}{4} + 3^{2s/d_x}\delta_*^2 \le \frac{L_h^2}{2},$$

$$\sum_{j=1}^{\infty} j^{2s/d_x}\langle g_\theta, \psi_j\rangle_X^2 \le 3^{2s/d_x}\delta_*^2 \le \frac{L_g^2}{4}.$$

Hence $h_\theta \in \mathcal{H}(s, L_h)$ and $g_\theta \in \mathcal{H}(s, L_g)$. By the Gaussian construction (A.8) in Step 2, each pair (h_θ, g_θ) defines a DGP $F_{n,\theta} \in \mathcal{F}(\rho_n)$.

It remains to verify the KL divergence condition (A.1). Since $h_\theta - g_\theta = \sqrt{\rho_n} v_0$ does not depend on θ, the conditional means of Y_i given (X_i, Z_i) under neighboring alternatives differ only by $2a\theta_j v_{J,j}(X_i)$. By (A.9),

$$\mathrm{KL}\left(F_{n,\theta}^{\otimes n},\; F_{n,\theta^{(j)}}^{\otimes n}\right) = \frac{2na^2}{\sigma^2} = \frac{2\delta_*^2}{\sigma^2} n J^{-1-2s/d_x} \le \frac{1}{2}$$

where the last inequality follows from the choices of J and δ_*.

Applying (A.2), we obtain

$$\inf_{\widehat{h}} \sup_{F_n \in \mathcal{F}(\rho_n)} E_{F_n^{\otimes n}} \left\| \widehat{h} - h_0 \right\|_X^2 \geq \frac{Ja^2}{4} = \frac{\delta_*^2}{4} J^{-2s/d_x} \asymp r_{n,\mathrm{OLS}}. \tag{A.10}$$

where δ_* is independent of n and ρ_n.

Step 4. The lower bound of order $\min\{\rho_n, r_{n,\mathrm{IV}}\}$

We next establish the lower bound of order $\min\{\rho_n, r_{n,\mathrm{IV}}\}$. Unlike Step 3, we hold the conditional mean function fixed and vary only the structural function. The smoothing property of T allows these structural functions to be separated in $L^2(X)$ while keeping the corresponding sample distributions close in KL divergence.

For $J \geq J_0 + 1$, set $a^2 = \delta_*^2 / J \min\{\rho_n, J^{-2s/d_x}\}$. For each $\theta \in \{-1, 1\}^J$, define

$$h_\theta = \sqrt{\rho_n - Ja^2}\, v_0 + a \sum_{j=1}^{J} \theta_j v_{J,j}, \qquad g_\theta = 0. \tag{A.11}$$

Since $Ja^2 \leq \delta_*^2 \rho_n \leq \rho_n / 2$, the square root is well defined. Moreover, by orthonormality,

$$h_\theta - g_\theta = h_\theta \in \mathcal{V}^\perp, \qquad \|g_\theta - h_\theta\|_X^2 = \rho_n - Ja^2 + Ja^2 = \rho_n.$$

Thus, all alternatives have the same magnitude of endogeneity ρ_n.

We next verify the smoothness restrictions. Since $g_\theta = 0$, it suffices to consider h_θ. By (A.3) and (A.4),

$$\begin{aligned} \sum_{j=1}^{\infty} j^{2s/d_x} \langle h_\theta, \psi_j \rangle_X^2 &\leq (\rho_n - Ja^2)(J_0 + 1)^{2s/d_x} + (3J)^{2s/d_x} Ja^2 \\ &\leq \frac{L_h^2}{4} + 3^{2s/d_x} \delta_*^2 \leq \frac{L_h^2}{2}. \end{aligned}$$

We now verify the KL divergence condition (A.1). For neighboring alternatives, $h_\theta - h_{\theta^{(j)}} = 2a\theta_j v_{J,j}$. Since $g_\theta = 0$, the linearity of $f \mapsto \mu_f$ implies that

$$M_\theta - M_{\theta^{(j)}} = \mu_{2a\theta_j v_{J,j}}.$$

Combining the Gaussian KL formula (A.9), the second-moment bound in Step 2, and the upper link bound (A.5), we obtain

$$\mathrm{KL}\left(F_{n,\theta}^{\otimes n},\ F_{n,\theta^{(j)}}^{\otimes n} \right) = \frac{n}{2\sigma^2} E\left[\mu_{2a\theta_j v_{J,j}}(X_i, Z_i)^2 \right] \leq \frac{4C}{\sigma^2} n a^2 \nu_{J+1}^2 \leq \frac{4C\delta_*^2}{\sigma^2} n \nu_{J+1}^2 J^{-1-2s/d_x}. \tag{A.12}$$

Unlike the KL bound in Step 3, this bound contains the additional factor ν_{J+1}^2, reflecting the smoothing effect of T on the perturbations.

It remains to choose J so that the KL divergence is bounded while the separation Ja^2 has the required order. In the mildly ill-posed case, take $J = \lceil n^{d_x/(2s+2\varsigma+d_x)} \rceil$. Since $\nu_{J+1}^2 \lesssim J^{-2\varsigma/d_x}$, we have that $n\nu_{J+1}^2 J^{-1-2s/d_x} \lesssim nJ^{-1-2(s+\varsigma)/d_x} \leq 1$ and $J^{-2s/d_x} \asymp n^{-2s/(2s+2\varsigma+d_x)} = r_{n,\mathrm{IV}}$. In the severely ill-posed case, take $J = \lceil (\log n)^{d_x/\varsigma} \rceil$. Since $\nu_{J+1}^2 \lesssim \exp\{-(J+1)^{\varsigma/d_x}\}$, we have that $n\nu_{J+1}^2 J^{-1-2s/d_x} \lesssim J^{-1-2s/d_x} \leq 1$, while $J^{-2s/d_x} \asymp (\log n)^{-2s/\varsigma} = r_{n,\mathrm{IV}}$.

In either case, the KL divergence between neighboring alternatives is bounded by a constant multiple of δ_*^2. By choosing δ_* sufficiently small, we obtain the KL divergence condition (A.1). Applying (A.2) therefore yields

$$\inf_{\widehat{h}} \sup_{F_n \in \mathcal{F}(\rho_n)} E_{F_n^{\otimes n}} \left\| \widehat{h} - h_0 \right\|_X^2 \geq \frac{Ja^2}{4} = \frac{\delta_*^2}{4} \min\{\rho_n, J^{-2s/d_x}\} \gtrsim \min\{\rho_n, r_{n,\mathrm{IV}}\}. \tag{A.13}$$

Combining (A.10) and (A.13), we obtain that

$$\inf_{\widehat{h}} \sup_{F_n \in \mathcal{F}(\rho_n)} E_{F_n^{\otimes n}} \left\| \widehat{h} - h_0 \right\|_X^2 \gtrsim \begin{cases} r_{n,\mathrm{OLS}}, & \rho_n \lesssim r_{n,\mathrm{OLS}}, \\ \rho_n, & r_{n,\mathrm{OLS}} \lesssim \rho_n \lesssim r_{n,\mathrm{IV}}, \\ r_{n,\mathrm{IV}}, & \rho_n \gtrsim r_{n,\mathrm{IV}}. \end{cases}$$

A.2 Sieve Construction and Choice of Pilot Dimension

This subsection specifies a deterministic pilot sieve dimension, verifies that it satisfies the conditions for the uniform asymptotic expansion in Appendix A.3, and provides a feasible counterpart for implementation.

Recall from Section 4.2, the population matrices G_X, G_Z, and S, the minimum singular value σ_{JK}, and the normalized sieve basis envelope ζ_J. Following Breunig and Chen (2024), we define

$$A = G_X^{1/2} \left(S' G_Z^{-1} S \right)^{-1} S' G_Z^{-1}.$$

Let $\underline{J} = \lfloor \sqrt{\log \log n} \rfloor$ and recall that $K = K(J)$ denotes the instrument sieve dimension associated with J. For a fixed constant $C_\Delta > 0$, define the deterministic pilot sieve dimension by

$$J_{\Delta,\max} = \max \left\{ J > \underline{J} : C_\Delta \, \zeta_J \sqrt{\frac{JK \log J}{n}} < \sigma_{JK}^2 \sqrt{\mathrm{tr}(AG_Z A')} \right\}, \tag{A.14}$$

Set $K_{\Delta,\max} = K(J_{\Delta,\max})$.

Let $\hat{A}$, $\hat{G}_Z$, and $\hat{\sigma}_{JK}$ denote the sample counterparts of A, G_Z, and σ_{JK}, respectively. In the

Monte Carlo simulation in Section 6, we use the feasible pilot sieve dimension

$$\widehat{J}_{\Delta,\max} = \max\left\{ J > \underline{J} : 0.45\, J\sqrt{\frac{J\log J}{n}} < \hat{\sigma}_{JK}^2 \sqrt{\operatorname{tr}\left(\hat{A}\hat{G}_Z\hat{A}'\right)} \right\}. \tag{A.15}$$

We first state high-level conditions on the pilot sieve dimension required for the uniform asymptotic expansions in Appendix A.3. We then verify that the deterministic choice in (A.14) satisfies these conditions under primitive restrictions. To state these conditions, define the population matrices from Section 4.2, evaluated at $(J_{\Delta,\max}, K_{\Delta,\max})$, by

$$G_{X,\Delta} := E\left[p^{J_{\Delta,\max}}(X_i)p^{J_{\Delta,\max}}(X_i)'\right], \qquad G_{Z,\Delta} := E\left[q^{K_{\Delta,\max}}(Z_i)q^{K_{\Delta,\max}}(Z_i)'\right],$$

$$S_\Delta := E\left[q^{K_{\Delta,\max}}(Z_i)p^{J_{\Delta,\max}}(X_i)'\right], \qquad A_\Delta := G_{X,\Delta}^{1/2}\left(S_\Delta' G_{Z,\Delta}^{-1} S_\Delta\right)^{-1} S_\Delta' G_{Z,\Delta}^{-1}.$$

Their sample analogues are denoted by $\hat{G}_{X,\Delta}$, $\hat{G}_{Z,\Delta}$, $\hat{S}_\Delta$, and $\hat{A}_\Delta$, respectively. We also write

$$\sigma_\Delta := \sigma_{J_{\Delta,\max},K_{\Delta,\max}}, \qquad \zeta_\Delta := \max\left\{\zeta_{X,J_{\Delta,\max}},\ \zeta_{Z,K_{\Delta,\max}}\right\}.$$

Assumption A.1 (Pilot sieve dimension)**.** *Let $J_{\Delta,\max}\to\infty$. Suppose that Assumption 4.4 (i) also holds at the pilot pair $(J_{\Delta,\max}, K_{\Delta,\max})$, so that*

$$\frac{\sigma_\Delta^{-2}\zeta_\Delta\sqrt{(K_{\Delta,\max}\log J_{\Delta,\max})/n}}{\left(J_{\Delta,\max}^{-1}\sum_{j=1}^{J_{\Delta,\max}}\nu_j^{-2}\right)^{1/2}} = O(1), \qquad \frac{\sigma_\Delta^{-1}\zeta_\Delta J_{\Delta,\max}^{-s/d_x}\sqrt{\log J_{\Delta,\max}}}{\left(J_{\Delta,\max}^{-1}\sum_{j=1}^{J_{\Delta,\max}}\nu_j^{-2}\right)^{1/2}} = O(1). \tag{A.16}$$

In addition, assume that

$$\max_{\substack{J\in\mathcal{J}_n\\ J\le J_{\Delta,\max}}} \frac{\sqrt{n}\,\nu_{J_{\Delta,\max}+1}J_{\Delta,\max}^{-s/d_x}}{\nu_J\left(\sum_{j=1}^{J}\nu_j^{-2}\right)^{1/2}} = o(1). \tag{A.17}$$

Remark A.1. Assumption A.1 is similar to those in Horowitz (2014a, Assumption 6), which restrict the growth of the pilot sieve dimension to control the instability of NPIV estimation. The conditions in (A.16) control estimation and approximation errors at the pilot dimension. The condition (A.17) ensures that the projected pilot approximation bias is asymptotically negligible relative to the leading stochastic error. ■

In Proposition B.5, we verify that the deterministic choice in (A.14) satisfies Assumption A.1 under the smoothness and sieve regularity conditions.

A.3 Uniform Asymptotic Expansions

This subsection proves Proposition 5.1 by establishing uniform asymptotic expansions for the sieve NP-OLS estimator $\hat{g}_J$, and NPIV-based estimate of endogeneity bias $\hat{\Delta}_J$.

We first define the leading stochastic terms mentioned in Proposition 5.1:

$$\Psi_{g,J}(x) = \frac{1}{n}\sum_{i=1}^{n} p^J(x)' G_X^{-1} p^J(X_i)\left(Y_i - \Pi_J g_0(X_i)\right), \tag{A.18}$$

and

$$\Psi_{\Delta,J}(x) = \Psi_{g,J}(x) - \Pi_J\left[\frac{1}{n}\sum_{i=1}^{n} p^{J_{\Delta,\max}}(\cdot)' G_{X,\Delta}^{-1/2} A_\Delta q^{K_{\Delta,\max}}(Z_i)\left(Y_i - \Xi_{J_{\Delta,\max}} h_0(X_i)\right)\right](x). \tag{A.19}$$

The proof proceeds in two steps. We first derive asymptotic expansions for $\hat{g}_J$ and the common NPIV pilot, with uniform control of their remainder terms. We then combine these expansions to obtain the uniform asymptotic expansion for $\hat{\Delta}_J$.

Step 1. Uniform expansion of the sieve NP-OLS estimator

Decompose that $\widehat{g}_J - \Pi_J g_0 = \Psi_{g,J} + r_{g,J}$ where $\Psi_{g,J}$ is defined in (A.18) and

$$r_{g,J}(x) = p^J(x)'\left(\widehat{G}_X^{-1} - G_X^{-1}\right)\left[\frac{1}{n}\sum_{i=1}^{n} p^J(X_i)\left(Y_i - \Pi_J g_0(X_i)\right)\right]. \tag{A.20}$$

This decomposition is similar to the linearization of nonparametric least squares estimators in Belloni, Chernozhukov, Chetverikov, and Kato (2015, Lemma 4.1).

We first verify that $\Psi_{g,J}$ is centered. Since $Y_i - \Pi_J g_0(X_i) = \varepsilon_i + g_0(X_i) - \Pi_J g_0(X_i)$, the first-order condition defining $\Pi_J g_0$ and $E[\varepsilon_i|X_i] = 0$ imply that

$$E\left[p^J(X_i)(Y_i - \Pi_J g_0(X_i))\right] = E\left[p^J(X_i)\varepsilon_i\right] + E\left[p^J(X_i)\left(g_0(X_i) - \Pi_J g_0(X_i)\right)\right] = 0,$$

which proves $E[\Psi_{g,J}] = 0$.

We next establish the rate of leading variance. Because $\|p^J(\cdot)' G_X^{-1} a\|_X^2 = a' G_X^{-1} a$ for any $a \in \mathbb{R}^J$, independence across observations and the preceding centering condition imply that

$$E\left\|\Psi_{g,J}\right\|_X^2 = \frac{1}{n} E\left[\left(Y_i - \Pi_J g_0(X_i)\right)^2 p^J(X_i)' G_X^{-1} p^J(X_i)\right]. \tag{A.21}$$

For the upper bound, again $Y_i - \Pi_J g_0(X_i) = \varepsilon_i + g_0(X_i) - \Pi_J g_0(X_i)$. Then

$$E\left[\varepsilon_i^2 p^J(X_i)' G_X^{-1} p^J(X_i)\right] \le \bar{\sigma}_\varepsilon^2 J$$

$$E\left[\left(g_0(X_i) - \Pi_J g_0(X_i)\right)^2 p^J(X_i)' G_X^{-1} p^J(X_i)\right] \le \zeta_{X,J}^2 \|g_0 - \Pi_J g_0\|_X^2 \lesssim \zeta_{X,J}^2 J^{-2s/d_x}.$$

Under Assumption 4.4 (i), the second term is $O(1/\log J) = o(J)$. Hence $E\|\Psi_{g,J}\|_X^2 \lesssim J/n$.

For the lower bound, the law of total variance and Assumption 4.2 (i) give

$$\mathrm{Var}(Y_i \mid X_i) = \mathrm{Var}(U_i \mid X_i) \geq E[\mathrm{Var}(U_i \mid X_i, Z_i) \mid X_i] \geq \underline{\sigma}^2.$$

Consequently, $E[(Y_i - \Pi_J g_0(X_i))^2 \mid X_i] \geq \underline{\sigma}^2$, and therefore

$$E\left\|\Psi_{g,J}\right\|_X^2 \geq \frac{\underline{\sigma}^2}{n} E[p^J(X_i)' G_X^{-1} p^J(X_i)] = \underline{\sigma}^2 \frac{J}{n}.$$

Combining the upper and lower bounds gives

$$E\left\|\Psi_{g,J}\right\|_X^2 \asymp \frac{J}{n}. \tag{A.22}$$

It remains to control (A.20). By Chen and Christensen (2015b, Lemma 2.1), together with a union bound over $\mathcal{J}_n$, we obtain that

$$\max_{J \in \mathcal{J}_n} \left\| G_X^{-1/2} \widehat{G}_X G_X^{-1/2} - I_J \right\| = o_p(1), \tag{A.23}$$

because the construction of $J_{\max}$ implies $\zeta_{X,J_{\max}} \sqrt{\log J_{\max}/n} = o(1)$. Hence, uniformly over $J \in \mathcal{J}_n$,

$$\left\| G_X^{1/2} \left(\widehat{G}_X^{-1} - G_X^{-1} \right) G_X^{1/2} \right\| = O_p\left(\zeta_{X,J} \sqrt{\frac{\log J}{n}} \right). \tag{A.24}$$

Moreover, using $E[U_i^4] < \infty$ and Jensen's inequality, $E[\varepsilon_i^4] < \infty$. A fourth-moment calculation for sums of independent random vectors gives

$$E\left[\left| \frac{n \left\| G_X^{-1/2} n^{-1} \sum_{i=1}^n p^J(X_i) \left(Y_i - \Pi_J g_0(X_i)\right) \right\|^2}{J} - \frac{1}{J} E\left[\left(Y_i - \Pi_J g_0(X_i)\right)^2 p^J(X_i)' G_X^{-1} p^J(X_i) \right] \right|^2 \right] \lesssim \frac{1}{J} + \frac{\zeta_{X,J}^2}{nJ}.$$

Since the grid is geometric, $\sum_{J \in \mathcal{J}_n} \frac{1}{J} \lesssim \frac{1}{\underline{J}} = o(1)$. It follows that

$$\max_{J \in \mathcal{J}_n} \frac{\left\| G_X^{-1/2} n^{-1} \sum_{i=1}^n p^J(X_i) \left(Y_i - \Pi_J g_0(X_i)\right) \right\|}{\sqrt{J/n}} = O_p(1). \tag{A.25}$$

We can now bound the remainder. By (A.20),

$$\begin{aligned} \left\| r_{g,J} \right\|_X &= \left\| G_X^{1/2} \left(\widehat{G}_X^{-1} - G_X^{-1} \right) \left[\frac{1}{n} \sum_{i=1}^n p^J(X_i) \left(Y_i - \Pi_J g_0(X_i)\right) \right] \right\| \\ &\leq \left\| G_X^{1/2} \left(\widehat{G}_X^{-1} - G_X^{-1} \right) G_X^{1/2} \right\| \times \left\| G_X^{-1/2} \frac{1}{n} \sum_{i=1}^n p^J(X_i) \left(Y_i - \Pi_J g_0(X_i)\right) \right\|. \end{aligned}$$

Using (A.24) and (A.25), we obtain

$$\max_{J\in\mathcal{J}_n}\frac{\|r_{g,J}\|_X}{\sqrt{J/n}}=O_p\left(\max_{J\in\mathcal{J}_n}\zeta_{X,J}\sqrt{\frac{\log J}{n}}\right)=o_p(1).$$

Finally, combining with (A.22), we have that

$$\max_{J\in\mathcal{J}_n}\frac{\|r_{g,J}\|_X}{\left(E\|\Psi_{g,J}\|_X^2\right)^{1/2}}=o_p(1),$$

which proves the claimed uniform expansion of the sieve NP-OLS estimator over $J\in\mathcal{J}_n$.

Step 2. Expansion of the common NPIV pilot

We next derive an asymptotic expansion for the common NPIV pilot at dimension $J_{\Delta,\max}$.

At the pilot dimension $J_{\Delta,\max}$, the sieve 2SLS estimator has the expansion (Chen and Christensen, 2018, Appendix G)

$$\widetilde{h}_{\mathrm{IV},J_{\Delta,\max}}-\Xi_{J_{\Delta,\max}}h_0=\frac{1}{n}\sum_{i=1}^n\xi_{\mathrm{IV},J_{\Delta,\max}}(\mathcal{D}_i)+r_{\mathrm{IV},J_{\Delta,\max}},\tag{A.26}$$

where

$$\xi_{\mathrm{IV},J_{\Delta,\max}}(\mathcal{D}_i)=p^{J_{\Delta,\max}}(\cdot)'G_{X,\Delta}^{-1/2}A_\Delta q^{K_{\Delta,\max}}(Z_i)\left(Y_i-\Xi_{J_{\Delta,\max}}h_0(X_i)\right)\tag{A.27}$$

and

$$r_{\mathrm{IV},J_{\Delta,\max}}=p^{J_{\Delta,\max}}(\cdot)'\left(\widehat{G}_{X,\Delta}^{-1/2}\widehat{A}_\Delta-G_{X,\Delta}^{-1/2}A_\Delta\right)\left[\frac{1}{n}\sum_{i=1}^n q^{K_{\Delta,\max}}(Z_i)\left(Y_i-\Xi_{J_{\Delta,\max}}h_0(X_i)\right)\right].\tag{A.28}$$

We first verify that the leading term (A.27) has expectation zero. By definition,

$$\Xi_{J_{\Delta,\max}}h_0\in\underset{h\in\mathcal{H}_{J_{\Delta,\max}}}{\arg\min}\left\|\Pi_{K_{\Delta,\max}}T(h-h_0)\right\|_Z^2.$$

For every $v(x)=p^{J_{\Delta,\max}}(x)'\delta\in\mathcal{H}_{J_{\Delta,\max}}$, the first-order condition (Chen and Christensen, 2018, Appendix G, Equation 28) gives

$$\left\langle\Pi_{K_{\Delta,\max}}T\left(\Xi_{J_{\Delta,\max}}h_0-h_0\right),\ \Pi_{K_{\Delta,\max}}Tv\right\rangle_Z=0.\tag{A.29}$$

Since $\Pi_{K_{\Delta,\max}}Tv(z)=q^{K_{\Delta,\max}}(z)'G_{Z,\Delta}^{-1}S_\Delta\delta$, and (A.29) holds for every $\delta\in\mathbb{R}^J$, we have that

$$S_\Delta'G_{Z,\Delta}^{-1}E\left[q^{K_{\Delta,\max}}(Z_i)\left(\Xi_{J_{\Delta,\max}}h_0(X_i)-h_0(X_i)\right)\right]=0.$$

Moreover, $Y_i = h_0(X_i) + U_i$ and $E[U_i \mid Z_i] = 0$ implies that $E[q^{K_{\Delta,\max}}(Z_i)U_i] = 0$ and

$$S'_\Delta G^{-1}_{Z,\Delta} E\left[q^{K_{\Delta,\max}}(Z_i)\left(Y_i - \Xi_{J_{\Delta,\max}} h_0(X_i)\right)\right] = 0.$$

By the definition of A_Δ, we obtain $E[\xi_{\mathrm{IV},J_{\Delta,\max}}(\mathcal{D}_i)] = 0$.

We next establish a bound for $\|\Pi_{K_{\Delta,\max}} Tr_{\mathrm{IV},J_{\Delta,\max}}\|_Z$, which will be used to control the projection of the pilot remainder onto lower-dimensional sieve spaces in Step 5.

Following the notation of Chen and Christensen (2018, Appendix G), for any matrix A, we define $A_l^- := (A'A)^- A'$, where the superscript $-$ denotes a generalized inverse. When A has full column rank, A_l^- is the usual left pseudoinverse and satisfies $A_l^- A = I$. By the definitions of A_Δ and $\hat{A}_\Delta$, we have that

$$G^{-1/2}_{X,\Delta} A_\Delta = \left(G^{-1/2}_{Z,\Delta} S_\Delta\right)_l^- G^{-1/2}_{Z,\Delta}, \qquad \hat{G}^{-1/2}_{X,\Delta} \hat{A}_\Delta = \left(\hat{G}^{-1/2}_{Z,\Delta} \hat{S}_\Delta\right)_l^- \hat{G}^{-1/2}_{Z,\Delta}.$$

Also notice that for any $v \in \mathbb{R}^{J_{\Delta,\max}}$, we have that $\|\Pi_{K_{\Delta,\max}} T\left[p^{J_{\Delta,\max}}(\cdot)'v\right]\|_Z = \|G^{-1/2}_{Z,\Delta} S_\Delta v\|$. Since $r_{\mathrm{IV},J_{\Delta,\max}} \in \mathcal{H}_{J_{\Delta,\max}}$, substituting these identities into (A.28) gives that

$$\begin{aligned}\left\|\Pi_{K_{\Delta,\max}} Tr_{\mathrm{IV},J_{\Delta,\max}}\right\|_Z \leq{}& \left\|G^{-1/2}_{Z,\Delta} S_\Delta \left[\left(\widehat{G}^{-1/2}_{Z,\Delta}\widehat{S}_\Delta\right)_l^- \widehat{G}^{-1/2}_{Z,\Delta} G^{1/2}_{Z,\Delta} - \left(G^{-1/2}_{Z,\Delta} S_\Delta\right)_l^-\right]\right\|_{\mathrm{op}} \\ &\times \left\|G^{-1/2}_{Z,\Delta} \frac{1}{n}\sum_{i=1}^n q^{K_{\Delta,\max}}(Z_i)\left(Y_i - \Xi_{J_{\Delta,\max}} h_0(X_i)\right)\right\|. \qquad \text{(A.30)}\end{aligned}$$

Proposition B.2 gives

$$\left\|G^{-1/2}_{Z,\Delta} S_\Delta \left[\left(\widehat{G}^{-1/2}_{Z,\Delta}\widehat{S}_\Delta\right)_l^- \widehat{G}^{-1/2}_{Z,\Delta} G^{1/2}_{Z,\Delta} - \left(G^{-1/2}_{Z,\Delta} S_\Delta\right)_l^-\right]\right\| = O_p\left(\sigma_\Delta^{-1}\zeta_\Delta\sqrt{\frac{\log J_{\Delta,\max}}{n}}\right). \qquad \text{(A.31)}$$

It remains to bound the sample moment in (A.30). We first decompose the normalized sample moment into a centered empirical component and its population mean:

$$\begin{aligned}&G^{-1/2}_{Z,\Delta} \frac{1}{n}\sum_{i=1}^n q^{K_{\Delta,\max}}(Z_i)\left(Y_i - \Xi_{J_{\Delta,\max}} h_0(X_i)\right) \\ &= G^{-1/2}_{Z,\Delta} \frac{1}{n}\sum_{i=1}^n \left\{q^{K_{\Delta,\max}}(Z_i)\left(Y_i - \Xi_{J_{\Delta,\max}} h_0(X_i)\right) - E\left[q^{K_{\Delta,\max}}(Z_i)\left(Y_i - \Xi_{J_{\Delta,\max}} h_0(X_i)\right)\right]\right\} \qquad \text{(A.32)} \\ &\quad + G^{-1/2}_{Z,\Delta} E\left[q^{K_{\Delta,\max}}(Z_i)\left(Y_i - \Xi_{J_{\Delta,\max}} h_0(X_i)\right)\right]. \qquad \text{(A.33)}\end{aligned}$$

We bound the two terms separately. First consider the centered empirical component (A.32). Because the observations are i.i.d. and the summands are centered, the cross terms vanish. Therefore,

$$E\left\|G_{Z,\Delta}^{-1/2}\frac{1}{n}\sum_{i=1}^{n}\left\{q^{K_{\Delta,\max}}(Z_i)\left(Y_i-\Xi_{J_{\Delta,\max}}h_0(X_i)\right)-E\left[q^{K_{\Delta,\max}}(Z_i)\left(Y_i-\Xi_{J_{\Delta,\max}}h_0(X_i)\right)\right]\right\}\right\|^2$$

$$=\frac{1}{n}E\left\|G_{Z,\Delta}^{-1/2}\left\{q^{K_{\Delta,\max}}(Z_i)\left(Y_i-\Xi_{J_{\Delta,\max}}h_0(X_i)\right)-E\left[q^{K_{\Delta,\max}}(Z_i)\left(Y_i-\Xi_{J_{\Delta,\max}}h_0(X_i)\right)\right]\right\}\right\|^2$$

$$\leq\frac{1}{n}E\left[\left(Y_i-\Xi_{J_{\Delta,\max}}h_0(X_i)\right)^2q^{K_{\Delta,\max}}(Z_i)'G_{Z,\Delta}^{-1}q^{K_{\Delta,\max}}(Z_i)\right]$$

$$\leq\frac{2}{n}E\left[U_i^2q^{K_{\Delta,\max}}(Z_i)'G_{Z,\Delta}^{-1}q^{K_{\Delta,\max}}(Z_i)\right]+\frac{2}{n}E\left[\left(h_0(X_i)-\Xi_{J_{\Delta,\max}}h_0(X_i)\right)^2q^{K_{\Delta,\max}}(Z_i)'G_{Z,\Delta}^{-1}q^{K_{\Delta,\max}}(Z_i)\right] \tag{A.34}$$

For the first term of (A.34), we have

$$E\left[U_i^2q^{K_{\Delta,\max}}(Z_i)'G_{Z,\Delta}^{-1}q^{K_{\Delta,\max}}(Z_i)\right]=E\left[E[U_i^2\mid Z_i]\,q^{K_{\Delta,\max}}(Z_i)'G_{Z,\Delta}^{-1}q^{K_{\Delta,\max}}(Z_i)\right]$$

$$\overset{(a)}{\leq}\bar{\sigma}_U^2E\left[q^{K_{\Delta,\max}}(Z_i)'G_{Z,\Delta}^{-1}q^{K_{\Delta,\max}}(Z_i)\right]\overset{(b)}{=}\bar{\sigma}_U^2K_{\Delta,\max}$$

where the step (a) follows from Assumption 4.3, and (b) follows from the definition of $G_{Z,\Delta}$.

For the second term of (A.34), we have

$$E\left[\left(h_0(X_i)-\Xi_{J_{\Delta,\max}}h_0(X_i)\right)^2q^{K_{\Delta,\max}}(Z_i)'G_{Z,\Delta}^{-1}q^{K_{\Delta,\max}}(Z_i)\right]\leq\zeta_\Delta^2\left\|h_0-\Xi_{J_{\Delta,\max}}h_0\right\|_X^2$$

$$\lesssim\zeta_\Delta^2J_{\Delta,\max}^{-2s/d_x}\overset{(a)}{\lesssim}\frac{\sigma_\Delta^2}{\log J_{\Delta,\max}}\left(\frac{1}{J_{\Delta,\max}}\sum_{j=1}^{J_{\Delta,\max}}\nu_j^{-2}\right)\overset{(b)}{\lesssim}\frac{1}{\log J_{\Delta,\max}}\lesssim K_{\Delta,\max}$$

where (a) follows from the second condition in (A.16). For step (b), by Breunig and Chen (2024, Page 2038), the right-hand inequality in (4.2) and the monotonicity of $\{\nu_j\}$ imply

$$\sigma_\Delta^2\leq\left\|\Pi_{K_{\Delta,\max}}T\psi_{J_{\Delta,\max}}\right\|_Z^2\lesssim\nu_{J_{\Delta,\max}}^2,\qquad\frac{1}{J_{\Delta,\max}}\sum_{j=1}^{J_{\Delta,\max}}\nu_j^{-2}\leq\nu_{J_{\Delta,\max}}^{-2}.$$

The step (c) holds since $J_{\Delta,\max}\to\infty$ and $K_{\Delta,\max}\geq J_{\Delta,\max}$.

Consequently,

$$E\left\|G_{Z,\Delta}^{-1/2}\frac{1}{n}\sum_{i=1}^{n}\left\{q^{K_{\Delta,\max}}(Z_i)\left(Y_i-\Xi_{J_{\Delta,\max}}h_0(X_i)\right)-E\left[q^{K_{\Delta,\max}}(Z_i)\left(Y_i-\Xi_{J_{\Delta,\max}}h_0(X_i)\right)\right]\right\}\right\|^2\lesssim\frac{K_{\Delta,\max}}{n}.$$

Markov's inequality therefore yields

$$\left\|G_{Z,\Delta}^{-1/2}\frac{1}{n}\sum_{i=1}^{n}\left\{q^{K_{\Delta,\max}}(Z_i)\left(Y_i-\Xi_{J_{\Delta,\max}}h_0(X_i)\right)-E\left[q^{K_{\Delta,\max}}(Z_i)\left(Y_i-\Xi_{J_{\Delta,\max}}h_0(X_i)\right)\right]\right\}\right\|=O_p\left(\sqrt{\frac{K_{\Delta,\max}}{n}}\right).$$

By the conditional moment restriction (2.2), the population component (A.33) satisfies

$$\begin{aligned}\left\|G_{Z,\Delta}^{-1/2}E\left[q^{K_{\Delta,\max}}(Z_i)\left(Y_i-\Xi_{J_{\Delta,\max}}h_0(X_i)\right)\right]\right\|&=\left\|\Pi_{K_{\Delta,\max}}T\left(h_0-\Xi_{J_{\Delta,\max}}h_0\right)\right\|_Z\\&\le\left\|\Pi_{K_{\Delta,\max}}T\left(h_0-\Pi_{J_{\Delta,\max}}h_0\right)\right\|_Z\lesssim\sigma_\Delta J_{\Delta,\max}^{-s/d_x}\end{aligned}$$

where the first inequality follows from the definition of $\Xi_{J_{\Delta,\max}}h_0$, and the final inequality follows from Assumption 4.4 (ii) and the sieve approximation bound.

Combining the bounds for (A.32) and (A.33), we obtain

$$\left\|G_{Z,\Delta}^{-1/2}\frac{1}{n}\sum_{i=1}^{n}q^{K_{\Delta,\max}}(Z_i)\left(Y_i-\Xi_{J_{\Delta,\max}}h_0(X_i)\right)\right\|=O_p\left(\sqrt{\frac{K_{\Delta,\max}}{n}}+\sigma_\Delta J_{\Delta,\max}^{-s/d_x}\right).\tag{A.35}$$

Finally, combining (A.31) and (A.35) gives that

$$\left\|\Pi_{K_{\Delta,\max}}Tr_{\mathrm{IV},J_{\Delta,\max}}\right\|_Z=O_p\left(\sigma_\Delta^{-1}\zeta_\Delta\sqrt{\frac{K_{\Delta,\max}\log J_{\Delta,\max}}{n^2}}+\zeta_\Delta J_{\Delta,\max}^{-s/d_x}\sqrt{\frac{\log J_{\Delta,\max}}{n}}\right).\tag{A.36}$$

Step 3. The population centering error induced by the common pilot

The common NPIV pilot is centered at the population sieve 2SLS projection $\Xi_{J_{\Delta,\max}}h_0$, rather than at the $L^2(X)$-projection $\Pi_{J_{\Delta,\max}}h_0$. We now bound the resulting centering error after projection onto any lower-dimensional sieve space $\mathcal{H}_J$, where $J\le J_{\Delta,\max}$.

Write

$$\Xi_{J_{\Delta,\max}}h_0-\Pi_{J_{\Delta,\max}}h_0\in\mathcal{H}_{J_{\Delta,\max}}.$$

The first-order condition in (A.29) evaluating in the direction $\Xi_{J_{\Delta,\max}}h_0-\Pi_{J_{\Delta,\max}}h_0$ gives that

$$0=\left\langle\Pi_{K_{\Delta,\max}}T\left(\Xi_{J_{\Delta,\max}}h_0-h_0\right),\ \Pi_{K_{\Delta,\max}}T\left(\Xi_{J_{\Delta,\max}}h_0-\Pi_{J_{\Delta,\max}}h_0\right)\right\rangle_Z.$$

Using $\Xi_{J_{\Delta,\max}}h_0-h_0=\left(\Xi_{J_{\Delta,\max}}h_0-\Pi_{J_{\Delta,\max}}h_0\right)-\left(h_0-\Pi_{J_{\Delta,\max}}h_0\right)$, we can rearrange the first-order

condition as

$$\left\|\Pi_{K_{\Delta,\max}} T\left(\Xi_{J_{\Delta,\max}} h_0 - \Pi_{J_{\Delta,\max}} h_0\right)\right\|_Z^2$$

$$= \left\langle \Pi_{K_{\Delta,\max}} T\left(h_0 - \Pi_{J_{\Delta,\max}} h_0\right),\ \Pi_{K_{\Delta,\max}} T\left(\Xi_{J_{\Delta,\max}} h_0 - \Pi_{J_{\Delta,\max}} h_0\right)\right\rangle_Z.$$

By the Cauchy-Schwarz inequality,

$$\left\|\Pi_{K_{\Delta,\max}} T\left(\Xi_{J_{\Delta,\max}} h_0 - \Pi_{J_{\Delta,\max}} h_0\right)\right\|_Z \le \left\|\Pi_{K_{\Delta,\max}} T\left(h_0 - \Pi_{J_{\Delta,\max}} h_0\right)\right\|_Z \le \left\|T\left(h_0 - \Pi_{J_{\Delta,\max}} h_0\right)\right\|_Z. \tag{A.37}$$

where the second inequality follows because $\Pi_{K_{\Delta,\max}}$ is an $L^2(Z)$-orthogonal projection.

We next bound the right-hand side of (A.37). Because $\{\psi\}_j$ is orthonormal and $\mathcal{H}_J = \mathrm{clsp}\{\psi_1, ..., \psi_J\}$, the $L^2(X)$-orthogonal projection of h_0 onto $\mathcal{H}_J$ is $\Pi_J h_0 = \sum_{j=1}^J \langle h_0, \psi_j\rangle_X \psi_j$ and

$$h_0 - \Pi_{J_{\Delta,\max}} h_0 = \sum_{j > J_{\Delta,\max}} \langle h_0, \psi_j\rangle_X \psi_j.$$

The upper link condition (4.2) and Assumption 4.1 imply

$$\left\|T\left(h_0 - \Pi_{J_{\Delta,\max}} h_0\right)\right\|_Z^2 \le C \sum_{j > J_{\Delta,\max}} \nu_j^2 \langle h_0, \psi_j\rangle_X^2 \le C\nu_{J_{\Delta,\max}+1}^2 \left\|h_0 - \Pi_{J_{\Delta,\max}} h_0\right\|_X^2 \lesssim \nu_{J_{\Delta,\max}+1}^2 J_{\Delta,\max}^{-2s/d_x}. \tag{A.38}$$

Following the argument in Breunig and Chen (2024, p. 2038), by Assumption 4.4 (iii), applied at $(J_{\Delta,\max}, K_{\Delta,\max})$, and the definition of τ_J in (2.4), we have, uniformly over $h \in \mathcal{H}_{J_{\Delta,\max}}$,

$$\left\|(I - \Pi_{K_{\Delta,\max}}) Th\right\|_Z \le o(1)\tau_{J_{\Delta,\max}}^{-1} \|h\|_X \le o(1)\|Th\|_Z.$$

Consequently, the reverse triangle inequality and the lower link condition (4.2) in Assumption 4.4 (iv) imply that, for all sufficiently large n,

$$\left\|\Pi_{K_{\Delta,\max}} Th\right\|_Z^2 \gtrsim \left\|Th\right\|_Z^2 \gtrsim \sum_{j=1}^{J_{\Delta,\max}} \nu_j^2 \langle h, \psi_j\rangle_X^2, \qquad h \in \mathcal{H}_{J_{\Delta,\max}}. \tag{A.39}$$

Combining (A.39), (A.37), and (A.38) gives

$$\left[\sum_{j=1}^{J_{\Delta,\max}} \nu_j^2 \left\langle \Xi_{J_{\Delta,\max}} h_0 - \Pi_{J_{\Delta,\max}} h_0, \psi_j\right\rangle_X^2\right]^{1/2} \lesssim \nu_{J_{\Delta,\max}+1} J_{\Delta,\max}^{-s/d_x}. \tag{A.40}$$

It remains to translate this weighted bound into the $L^2(X)$ norm relevant for the lower-dimensional

sieve space $\mathcal{H}_J$. Since $J \leq J_{\Delta,\max}$ and $\Pi_J(h_0 - \Pi_{J_{\Delta,\max}} h_0) = 0$, we obtain

$$\begin{aligned}\left\|\Pi_J\left(\Xi_{J_{\Delta,\max}} h_0 - h_0\right)\right\|_X =& \left\|\Pi_J\left(\Xi_{J_{\Delta,\max}} h_0 - \Pi_{J_{\Delta,\max}} h_0\right)\right\|_X = \sum_{j=1}^{J} \left\langle \Xi_{J_{\Delta,\max}} h_0 - \Pi_{J_{\Delta,\max}} h_0, \psi_j \right\rangle_X^2 \\ \leq & \nu_J^{-1} \left[\sum_{j=1}^{J_{\Delta,\max}} \nu_j^2 \left\langle \Xi_{J_{\Delta,\max}} h_0 - \Pi_{J_{\Delta,\max}} h_0, \psi_j \right\rangle_X^2\right]^{1/2} \lesssim \frac{\nu_{J_{\Delta,\max}+1}}{\nu_J} J_{\Delta,\max}^{-s/d_x}.\end{aligned} \tag{A.41}$$

Step 4. Rate of the Variance of $\Psi_{\Delta,J}$

Since $E[\Psi_{g,J}] = 0$ and $E[\xi_{\mathrm{IV},J_{\Delta,\max}}(\mathcal{D}_i)] = 0$, it follows that $E[\Psi_{\Delta,J}] = 0$. We now establish the order of the variance of $\Psi_{\Delta,J}$. We begin with the projected NPIV component. Let

$$\psi^{J_{\Delta,\max}}(x) = \left(\psi_1(x), \ldots, \psi_{J_{\Delta,\max}}(x)\right)'.$$

Since both $G_{X,\Delta}^{-1/2} p^{J_{\Delta,\max}}$ and $\psi^{J_{\Delta,\max}}$ are orthonormal bases of $\mathcal{H}_{J_{\Delta,\max}}$, there exists an orthogonal matrix $O_\Delta \in \mathbb{R}^{J_{\Delta,\max} \times J_{\Delta,\max}}$ such that

$$G_{X,\Delta}^{-1/2} p^{J_{\Delta,\max}}(x) = O_\Delta \psi^{J_{\Delta,\max}}(x).$$

Using the definition of A_Δ, we obtain

$$A_\Delta G_{Z,\Delta} A_\Delta' = O_\Delta \Big(E[\psi^{J_{\Delta,\max}}(X_i) q^{K_{\Delta,\max}}(Z_i)'] G_{Z,\Delta}^{-1} \times E[q^{K_{\Delta,\max}}(Z_i) \psi^{J_{\Delta,\max}}(X_i)'] \Big)^{-1} O_\Delta'.$$

Projection onto $\mathcal{H}_J$ retains the first J coordinates in the basis $\{\psi_j\}$. Therefore,

$$\begin{aligned} & E \left\| \Pi_J \left[p^{J_{\Delta,\max}}(\cdot)' G_{X,\Delta}^{-1/2} A_\Delta q^{K_{\Delta,\max}}(Z_i) \right] \right\|_X^2 \\ = & \sum_{j=1}^{J} \left[\Big(E[\psi^{J_{\Delta,\max}}(X_i) q^{K_{\Delta,\max}}(Z_i)'] G_{Z,\Delta}^{-1} E[q^{K_{\Delta,\max}}(Z_i) \psi^{J_{\Delta,\max}}(X_i)'] \Big)^{-1} \right]_{jj}. \end{aligned} \tag{A.42}$$

To evaluate the right-hand side, take any $h = \psi^{J_{\Delta,\max}\prime} a \in \mathcal{H}_{J_{\Delta,\max}}$, then

$$\left\| \Pi_{K_{\Delta,\max}} T h \right\|_Z^2 = a' E\Big[\psi^{J_{\Delta,\max}}(X_i) q^{K_{\Delta,\max}}(Z_i)'\Big] G_{Z,\Delta}^{-1} \times E\Big[q^{K_{\Delta,\max}}(Z_i) \psi^{J_{\Delta,\max}}(X_i)'\Big] a.$$

Thus, the matrix in (A.42) represents the quadratic form $h \mapsto \|\Pi_{K_{\Delta,\max}} T h\|_Z^2$ on $\mathcal{H}_{J_{\Delta,\max}}$.

By the exact link condition and the instrument-sieve approximation condition, there exist con-

stants $0 < c \leq C < \infty$ such that, for all sufficiently large n,

$$c \sum_{j=1}^{J_{\Delta,\max}} \nu_j^2 a_j^2 \leq \left\| \Pi_{K_{\Delta,\max}} Th \right\|_Z^2 \leq C \sum_{j=1}^{J_{\Delta,\max}} \nu_j^2 a_j^2$$

uniformly over $a \in \mathbb{R}^{J_{\Delta,\max}}$. Equivalently, in the Loewner order,

$$\begin{aligned} c \operatorname{diag}\left(\nu_1^2, \ldots, \nu_{J_{\Delta,\max}}^2\right) &\preceq E[\psi^{J_{\Delta,\max}}(X_i) q^{K_{\Delta,\max}}(Z_i)'] G_{Z,\Delta}^{-1} \times E[q^{K_{\Delta,\max}}(Z_i) \psi^{J_{\Delta,\max}}(X_i)'] \\ &\preceq C \operatorname{diag}\left(\nu_1^2, \ldots, \nu_{J_{\Delta,\max}}^2\right). \end{aligned}$$

Since these matrices are positive definite, inversion reverses the Loewner order. Hence

$$\begin{aligned} C^{-1} \operatorname{diag}\left(\nu_1^{-2}, \ldots, \nu_{J_{\Delta,\max}}^{-2}\right) &\preceq \left(E[\psi^{J_{\Delta,\max}}(X_i) q^{K_{\Delta,\max}}(Z_i)'] G_{Z,\Delta}^{-1} \times E[q^{K_{\Delta,\max}}(Z_i) \psi^{J_{\Delta,\max}}(X_i)'] \right)^{-1} \\ &\preceq c^{-1} \operatorname{diag}\left(\nu_1^{-2}, \ldots, \nu_{J_{\Delta,\max}}^{-2}\right). \end{aligned}$$

Taking the sum of the first J diagonal elements and using (A.42) gives

$$E \left\| \Pi_J \left[p^{J_{\Delta,\max}}(\cdot)' G_{X,\Delta}^{-1/2} A_\Delta q^{K_{\Delta,\max}}(Z_i) \right] \right\|_X^2 \asymp \sum_{j=1}^{J} \nu_j^{-2}. \tag{A.43}$$

We now establish the corresponding order for $E\|\Psi_{\Delta,J}\|_X^2$. Recall from (A.19) that

$$\begin{aligned} \Psi_{\Delta,J}(x) = \frac{1}{n} \sum_{i=1}^{n} \Bigg\{ & p^J(x)' G_X^{-1} p^J(X_i) \big(Y_i - \Pi_J g_0(X_i)\big) \\ & - \Pi_J \left[p^{J_{\Delta,\max}}(\cdot)' G_{X,\Delta}^{-1/2} A_\Delta q^{K_{\Delta,\max}}(Z_i) \big(Y_i - \Xi_{J_{\Delta,\max}} h_0(X_i)\big) \right](x) \Bigg\}. \end{aligned}$$

Because the summands are i.i.d. and mean zero,

$$\begin{aligned} E\left\| \Psi_{\Delta,J} \right\|_X^2 = \frac{1}{n} E \Bigg\| & p^J(\cdot)' G_X^{-1} p^J(X_i) \big(Y_i - \Pi_J g_0(X_i)\big) \\ & - \Pi_J \left[p^{J_{\Delta,\max}}(\cdot)' G_{X,\Delta}^{-1/2} A_\Delta q^{K_{\Delta,\max}}(Z_i) \big(Y_i - \Xi_{J_{\Delta,\max}} h_0(X_i)\big) \right] \Bigg\|_X^2. \end{aligned}$$

For the upper bound, the inequality $\|a - b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, the conditional second-moment

bounds, (A.22), and (A.43) imply

$$E\left\|\Psi_{\Delta,J}\right\|_X^2 \lesssim \frac{1}{n}\left(J+\sum_{j=1}^{J}\nu_j^{-2}\right) \lesssim \frac{1}{n}\sum_{j=1}^{J}\nu_j^{-2}$$

where the second inequality holds because $\nu_j \downarrow 0$, which implies $J = o\left(\sum_{j=1}^{J}\nu_j^{-2}\right)$ uniformly over $J \geq \underline{J}$ as $\underline{J} \to \infty$.

For the lower bound, write

$$Y_i - \Pi_J g_0(X_i) = \big(Y_i - E[Y_i \mid X_i, Z_i]\big) + \big(E[Y_i \mid X_i, Z_i] - \Pi_J g_0(X_i)\big),$$

and

$$Y_i - \Xi_{J_{\Delta,\max}} h_0(X_i) = \big(Y_i - E[Y_i \mid X_i, Z_i]\big) + \big(E[Y_i \mid X_i, Z_i] - \Xi_{J_{\Delta,\max}} h_0(X_i)\big).$$

Substituting these decompositions into the ith summand of $\Psi_{\Delta,J}$, and using the linearity of Π_J, gives

$$\begin{aligned}
&\left\{p^J(\cdot)'G_X^{-1}p^J(X_i) - \Pi_J\left[p^{J_{\Delta,\max}}(\cdot)'G_{X,\Delta}^{-1/2}A_\Delta q^{K_{\Delta,\max}}(Z_i)\right]\right\}\big(Y_i - E[Y_i \mid X_i, Z_i]\big) \\
&+ p^J(\cdot)'G_X^{-1}p^J(X_i)\big(E[Y_i \mid X_i, Z_i] - \Pi_J g_0(X_i)\big) \\
&- \Pi_J\left[p^{J_{\Delta,\max}}(\cdot)'G_{X,\Delta}^{-1/2}A_\Delta q^{K_{\Delta,\max}}(Z_i)\right]\big(E[Y_i \mid X_i, Z_i] - \Xi_{J_{\Delta,\max}} h_0(X_i)\big).
\end{aligned}$$

Conditional on (X_i, Z_i), the first term has mean zero, whereas the remaining two terms are measurable with respect to (X_i, Z_i). The conditional covariance between the first term and the remaining terms is zero. Since $\mathrm{Var}(Y_i \mid X_i, Z_i) = \mathrm{Var}(U_i \mid X_i, Z_i) \geq \underline{\sigma}^2$ by Assumption 4.2(i), we obtain

$$nE\left\|\Psi_{\Delta,J}\right\|_X^2 \geq \underline{\sigma}^2 E\left\|p^J(\cdot)'G_X^{-1}p^J(X_i) - \Pi_J\left[p^{J_{\Delta,\max}}(\cdot)'G_{X,\Delta}^{-1/2}A_\Delta q^{K_{\Delta,\max}}(Z_i)\right]\right\|_X^2. \tag{A.44}$$

We evaluate the three terms in the squared norm separately. First,

$$E\left\|p^J(\cdot)'G_X^{-1}p^J(X_i)\right\|_X^2 = E\left[p^J(X_i)'G_X^{-1}p^J(X_i)\right] = J.$$

Second, by (A.43),

$$E\left\|\Pi_J\left[p^{J_{\Delta,\max}}(\cdot)'G_{X,\Delta}^{-1/2}A_\Delta q^{K_{\Delta,\max}}(Z_i)\right]\right\|_X^2 \asymp \sum_{j=1}^{J}\nu_j^{-2}.$$

Finally, the cross term equals J. Indeed, the defining left-inverse property of A_Δ, together with the projection onto the first J coordinates, gives

$$E\left\langle p^J(\cdot)'G_X^{-1}p^J(X_i),\ \Pi_J\left[p^{J_{\Delta,\max}}(\cdot)'G_{X,\Delta}^{-1/2}A_\Delta q^{K_{\Delta,\max}}(Z_i)\right]\right\rangle_X = J.$$

Consequently,

$$E\left\|p^J(\cdot)'G_X^{-1}p^J(X_i) - \Pi_J\left[p^{J_{\Delta,\max}}(\cdot)'G_{X,\Delta}^{-1/2}A_\Delta q^{K_{\Delta,\max}}(Z_i)\right]\right\|_X^2$$

$$= E\left\|\Pi_J\left[p^{J_{\Delta,\max}}(\cdot)'G_{X,\Delta}^{-1/2}A_\Delta q^{K_{\Delta,\max}}(Z_i)\right]\right\|_X^2 - J \asymp \sum_{j=1}^{J}\nu_j^{-2} - J \gtrsim \sum_{j=1}^{J}\nu_j^{-2}.$$

Together with the upper bound, we conclude that

$$E\left\|\Psi_{\Delta,J}\right\|_X^2 \asymp \frac{1}{n}\sum_{j=1}^{J}\nu_j^{-2} \tag{A.45}$$

uniformly over $J \in \mathcal{J}_n$ with $J \le J_{\Delta,\max}$.

Step 5. Uniform asymptotic expansion of $\widehat{\Delta}_J - \Delta_J$

We now establish the uniform expansion for $\widehat{\Delta}_J - \Delta_J$. Because the sieve spaces are nested and $\widehat{\Pi}_J$ is the empirical least-squares projection, we have that $\widehat{\Pi}_J\widehat{g}_{J_{\Delta,\max}} = \widehat{g}_J$, for $J \le J_{\Delta,\max}$.

Recall that $\widehat{\Delta}_J = \widehat{g}_J - \widehat{\Pi}_J\widetilde{h}_{\mathrm{IV},J_{\Delta,\max}}$. Subtracting $\Delta_J = \Pi_J g_0 - \Pi_J h_0$ and using (A.26), we obtain that

$$\begin{aligned}\widehat{\Delta}_J - \Delta_J &= (\widehat{g}_J - \Pi_J g_0) - \widehat{\Pi}_J\left(\widetilde{h}_{\mathrm{IV},J_{\Delta,\max}} - \Xi_{J_{\Delta,\max}}h_0\right) - \left[\widehat{\Pi}_J\Xi_{J_{\Delta,\max}}h_0 - \Pi_J h_0\right]\\ &:= \Psi_{\Delta,J} + r_{\Delta,J}\end{aligned}$$

where

$$\Psi_{\Delta,J} = \Psi_{g,J} - \Pi_J\left[\frac{1}{n}\sum_{i=1}^{n}\xi_{\mathrm{IV},J_{\Delta,\max}}(\mathcal{D}_i)\right],$$

and

$$\begin{aligned}r_{\Delta,J} = r_{g,J} - \Pi_J r_{\mathrm{IV},J_{\Delta,\max}} - \left(\widehat{\Pi}_J - \Pi_J\right)\left[\frac{1}{n}\sum_{i=1}^{n}\xi_{\mathrm{IV},J_{\Delta,\max}}(\mathcal{D}_i)\right] - \left(\widehat{\Pi}_J - \Pi_J\right)r_{\mathrm{IV},J_{\Delta,\max}}\\ - (\widehat{\Pi}_J - \Pi_J)\Xi_{J_{\Delta,\max}}h_0 - \Pi_J\left(\Xi_{J_{\Delta,\max}}h_0 - h_0\right).\end{aligned} \tag{A.46}$$

The NP-OLS remainder term $r_{g,J}$ has already been controlled in Step 1. We now consider the five remaining terms in the order in which they appear in (A.46).

We first bound $\Pi_J r_{\mathrm{IV},J_{\Delta,\max}}$.

Since $r_{\mathrm{IV},J_{\Delta,\max}} \in \mathcal{H}_{J_{\Delta,\max}}$, (A.39) and the monotonicity of $\{\nu_j\}$ imply that, for $J \le J_{\Delta,\max}$,

$$\left\|\Pi_J r_{\mathrm{IV},J_{\Delta,\max}}\right\|_X^2 = \sum_{j=1}^{J} \left\langle r_{\mathrm{IV},J_{\Delta,\max}}, \psi_j \right\rangle_X^2 \le \nu_J^{-2} \sum_{j=1}^{J_{\Delta,\max}} \nu_j^2 \left\langle r_{\mathrm{IV},J_{\Delta,\max}}, \psi_j \right\rangle_X^2 \lesssim \nu_J^{-2} \left\|\Pi_{K_{\Delta,\max}} T r_{\mathrm{IV},J_{\Delta,\max}}\right\|_Z^2. \tag{A.47}$$

Combining (A.47) with (A.36), the variance bound (A.45), and the pilot sieve restrictions (A.16) gives, uniformly over $J \in \mathcal{J}_n$ with $J \le J_{\Delta,\max}$,

$$\frac{\left\|\Pi_J r_{\mathrm{IV},J_{\Delta,\max}}\right\|_X}{\left(E\|\Psi_{\Delta,J}\|_X^2\right)^{1/2}} = O_p\left(\frac{\nu_J^{-1}}{\left(\sum_{j=1}^{J} \nu_j^{-2}\right)^{1/2}} \left[\sigma_\Delta^{-1} \zeta_\Delta \sqrt{\frac{K_{\Delta,\max} \log J_{\Delta,\max}}{n}} + \zeta_\Delta J_{\Delta,\max}^{-s/d_x} \sqrt{\log J_{\Delta,\max}} \right] \right). \tag{A.48}$$

We show that the right-hand side is $o_p(1)$. In the mildly ill-posed case, $\nu_j \asymp j^{-\varsigma/d_x}$, so

$$\frac{\nu_J^{-1}}{\left(\sum_{j=1}^{J} \nu_j^{-2}\right)^{1/2}} \asymp J^{-1/2}.$$

Moreover, again by the arguments in Breunig and Chen (2024, page 2038), we have that

$$\sigma_\Delta^{-1} \asymp \nu_{J_{\Delta,\max}}^{-1} \asymp \left(J_{\Delta,\max}^{-1} \sum_{j=1}^{J_{\Delta,\max}} \nu_j^{-2} \right)^{1/2}.$$

Together with (A.16), this implies that the expression in square brackets in (A.48) is $O(1)$. Since $J \ge \underline{J} \to \infty$, the bound is $o_p(1)$ uniformly over the candidate dimensions.

In the severely ill-posed case, $\nu_j \asymp \exp(-j^{\varsigma/d_x}/2)$, by monotonicity of $\{\nu_j\}$,

$$\frac{\nu_J^{-1}}{\left(\sum_{j=1}^{J} \nu_j^{-2}\right)^{1/2}} \le 1.$$

Furthermore, it follows that

$$\nu_{J_{\Delta,\max}}^2 J_{\Delta,\max}^{-1} \sum_{j=1}^{J_{\Delta,\max}} \nu_j^{-2} \asymp \frac{\exp(-J_{\Delta,\max}^{\varsigma/d_x})}{J_{\Delta,\max}} \sum_{j=1}^{J_{\Delta,\max}} \exp(j^{\varsigma/d_x}) \asymp J_{\Delta,\max}^{-\min\{\varsigma/d_x,1\}} \to 0. \tag{A.49}$$

Since $\sigma_\Delta \asymp \nu_{J_{\Delta,\max}}$, (A.16) and (A.49) imply that both terms in square brackets in (A.48) are

$o(1)$. The bound is therefore $o_p(1)$ uniformly over the candidate dimensions in this case as well. Thus

$$\max_{\substack{J\in\mathcal{J}_n\\ J\le J_{\Delta,\max}}} \frac{\left\|\Pi_J r_{\mathrm{IV},J_{\Delta,\max}}\right\|_X}{\left(E\|\Psi_{\Delta,J}\|_X^2\right)^{1/2}} = o_p(1). \tag{A.50}$$

We next control the replacement of Π_J by $\widehat{\Pi}_J$.

The same Gram matrix concentration argument used to establish (A.23) now applied at the pilot dimension, gives

$$\sup_{\substack{J\in\mathcal{J}_n\\ J\le J_{\Delta,\max}}} \sup_{\substack{f\in\mathcal{H}_{J_{\Delta,\max}}\\ f\neq 0}} \frac{\|(\widehat{\Pi}_J-\Pi_J)f\|_X}{\|f\|_X} = O_p\left(\zeta_{X,J_{\Delta,\max}}\sqrt{\frac{\log J_{\Delta,\max}}{n}}\right). \tag{A.51}$$

To verify (A.51), let $f\in\mathcal{H}_{J_{\Delta,\max}}$ and set $\widehat{\Pi}_J f-\Pi_J f\in\mathcal{H}_J$. The first-order conditions for the empirical least squares projection imply that, for every $h\in\mathcal{H}_J$,

$$\left\langle \widehat{\Pi}_J f-\Pi_J f, h\right\rangle_{X,n} = \left\langle f-\Pi_J f, h\right\rangle_{X,n}. \tag{A.52}$$

The corresponding population inner product on the right-hand side is zero. Uniform concentration of the empirical inner product over $\mathcal{H}_{J_{\Delta,\max}}$, together with uniform equivalence between $\|\cdot\|_{X,n}$ and $\|\cdot\|_X$, gives

$$\left\|\widehat{\Pi}_J f-\Pi_J f\right\|_X \lesssim_p \zeta_{X,J_{\Delta,\max}}\sqrt{\frac{\log J_{\Delta,\max}}{n}}\left\|f-\Pi_J f\right\|_X.$$

Since $\|f-\Pi_J f\|_X\le\|f\|_X$ uniformly over $J\le J_{\Delta,\max}$, this proves (A.51).

We first apply (A.51) to the leading pilot score. By (A.29), (A.60), and the second condition in (A.16), together with (A.43), we obtain that

$$\left\|\frac{1}{n}\sum_{i=1}^n \xi_{\mathrm{IV},J_{\Delta,\max}}(\mathcal{D}_i)\right\|_X = O_p\left[\left(\frac{1}{n}\sum_{j=1}^{J_{\Delta,\max}}\nu_j^{-2}\right)^{1/2}\right]. \tag{A.53}$$

Combining (A.51), (A.53), and (A.45) yields

$$\frac{\left\|(\widehat{\Pi}_J-\Pi_J)n^{-1}\sum_{i=1}^n \xi_{\mathrm{IV},J_{\Delta,\max}}(\mathcal{D}_i)\right\|_X}{\left(E\|\Psi_{\Delta,J}\|_X^2\right)^{1/2}} = O_p\left[\zeta_{X,J_{\Delta,\max}}\sqrt{\frac{\log J_{\Delta,\max}}{n}}\frac{\left(\sum_{j=1}^{J_{\Delta,\max}}\nu_j^{-2}\right)^{1/2}}{\left(\sum_{j=1}^{J}\nu_j^{-2}\right)^{1/2}}\right]. \tag{A.54}$$

Since $\{\nu_j\}$ is decreasing and $\sigma_\Delta\asymp\nu_{J_{\Delta,\max}}$,

$$\sum_{j=1}^{J_{\Delta,\max}}\nu_j^{-2}\le J_{\Delta,\max}\nu_{J_{\Delta,\max}}^{-2}\asymp J_{\Delta,\max}\sigma_\Delta^{-2},$$

and $K_{\Delta,\max} \asymp J_{\Delta,\max}$, the right-hand side of (A.54) is bounded, up to a constant, by

$$\frac{\sigma_\Delta^{-1}\zeta_\Delta\sqrt{K_{\Delta,\max}\log J_{\Delta,\max}/n}}{\left(\sum_{j=1}^{J}\nu_j^{-2}\right)^{1/2}}. \tag{A.55}$$

The first pilot sieve condition in (A.16) implies

$$\sigma_\Delta^{-1}\zeta_\Delta\sqrt{\frac{K_{\Delta,\max}\log J_{\Delta,\max}}{n}} \lesssim \sigma_\Delta\left(J_{\Delta,\max}^{-1}\sum_{j=1}^{J_{\Delta,\max}}\nu_j^{-2}\right)^{1/2}.$$

Under mild ill-posedness, the numerator in (A.55) is $O(1)$, while the denominator diverges uniformly because $J \geq \underline{J} \to \infty$. Under severe ill-posedness, the numerator itself is $o(1)$ by (A.49). Consequently,

$$\max_{\substack{J\in\mathcal{J}_n\\ J\leq J_{\Delta,\max}}}\frac{\left\|(\widehat{\Pi}_J-\Pi_J)n^{-1}\sum_{i=1}^{n}\xi_{\mathrm{IV},J_{\Delta,\max}}(\mathcal{D}_i)\right\|_X}{\left(E\|\Psi_{\Delta,J}\|_X^2\right)^{1/2}} = o_p(1). \tag{A.56}$$

We next consider $(\widehat{\Pi}_J-\Pi_J)r_{\mathrm{IV},J_{\Delta,\max}}$. Since $r_{\mathrm{IV},J_{\Delta,\max}} \in \mathcal{H}_{J_{\Delta,\max}}$, the definition of σ_Δ gives

$$\left\|r_{\mathrm{IV},J_{\Delta,\max}}\right\|_X \leq \sigma_\Delta^{-1}\left\|\Pi_{K_{\Delta,\max}}Tr_{\mathrm{IV},J_{\Delta,\max}}\right\|_Z. \tag{A.57}$$

Applying (A.51) gives that

$$\left\|(\widehat{\Pi}_J-\Pi_J)r_{\mathrm{IV},J_{\Delta,\max}}\right\|_X \lesssim_p \zeta_{X,J_{\Delta,\max}}\sqrt{\frac{\log J_{\Delta,\max}}{n}}\sigma_\Delta^{-1}\times\left\|\Pi_{K_{\Delta,\max}}Tr_{\mathrm{IV},J_{\Delta,\max}}\right\|_Z.$$

By comparison, (A.47) gives

$$\|\Pi_J r_{\mathrm{IV},J_{\Delta,\max}}\|_X \lesssim \nu_J^{-1}\left\|\Pi_{K_{\Delta,\max}}Tr_{\mathrm{IV},J_{\Delta,\max}}\right\|_Z.$$

Thus the bound for $(\widehat{\Pi}_J-\Pi_J)r_{\mathrm{IV},J_{\Delta,\max}}$ differs from the bound already used for $\Pi_J r_{\mathrm{IV},J_{\Delta,\max}}$ by at most the multiplicative factor

$$\nu_J\sigma_\Delta^{-1}\zeta_{X,J_{\Delta,\max}}\sqrt{\frac{\log J_{\Delta,\max}}{n}}. \tag{A.58}$$

By the first pilot sieve condition in (A.16) and $\zeta_{X,J_{\Delta,\max}} \leq \zeta_\Delta$,

$$\sigma_\Delta^{-1}\zeta_{X,J_{\Delta,\max}}\sqrt{\frac{\log J_{\Delta,\max}}{n}} \lesssim \frac{\sigma_\Delta}{\sqrt{K_{\Delta,\max}}}\left(J_{\Delta,\max}^{-1}\sum_{j=1}^{J_{\Delta,\max}}\nu_j^{-2}\right)^{1/2} = O\left(K_{\Delta,\max}^{-1/2}\right) = o(1),$$

Since $\{\nu_J\}$ is bounded above, the factor in (A.58) is $o(1)$ uniformly over $J \leq J_{\Delta,\max}$.

Combining this result with (A.50) yields

$$\max_{\substack{J\in\mathcal{J}_n\\ J\le J_{\Delta,\max}}} \frac{\|(\widehat{\Pi}_J-\Pi_J)r_{\mathrm{IV},J_{\Delta,\max}}\|_X}{\left(E\|\Psi_{\Delta,J}\|_X^2\right)^{1/2}} = o_p(1). \tag{A.59}$$

We next consider $(\widehat{\Pi}_J-\Pi_J)\Xi_{J_{\Delta,\max}}h_0$.

First, Assumption 4.4 (ii) and the optimality of $\Xi_{J_{\Delta,\max}}h_0$ imply

$$\left\|\Xi_{J_{\Delta,\max}}h_0-h_0\right\|_X \lesssim J_{\Delta,\max}^{-s/d_x}. \tag{A.60}$$

Indeed,

$$\begin{aligned}
\sigma_\Delta\left\|\Xi_{J_{\Delta,\max}}h_0-\Pi_{J_{\Delta,\max}}h_0\right\|_X &\le \left\|\Pi_{K_{\Delta,\max}}T\left(\Xi_{J_{\Delta,\max}}h_0-\Pi_{J_{\Delta,\max}}h_0\right)\right\|_Z\\
&\le \left\|\Pi_{K_{\Delta,\max}}T\left(\Xi_{J_{\Delta,\max}}h_0-h_0\right)\right\|_Z + \left\|\Pi_{K_{\Delta,\max}}T\left(h_0-\Pi_{J_{\Delta,\max}}h_0\right)\right\|_Z\\
&\lesssim \sigma_\Delta J_{\Delta,\max}^{-s/d_x}.
\end{aligned}$$

Consequently, for $J\le J_{\Delta,\max}$,

$$\left\|\Xi_{J_{\Delta,\max}}h_0-\Pi_J\Xi_{J_{\Delta,\max}}h_0\right\|_X \le \left\|\Xi_{J_{\Delta,\max}}h_0-\Pi_Jh_0\right\|_X \lesssim J^{-s/d_x}. \tag{A.61}$$

The first-order conditions for the empirical least squares projection imply that

$$\left(\widehat{\Pi}_J-\Pi_J\right)\Xi_{J_{\Delta,\max}}h_0 = p^J(\cdot)'\widehat{G}_X^{-1}\frac{1}{n}\sum_{i=1}^n p^J(X_i)\left[\Xi_{J_{\Delta,\max}}h_0(X_i)-\Pi_J\Xi_{J_{\Delta,\max}}h_0(X_i)\right]. \tag{A.62}$$

On the event on which the empirical Gram matrices are uniformly well conditioned, (A.62) and (A.61) imply

$$E\left\|(\widehat{\Pi}_J-\Pi_J)\Xi_{J_{\Delta,\max}}h_0\right\|_X^2 \lesssim \frac{\zeta_{X,J}^2J^{-2s/d_x}}{n}. \tag{A.63}$$

The complement of this event has probability $o(1)$ and therefore does not affect the desired uniform probability bound.

Using (A.45),

$$\frac{E\left\|(\widehat{\Pi}_J-\Pi_J)\Xi_{J_{\Delta,\max}}h_0\right\|_X^2}{E\|\Psi_{\Delta,J}\|_X^2} \lesssim \frac{\zeta_{X,J}^2J^{-2s/d_x}}{\sum_{j=1}^J\nu_j^{-2}}. \tag{A.64}$$

To bound the right-hand side, use $\zeta_{X,J} \leq \zeta_J$ to obtain

$$\frac{\zeta_{X,J}^2 J^{-2s/d_x}}{\sum_{j=1}^J \nu_j^{-2}} \leq \left[\frac{\sigma_{JK}^{-1}\zeta_J J^{-s/d_x}\sqrt{\log J}}{\left(J^{-1}\sum_{j=1}^J \nu_j^{-2}\right)^{1/2}}\right]^2 \frac{\sigma_{JK}^2}{J\log J} \lesssim \frac{1}{J\log J},$$

where the last inequality follows from Assumption 4.4 (i) and the boundedness of σ_{JK}.

Since $\mathcal{J}_n$ is a geometric grid and $\underline{J} \to \infty$,

$$\sum_{\substack{J\in\mathcal{J}_n \\ J\leq J_{\Delta,\max}}} \frac{1}{J\log J} \lesssim \frac{1}{\underline{J}\log \underline{J}} = o(1).$$

Markov's inequality and a union bound therefore give

$$\max_{\substack{J\in\mathcal{J}_n \\ J\leq J_{\Delta,\max}}} \frac{\left\|(\widehat{\Pi}_J - \Pi_J)\Xi_{J_{\Delta,\max}} h_0\right\|_X}{\left(E\|\Psi_{\Delta,J}\|_X^2\right)^{1/2}} = o_p(1). \tag{A.65}$$

By (A.41) and (A.45),

$$\max_{\substack{J\in\mathcal{J}_n \\ J\leq J_{\Delta,\max}}} \frac{\left\|\Pi_J\left(\Xi_{J_{\Delta,\max}} h_0 - h_0\right)\right\|_X}{\left(E\|\Psi_{\Delta,J}\|_X^2\right)^{1/2}} \lesssim \max_{\substack{J\in\mathcal{J}_n \\ J\leq J_{\Delta,\max}}} \frac{\sqrt{n}\,\nu_{J_{\Delta,\max}+1} J_{\Delta,\max}^{-s/d_x}}{\nu_J\left(\sum_{j=1}^J \nu_j^{-2}\right)^{1/2}} = o(1)$$

by Assumption A.1.

Finally, the uniform NP-OLS remainder bound established in Step 1, together with (A.50), (A.56), (A.59), (A.65), and the preceding display, controls all six terms in (A.46). Therefore,

$$\max_{\substack{J\in\mathcal{J}_n \\ J\leq J_{\Delta,\max}}} \frac{\|r_{\Delta,J}\|_X}{\left(E\|\Psi_{\Delta,J}\|_X^2\right)^{1/2}} = o_p(1). \tag{A.66}$$

This completes the proof of the uniform asymptotic expansion of $\widehat{\Delta}_J - \Delta_J$ in Proposition 5.1.

A.4 Uniform Consistency of Variance Estimators

We now construct estimators of the asymptotic variance terms $V_{g,J}$ and $V_{\Delta,J}$ and establish their uniform consistency over the candidate sieve dimensions.

We first consider the sieve NP-OLS component. Recall from (A.21) that

$$V_{g,J} = E\left\|\Psi_{g,J}\right\|_X^2 = \frac{1}{n}E\left[(Y_i - \Pi_J g_0(X_i))^2\, p^J(X_i)' G_X^{-1} p^J(X_i)\right].$$

We estimate $V_{g,J}$ by

$$\widehat{V}_{g,J} = \frac{1}{n^2}\sum_{i=1}^{n} p^J(X_i)'\widehat{G}_X^{-1}p^J(X_i)\left(Y_i - \widehat{g}_J(X_i)\right)^2. \tag{A.67}$$

For the endogeneity component, define that

$$\xi_{\Delta,J}(\mathcal{D}_i) = p^J(\cdot)'G_X^{-1}p^J(X_i)\left(Y_i - \Pi_J g_0(X_i)\right) - \Pi_J\left[\xi_{\mathrm{IV},J_{\Delta,\max}}(\mathcal{D}_i)\right]. \tag{A.68}$$

By the definition of $\Psi_{\Delta,J}$,

$$\Psi_{\Delta,J} = \frac{1}{n}\sum_{i=1}^{n}\xi_{\Delta,J}(\mathcal{D}_i), \qquad E\left[\xi_{\Delta,J}(\mathcal{D}_i)\right] = 0.$$

Since the observations are independent,

$$V_{\Delta,J} = E\left\|\Psi_{\Delta,J}\right\|_X^2 = \frac{1}{n}E\left\|\xi_{\Delta,J}(\mathcal{D}_i)\right\|_X^2. \tag{A.69}$$

The corresponding plug-in estimate for $\xi_{\Delta,J}(\mathcal{D}_i)$ is

$$\begin{aligned}\widehat{\xi}_{\Delta,J}(\mathcal{D}_i) &= p^J(\cdot)'\widehat{G}_X^{-1}p^J(X_i)\left(Y_i - \widehat{g}_J(X_i)\right)\\ &\quad - \widehat{\Pi}_J\Big[p^{J_{\Delta,\max}}(\cdot)'\widehat{G}_{X,\Delta}^{-1/2}\widehat{A}_\Delta q^{K_{\Delta,\max}}(Z_i)\left(Y_i - \widetilde{h}_{\mathrm{IV},J_{\Delta,\max}}(X_i)\right)\Big].\end{aligned} \tag{A.70}$$

The resulting variance estimator for $V_{\Delta,J}$ is

$$\widehat{V}_{\Delta,J} = \frac{1}{n^2}\sum_{i=1}^{n}\left\|\widehat{\xi}_{\Delta,J}(\mathcal{D}_i)\right\|_{X,n}^2, \qquad J \le J_{\Delta,\max}. \tag{A.71}$$

Proposition A.1 (Uniform consistency of variance estimators)**.** *Suppose that Assumptions 2.1, 4.1 to 4.4 and 5.1 hold. Then, uniformly over $F_n \in \mathcal{F}(\rho_n)$,*

$$\max_{J\in\mathcal{J}_n}\left|\frac{\widehat{V}_{g,J}}{V_{g,J}} - 1\right| = o_p(1), \tag{A.72}$$

$$\max_{\substack{J\in\mathcal{J}_n\\ J\le J_{\Delta,\max}}}\left|\frac{\widehat{V}_{\Delta,J}}{V_{\Delta,J}} - 1\right| = o_p(1). \tag{A.73}$$

Proof of Proposition A.1. We first establish the uniform consistency of $\hat{V}_{g,J}$. By (A.21) and (A.22), $V_{g,J} \asymp J/n$ uniformly over $J \in \mathcal{J}_n$. It suffices to show that

$$\max_{J\in\mathcal{J}_n}\frac{1}{J}\left|\frac{1}{n}\sum_{i=1}^{n}p^J(X_i)'G_X^{-1}p^J(X_i)\left(Y_i - \widehat{g}_J(X_i)\right)^2 - E\left[p^J(X_i)'G_X^{-1}p^J(X_i)\left(Y_i - \Pi_J g_0(X_i)\right)^2\right]\right| = o_p(1).$$

We proceed in three steps. First, we show that

$$\max_{J\in\mathcal{J}_n}\frac{1}{J}\left|\frac{1}{n}\sum_{i=1}^{n}p^J(X_i)'G_X^{-1}p^J(X_i)\left(Y_i-\Pi_J g_0(X_i)\right)^2-E\left[p^J(X_i)'G_X^{-1}p^J(X_i)\left(Y_i-\Pi_J g_0(X_i)\right)^2\right]\right|=o_p(1). \tag{A.74}$$

Since $\underline{J}\to\infty$, $\max_{J\in\mathcal{J}_n}\|g_0-\Pi_J g_0\|_\infty=o(1)$. Moreover, $Y_i-\Pi_J g_0(X_i)=\varepsilon_i+g_0(X_i)-\Pi_J g_0(X_i)$. Using $|a+b|^4\le 8|a|^4+8|b|^4$ and Assumption 5.1 evaluated at $k=4$,

$$\begin{aligned}E\left[\left|Y_i-\Pi_J g_0(X_i)\right|^4\;\middle|\;X_i\right]&\le 8E\left[\left|\varepsilon_i\right|^4\middle|X_i\right]+8\left|g_0(X_i)-\Pi_J g_0(X_i)\right|^4\\&\le 96\bar{\sigma}_\varepsilon^2C_\varepsilon^2+8\left\|g_0-\Pi_J g_0\right\|_\infty^4\lesssim 1\end{aligned}$$

uniformly over $J\in\mathcal{J}_n$.

Since $p^J(X_i)'G_X^{-1}p^J(X_i)\le\zeta_{X,J}^2$ and $E\left[p^J(X_i)'G_X^{-1}p^J(X_i)\right]=J$, we obtain

$$E\left[\left\{p^J(X_i)'G_X^{-1}p^J(X_i)\right\}^2\left(Y_i-\Pi_J g_0(X_i)\right)^4\right]\le\zeta_{X,J}^2E\left[p^J(X_i)'G_X^{-1}p^J(X_i)\left(Y_i-\Pi_J g_0(X_i)\right)^4\right]\lesssim\zeta_{X,J}^2J. \tag{A.75}$$

Chebyshev's inequality and a union bound therefore give, for every $c>0$,

$$\begin{aligned}&P\left(\max_{J\in\mathcal{J}_n}\frac{1}{J}\left|\frac{1}{n}\sum_{i=1}^{n}p^J(X_i)'G_X^{-1}p^J(X_i)\left(Y_i-\Pi_J g_0(X_i)\right)^2-E\left[p^J(X_i)'G_X^{-1}p^J(X_i)\left(Y_i-\Pi_J g_0(X_i)\right)^2\right]\right|>c\right)\\&\lesssim\frac{1}{c^2}\sum_{J\in\mathcal{J}_n}\frac{\zeta_{X,J}^2}{nJ}\lesssim\frac{1}{c^2}\begin{cases}|\mathcal{J}_n|/n, & \zeta_{X,J}=O(\sqrt{J})\\ J_{\max}/n, & \zeta_{X,J}=O(J)\end{cases}=o(1).\end{aligned}$$

For the second case, the last inequality uses the geometric spacing of $\mathcal{J}_n$, which implies that $\sum_{J\in\mathcal{J}_n}J\lesssim J_{\max}$. This proves (A.74).

Second, we show that replacing $\Pi_J g_0$ with $\widehat{g}_J$ has an asymptotically negligible effect on the variance estimator. By (A.23), (A.25), and the rate condition used to establish (A.23),

$$\begin{aligned}&\max_{J\in\mathcal{J}_n}\frac{1}{J}\frac{1}{n}\sum_{i=1}^{n}p^J(X_i)'G_X^{-1}p^J(X_i)\left(\widehat{g}_J(X_i)-\Pi_J g_0(X_i)\right)^2\\&\le\left(\max_{J\in\mathcal{J}_n}\frac{\zeta_{X,J}^2}{n}\right)\left(\max_{J\in\mathcal{J}_n}\frac{\|\widehat{g}_J-\Pi_J g_0\|_{X,n}^2}{J/n}\right)=o(1)\,O_p(1)=o_p(1).\end{aligned} \tag{A.76}$$

Applying the inequality $\left|a^2-b^2\right|\le 2|b||a-b|+|a-b|^2$, followed by the Cauchy–Schwarz

inequality, (A.74), and (A.76), gives

$$\max_{J\in\mathcal{J}_n}\frac{1}{J}\left|\frac{1}{n}\sum_{i=1}^{n}p^J(X_i)'G_X^{-1}p^J(X_i)\left[(Y_i-\widehat{g}_J(X_i))^2-(Y_i-\Pi_J g_0(X_i))^2\right]\right|$$

$$\leq 2\left\{\max_{J\in\mathcal{J}_n}\frac{1}{J}\frac{1}{n}\sum_{i=1}^{n}p^J(X_i)'G_X^{-1}p^J(X_i)\left(Y_i-\Pi_J g_0(X_i)\right)^2\right\}^{1/2}$$

$$\times\left\{\max_{J\in\mathcal{J}_n}\frac{1}{J}\frac{1}{n}\sum_{i=1}^{n}p^J(X_i)'G_X^{-1}p^J(X_i)\left(\widehat{g}_J(X_i)-\Pi_J g_0(X_i)\right)^2\right\}^{1/2}$$

$$+\max_{J\in\mathcal{J}_n}\frac{1}{J}\frac{1}{n}\sum_{i=1}^{n}p^J(X_i)'G_X^{-1}p^J(X_i)\left(\widehat{g}_J(X_i)-\Pi_J g_0(X_i)\right)^2.$$

The first factor is $O_p(1)$ by (A.74), (A.21), and (A.22), while the remaining terms are $o_p(1)$ by (A.76). Therefore,

$$\max_{J\in\mathcal{J}_n}\frac{1}{J}\left|\frac{1}{n}\sum_{i=1}^{n}p^J(X_i)'G_X^{-1}p^J(X_i)\left[(Y_i-\widehat{g}_J(X_i))^2-(Y_i-\Pi_J g_0(X_i))^2\right]\right|=o_p(1). \tag{A.77}$$

Finally, we show that replacing G_X^{-1} with $\widehat{G}_X^{-1}$ has an asymptotically negligible effect on the variance estimator. It follows that

$$\max_{J\in\mathcal{J}_n}\frac{1}{J}\left|\frac{1}{n}\sum_{i=1}^{n}p^J(X_i)'\left(\widehat{G}_X^{-1}-G_X^{-1}\right)p^J(X_i)\left(Y_i-\widehat{g}_J(X_i)\right)^2\right|$$

$$\leq\left\{\max_{J\in\mathcal{J}_n}\left\|G_X^{1/2}\widehat{G}_X^{-1}G_X^{1/2}-I_J\right\|\right\}\times\left\{\max_{J\in\mathcal{J}_n}\frac{1}{J}\frac{1}{n}\sum_{i=1}^{n}p^J(X_i)'G_X^{-1}p^J(X_i)\left(Y_i-\widehat{g}_J(X_i)\right)^2\right\}=o_p(1),$$

where the first item is $o_p(1)$ by (A.23), and the second item is $O_p(1)$ by (A.74) and (A.77), together with (A.21) and (A.22). Combining the three steps,

$$\max_{J\in\mathcal{J}_n}\frac{1}{J}\left|\frac{1}{n}\sum_{i=1}^{n}p^J(X_i)'\widehat{G}_X^{-1}p^J(X_i)\left(Y_i-\widehat{g}_J(X_i)\right)^2-E\left[p^J(X_i)'G_X^{-1}p^J(X_i)\left(Y_i-\Pi_J g_0(X_i)\right)^2\right]\right|=o_p(1).$$

Combining this result with (A.74), (A.77), (A.21), and (A.22) gives

$$\max_{J\in\mathcal{J}_n}\left|\frac{\widehat{V}_{g,J}}{V_{g,J}}-1\right|=o_p(1).$$

We then establish the uniform consistency of $\hat{V}_{\Delta,J}$. By (A.45) and (A.69),

$$E\left\|\xi_{\Delta,J}(\mathcal{D}_i)\right\|_X^2 \asymp \sum_{j=1}^{J} \nu_j^{-2}. \tag{A.78}$$

We first establish a uniform law of large numbers for $\|\xi_{\Delta,J}(\mathcal{D}_i)\|_X^2$. The Loewner-order argument used to derive (A.43) also gives

$$\left\|\Pi_J\left[p^{J_{\Delta,\max}}(\cdot)' G_{X,\Delta}^{-1/2} A_\Delta q^{K_{\Delta,\max}}(z)\right]\right\|_X^2 \lesssim \nu_J^{-2}\left\|G_{Z,\Delta}^{-1/2} q^{K_{\Delta,\max}}(z)\right\|^2. \tag{A.79}$$

In particular,

$$\left\|\Pi_J\left[p^{J_{\Delta,\max}}(\cdot)' G_{X,\Delta}^{-1/2} A_\Delta q^{K_{\Delta,\max}}(z)\right]\right\|_X^2 \lesssim \nu_J^{-2}\zeta_\Delta^2. \tag{A.80}$$

By (A.75),

$$E\left\|p^J(\cdot)' G_X^{-1} p^J(X_i)\left(Y_i - \Pi_J g_0(X_i)\right)\right\|_X^4 = E\left[\left\{p^J(X_i)' G_X^{-1} p^J(X_i)\right\}^2 \left(Y_i - \Pi_J g_0(X_i)\right)^4\right] \lesssim \zeta_{X,J}^2 J.$$

Similarly, combining (A.43) and (A.79) with the conditional fourth-moment bound in (A.94) yields

$$E\left\|\Pi_J\left[\xi_{\mathrm{IV},J_{\Delta,\max}}(\mathcal{D}_i)\right]\right\|_X^4 \lesssim \zeta_\Delta^2 \nu_J^{-2} \sum_{j=1}^{J} \nu_j^{-2}.$$

Since $\|a-b\|^4 \le 8\|a\|^4 + 8\|b\|^4$,

$$E\left\|\xi_{\Delta,J}(\mathcal{D}_i)\right\|_X^4 \lesssim \zeta_{X,J}^2 J + \zeta_\Delta^2 \nu_J^{-2} \sum_{j=1}^{J} \nu_j^{-2}. \tag{A.81}$$

It follows from Chebyshev's inequality, a union bound, (A.78), and (A.81) that, for every $c > 0$,

$$P\left(\max_{\substack{J\in\mathcal{J}_n \\ J\le J_{\Delta,\max}}} \left|\frac{n^{-1}\sum_{i=1}^n \|\xi_{\Delta,J}(\mathcal{D}_i)\|_X^2}{E\|\xi_{\Delta,J}(\mathcal{D}_i)\|_X^2} - 1\right| > c\right) \lesssim \frac{1}{c^2} \sum_{\substack{J\in\mathcal{J}_n \\ J\le J_{\Delta,\max}}} \left\{\frac{\zeta_{X,J}^2 J}{n\left(\sum_{j=1}^J \nu_j^{-2}\right)^2} + \frac{\zeta_\Delta^2}{n}\frac{\nu_J^{-2}}{\sum_{j=1}^J \nu_j^{-2}}\right\}.$$

Since $\sum_{j=1}^J \nu_j^{-2} \gtrsim J$, the first sum is bounded by $\sum_{J\in\mathcal{J}_n} \frac{\zeta_{X,J}^2}{nJ} = o(1)$ as in Step 1 of the proof for $\widehat{V}_{g,J}$. For the second sum, $\nu_J^{-2} \le \sum_{j=1}^J \nu_j^{-2}$ and the geometric grid contains $O(\log J_{\Delta,\max})$ indices below $J_{\Delta,\max}$. Therefore

$$\frac{\zeta_\Delta^2}{n} \sum_{\substack{J\in\mathcal{J}_n \\ J\le J_{\Delta,\max}}} \frac{\nu_J^{-2}}{\sum_{j=1}^J \nu_j^{-2}} \lesssim \frac{\zeta_\Delta^2 \log J_{\Delta,\max}}{n} = o(1).$$

Hence

$$\max_{\substack{J\in\mathcal{J}_n\\ J\le J_{\Delta,\max}}}\left|\frac{n^{-1}\sum_{i=1}^n\|\xi_{\Delta,J}(\mathcal{D}_i)\|_X^2}{E\|\xi_{\Delta,J}(\mathcal{D}_i)\|_X^2}-1\right|=o_p(1). \tag{A.82}$$

We next show that replacing $\xi_{\Delta,J}(\mathcal{D}_i)$ by $\widehat{\xi}_{\Delta,J}(\mathcal{D}_i)$ is uniformly negligible. From (A.68) and (A.70), $\widehat{\xi}_{\Delta,J}(\mathcal{D}_i)-\xi_{\Delta,J}(\mathcal{D}_i)$ is the sum of the following four terms:

$$p^J(\cdot)'\left(\widehat{G}_X^{-1}-G_X^{-1}\right)p^J(X_i)\left(Y_i-\Pi_J g_0(X_i)\right)$$

$$-p^J(\cdot)'\widehat{G}_X^{-1}p^J(X_i)\left(\widehat{g}_J(X_i)-\Pi_J g_0(X_i)\right)$$

$$-\left\{\widehat{\Pi}_J\left[p^{J_{\Delta,\max}}(\cdot)'\widehat{G}_{X,\Delta}^{-1/2}\widehat{A}_\Delta q^{K_{\Delta,\max}}(Z_i)\right]-\Pi_J\left[p^{J_{\Delta,\max}}(\cdot)'G_{X,\Delta}^{-1/2}A_\Delta q^{K_{\Delta,\max}}(Z_i)\right]\right\}\times\left(Y_i-\Xi_{J_{\Delta,\max}}h_0(X_i)\right)$$

$$+\widehat{\Pi}_J\left[p^{J_{\Delta,\max}}(\cdot)'\widehat{G}_{X,\Delta}^{-1/2}\widehat{A}_\Delta q^{K_{\Delta,\max}}(Z_i)\right]\times\left(\widetilde{h}_{\mathrm{IV},J_{\Delta,\max}}(X_i)-\Xi_{J_{\Delta,\max}}h_0(X_i)\right).$$

We control these four terms in order.

For the first term, (A.24) and (A.74) imply

$$\max_{\substack{J\in\mathcal{J}_n\\ J\le J_{\Delta,\max}}}\frac{n^{-1}\sum_{i=1}^n\left\|p^J(\cdot)'\left(\widehat{G}_X^{-1}-G_X^{-1}\right)p^J(X_i)\left(Y_i-\Pi_J g_0(X_i)\right)\right\|_X^2}{\sum_{j=1}^J\nu_j^{-2}}\lesssim_p\max_{J\in\mathcal{J}_n}\left\{\zeta_{X,J}^2\frac{\log J}{n}\frac{J}{\sum_{j=1}^J\nu_j^{-2}}\right\}=o_p(1).$$

For the second term, uniform conditioning of $\widehat{G}_X$ and (A.76) give

$$\max_{\substack{J\in\mathcal{J}_n\\ J\le J_{\Delta,\max}}}\frac{n^{-1}\sum_{i=1}^n\left\|p^J(\cdot)'\widehat{G}_X^{-1}p^J(X_i)\left(\widehat{g}_J(X_i)-\Pi_J g_0(X_i)\right)\right\|_X^2}{\sum_{j=1}^J\nu_j^{-2}}\lesssim_p\max_{J\in\mathcal{J}_n}\frac{\zeta_{X,J}^2J/n}{\sum_{j=1}^J\nu_j^{-2}}\le\max_{J\in\mathcal{J}_n}\frac{\zeta_{X,J}^2}{n}=o_p(1).$$

We next consider the third term. Decompose that

$$\widehat{\Pi}_J\left[p^{J_{\Delta,\max}}(\cdot)'\widehat{G}_{X,\Delta}^{-1/2}\widehat{A}_\Delta q^{K_{\Delta,\max}}(z)\right]-\Pi_J\left[p^{J_{\Delta,\max}}(\cdot)'G_{X,\Delta}^{-1/2}A_\Delta q^{K_{\Delta,\max}}(z)\right]$$
$$=\Pi_J\left[p^{J_{\Delta,\max}}(\cdot)'\left(\widehat{G}_{X,\Delta}^{-1/2}\widehat{A}_\Delta-G_{X,\Delta}^{-1/2}A_\Delta\right)q^{K_{\Delta,\max}}(z)\right]+\left(\widehat{\Pi}_J-\Pi_J\right)\left[p^{J_{\Delta,\max}}(\cdot)'\widehat{G}_{X,\Delta}^{-1/2}\widehat{A}_\Delta q^{K_{\Delta,\max}}(z)\right].$$

Equation (A.31), the comparison used in the proof of (A.43), and (A.51) yield

$$\left\|\widehat{\Pi}_J\left[p^{J_{\Delta,\max}}(\cdot)'\widehat{G}_{X,\Delta}^{-1/2}\widehat{A}_\Delta q^{K_{\Delta,\max}}(z)\right]-\Pi_J\left[p^{J_{\Delta,\max}}(\cdot)'G_{X,\Delta}^{-1/2}A_\Delta q^{K_{\Delta,\max}}(z)\right]\right\|_X$$
$$\lesssim_p\nu_J^{-1}\sigma_\Delta^{-1}\zeta_\Delta\sqrt{\frac{\log J_{\Delta,\max}}{n}}\left\|G_{Z,\Delta}^{-1/2}q^{K_{\Delta,\max}}(z)\right\|. \tag{A.83}$$

The bound is uniform over $J \in \mathcal{J}_n$, $J \le J_{\Delta,\max}$.

Moreover, by the conditional second-moment bound,

$$\frac{1}{n}\sum_{i=1}^{n}\left\|G_{Z,\Delta}^{-1/2}q^{K_{\Delta,\max}}(Z_i)\right\|^2\left(Y_i-\Xi_{J_{\Delta,\max}}h_0(X_i)\right)^2 = O_p(K_{\Delta,\max}).$$

It follows from (A.83) that the normalized empirical second moment of the third term is bounded by

$$O_p\left[\max_{\substack{J\in\mathcal{J}_n\\ J\le J_{\Delta,\max}}}\frac{\nu_J^{-2}}{\sum_{j=1}^{J}\nu_j^{-2}}\sigma_\Delta^{-2}\zeta_\Delta^2\frac{K_{\Delta,\max}\log J_{\Delta,\max}}{n}\right]. \tag{A.84}$$

By the first condition in (A.16),

$$\sigma_\Delta^{-2}\zeta_\Delta^2\frac{K_{\Delta,\max}\log J_{\Delta,\max}}{n} = O\left(\sigma_\Delta^2 J_{\Delta,\max}^{-1}\sum_{j=1}^{J_{\Delta,\max}}\nu_j^{-2}\right).$$

In the mildly ill-posed case,

$$\sigma_\Delta^2 J_{\Delta,\max}^{-1}\sum_{j=1}^{J_{\Delta,\max}}\nu_j^{-2}\asymp 1, \qquad \frac{\nu_J^{-2}}{\sum_{j=1}^{J}\nu_j^{-2}}\asymp\frac{1}{J}.$$

Since $J\ge \underline{J}\to\infty$, (A.84) is $o_p(1)$. In the severely ill-posed case,

$$\sigma_\Delta^2 J_{\Delta,\max}^{-1}\sum_{j=1}^{J_{\Delta,\max}}\nu_j^{-2} = o(1)$$

by (A.49), while $\nu_J^{-2}/\sum_{j=1}^{J}\nu_j^{-2}\le 1$. Hence (A.84) is again $o_p(1)$.

It remains to control the fourth term. The exact sieve 2SLS formula, uniform conditioning of the pilot matrices, and (A.35) give

$$\left\|\widetilde{h}_{\mathrm{IV},J_{\Delta,\max}}-\Xi_{J_{\Delta,\max}}h_0\right\|_{X,n}^2 = O_p\left(\sigma_\Delta^{-2}\frac{K_{\Delta,\max}}{n}+J_{\Delta,\max}^{-2s/d_x}\right). \tag{A.85}$$

Furthermore, the first condition in (A.16), together with $\sigma_\Delta\asymp\nu_{J_{\Delta,\max}}$ and the monotonicity of $\{\nu_j\}$, implies that

$$\sigma_\Delta^{-1}\zeta_\Delta\sqrt{\frac{\log J_{\Delta,\max}}{n}} = o(1).$$

Consequently, (A.79) and (A.83) imply that

$$\left\|\widehat{\Pi}_J\left[p^{J_{\Delta,\max}}(\cdot)'\widehat{G}_{X,\Delta}^{-1/2}\widehat{A}_\Delta q^{K_{\Delta,\max}}(z)\right]\right\|_X \lesssim_p \nu_J^{-1}\left\|G_{Z,\Delta}^{-1/2}q^{K_{\Delta,\max}}(z)\right\|$$

uniformly over the relevant J and z. Therefore,

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\widehat{\Pi}_J\left[p^{J_{\Delta,\max}}(\cdot)'\widehat{G}_{X,\Delta}^{-1/2}\widehat{A}_\Delta q^{K_{\Delta,\max}}(Z_i)\right]\times\left(\widetilde{h}_{\mathrm{IV},J_{\Delta,\max}}(X_i)-\Xi_{J_{\Delta,\max}}h_0(X_i)\right)\right\|_X^2$$

$$\lesssim_p \nu_J^{-2}\zeta_\Delta^2\left\|\widetilde{h}_{\mathrm{IV},J_{\Delta,\max}}-\Xi_{J_{\Delta,\max}}h_0\right\|_{X,n}^2.$$

Using (A.85), the preceding quantity divided by $\sum_{j=1}^{J}\nu_j^{-2}$ is bounded by

$$O_p\left[\frac{\nu_J^{-2}}{\sum_{j=1}^{J}\nu_j^{-2}}\zeta_\Delta^2\left(\sigma_\Delta^{-2}\frac{K_{\Delta,\max}}{n}+J_{\Delta,\max}^{-2s/d_x}\right)\right]. \tag{A.86}$$

The conditions in (A.16) give that

$$\zeta_\Delta^2\sigma_\Delta^{-2}\frac{K_{\Delta,\max}}{n}=O\left(\frac{\sigma_\Delta^2 J_{\Delta,\max}^{-1}\sum_{j=1}^{J_{\Delta,\max}}\nu_j^{-2}}{\log J_{\Delta,\max}}\right),\quad \zeta_\Delta^2 J_{\Delta,\max}^{-2s/d_x}=O\left(\frac{\sigma_\Delta^2 J_{\Delta,\max}^{-1}\sum_{j=1}^{J_{\Delta,\max}}\nu_j^{-2}}{\log J_{\Delta,\max}}\right).$$

The same calculations used for (A.84) therefore show that (A.86) is $o_p(1)$ uniformly.

Combining the four preceding bounds and using $\|a_1+\cdots+a_4\|^2\le 4\sum_{\ell=1}^{4}\|a_\ell\|^2$ gives

$$\max_{\substack{J\in\mathcal{J}_n\\ J\le J_{\Delta,\max}}}\frac{n^{-1}\sum_{i=1}^{n}\left\|\widehat{\xi}_{\Delta,J}(\mathcal{D}_i)-\xi_{\Delta,J}(\mathcal{D}_i)\right\|_X^2}{E\left\|\xi_{\Delta,J}(\mathcal{D}_i)\right\|_X^2}=o_p(1). \tag{A.87}$$

For arbitrary elements a and b of a Hilbert space, $\left|\|a\|^2-\|b\|^2\right|\le 2\|b\|\|a-b\|+\|a-b\|^2$. Applying this inequality observation by observation and then the Cauchy–Schwarz inequality gives

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left\|\widehat{\xi}_{\Delta,J}(\mathcal{D}_i)\right\|_X^2-\frac{1}{n}\sum_{i=1}^{n}\left\|\xi_{\Delta,J}(\mathcal{D}_i)\right\|_X^2\right|$$

$$\le 2\left[\frac{1}{n}\sum_{i=1}^{n}\left\|\xi_{\Delta,J}(\mathcal{D}_i)\right\|_X^2\right]^{1/2}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|\widehat{\xi}_{\Delta,J}(\mathcal{D}_i)-\xi_{\Delta,J}(\mathcal{D}_i)\right\|_X^2\right]^{1/2}+\frac{1}{n}\sum_{i=1}^{n}\left\|\widehat{\xi}_{\Delta,J}(\mathcal{D}_i)-\xi_{\Delta,J}(\mathcal{D}_i)\right\|_X^2.$$

Combining (A.82) and (A.87), we obtain

$$\max_{\substack{J\in\mathcal{J}_n\\ J\le J_{\Delta,\max}}}\left|\frac{n^{-1}\sum_{i=1}^{n}\left\|\widehat{\xi}_{\Delta,J}(\mathcal{D}_i)\right\|_X^2}{E\left\|\xi_{\Delta,J}(\mathcal{D}_i)\right\|_X^2}-1\right|=o_p(1). \tag{A.88}$$

Finally, $\widehat{\xi}_{\Delta,J}(\mathcal{D}_i)\in\mathcal{H}_J\subseteq\mathcal{H}_{J_{\Delta,\max}}$. The Gram-matrix concentration used to establish (A.51)

gives

$$\sup_{f\in\mathcal{H}_{J_{\Delta,\max}},\, f\neq 0}\left|\frac{\|f\|_{X,n}^2}{\|f\|_X^2}-1\right|=o_p(1).$$

This inequality is pathwise and therefore also applies to the random functions $\widehat{\xi}_{\Delta,J}(\mathcal{D}_i)$. Hence

$$\frac{1}{n}\sum_{i=1}^n\left\|\widehat{\xi}_{\Delta,J}(\mathcal{D}_i)\right\|_{X,n}^2=\{1+o_p(1)\}\,\frac{1}{n}\sum_{i=1}^n\left\|\widehat{\xi}_{\Delta,J}(\mathcal{D}_i)\right\|_X^2$$

uniformly over $J\in\mathcal{J}_n$, $J\le J_{\Delta,\max}$. Together with (A.88), (A.69), and (A.71), this yields

$$\max_{\substack{J\in\mathcal{J}_n\\ J\le J_{\Delta,\max}}}\left|\frac{\widehat{V}_{\Delta,J}}{V_{\Delta,J}}-1\right|=o_p(1).$$

■

A.5 Concentration Arguments

This subsection establishes concentration inequalities for the estimated endogeneity bias $\hat{\Delta}_J$ and its increments across sieve dimensions. These inequalities are needed to control the stochastic error arising from data-driven selection of the shrinkage parameter and sieve dimension.

The main tool is the following Hilbert-space Bernstein inequality, which is used to control the leading stochastic terms in the estimation errors.

Lemma A.1 (Hilbert-space Bernstein inequality, Theorem 3.3, Pinelis (1994, 1999))**.** *Let $\{\mathscr{F}_i\}_{i=0}^n$ be a filtration, and let $\{W_i\}_{i=1}^n$ be an $\mathbb{H}$-valued martingale-difference sequence adapted to this filtration, where $\mathbb{H}$ is a real separable Hilbert space. Thus, for every $i=1,\dots,n$, the random element W_i is $\mathscr{F}_i$-measurable and $E\left[W_i\mid\mathscr{F}_{i-1}\right]=0$, a.s. Suppose that there exists $B>0$ and $\Gamma>0$ such that for every integer $k\ge 2$,*

$$\left\|\sum_{i=1}^n E\left[\,\left\|W_i\right\|_{\mathbb{H}}^k\;\middle|\;\mathscr{F}_{i-1}\right]\right\|_\infty\le\frac{k!}{2}\Gamma^{k-2}B^2.$$

Then for every $r\ge 0$, we have that

$$P\left(\max_{0\le\ell\le n}\left\|\sum_{i=1}^\ell W_i\right\|_{\mathbb{H}}\ge r\right)\le 2\exp\left\{-\frac{r^2}{B^2+\Gamma r+B\sqrt{B^2+2\Gamma r}}\right\}.\tag{A.89}$$

Assumption A.2. *In addition to Assumption A.1, we also assume that*

$$\sigma_\Delta^{-1}\zeta_\Delta\sqrt{\frac{\log n}{n}}=o(1),\qquad\zeta_\Delta\sqrt{\frac{\log n}{n}}\left(\frac{\sum_{j=1}^{J_{\Delta,\max}}\nu_j^{-2}}{\sum_{j=1}^{J}\nu_j^{-2}}\right)^{1/2}=o(1),$$

and

$$\max_{\substack{J\in\mathcal{J}_n\\ J\le J_{\Delta,\max}}} \frac{\nu_J^{-1}}{\left(\sum_{j=1}^J \nu_j^{-2}\right)^{1/2}}\left[\sigma_\Delta^{-1}\zeta_\Delta\sqrt{\frac{K_{\Delta,\max}\log J_{\Delta,\max}}{n}}+\zeta_\Delta J_{\Delta,\max}^{-s/d_x}\sqrt{\log J_{\Delta,\max}}\right]=o(1).$$

Proposition A.2. *Suppose that Assumptions A.1, A.2, 2.1, 4.1 to 4.4 and 5.1 hold. In addition, suppose that (B.11) or (B.12) holds. Let $\eta>0$ be fixed. For each $J\in\mathcal{J}_n$ with $J\le J_{\Delta,\max}$, define*

$$t_J=\log\left(e+3\frac{V_{\Delta,J}}{V_{g,J}}\right),\quad \kappa_{n,J}=\left(2+\frac{\eta}{4}\right)\left\{t_J+\log(2|\mathcal{J}_n|)\right\}.$$

Then, for all sufficiently large n, uniformly over $F_n\in\mathcal{F}(\rho_n)$ and $J\in\mathcal{J}_n$ with $J\le J_{\Delta,\max}$,

$$P\left(\left\|\widehat{\Delta}_J-\Delta_J\right\|_{X,n}^2>\kappa_{n,J}V_{\Delta,J}\right)\le\frac{\exp(-t_J)}{|\mathcal{J}_n|},\tag{A.90}$$

and

$$E\left[\left\|\widehat{\Delta}_J-\Delta_J\right\|_{X,n}^2-\kappa_{n,J}V_{\Delta,J}\right]_+\le\frac{3\exp(-t_J)}{|\mathcal{J}_n|}V_{\Delta,J}\le\frac{V_{g,J}}{|\mathcal{J}_n|}.\tag{A.91}$$

Proof of Proposition A.2. Recall that

$$\widehat{\Delta}_J-\Delta_J=\Psi_{\Delta,J}+r_{\Delta,J},\qquad \Psi_{\Delta,J}=\frac{1}{n}\sum_{i=1}^n\xi_{\Delta,J}(\mathcal{D}_i).\tag{A.92}$$

where $\xi_{\Delta,J}(\mathcal{D}_i)$ is defined in (A.68) and satisfies that $E[\xi_{\Delta,J}(\mathcal{D}_i)]=0$.

Throughout the proof, we write $\tilde{\kappa}_{n,J}:=t_J+\log(2|\mathcal{J}_n|)$ so that $\kappa_{n,J}=(2+\eta/4)\tilde{\kappa}_{n,J}$.

Step 1. Concentration of the leading term

We first establish a Bernstein moment bound for $\xi_{\Delta,J}(\mathcal{D}_i)$. The smoothness and pilot conditions imply that the approximation terms entering $Y_i-\Pi_J g_0(X_i)$ and $Y_i-\Xi_{J_{\Delta,\max}}h_0(X_i)$ are uniformly bounded over the relevant sieve dimensions. Hence Assumption 5.1 implies that there are finite constants C_1,C_2, independent of n, J, and F_n, such that, for every integer $k\ge 2$,

$$E\left[\left|Y_i-\Pi_J g_0(X_i)\right|^k\middle|X_i\right]\le\frac{k!}{2}C_1C_2^{k-2},\tag{A.93}$$

$$E\left[\left|Y_i-\Xi_{J_{\Delta,\max}}h_0(X_i)\right|^k\middle|X_i,Z_i\right]\le\frac{k!}{2}C_1C_2^{k-2}.\tag{A.94}$$

For the NP-OLS component, the definition of G_X gives

$$\left\|p^J(\cdot)'G_X^{-1}p^J(X_i)\right\|_X=\left\|G_X^{-1/2}p^J(X_i)\right\|\le\zeta_{X,J}.\tag{A.95}$$

For the NPIV component, by (A.80), we have that

$$\left\|\Pi_J\left[p^{J_{\Delta,\max}}(\cdot)'G_{X,\Delta}^{-1/2}A_\Delta q^{K_{\Delta,\max}}(Z_i)\right]\right\|_X \lesssim \nu_J^{-1}\zeta_\Delta. \tag{A.96}$$

For $k \geq 3$, by the triangle inequality and $(a+b)^k \leq 2^{k-1}(a^k+b^k)$, we have that

$$\begin{aligned}
E\left\|\xi_{\Delta,J}(\mathcal{D}_i)\right\|_X^k &\leq 2^{k-1}E\left[\left\|p^J(\cdot)'G_X^{-1}p^J(X_i)\right\|_X^k E\left[\left|Y_i-\Pi_J g_0(X_i)\right|^k \mid X_i\right]\right] \\
&\quad + 2^{k-1}E\left[\left\|\Pi_J\left[p^{J_{\Delta,\max}}(\cdot)'G_{X,\Delta}^{-1/2}A_\Delta q^{K_{\Delta,\max}}(Z_i)\right]\right\|_X^k E\left[\left|Y_i-\Xi_{J_{\Delta,\max}}h_0(X_i)\right|^k \mid X_i,Z_i\right]\right] \\
&\overset{(a)}{\leq} 2^{k-1}\frac{k!}{2}C_1C_2^{k-2}\left\{E\left\|p^J(\cdot)'G_X^{-1}p^J(X_i)\right\|_X^k + E\left\|\Pi_J\left[p^{J_{\Delta,\max}}(\cdot)'G_{X,\Delta}^{-1/2}A_\Delta q^{K_{\Delta,\max}}(Z_i)\right]\right\|_X^k\right\} \\
&\overset{(b)}{\leq} \frac{k!}{2}\left(C\zeta_\Delta\nu_J^{-1}\right)^{k-2}C\left\{E\left\|p^J(\cdot)'G_X^{-1}p^J(X_i)\right\|_X^2 + E\left\|\Pi_J\left[p^{J_{\Delta,\max}}(\cdot)'G_{X,\Delta}^{-1/2}A_\Delta q^{K_{\Delta,\max}}(Z_i)\right]\right\|_X^2\right\} \\
&\overset{(c)}{\lesssim} \frac{k!}{2}\left(C\zeta_\Delta\nu_J^{-1}\right)^{k-2}C\left[J+\sum_{j=1}^J\nu_j^{-2}\right] \lesssim \frac{k!}{2}\left(C\zeta_\Delta\nu_J^{-1}\right)^{k-2}\sum_{j=1}^J\nu_j^{-2}
\end{aligned}$$

where (a) is obtained from (A.93) and (A.94); (b) follows from (A.95) and (A.96), (c) follows from $E\|p^J(\cdot)'G_X^{-1}p^J(X_i)\|_X^2 = J$ and (A.43). The last inequality uses $J \lesssim \sum_{j=1}^J \nu_j^{-2}$.

Moreover, the variance calculation already established in (A.45) gives

$$E\left\|\xi_{\Delta,J}(\mathcal{D}_i)\right\|_X^2 = nV_{\Delta,J} \asymp \sum_{j=1}^J \nu_j^{-2}. \tag{A.97}$$

Combining the last two displays, and enlarging the fixed constant C if necessary, gives, for all integers $k \geq 2$,

$$E\left\|\xi_{\Delta,J}(\mathcal{D}_i)\right\|_X^k \leq \frac{k!}{2}\mathtt{b}_{\Delta,J}^{k-2}E\left\|\xi_{\Delta,J}(\mathcal{D}_i)\right\|_X^2, \qquad \text{where } \mathtt{b}_{\Delta,J} := C\zeta_\Delta\nu_J^{-1}. \tag{A.98}$$

The case $k=2$ holds with equality.

Apply Lemma A.1 with $W_i = n^{-1}\xi_{\Delta,J}(\mathcal{D}_i)$ and $\mathbb{H} = \mathcal{H}_J$. By independence and centering, $\{W_i\}_{i=1}^n$ is a martingale-difference sequence with respect to the natural filtration. Then by (A.98),

$$\sum_{i=1}^n E\left\|W_i\right\|_X^k \leq \frac{k!}{2}\left(\frac{\mathtt{b}_{\Delta,J}}{n}\right)^{k-2}V_{\Delta,J}.$$

Substituting $B^2 = V_{\Delta,J}$ and $\Gamma = \mathtt{b}_{\Delta,J}/n$ into (A.89), and taking $r = \sqrt{2V_{\Delta,J}t} + (\mathtt{b}_{\Delta,J}/n)t$, we obtain, for $t \geq 0$,

$$P\left(\left\| \Psi_{\Delta,J} \right\|_X > \sqrt{2V_{\Delta,J}t} + \frac{\mathtt{b}_{\Delta,J}}{n} t \right) \leq 2\exp(-t). \tag{A.99}$$

We next show that the linear term in the Bernstein bound is negligible at the concentration level $\tilde{\kappa}_{n,J}$. By the definition $\tilde{\kappa}_{n,J} = t_J + \log(2|\mathcal{J}_n|)$ and the variance orders $V_{\Delta,J} \asymp n^{-1}\sum_{j=1}^{J} \nu_j^{-2}$ and $V_{g,J} \asymp J/n$, we have

$$\max_{\substack{J \in \mathcal{J}_n \\ J \leq J_{\Delta,\max}}} \tilde{\kappa}_{n,J} = O(\log n). \tag{A.100}$$

By (A.97) and (A.98), together with the monotonicity of ν_j, we have

$$\max_{\substack{J \in \mathcal{J}_n \\ J \leq J_{\Delta,\max}}} \frac{\mathtt{b}_{\Delta,J}^2 \tilde{\kappa}_{n,J}}{n^2 V_{\Delta,J}} \lesssim \frac{\zeta_\Delta^2 \nu_J^{-2} \log n}{n\sum_{j=1}^{J} \nu_j^{-2}} \leq \frac{\zeta_\Delta^2 \log n}{n} = o(1). \tag{A.101}$$

where the last bound follows from Assumption A.2.

For $t \geq 0$, define

$$u_{\Delta,J}(t) = \left[\sqrt{2V_{\Delta,J}t} + \frac{\mathtt{b}_{\Delta,J}}{n} t \right]^2. \tag{A.102}$$

It follows from (A.101) that

$$\sup_{\substack{J \in \mathcal{J}_n \\ J \leq J_{\Delta,\max}}} \sup_{1 \leq t \leq C\log n} \frac{u_{\Delta,J}(t)}{tV_{\Delta,J}} = 2 + o(1). \tag{A.103}$$

Finally, we derive an integrated-tail bound for the squared norm of the leading term. Since $u_{\Delta,J}(t)$ is increasing, the tail-integration identity and (A.99) give, for $t \geq 1$,

$$\begin{aligned}
&E\left[\left\| \Psi_{\Delta,J} \right\|_X^2 - u_{\Delta,J}(t) \right]_+ \leq 2\int_t^\infty \exp(-z) u'_{\Delta,J}(z)\,\mathrm{d}z \\
&= 2\int_t^\infty \exp(-z)\left[2V_{\Delta,J} + 3\sqrt{2V_{\Delta,J}}\frac{\mathtt{b}_{\Delta,J}}{n}\sqrt{z} + 2\left(\frac{\mathtt{b}_{\Delta,J}}{n}\right)^2 z \right] \mathrm{d}z \\
&\leq \exp(-t)\left[4V_{\Delta,J} + 6\sqrt{2V_{\Delta,J}}\frac{\mathtt{b}_{\Delta,J}}{n}\left(\sqrt{t} + \frac{1}{2\sqrt{t}}\right) + 4\left(\frac{\mathtt{b}_{\Delta,J}}{n}\right)^2 (t+1) \right] \\
&\leq \exp(-t) V_{\Delta,J}\left[4 + C\frac{\mathtt{b}_{\Delta,J}}{n}\sqrt{\frac{t}{V_{\Delta,J}}} + C\frac{\mathtt{b}_{\Delta,J}^2 t}{n^2 V_{\Delta,J}} \right].
\end{aligned} \tag{A.104}$$

Step 2. Exponential control of the linearization remainder

Recall from (A.46) that

$$r_{\Delta,J} = r_{g,J} - \Pi_J r_{\mathrm{IV},J_{\Delta,\max}} - \left(\widehat{\Pi}_J - \Pi_J\right)\left[\frac{1}{n}\sum_{i=1}^{n}\xi_{\mathrm{IV},J_{\Delta,\max}}(\mathcal{D}_i)\right] - \left(\widehat{\Pi}_J - \Pi_J\right) r_{\mathrm{IV},J_{\Delta,\max}}$$

$$- (\widehat{\Pi}_J - \Pi_J)\Xi_{J_{\Delta,\max}} h_0 - \Pi_J\left(\Xi_{J_{\Delta,\max}} h_0 - h_0\right). \tag{A.105}$$

For $t \geq 1$, define

$$\varrho_{\Delta,n}(t) = \zeta_\Delta \sqrt{\frac{\log J_{\Delta,\max} + t}{n}} + \zeta_\Delta^2 \frac{\log J_{\Delta,\max} + t}{n}, \tag{A.106}$$

$$\varrho_{g,J,n}(t) = \sqrt{\frac{J+t}{n}} + \zeta_{X,J}\frac{t}{n}, \tag{A.107}$$

$$\varrho_{\mathrm{IV},\Delta,n}(t) = \sqrt{\frac{K_{\Delta,\max} + t}{n}} + \zeta_\Delta \frac{t}{n} + \sigma_\Delta J_{\Delta,\max}^{-s/d_x}. \tag{A.108}$$

Fix $J \in \mathcal{J}_n$ with $J \leq J_{\Delta,\max}$ and $1 \leq t \leq C\log n$. By Propositions B.1 to B.4, the corresponding concentration inequalities hold simultaneously on an event with probability at least $1 - C\exp(-t)$. We work on this event throughout the remainder of the argument. In particular,

$$\max_{\substack{J\in\mathcal{J}_n \\ J\leq J_{\Delta,\max}}} \left\| G_X^{-1/2}\widehat{G}_X G_X^{-1/2} - I_J \right\|_{\mathrm{op}} \lesssim \varrho_{\Delta,n}(t). \tag{A.109}$$

We bound the six terms in (A.105) separately. First, by (A.20), (B.2), and (B.4),

$$\left\| r_{g,J} \right\|_X \leq \left\| G_X^{1/2}\left(\widehat{G}_X^{-1} - G_X^{-1}\right) G_X^{1/2} \right\|_{\mathrm{op}} \left\| G_X^{-1/2}\frac{1}{n}\sum_{i=1}^{n} p^J(X_i)\left(Y_i - \Pi_J g_0(X_i)\right) \right\| \lesssim \varrho_{\Delta,n}(t)\varrho_{g,J,n}(t). \tag{A.110}$$

Next, we control the projected NPIV pilot remainder $\Pi_J r_{\mathrm{IV},J_{\Delta,\max}}$. Combining (A.30) with (B.3) and (B.5) gives

$$\left\| \Pi_{K_{\Delta,\max}} T r_{\mathrm{IV},J_{\Delta,\max}} \right\|_Z \lesssim \sigma_\Delta^{-1}\varrho_{\Delta,n}(t)\varrho_{\mathrm{IV},\Delta,n}(t).$$

Since $r_{\mathrm{IV},J_{\Delta,\max}} \in \mathcal{H}_{J_{\Delta,\max}}$, applying (A.47) gives

$$\left\| \Pi_J r_{\mathrm{IV},J_{\Delta,\max}} \right\|_X \lesssim \nu_J^{-1}\left\| \Pi_{K_{\Delta,\max}} T r_{\mathrm{IV},J_{\Delta,\max}} \right\|_Z \lesssim \nu_J^{-1}\sigma_\Delta^{-1}\varrho_{\Delta,n}(t)\varrho_{\mathrm{IV},\Delta,n}(t). \tag{A.111}$$

We next control the two terms involving $\widehat{\Pi}_J - \Pi_J$. By (A.52) and the Gram-matrix concentration bound in Proposition B.1,

$$\sup_{\substack{J\in\mathcal{J}_n \\ J\le J_{\Delta,\max}}} \sup_{\substack{f\in\mathcal{H}_{J_{\Delta,\max}} \\ f\neq 0}} \frac{\|(\widehat{\Pi}_J - \Pi_J)f\|_X}{\|f\|_X} \lesssim \varrho_{\Delta,n}(t). \tag{A.112}$$

Since $n^{-1}\sum_{i=1}^n \xi_{\mathrm{IV},J_{\Delta,\max}}(\mathcal{D}_i)$ belongs to $\mathcal{H}_{J_{\Delta,\max}}$, combining (A.112) with (B.9) gives

$$\left\| \left(\widehat{\Pi}_J - \Pi_J\right) \left[\frac{1}{n}\sum_{i=1}^n \xi_{\mathrm{IV},J_{\Delta,\max}}(\mathcal{D}_i)\right] \right\|_X \lesssim \varrho_{\Delta,n}(t) \left[\sqrt{\frac{t}{n}\sum_{j=1}^{J_{\Delta,\max}} \nu_j^{-2}} + \sigma_\Delta^{-1}\zeta_\Delta \frac{t}{n}\right]. \tag{A.113}$$

Similarly, since $r_{\mathrm{IV},J_{\Delta,\max}} \in \mathcal{H}_{J_{\Delta,\max}}$, (A.112) and (A.57) imply that

$$\begin{aligned} \left\| \left(\widehat{\Pi}_J - \Pi_J\right) r_{\mathrm{IV},J_{\Delta,\max}} \right\|_X &\lesssim \varrho_{\Delta,n}(t) \left\| r_{\mathrm{IV},J_{\Delta,\max}} \right\|_X \lesssim \sigma_\Delta^{-1}\varrho_{\Delta,n}(t) \left\| \Pi_{K_{\Delta,\max}} T r_{\mathrm{IV},J_{\Delta,\max}} \right\|_Z \\ &\lesssim \sigma_\Delta^{-2}\varrho_{\Delta,n}(t)^2 \varrho_{\mathrm{IV},\Delta,n}(t). \end{aligned} \tag{A.114}$$

For the empirical projection of the population pilot target, Proposition B.4 directly establishes

$$\left\| \left(\widehat{\Pi}_J - \Pi_J\right) \Xi_{J_{\Delta,\max}} h_0 \right\|_X \lesssim \zeta_{X,J} J^{-s/d_x} \sqrt{\frac{t}{n}} + \zeta_{X,J}\frac{t}{n}. \tag{A.115}$$

Finally, the population centering bound (A.41) gives

$$\left\| \Pi_J \left(\Xi_{J_{\Delta,\max}} h_0 - h_0\right) \right\|_X \lesssim \nu_J^{-1} \nu_{J_{\Delta,\max}+1} J_{\Delta,\max}^{-s/d_x}. \tag{A.116}$$

Let $\mathfrak{q}_{\Delta,J,n}(t)$ be a sufficiently large constant multiple of the sum of the right-hand sides of Equations (A.110), (A.111) and (A.113) to (A.116). The triangle inequality and a union bound give

$$P\Big(\left\| r_{\Delta,J} \right\|_X > \mathfrak{q}_{\Delta,J,n}(t) \Big) \le C\exp(-t), \qquad 1 \le t \le C\log n. \tag{A.117}$$

We now show that the remainder radius $\mathfrak{q}_{\Delta,J,n}(\tilde{\kappa}_{n,J})$ is negligible relative to the stochastic scale of the leading term $\sqrt{\tilde{\kappa}_{n,J} V_{\Delta,J}}$. Recall from (A.97) that

$$V_{\Delta,J} \asymp \frac{1}{n}\sum_{j=1}^J \nu_j^{-2}.$$

Throughout the argument, let $1 \le t \le C\log n$. The first condition in Assumption A.2 implies that

$$\sigma_\Delta^{-1}\varrho_{\Delta,n}(t) = o(1), \qquad \varrho_{\Delta,n}(t) \lesssim \zeta_\Delta \sqrt{\frac{\log J_{\Delta,\max} + t}{n}},$$

uniformly over this range.

First, we consider the right-hand sides of Equations (A.110) and (A.113). Since $\sum_{j=1}^{J}\nu_j^{-2} \gtrsim J$, the bounds (A.110) and (A.113) give

$$\frac{\varrho_{\Delta,n}(t)\varrho_{g,J,n}(t)}{\sqrt{tV_{\Delta,J}}} \lesssim \varrho_{\Delta,n}(t)\left[\sqrt{\frac{J+t}{t\sum_{j=1}^{J}\nu_j^{-2}}} + \frac{\zeta_{X,J}\sqrt{t/n}}{\left(\sum_{j=1}^{J}\nu_j^{-2}\right)^{1/2}}\right] = o(1)$$

$$\frac{\varrho_{\Delta,n}(t)\left[\sqrt{\frac{t}{n}\sum_{j=1}^{J_{\Delta,\max}}\nu_j^{-2}} + \sigma_\Delta^{-1}\zeta_\Delta\frac{t}{n}\right]}{\sqrt{tV_{\Delta,J}}} \lesssim \varrho_{\Delta,n}(t)\left(\frac{\sum_{j=1}^{J_{\Delta,\max}}\nu_j^{-2}}{\sum_{j=1}^{J}\nu_j^{-2}}\right)^{1/2} + \sigma_\Delta^{-1}\varrho_{\Delta,n}(t)\frac{\zeta_\Delta\sqrt{t/n}}{\left(\sum_{j=1}^{J}\nu_j^{-2}\right)^{1/2}} = o(1).$$

The first equality follows from $\varrho_{\Delta,n}(t) = o(1)$ and $\zeta_\Delta\sqrt{\log n/n} = o(1)$. The second follows from the first two conditions in Assumption A.2, using $J \geq \underline{J}$.

Next, consider the right-hand side of (A.111). The definition of $\varrho_{\mathrm{IV},\Delta,n}(t)$, and the bound for $\varrho_{\Delta,n}(t)$ above,

$$\frac{\nu_J^{-1}\sigma_\Delta^{-1}\varrho_{\Delta,n}(t)\varrho_{\mathrm{IV},\Delta,n}(t)}{\sqrt{tV_{\Delta,J}}} \lesssim \frac{\nu_J^{-1}}{\left(\sum_{j=1}^{J}\nu_j^{-2}\right)^{1/2}}\left[\sigma_\Delta^{-1}\zeta_\Delta\sqrt{\frac{K_{\Delta,\max}\log J_{\Delta,\max}}{n}} + \zeta_\Delta J_{\Delta,\max}^{-s/d_x}\sqrt{\log J_{\Delta,\max}}\right] + o(1) = o(1),$$

where we use

$$\frac{(\log J_{\Delta,\max}+t)(K_{\Delta,\max}+t)}{t} \lesssim K_{\Delta,\max}\log J_{\Delta,\max} + t.$$

The last equality follows from the third condition in Assumption A.2. The terms absorbed into $o(1)$ are controlled by $\sigma_\Delta^{-1}\zeta_\Delta\sqrt{\log n/n} = o(1)$.

For the right-hand side of (A.114), note that the ratio of the bounds in (A.114) and (A.111) is

$$\nu_J\sigma_\Delta^{-1}\varrho_{\Delta,n}(t) = o(1),$$

since ν_J is uniformly bounded. Thus, this term is also negligible relative to $\sqrt{tV_{\Delta,J}}$.

For the empirical projection of the population pilot target, (A.115) gives

$$\frac{\zeta_{X,J}J^{-s/d_x}\sqrt{t/n} + \zeta_{X,J}t/n}{\sqrt{tV_{\Delta,J}}} \lesssim \frac{\zeta_{X,J}J^{-s/d_x}}{\left(\sum_{j=1}^{J}\nu_j^{-2}\right)^{1/2}} + \frac{\zeta_{X,J}\sqrt{t/n}}{\left(\sum_{j=1}^{J}\nu_j^{-2}\right)^{1/2}} = o(1).$$

Here the first term is negligible by the sieve basis-envelope and smoothness restrictions, together with $\underline{J} \to \infty$. The second is negligible because $\zeta_{X,J} \leq \zeta_\Delta$ and $\zeta_\Delta\sqrt{\log n/n} = o(1)$.

Finally, the deterministic centering term in (A.116) satisfies

$$\frac{\nu_J^{-1}\nu_{J_{\Delta,\max}+1}J_{\Delta,\max}^{-s/d_x}}{\sqrt{tV_{\Delta,J}}} \lesssim \frac{\sqrt{n}\,\nu_{J_{\Delta,\max}+1}J_{\Delta,\max}^{-s/d_x}}{\nu_J\left(\sum_{j=1}^J \nu_j^{-2}\right)^{1/2}} = o(1)$$

by (A.17), since $t \geq 1$.

Combining these bounds yields

$$\max_{\substack{J\in\mathcal{J}_n \\ J\leq J_{\Delta,\max}}} \frac{\mathtt{q}_{\Delta,J,n}(\tilde{\kappa}_{n,J})}{\sqrt{\tilde{\kappa}_{n,J}V_{\Delta,J}}} = o(1). \tag{A.118}$$

Step 3. Concentration of the full estimation error

Combining (A.92), (A.99), and (A.117), we have that

$$P\left(\left\|\widehat{\Delta}_J - \Delta_J\right\|_X > \sqrt{2V_{\Delta,J}t} + \frac{\mathtt{b}_{\Delta,J}}{n}t + \mathtt{q}_{\Delta,J,n}(t)\right) \leq \exp(-t) \tag{A.119}$$

uniformly over the relevant J and t.

By (A.101) and (A.118),

$$\max_{\substack{J\in\mathcal{J}_n \\ J\leq J_{\Delta,\max}}} \frac{\left[\sqrt{2V_{\Delta,J}\tilde{\kappa}_{n,J}} + (\mathtt{b}_{\Delta,J}/n)\tilde{\kappa}_{n,J} + \mathtt{q}_{\Delta,J,n}(\tilde{\kappa}_{n,J})\right]^2}{\tilde{\kappa}_{n,J}V_{\Delta,J}} = 2 + o(1). \tag{A.120}$$

Moreover, $\widehat{\Delta}_J - \Delta_J \in \mathcal{H}_J$. Hence

$$\left\|\widehat{\Delta}_J - \Delta_J\right\|_{X,n}^2 \leq \left[1 + \left\|G_X^{-1/2}\widehat{G}_X G_X^{-1/2} - I_J\right\|_{\mathrm{op}}\right]\left\|\widehat{\Delta}_J - \Delta_J\right\|_X^2. \tag{A.121}$$

By (A.109) and Assumption A.2, the multiplicative factor in (A.121) is $1 + o(1)$ at $t = \tilde{\kappa}_{n,J}$. Therefore, for all sufficiently large n,

$$P\left(\left\|\widehat{\Delta}_J - \Delta_J\right\|_{X,n}^2 > \left(2 + \frac{\eta}{4}\right)\tilde{\kappa}_{n,J}V_{\Delta,J}\right) \leq 2\exp(-\tilde{\kappa}_{n,J}) = \frac{\exp(-t_J)}{|\mathcal{J}_n|},$$

which proves (A.90).

We next establish (A.91). The integrated-tail argument in (A.104), together with (A.117), (A.118), and (A.121), gives

$$E\left[\left\|\widehat{\Delta}_J - \Delta_J\right\|_{X,n}^2 - \left(2 + \frac{\eta}{4}\right)\tilde{\kappa}_{n,J}V_{\Delta,J}\right]_+ \leq 6\exp(-\tilde{\kappa}_{n,J})V_{\Delta,J} \tag{A.122}$$

for all sufficiently large n, uniformly over the relevant J.

By the definition of $\tilde{\kappa}_{n,J}$, $6\exp(-\tilde{\kappa}_{n,J}) = 3\exp(-t_J)/|\mathcal{J}_n|$. And

$$3\exp(-t_J)V_{\Delta,J} = \frac{3V_{\Delta,J}}{e + 3V_{\Delta,J}/V_{g,J}} \leq V_{g,J},$$

which proves (A.91). ■

Proposition A.3. *Suppose that the assumptions of Proposition A.2 hold, and let $\eta > 0$ be fixed. For each non-maximal $J \in \mathcal{J}_n$ whose immediate successor J^+ satisfies $J^+ \leq J_{\Delta,\max}$, define*

$$V_{\Delta;J,J^+} = E\left\|\Psi_{\Delta,J^+} - \Psi_{\Delta,J}\right\|_X^2. \tag{A.123}$$

Let

$$t_{\Delta;J,J^+} = \log\left(e + \frac{3V_{\Delta;J,J^+}}{V_{g,J}}\right), \quad \kappa_{n;J,J^+} = \left(2 + \frac{\eta}{4}\right)\left\{t_{\Delta;J,J^+} + \log(2|\mathcal{J}_n|)\right\}. \tag{A.124}$$

Then, for all sufficiently large n, uniformly over $F_n \in \mathcal{F}(\rho_n)$ and the stated pairs (J, J^+),

$$P\left(\left\|\left(\widehat{\Delta}_{J^+} - \Delta_{J^+}\right) - \left(\widehat{\Delta}_J - \Delta_J\right)\right\|_{X,n}^2 > \kappa_{n;J,J^+}V_{\Delta;J,J^+}\right) \leq \frac{\exp(-t_{\Delta;J,J^+})}{|\mathcal{J}_n|}, \tag{A.125}$$

and

$$E\left[\left\|\left(\widehat{\Delta}_{J^+} - \Delta_{J^+}\right) - \left(\widehat{\Delta}_J - \Delta_J\right)\right\|_{X,n}^2 - \kappa_{n;J,J^+}V_{\Delta;J,J^+}\right]_+ \leq \frac{3\exp(-t_{\Delta;J,J^+})}{|\mathcal{J}_n|}V_{\Delta;J,J^+} \leq \frac{V_{g,J}}{|\mathcal{J}_n|}. \tag{A.126}$$

Proof of Proposition A.3. Fix a pair (J, J^+) as stated in the proposition. The leading term of the successor increment is

$$\Psi_{\Delta;J,J^+} = \Psi_{\Delta,J^+} - \Psi_{\Delta,J} = \frac{1}{n}\sum_{i=1}^n \left[\xi_{\Delta,J^+}(\mathcal{D}_i) - \xi_{\Delta,J}(\mathcal{D}_i)\right].$$

Because the same deterministic NPIV pilot is used at every sieve dimension, the NPIV component of the increment is obtained by applying $\Pi_{J^+} - \Pi_J$ to the common pilot score. The population matrix comparison used in the proof of (A.45) therefore yields

$$V_{\Delta;J,J^+} \asymp \frac{1}{n}\sum_{j=J+1}^{J^+} \nu_j^{-2}. \tag{A.127}$$

We next establish concentration of the leading term. Applying the moment argument in Step 1

to $\xi_{\Delta,J^+}(\mathcal{D}_i) - \xi_{\Delta,J}(\mathcal{D}_i)$ gives, for every integer $k \geq 2$,

$$E\left\|\xi_{\Delta,J^+}(\mathcal{D}_i) - \xi_{\Delta,J}(\mathcal{D}_i)\right\|_X^k \leq \frac{k!}{2}\left(C\zeta_\Delta \nu_{J^+}^{-1}\right)^{k-2} E\left\|\xi_{\Delta,J^+}(\mathcal{D}_i) - \xi_{\Delta,J}(\mathcal{D}_i)\right\|_X^2. \tag{A.128}$$

Since the summands are independent and centered, Lemma A.1 implies that, for every $t \geq 0$,

$$P\left(\left\|\Psi_{\Delta;J,J^+}\right\|_X > \sqrt{2V_{\Delta;J,J^+}t} + C\zeta_\Delta \nu_{J^+}^{-1}\frac{t}{n}\right) \leq 2\exp(-t). \tag{A.129}$$

We next control the linearization remainder. By the expansions established in Appendix A.1,

$$\left(\widehat{\Delta}_{J^+} - \Delta_{J^+}\right) - \left(\widehat{\Delta}_J - \Delta_J\right) = \Psi_{\Delta;J,J^+} + \left(r_{\Delta,J^+} - r_{\Delta,J}\right).$$

Since $\mathcal{J}_n$ is a geometric grid and $\{\nu_j\}$ is decreasing, (A.127) implies that $V_{\Delta,J} + V_{\Delta,J^+} \lesssim V_{\Delta;J,J^+}$. Consequently, the uniform remainder bounds in Step 2 imply that

$$\left\|r_{\Delta,J^+} - r_{\Delta,J}\right\|_X \leq \left\|r_{\Delta,J^+}\right\|_X + \left\|r_{\Delta,J}\right\|_X = o\left(\sqrt{tV_{\Delta;J,J^+}}\right)$$

uniformly at the relevant concentration levels.

Finally, applying the norm comparison (A.121) on $\mathcal{H}_{J^+}$ and repeating the argument of Step 3 with the concentration level $t = t_{\Delta;J,J^+} + \log(2|\mathcal{J}_n|)$ establishes (A.125). The corresponding integrated-tail argument gives (A.126). In particular,

$$3\exp(-t_{\Delta;J,J^+})V_{\Delta;J,J^+} = \frac{3V_{\Delta;J,J^+}}{e + 3V_{\Delta;J,J^+}/V_{g,J}} \leq V_{g,J},$$

which completes the argument. ■

A.6 Proof of Theorem 5.1

The proof establishes the oracle inequality by comparing the data-driven choice $(\hat{\lambda}, \hat{J})$ with two alternative choices at a fixed sieve dimension $J \in \mathcal{J}_n$. We first consider $(\hat{\lambda}(J), J)$, where $\hat{\lambda}(J)$ minimizes the empirical criterion $\hat{R}_n(\lambda, J)$ over $\lambda \in [0,1]$ for a given J. This comparison yields an upper bound in terms of the variance associated with estimating the endogeneity bias Δ and its sieve approximation error. We then compare $(\hat{\lambda}, \hat{J})$ with $(0, J)$, which corresponds to NP-OLS and yields an alternative bound in terms of the magnitude of endogeneity ρ_n.

The main challenge is to control the stochastic fluctuations arising from the data-dependent selection of $(\hat{\lambda}, \hat{J})$. We use concentration inequalities to control the upward stochastic fluctuations for both NP-OLS and the estimated endogeneity bias component. Crucially, the downward adjustment allows us to control the latter fluctuations at the NP-OLS variance scale. Combining the two comparisons and taking expectations then yields the desired oracle inequality, from which the

adaptive rates of convergence follow by optimizing over $J \in \mathcal{J}_n$.

Remark A.2. The proof strategy builds on the use of concentration inequalities to establish oracle inequalities in adaptive estimation (Birgé and Massart, 1997; Barron, Birgé, and Massart, 1999; Massart, 2007). Since the selected tuning parameters depend on the data, the stochastic terms in the basic inequality must be controlled uniformly over the candidate tuning parameters. Concentration inequalities provide such uniform control, allowing us to bound the stochastic remainders and derive an oracle inequality for the risk of the selected estimator. ■

Remark A.3. The one-sided downward adjustment in our approach is related to the risk hull method of Cavalier and Golubev (2006), which controls stochastic fluctuations uniformly over candidate regularization parameters, and to Loubes and Marteau (2009); Horowitz (2014a) who use logarithmically inflated variance terms in adaptive NPIV estimation. The key distinction is that our adjustment controls the upward stochastic fluctuations of the estimated endogeneity bias and is calibrated to bound the resulting remainder at the NP-OLS variance scale. ■

Recall that the structural function satisfies $h_0(x) = g_0(x) - \Delta(x)$, where $g_0(x) = E[Y \mid X = x]$ is the conditional mean function and $\Delta(x) = E[U \mid X = x]$ is the endogeneity bias. The magnitude of endogeneity is denoted as $\rho_n = \|\Delta\|_X^2$. Let $\varepsilon = Y - g_0(X)$ so that $E[\varepsilon \mid X] = 0$.

For each sieve dimension J in the geometric grid $\mathcal{J}_n$, let $\widehat{g}_J$ denote the sieve NP-OLS estimator of g_0. Let $\widehat{\Pi}_J$ be the empirical least-squares projection onto the sieve space $\mathcal{H}_J$. For the moment, we consider the deterministic pilot dimension $J_{\Delta,\max}$ and the corresponding common-pilot estimator $\widetilde{\Delta}_{J_{\Delta,\max}}$. The estimated endogeneity bias is constructed as

$$\widehat{\Delta}_J = \widehat{\Pi}_{J \wedge J_{\Delta,\max}} \widetilde{\Delta}_{J_{\Delta,\max}}.$$

The shrinkage estimator is

$$\widehat{h}_{\lambda,J} = \widehat{g}_J - \lambda \widehat{\Delta}_J, \qquad \lambda \in [0,1].$$

Recall from (5.8) that the empirical selection criterion is

$$\widehat{R}_n(\lambda, J) := 2\widehat{V}_{g,J} - \left\|\hat{g}_J\right\|_{X,n}^2 + \lambda^2 \widehat{V}_{\Delta, J \wedge J_{\Delta,\max}} - (2\lambda - \lambda^2)\widehat{B}_{\Delta,J} \tag{A.130}$$

where

$$\widehat{B}_{\Delta,J} = \left[\left\|\hat{\Delta}_J\right\|_{X,n}^2 - \hat{\kappa}_{n,J} \widehat{V}_{\Delta, J \wedge J_{\Delta,\max}} \right]_+. \tag{A.131}$$

For each fixed J, the criterion is quadratic in λ. For each $J \in \mathcal{J}_n$, minimization over $\lambda \in [0,1]$ yields that

$$\hat{\lambda}(J) = \frac{\widehat{B}_{\Delta,J}}{\widehat{B}_{\Delta,J} + \widehat{V}_{\Delta, J \wedge J_{\Delta,\max}}}. \tag{A.132}$$

We also define that

$$\mathcal{C}_{\Delta,J} = \frac{\widehat{B}_{\Delta,J}^2}{\widehat{B}_{\Delta,J} + \widehat{V}_{\Delta, J \wedge J_{\Delta,\max}}}. \tag{A.133}$$

The quantity $\mathcal{C}_{\Delta,J}$ is the reduction in the empirical criterion obtained by optimizing the shrinkage parameter at dimension J, relative to choosing $\lambda = 0$.

Consequently, the joint minimization is equivalent to

$$\widehat{J} \in \arg\min_{J \in \mathcal{J}_n} \left\{ 2\widehat{V}_{g,J} - \left\|\widehat{g}_J\right\|_{X,n}^2 - \mathcal{C}_{\Delta,J} \right\}, \qquad \hat{\lambda} = \hat{\lambda}(\widehat{J}). \tag{A.134}$$

Fix an arbitrary candidate $J \in \mathcal{J}_n$. Since $\hat{g}_J$ is the empirical least square projection of $Y = g_0 + \varepsilon$ onto the sieve space $\mathcal{H}_J$, the projection identity gives that $\langle Y - \hat{g}_J, \hat{g}_J \rangle_{X,n} = 0$, then $\langle g_0, \hat{g}_J \rangle_{X,n} = \|\hat{g}_J\|_{X,n}^2 - \langle \varepsilon, \hat{g}_J \rangle_{X,n}$. Expanding $\|\hat{g}_J - g_0\|_{X,n}^2$ and substituting this identity, we obtain that

$$\begin{aligned} \left\|\hat{g}_J - g_0\right\|_{X,n}^2 &= \left\|\hat{g}_J\right\|_{X,n}^2 - 2\left\langle g_0, \hat{g}_J \right\rangle_{X,n} + \left\|g_0\right\|_{X,n}^2 \\ &= -\left\|\hat{g}_J\right\|_{X,n}^2 + 2\left\langle \varepsilon, \hat{g}_J \right\rangle_{X,n} + \left\|g_0\right\|_{X,n}^2 \end{aligned}$$

Rearranging gives that

$$-\left\|\hat{g}_J\right\|_{X,n}^2 = \left\|\hat{g}_J - g_0\right\|_{X,n}^2 - 2\left\langle \varepsilon, \hat{g}_J \right\rangle_{X,n} - \left\|g_0\right\|_{X,n}^2. \tag{A.135}$$

Define the centered stochastic term $\mathcal{Z}_J := 2\langle \varepsilon, \hat{g}_J - g_0 \rangle_{X,n} - 2\widehat{V}_{g,J}$. Substituting (A.135) into the profile criterion (A.134) gives that

$$\begin{aligned} 2\widehat{V}_{g,J} - \left\|\widehat{g}_J\right\|_{X,n}^2 - \mathcal{C}_{\Delta,J} &= \left\|\hat{g}_J - g_0\right\|_{X,n}^2 - 2\left\langle \varepsilon, \hat{g}_J \right\rangle_{X,n} + 2\widehat{V}_{g,J} - \left\|g_0\right\|_{X,n}^2 - \mathcal{C}_{\Delta,J} \\ &= \left\|\hat{g}_J - g_0\right\|_{X,n}^2 - \mathcal{Z}_J - \mathcal{C}_{\Delta,J} - 2\left\langle \varepsilon, g_0 \right\rangle_{X,n} - \left\|g_0\right\|_{X,n}^2 \end{aligned}$$

where the last two terms do not depend on J. Since $\hat{J}$ minimizes the profiled criterion, comparison with any $J \in \mathcal{J}_n$ yields

$$\left\|\hat{g}_{\hat{J}} - g_0\right\|_{X,n}^2 \leq \left\|\hat{g}_J - g_0\right\|_{X,n}^2 + \mathcal{Z}_{\hat{J}} - \mathcal{Z}_J + \mathcal{C}_{\Delta,\hat{J}} - \mathcal{C}_{\Delta,J}. \tag{A.136}$$

The inequality (A.136) compares the estimation error of NP-OLS at the selected $\hat{J}$ with that at an arbitrary J. To obtain an expected-risk bound, we must control the stochastic fluctuation $\mathcal{Z}_{\hat{J}} - \mathcal{Z}_J$. Applying the penalized least squares arguments of Baraud (2000) and Massart (2007, Section 4.2) to our setting, we obtain that, for any fixed $\gamma \in (0,1)$,

$$E\left[\mathcal{Z}_{\hat{J}} - \mathcal{Z}_J\right] \leq \gamma E\left\|\widehat{g}_{\hat{J}} - g_0\right\|_{X,n}^2 + C_\gamma \left[V_{g,J} + \left\|g_0 - \Pi_J g_0\right\|_X^2 \right]. \tag{A.137}$$

Comparison of endogeneity components

We next compare the endogeneity component $\|\hat{\lambda}\widehat{\Delta}_{\hat{J}} - \Delta\|_{X,n}^2$ across sieve dimensions. The key property is that the estimators $\{\widehat{\Delta}_J : J \in \mathcal{J}_n\}$ are obtained by projecting a common pilot estimator

$\widetilde{\Delta}_{J_{\Delta,\max}}$ onto nested sieve spaces.

For any $J < J'$ in $\mathcal{J}_n$, the nesting of the sieve spaces implies

$$\begin{aligned}\widehat{\Pi}_{J\wedge J_{\Delta,\max}}\widehat{\Delta}_{J'} &= \widehat{\Pi}_{J\wedge J_{\Delta,\max}}\widehat{\Pi}_{J'\wedge J_{\Delta,\max}}\widetilde{\Delta}_{J_{\Delta,\max}} \\ &= \widehat{\Pi}_{J\wedge J_{\Delta,\max}}\widetilde{\Delta}_{J_{\Delta,\max}} = \widehat{\Delta}_J.\end{aligned} \tag{A.138}$$

Thus, $\widehat{\Delta}_J$ is the empirical orthogonal projection of $\widehat{\Delta}_{J'}$ onto the smaller sieve space. In particular, $\widehat{\Delta}_{J'} - \widehat{\Delta}_J$ is empirically orthogonal to $\widehat{\Delta}_J$. The Pythagorean identity therefore gives

$$\left\|\widehat{\Delta}_{J'} - \widehat{\Delta}_J\right\|_{X,n}^2 = \left\|\widehat{\Delta}_{J'}\right\|_{X,n}^2 - \left\|\widehat{\Delta}_J\right\|_{X,n}^2. \tag{A.139}$$

The same orthogonality argument applies to the estimation error of the common pilot. For $J < J'$,

$$\begin{aligned}&\left\|\widehat{\Pi}_{J'\wedge J_{\Delta,\max}}(\widetilde{\Delta}_{J_{\Delta,\max}} - \Delta)\right\|_{X,n}^2 - \left\|\widehat{\Pi}_{J\wedge J_{\Delta,\max}}(\widetilde{\Delta}_{J_{\Delta,\max}} - \Delta)\right\|_{X,n}^2 \\ &= \left\|\left(\widehat{\Pi}_{J'\wedge J_{\Delta,\max}} - \widehat{\Pi}_{J\wedge J_{\Delta,\max}}\right)(\widetilde{\Delta}_{J_{\Delta,\max}} - \Delta)\right\|_{X,n}^2 \geq 0.\end{aligned}$$

Hence, the squared empirical norm of the projected pilot estimation error is nondecreasing in J. By contrast, the squared approximation error $\|\Delta - \widehat{\Pi}_{J\wedge J_{\Delta,\max}}\Delta\|_{X,n}^2$ is nonincreasing in J, because enlarging the sieve space cannot increase the empirical least squares approximation error.

These properties yield an exact decomposition of the endogeneity estimation error. Since $\widehat{\Delta}_J = \widehat{\Pi}_{J\wedge J_{\Delta,\max}}\widetilde{\Delta}_{J_{\Delta,\max}}$, empirical orthogonality implies

$$\left\|\widehat{\Delta}_J - \Delta\right\|_{X,n}^2 = \left\|\widehat{\Pi}_{J\wedge J_{\Delta,\max}}(\widetilde{\Delta}_{J_{\Delta,\max}} - \Delta)\right\|_{X,n}^2 + \left\|\Delta - \widehat{\Pi}_{J\wedge J_{\Delta,\max}}\Delta\right\|_{X,n}^2. \tag{A.140}$$

The first component measures the estimation error of the pilot projected onto the sieve space, while the second measures the approximation error of Δ. As J increases, the first component is nondecreasing and the second is nonincreasing. Both remain unchanged once $J \geq J_{\Delta,\max}$.

More generally, for any $\lambda \in [0,1]$, the same orthogonality argument gives

$$\left\|\lambda\widehat{\Delta}_J - \Delta\right\|_{X,n}^2 = \left\|\widehat{\Pi}_{J\wedge J_{\Delta,\max}}(\lambda\widetilde{\Delta}_{J_{\Delta,\max}} - \Delta)\right\|_{X,n}^2 + \left\|\Delta - \widehat{\Pi}_{J\wedge J_{\Delta,\max}}\Delta\right\|_{X,n}^2.$$

Together with the nesting identities above, it provides the basis for comparing the endogeneity components at the selected $(\hat{\lambda}, \hat{J})$ and an arbitrary candidate (λ, J).

The orthogonal decomposition (A.140) separates the estimation error of $\widehat{\Delta}_J$ into a projected pilot estimation error and a sieve approximation error. To control the former, we introduce a

positive-part remainder. For any fixed $\eta > 0$, choose $\delta = \eta/(4(8+\eta))$. Then $0 < \delta < 1/4$ and

$$(1-\delta)\left(2+\frac{\eta}{4}\right) = 2 + \frac{3\eta}{16} > 2. \tag{A.141}$$

Define that

$$\mathcal{U}_J = \left[\left\|\widehat{\Delta}_J - \widehat{\Pi}_{J\wedge J_{\Delta,\max}}\Delta\right\|_{X,n}^2 - (1-\delta)\widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}}\right]_+. \tag{A.142}$$

By definition,

$$\left\|\widehat{\Delta}_J - \widehat{\Pi}_{J\wedge J_{\Delta,\max}}\Delta\right\|_{X,n}^2 \le (1-\delta)\widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}} + \mathcal{U}_J. \tag{A.143}$$

We next establish two deterministic inequalities for the profiled shrinkage parameter $\hat{\lambda}(J)$ and the profiled gain $\mathcal{C}_{\Delta,J}$. For every $J \in \mathcal{J}_n$,

$$\left(1-\hat{\lambda}(J)\right)^2 \left\|\widehat{\Delta}_J\right\|_{X,n}^2 \le \widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}}, \tag{A.144}$$

$$\left\|\widehat{\Delta}_J\right\|_{X,n}^2 - \mathcal{C}_{\Delta,J} \le 2\widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}}. \tag{A.145}$$

To establish these inequalities, consider first the case $\hat{B}_{\Delta,J} = 0$. We have that $\hat{\lambda}(J) = 0$ and $\mathcal{C}_{\Delta,J} = 0$. Moreover, $\|\widehat{\Delta}_J\|_{X,n}^2 \le \widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}}$, so both inequalities follow immediately.

Now suppose that $\hat{B}_{\Delta,J} > 0$. By definition, $\|\widehat{\Delta}_J\|_{X,n}^2 = \widehat{B}_{\Delta,J} + \widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}}$. Substituting the expression (A.132) and using $\hat{\kappa}_{n,J} \ge 1$, we have that

$$\left(1-\hat{\lambda}(J)\right)\left\|\widehat{\Delta}_J\right\|_{X,n}^2 = \frac{\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}}}{\widehat{B}_{\Delta,J} + \widehat{V}_{\Delta,J\wedge J_{\Delta,\max}}}\left(\widehat{B}_{\Delta,J} + \widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}}\right) \le \hat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}}.$$

Since $0 \le 1-\hat{\lambda}(J) \le 1$, the first inequality (A.144) follows.

For the second inequality (A.145), recall that $\mathcal{C}_{\Delta,J} = \hat{\lambda}(J)\widehat{B}_{\Delta,J}$. It follows that

$$\begin{aligned}\left\|\widehat{\Delta}_J\right\|_{X,n}^2 - \mathcal{C}_{\Delta,J} &= \widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}} + (1-\hat{\lambda}(J))\widehat{B}_{\Delta,J} \\ &\le \widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}} + \widehat{V}_{\Delta,J\wedge J_{\Delta,\max}} \le 2\widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}}\end{aligned}$$

as desired.

The inequalities (A.144) and (A.145) allow to compare the endogeneity component at the selected $(\hat{\lambda}, \hat{J})$ with that at an arbitrary candidate $J \in \mathcal{J}_n$. Specifically, we seek to bound $\|\hat{\lambda}\widehat{\Delta}_{\hat{J}} - \Delta\|_{X,n}^2$ in terms of the variance penalty $\hat{\kappa}_{n,J}\hat{V}_{\Delta,J\wedge J_{\Delta,\max}}$, the approximation error $\|\Delta - \widehat{\Pi}_{J\wedge J_{\Delta,\max}}\Delta\|_{X,n}^2$, and the positive-part remainders $\mathcal{U}_J$ and $\mathcal{U}_{\hat{J}}$. The comparison also involves the difference $\mathcal{C}_{\Delta,J} - \mathcal{C}_{\Delta,\hat{J}}$, which cancels with the corresponding terms in (A.136). We establish the comparison separately for

$\hat{J} \leq J$ and $\hat{J} > J$.

Case 1: $\hat{J} \leq J$.

When $\hat{J} \leq J$, the nesting identity (A.138) implies that $\widehat{\Delta}_{\hat{J}}$ is the empirical orthogonal projection of $\widehat{\Delta}_J$ onto the smaller sieve space. Consequently, $\widehat{\Delta}_J - \widehat{\Delta}_{\hat{J}}$ is empirically orthogonal to $\widehat{\Delta}_{\hat{J}}$. Decomposing the difference between $\hat{\lambda}\widehat{\Delta}_{\hat{J}}$ and $\widehat{\Delta}_J$, the Pythagorean identity gives

$$\left\|\hat{\lambda}\widehat{\Delta}_{\hat{J}} - \widehat{\Delta}_J\right\|_{X,n}^2 = \left\|\widehat{\Delta}_J - \widehat{\Delta}_{\hat{J}}\right\|_{X,n}^2 + (1-\hat{\lambda})^2 \left\|\widehat{\Delta}_{\hat{J}}\right\|_{X,n}^2 = \left\|\widehat{\Delta}_J\right\|_{X,n}^2 - (2\hat{\lambda} - \hat{\lambda}^2)\left\|\widehat{\Delta}_{\hat{J}}\right\|_{X,n}^2$$

$$\overset{(a)}{\leq} \left\|\widehat{\Delta}_J\right\|_{X,n}^2 - \mathcal{C}_{\Delta,\hat{J}} \overset{(b)}{\leq} \mathcal{C}_{\Delta,J} - \mathcal{C}_{\Delta,\hat{J}} + 2\widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}}. \tag{A.146}$$

The inequality (a) follows from

$$\mathcal{C}_{\Delta,\hat{J}} = \hat{\lambda}\widehat{B}_{\Delta,\hat{J}} \leq \hat{\lambda}\left\|\widehat{\Delta}_{\hat{J}}\right\|_{X,n}^2 \leq (2\hat{\lambda} - \hat{\lambda}^2)\left\|\widehat{\Delta}_{\hat{J}}\right\|_{X,n}^2,$$

where the last inequality uses $\hat{\lambda} \in [0,1]$. The inequality (b) follows from (A.145).

We decompose the estimation error of the endogeneity component by adding and subtracting $\widehat{\Delta}_J$. The squared triangle inequality and empirical orthogonality give

$$\left\|\hat{\lambda}\widehat{\Delta}_{\hat{J}} - \Delta\right\|_{X,n}^2 \leq 2\left\|\hat{\lambda}\widehat{\Delta}_{\hat{J}} - \widehat{\Delta}_J\right\|_{X,n}^2 + 2\left\|\widehat{\Delta}_J - \Delta\right\|_{X,n}^2$$

$$\overset{(a)}{\leq} 2(\mathcal{C}_{\Delta,J} - \mathcal{C}_{\Delta,\hat{J}}) + 4\widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}} + 2\left\|\widehat{\Delta}_J - \widehat{\Pi}_{J\wedge J_{\Delta,\max}}\Delta\right\|_{X,n}^2 + 2\left\|\Delta - \widehat{\Pi}_{J\wedge J_{\Delta,\max}}\Delta\right\|_{X,n}^2$$

$$\overset{(b)}{\leq} 2(\mathcal{C}_{\Delta,J} - \mathcal{C}_{\Delta,\hat{J}}) + C_\eta\left[\widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}} + \left\|\Delta - \widehat{\Pi}_{J\wedge J_{\Delta,\max}}\Delta\right\|_{X,n}^2 + \mathcal{U}_J\right]. \tag{A.147}$$

The inequality (a) follows from (A.146) and the orthogonal error decomposition (A.140). The inequality (b) follows from (A.143).

Moreover, (A.146) implies $\mathcal{C}_{\Delta,J} - \mathcal{C}_{\Delta,\hat{J}} \geq -2\widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}}$.

Case 2: $\hat{J} > J$.

When the selected sieve dimension is larger, the additional sieve direction introduce sampling variation that must be controlled. We begin with an exact identity for the profile criterion.

$$\mathcal{C}_{\Delta,J} - \mathcal{C}_{\Delta,\hat{J}}$$
$$= (\widehat{B}_{\Delta,J} + \widehat{V}_{\Delta,J\wedge J_{\Delta,\max}})(\hat{\lambda} - \hat{\lambda}(J))^2 + \hat{\lambda}^2(\widehat{V}_{\Delta,\hat{J}\wedge J_{\Delta,\max}} - \widehat{V}_{\Delta,J\wedge J_{\Delta,\max}}) - (2\hat{\lambda} - \hat{\lambda}^2)(\widehat{B}_{\Delta,\hat{J}} - \widehat{B}_{\Delta,J}). \tag{A.148}$$

When $\hat{\lambda} > 0$, the definition of $\hat{B}_{\Delta,J}$ together with (A.139) gives that

$$\begin{aligned}\widehat{B}_{\Delta,\hat{J}} - \widehat{B}_{\Delta,J} &\le \left\|\widehat{\Delta}_{\hat{J}}\right\|_{X,n}^2 - \left\|\widehat{\Delta}_J\right\|_{X,n}^2 - (\widehat{\kappa}_{n,\hat{J}}\widehat{V}_{\Delta,\hat{J}\wedge J_{\Delta,\max}} - \widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}}) \\ &= \left\|\widehat{\Delta}_{\hat{J}} - \widehat{\Delta}_J\right\|_{X,n}^2 - (\widehat{\kappa}_{n,\hat{J}}\widehat{V}_{\Delta,\hat{J}\wedge J_{\Delta,\max}} - \widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}}).\end{aligned}$$

When $\hat{\lambda} = 0$, the corresponding difference in (A.148) is multiplied by zero, so the argument below also applies.

Since $\hat{\lambda}^2\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}} + (2\hat{\lambda} - \hat{\lambda}^2)\widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}} \le 2\widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}}$, we obtain

$$\begin{aligned}&(\widehat{B}_{\Delta,J} + \widehat{V}_{\Delta,J\wedge J_{\Delta,\max}})(\hat{\lambda} - \hat{\lambda}(J))^2 + (2\hat{\lambda} - \hat{\lambda}^2)\widehat{\kappa}_{n,\hat{J}}\widehat{V}_{\Delta,\hat{J}\wedge J_{\Delta,\max}} \\ &\le \mathcal{C}_{\Delta,J} - \mathcal{C}_{\Delta,\hat{J}} + 2\widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}} + (2\hat{\lambda} - \hat{\lambda}^2)\left\|\widehat{\Delta}_{\hat{J}} - \widehat{\Delta}_J\right\|_{X,n}^2. \qquad \text{(A.149)}\end{aligned}$$

We next bound the increment $\|\hat{\Delta}_{\hat{J}} - \hat{\Delta}_J\|_{X,n}^2$ using the nested projection structure. By the common pilot construction,

$$\widehat{\Delta}_{\hat{J}} - \widehat{\Delta}_J = (\widehat{\Pi}_{\hat{J}\wedge J_{\Delta,\max}} - \widehat{\Pi}_{J\wedge J_{\Delta,\max}})\Delta + (\widehat{\Pi}_{\hat{J}\wedge J_{\Delta,\max}} - \widehat{\Pi}_{J\wedge J_{\Delta,\max}})(\widetilde{\Delta}_{J_{\Delta,\max}} - \Delta).$$

The squared norms of these two terms are bounded by $\|\Delta - \widehat{\Pi}_{J\wedge J_{\Delta,\max}}\Delta\|_{X,n}^2$ and $\|\widehat{\Delta}_{\hat{J}} - \widehat{\Pi}_{\hat{J}\wedge J_{\Delta,\max}}\Delta\|_{X,n}^2$, respectively. Hence Young's inequality and (A.143) give

$$\begin{aligned}\left\|\widehat{\Delta}_{\hat{J}} - \widehat{\Delta}_J\right\|_{X,n}^2 &\le (1+\delta/2)\left\|\widehat{\Delta}_{\hat{J}} - \widehat{\Pi}_{\hat{J}\wedge J_{\Delta,\max}}\Delta\right\|_{X,n}^2 + (1+2/\delta)\left\|\Delta - \widehat{\Pi}_{J\wedge J_{\Delta,\max}}\Delta\right\|_{X,n}^2 \\ &\le (1-\delta/2)\widehat{\kappa}_{n,\hat{J}}\widehat{V}_{\Delta,\hat{J}\wedge J_{\Delta,\max}} + 2\mathcal{U}_{\hat{J}} + (1+2/\delta)\left\|\Delta - \widehat{\Pi}_{J\wedge J_{\Delta,\max}}\Delta\right\|_{X,n}^2. \qquad \text{(A.150)}\end{aligned}$$

Substitution into (A.149) and using $0 \le 2\hat{\lambda} - \hat{\lambda}^2 \le 1$ gives that

$$\begin{aligned}&(\widehat{B}_{\Delta,J} + \widehat{V}_{\Delta,J\wedge J_{\Delta,\max}})(\hat{\lambda} - \hat{\lambda}(J))^2 + \frac{\delta}{2}(2\hat{\lambda} - \hat{\lambda}^2)\widehat{\kappa}_{n,\hat{J}}\widehat{V}_{\Delta,\hat{J}\wedge J_{\Delta,\max}} \\ &\le \mathcal{C}_{\Delta,J} - \mathcal{C}_{\Delta,\hat{J}} + 2\widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}} + 2\mathcal{U}_{\hat{J}} + (1+2/\delta)\left\|\Delta - \widehat{\Pi}_{J\wedge J_{\Delta,\max}}\Delta\right\|_{X,n}^2. \qquad \text{(A.151)}\end{aligned}$$

We now use this bound (A.151) to compare $\hat{\lambda}\hat{\Delta}_{\hat{J}}$ with $\hat{\lambda}(J)\hat{\Delta}_J$. By empirical orthogonality,

$$\left\|\hat{\lambda}\widehat{\Delta}_{\hat{J}} - \hat{\lambda}(J)\widehat{\Delta}_J\right\|_{X,n}^2 = (\hat{\lambda} - \hat{\lambda}(J))^2\left\|\widehat{\Delta}_J\right\|_{X,n}^2 + \hat{\lambda}^2\left\|\widehat{\Delta}_{\hat{J}} - \widehat{\Delta}_J\right\|_{X,n}^2.$$

Since $\|\widehat{\Delta}_J\|_{X,n}^2 \le \widehat{B}_{\Delta,J} + \widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}}$ and $\hat{\lambda}^2 \le 2\hat{\lambda} - \hat{\lambda}^2$, (A.150)–(A.151) imply that

$$\left\|\hat{\lambda}\widehat{\Delta}_{\hat{J}} - \hat{\lambda}(J)\widehat{\Delta}_J\right\|_{X,n}^2 \le (1+2/\delta)\left(\mathcal{C}_{\Delta,J} - \mathcal{C}_{\Delta,\hat{J}}\right) + C_\eta\left[\widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}} + \left\|\Delta - \widehat{\Pi}_{J\wedge J_{\Delta,\max}}\Delta\right\|_{X,n}^2 + \mathcal{U}_{\hat{J}}\right]. \tag{A.152}$$

It remains to bound the estimation error of $\hat{\lambda}(J)\widehat{\Delta}_J$. Adding and subtracting $\widehat{\Delta}_J$ and applying the squared triangle inequality, we obtain

$$\begin{aligned}\left\|\hat{\lambda}(J)\widehat{\Delta}_J - \Delta\right\|_{X,n}^2 &\le 2(1-\hat{\lambda}(J))^2\left\|\widehat{\Delta}_J\right\|_{X,n}^2 + 2\left\|\widehat{\Delta}_J - \Delta\right\|_{X,n}^2 \\ &\le C\left[\widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}} + \left\|\Delta - \widehat{\Pi}_{J\wedge J_{\Delta,\max}}\Delta\right\|_{X,n}^2 + \mathcal{U}_J\right].\end{aligned} \tag{A.153}$$

The second inequality follows from (A.144), (A.140), and (A.143).

Combining Equations (A.152) and (A.153) with

$$\left\|\hat{\lambda}\widehat{\Delta}_{\hat{J}} - \Delta\right\|_{X,n}^2 \le 2\left\|\hat{\lambda}\widehat{\Delta}_{\hat{J}} - \hat{\lambda}(J)\widehat{\Delta}_J\right\|_{X,n}^2 + 2\left\|\hat{\lambda}(J)\widehat{\Delta}_J - \Delta\right\|_{X,n}^2$$

gives

$$\begin{aligned}\left\|\hat{\lambda}\widehat{\Delta}_{\hat{J}} - \Delta\right\|_{X,n}^2 &\le (2+4/\delta)(\mathcal{C}_{\Delta,J} - \mathcal{C}_{\Delta,\hat{J}}) \\ &\quad + C_\eta\left[\widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}} + \left\|\Delta - \widehat{\Pi}_{J\wedge J_{\Delta,\max}}\Delta\right\|_{X,n}^2 + \mathcal{U}_J + \mathcal{U}_{\hat{J}}\right].\end{aligned} \tag{A.154}$$

Together with (A.147), this proves the same inequality for both $\hat{J} \le J$ and $\hat{J} > J$.

Combine two oracle comparisons

We combine the NP-OLS basic inequality (A.136) with the endogeneity component bound (A.154). Multiply (A.136) by $(2+4/\delta)$ and add (A.154). The terms $\mathcal{C}_{\Delta,\hat{J}} - \mathcal{C}_{\Delta,J}$ and $\mathcal{C}_{\Delta,J} - \mathcal{C}_{\Delta,\hat{J}}$ cancel exactly. Thus

$$\begin{aligned}&(2+4/\delta)\left\|\widehat{g}_{\hat{J}} - g_0\right\|_{X,n}^2 + \left\|\hat{\lambda}\widehat{\Delta}_{\hat{J}} - \Delta\right\|_{X,n}^2 \\ &\le (2+4/\delta)\left\|\widehat{g}_J - g_0\right\|_{X,n}^2 + (2+4/\delta)(\mathcal{Z}_{\hat{J}} - \mathcal{Z}_J) \\ &\quad + C_\eta\left[\widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}} + \left\|\Delta - \widehat{\Pi}_{J\wedge J_{\Delta,\max}}\Delta\right\|_{X,n}^2 + \mathcal{U}_J + \mathcal{U}_{\hat{J}}\right].\end{aligned} \tag{A.155}$$

This is the first pathwise oracle comparison. It bounds the estimation errors of both components at the selected $(\hat{\lambda}, \hat{J})$ by the NP-OLS estimation error, the inflated variance penalty, and the approximation error at an arbitrary fixed dimension J. The dependence on the selected dimension on the right-hand side is confined to $\mathcal{Z}_{\hat{J}} - \mathcal{Z}_J$ and $\mathcal{U}_{\hat{J}}$, whose expectations will be controlled below.

We next derive a complementary bound by comparing the selected estimator with the candidate $(0, J)$, which leaves the endogeneity bias uncorrected. Since $(\hat{\lambda}, \hat{J})$ minimizes the empirical selection criterion, comparison with $(0, J)$ gives that

$$\|\widehat{g}_{\hat{J}} - g_0\|_{X,n}^2 \leq \|\widehat{g}_J - g_0\|_{X,n}^2 + \mathcal{Z}_{\hat{J}} - \mathcal{Z}_J + \mathcal{C}_{\Delta,\hat{J}}. \tag{A.156}$$

We aim to claim the pathwise inequality

$$\left\|\hat{\lambda}\widehat{\Delta}_{\hat{J}} - \Delta\right\|_{X,n}^2 + (2 + 4/\delta)\mathcal{C}_{\Delta,\hat{J}} \leq C_\eta \left[\|\Delta\|_{X,n}^2 + \mathcal{U}_{\hat{J}}\right]. \tag{A.157}$$

Consider first the case $\widehat{B}_{\Delta,\hat{J}} = 0$, then $\hat{\lambda} = 0$ and (A.157) holds immediately.

Now consider $\widehat{B}_{\Delta,\hat{J}} > 0$. By definition, $\|\hat{\Delta}_{\hat{J}}\|_{X,n}^2 > \hat{\kappa}_{n,\hat{J}}\hat{V}_{\Delta,\hat{J}\wedge J_{\Delta,\max}}$. Young's inequality, the positive-part bound (A.143), and the contraction property of empirical orthogonal projection imply

$$\begin{aligned}\left\|\widehat{\Delta}_{\hat{J}}\right\|_{X,n}^2 &\leq (1 + \delta/2)\left\|\widehat{\Delta}_{\hat{J}} - \widehat{\Pi}_{\hat{J}\wedge J_{\Delta,\max}}\Delta\right\|_{X,n}^2 + (1 + 2/\delta)\left\|\widehat{\Pi}_{\hat{J}\wedge J_{\Delta,\max}}\Delta\right\|_{X,n}^2 \\ &\leq (1 - \delta/2)\widehat{\kappa}_{n,\hat{J}}\widehat{V}_{\Delta,\hat{J}\wedge J_{\Delta,\max}} + 2\mathcal{U}_{\hat{J}} + (1 + 2/\delta)\|\Delta\|_{X,n}^2. \end{aligned} \tag{A.158}$$

Combining the strict inequality above with (A.158) and rearranging yields

$$\frac{\delta}{2}\widehat{\kappa}_{n,\hat{J}}\widehat{V}_{\Delta,\hat{J}\wedge J_{\Delta,\max}} \leq 2\mathcal{U}_{\hat{J}} + (1 + 2/\delta)\|\Delta\|_{X,n}^2.$$

Since $\delta > 0$ depends only on the fixed constant η, it follows that $\hat{\kappa}_{n,\hat{J}}\hat{V}_{\Delta,\hat{J}\wedge J_{\Delta,\max}} \leq C_\eta[\|\Delta\|_{X,n}^2 + \mathcal{U}_{\hat{J}}]$. Substituting this bound into (A.158) and enlarging C_η if necessary, we obtain

$$\left\|\widehat{\Delta}_{\hat{J}}\right\|_{X,n}^2 \leq C_\eta \left[\|\Delta\|_{X,n}^2 + \mathcal{U}_{\hat{J}}\right].$$

Since $0 \leq \hat{\lambda} \leq 1$ and (A.133) implies that $0 \leq \mathcal{C}_{\Delta,\hat{J}} \leq \|\widehat{\Delta}_{\hat{J}}\|_{X,n}^2$, the triangle inequality gives

$$\left\|\hat{\lambda}\widehat{\Delta}_{\hat{J}} - \Delta\right\|_{X,n}^2 + (2 + 4/\delta)\mathcal{C}_{\Delta,\hat{J}} \leq (4 + 4/\delta)\left\|\widehat{\Delta}_{\hat{J}}\right\|_{X,n}^2 + 2\|\Delta\|_{X,n}^2 \leq C_\eta \left[\|\Delta\|_{X,n}^2 + \mathcal{U}_{\hat{J}}\right],$$

which establishes (A.157).

We can now complete the second oracle comparison. Multiplying (A.156) by $(2 + 4/\delta)$ and

adding (A.157) again cancels the selected profiled gain $\mathcal{C}_{\Delta,\hat{J}}$. Consequently,

$$
\begin{aligned}
&(2+4/\delta)\left\|\widehat{g}_{\hat{J}}-g_0\right\|_{X,n}^2+\left\|\hat{\lambda}\widehat{\Delta}_{\hat{J}}-\Delta\right\|_{X,n}^2 \\
&\leq (2+4/\delta)\left\|\widehat{g}_J-g_0\right\|_{X,n}^2+(2+4/\delta)(\mathcal{Z}_{\hat{J}}-\mathcal{Z}_J)+C_\eta\left[\|\Delta\|_{X,n}^2+\mathcal{U}_{\hat{J}}\right]. \qquad \text{(A.159)}
\end{aligned}
$$

Oracle risk bound

We now take expectations in (A.155) and (A.159). Fix $J \in \mathcal{J}_n$. For a given sieve dimension J, the usual risk bound for the sieve NP-OLS gives that

$$
E\left\|\widehat{g}_J-g_0\right\|_{X,n}^2 \lesssim V_{g,J}+\left\|g_0-\Pi_J g_0\right\|_X^2. \qquad \text{(A.160)}
$$

Next, since $\widehat{\Pi}_{J\wedge J_{\Delta,\max}}$ is the empirical least-squares projection onto $\mathcal{H}_{J\wedge J_{\Delta,\max}}$, the sieve approximation error of Δ satisfies

$$
E\left\|\Delta-\widehat{\Pi}_{J\wedge J_{\Delta,\max}}\Delta\right\|_{X,n}^2 \leq E\left\|\Delta-\Delta_{J\wedge J_{\Delta,\max}}\right\|_{X,n}^2=\left\|\Delta-\Delta_{J\wedge J_{\Delta,\max}}\right\|_X^2. \qquad \text{(A.161)}
$$

The uniform consistency of variance estimators and moment bounds of the error terms give that

$$
E\left[\widehat{\kappa}_{n,J}\widehat{V}_{\Delta,J\wedge J_{\Delta,\max}}\right] \lesssim \kappa_{n,J}V_{\Delta,J\wedge J_{\Delta,\max}}. \qquad \text{(A.162)}
$$

Applying the concentration inequalities developed in Section A.5, together with the corresponding empirical centering gives that

$$
E[\mathcal{U}_J+\mathcal{U}_{\hat{J}}] \lesssim V_{g,J}, \qquad \text{for } J \in \mathcal{J}_n. \qquad \text{(A.163)}
$$

Taking expectations in (A.155) and applying (A.137) with $\gamma = 1/2$, we obtain, by (A.160), (A.161), (A.162), and (A.163), that

$$
E\left[\left\|\widehat{g}_{\hat{J}}-g_0\right\|_{X,n}^2+\left\|\hat{\lambda}\widehat{\Delta}_{\hat{J}}-\Delta\right\|_{X,n}^2\right] \lesssim V_{g,J}+\left\|g_0-\Pi_J g_0\right\|_X^2+\kappa_{n,J}V_{\Delta,J\wedge J_{\Delta,\max}}+\left\|\Delta-\Delta_{J\wedge J_{\Delta,\max}}\right\|_X^2. \qquad \text{(A.164)}
$$

This inequality bounds the risk of the selected estimator by the NP-OLS estimation error and the additional endogeneity component term at a fixed sieve dimension J.

Similarly, taking expectations in (A.159) and applying (A.137) with $\gamma = 1/2$, we obtain, by (A.160), (A.163), and $E\|\Delta\|_{X,n}^2 = \rho_n$, that

$$
E\left[\left\|\widehat{g}_{\hat{J}}-g_0\right\|_{X,n}^2+\left\|\hat{\lambda}\widehat{\Delta}_{\hat{J}}-\Delta\right\|_{X,n}^2\right] \lesssim V_{g,J}+\left\|g_0-\Pi_J g_0\right\|_X^2+\rho_n \qquad \text{(A.165)}
$$

Since the left-hand sides of (A.164) and (A.165) are identical, taking the smaller of their right-hand sides gives

$$E\left[\left\|\widehat{g}_{\hat{J}}-g_0\right\|_{X,n}^2+\left\|\hat{\lambda}\widehat{\Delta}_{\hat{J}}-\Delta\right\|_{X,n}^2\right]$$

$$\lesssim V_{g,J}+\left\|g_0-\Pi_J g_0\right\|_X^2+\min\left\{\rho_n,\ \kappa_{n,J}V_{\Delta,J\wedge J_{\Delta,\max}}+\left\|\Delta-\Delta_{J\wedge J_{\Delta,\max}}\right\|_X^2\right\}. \tag{A.166}$$

The minimum reflects the trade-off between leaving endogeneity uncorrected and estimating its component using NPIV. The inflation term $\kappa_{n,J}$ enters only the variance term associated with estimating the endogeneity bias.

Adaptation result

We now translate (A.166) into a risk bound for the structural function. Since $h_0=g_0-\Delta$ and $\widehat{h}_{\hat{\lambda},\hat{J}}=\widehat{g}_{\hat{J}}-\hat{\lambda}\widehat{\Delta}_{\hat{J}}$, the squared triangle inequality gives

$$\left\|\widehat{h}_{\hat{\lambda},\hat{J}}-h_0\right\|_{X,n}^2\le 2\left\|\widehat{g}_{\hat{J}}-g_0\right\|_{X,n}^2+2\left\|\hat{\lambda}\widehat{\Delta}_{\hat{J}}-\Delta\right\|_{X,n}^2.$$

Thus, (A.166) provides a bound on the expected empirical risk of the selected estimator.

To obtain the population risk, we use the population projection $\Pi_J h_0$ as an intermediate approximation. Fix the same deterministic sieve dimension $J\in\mathcal{J}_n$. Since the sieve spaces are nested, $\widehat{h}_{\hat{\lambda},\hat{J}}$ and $\Pi_J h_0$ both belong to $\mathcal{H}_{J\vee\hat{J}}$.

We can therefore apply the uniform Gram-matrix concentration bound (A.23) to compare the population and empirical norms of the difference $\widehat{h}_{\hat{\lambda},\hat{J}}-\Pi_J h_0$. On the event where the empirical Gram matrices are uniformly well conditioned, the two-sided norm comparison gives, for all sufficiently large n,

$$\left\|\widehat{h}_{\hat{\lambda},\hat{J}}-\Pi_J h_0\right\|_X^2\le 2\left\|\widehat{h}_{\hat{\lambda},\hat{J}}-\Pi_J h_0\right\|_{X,n}^2.$$

Applying the squared triangle inequality first in the population norm and then in the empirical norm yields

$$\left\|\widehat{h}_{\hat{\lambda},\hat{J}}-h_0\right\|_X^2\le 8\left\|\widehat{h}_{\hat{\lambda},\hat{J}}-h_0\right\|_{X,n}^2+8\left\|h_0-\Pi_J h_0\right\|_{X,n}^2+2\left\|h_0-\Pi_J h_0\right\|_X^2.$$

This inequality bounds the population estimation error by its empirical counterpart and the sieve approximation error of h_0. The approximation error is evaluated at the fixed dimension J. Therefore, $E\left\|h_0-\Pi_J h_0\right\|_{X,n}^2=\left\|h_0-\Pi_J h_0\right\|_X^2$. Taking expectations, controlling the complementary

Gram-matrix event, and applying (A.166), we obtain

$$E\left\|\widehat{h}_{\hat{\lambda},\hat{J}}-h_0\right\|_X^2 \lesssim V_{g,J}+\left\|g_0-\Pi_J g_0\right\|_X^2+\left\|h_0-\Pi_J h_0\right\|_X^2+\min\left\{\rho_n,\ \kappa_{n,J}V_{\Delta,J\wedge J_{\Delta,\max}}+\left\|\Delta-\Delta_{J\wedge J_{\Delta,\max}}\right\|_X^2\right\}.$$

Under Assumption 4.1, the sieve approximation errors satisfy $\|g_0-\Pi_J g_0\|_X^2+\|h_0-\Pi_J h_0\|_X^2 \lesssim J^{-2s/d_x}$. Furthermore, since $\Delta=g_0-h_0$, it follows that $\|\Delta-\Delta_{J\wedge J_{\Delta,\max}}\|_X^2 \lesssim (J\wedge J_{\Delta,\max})^{-2s/d_x}$. Since $J\in\mathcal{J}_n$ is deterministic, we conclude that

$$E\left\|\widehat{h}_{\hat{\lambda},\hat{J}}-h_0\right\|_X^2 \lesssim \inf_{J\in\mathcal{J}_n}\left[V_{g,J}+J^{-2s/d_x}+\min\left\{\rho_n,\ \kappa_{n,J}V_{\Delta,J\wedge J_{\Delta,\max}}+(J\wedge J_{\Delta,\max})^{-2s/d_x}\right\}\right]. \quad \text{(A.167)}$$

By the definition of $\kappa_{n,J}$ and the variance bounds (A.22) and (A.45),

$$\kappa_{n,J}\lesssim \log\left(e+\frac{1}{J}\sum_{j=1}^{J}\nu_j^{-2}\right)+\log(2|\mathcal{J}_n|).$$

Under mild ill-posedness $\nu_j\asymp j^{-\varsigma/d_x}$ and severe ill-posedness $\nu_j\asymp\exp(-j^{\varsigma/d_x}/2)$, respectively, this yields

$$\kappa_{n,J}\lesssim\begin{cases}\log(e+J)+\log(2|\mathcal{J}_n|), & \text{mild ill-posedness,}\\ 1+J^{\varsigma/d_x}+\log(2|\mathcal{J}_n|), & \text{severe ill-posedness.}\end{cases}$$

Since the geometric grid satisfies $|\mathcal{J}_n|\lesssim\log n$, substituting these bounds into (A.167) and minimizing over J gives the rates in Theorem 5.1.

Appendix B More Technical Lemmas

Proposition B.1. *Suppose that Assumptions A.1, 4.1 and 4.4 hold. For every $t \geq 1$, uniformly over $F_n \in \mathcal{F}(\rho_n)$, the following inequalities hold simultaneously with probability at least $1 - C\exp(-t)$:*

$$\max_{\substack{J \in \mathcal{J}_n \\ J \leq J_{\Delta,\max}}} \left\| G_X^{-1/2} \widehat{G}_X G_X^{-1/2} - I_J \right\|_{\text{op}} \lesssim \varrho_{\Delta,n}(t)$$

$$\left\| G_{Z,\Delta}^{-1/2} \widehat{G}_{Z,\Delta} G_{Z,\Delta}^{-1/2} - I_{K_{\Delta,\max}} \right\|_{\text{op}} \lesssim \varrho_{\Delta,n}(t), \qquad \left\| G_{Z,\Delta}^{-1/2} (\widehat{S}_\Delta - S_\Delta) G_{X,\Delta}^{-1/2} \right\|_{\text{op}} \lesssim \varrho_{\Delta,n}(t) \tag{B.1}$$

where

$$\varrho_{\Delta,n}(t) = \zeta_\Delta \sqrt{\frac{\log J_{\Delta,\max} + t}{n}} + \zeta_\Delta^2 \frac{\log J_{\Delta,\max} + t}{n}.$$

Proof of Proposition B.1. It can be derived from the matrix Bernstein inequalities in Lemma F.7 of Chen and Christensen (2018). ■

Proposition B.2. *Suppose that Assumptions A.1, A.2, 4.1 and 4.4 hold. Then, uniformly over $F_n \in \mathcal{F}(\rho_n)$, the following inequalities hold simultaneously with probability at least $1 - C\exp(-t)$:*

$$\max_{\substack{J \in \mathcal{J}_n \\ J \leq J_{\Delta,\max}}} \left\| G_X^{1/2} (\widehat{G}_X^{-1} - G_X^{-1}) G_X^{1/2} \right\|_{\text{op}} \lesssim \varrho_{\Delta,n}(t) \tag{B.2}$$

$$\left\| G_{Z,\Delta}^{-1/2} S_\Delta \left[\left(\widehat{G}_{Z,\Delta}^{-1/2} \widehat{S}_\Delta \right)_l^- \widehat{G}_{Z,\Delta}^{-1/2} G_{Z,\Delta}^{1/2} - \left(G_{Z,\Delta}^{-1/2} S_\Delta \right)_l^- \right] \right\|_{\text{op}} \lesssim \sigma_\Delta^{-1} \varrho_{\Delta,n}(t). \tag{B.3}$$

Proof of Proposition B.2. It follows by adapting the matrix perturbation argument from Lemma F.10 of Chen and Christensen (2018), together with Proposition B.1. ■

Proposition B.3. *Suppose that Assumptions A.1, 4.1, 4.3, 4.4 and 5.1 hold. For every $J \in \mathcal{J}_n$ with $J \leq J_{\Delta,\max}$ and every $t \geq 1$, uniformly over $F_n \in \mathcal{F}(\rho_n)$, the following inequalities hold simultaneously with probability at least $1 - C\exp(-t)$:*

$$\left\| G_X^{-1/2} \frac{1}{n} \sum_{i=1}^n p^J(X_i) \left(Y_i - \Pi_J g_0(X_i) \right) \right\| \lesssim \varrho_{g,J,n}(t), \tag{B.4}$$

$$\left\| G_{Z,\Delta}^{-1/2} \frac{1}{n} \sum_{i=1}^n q^{K_{\Delta,\max}}(Z_i) \left(Y_i - \Xi_{J_{\Delta,\max}} h_0(X_i) \right) \right\| \lesssim \varrho_{\text{IV},\Delta,n}(t), \tag{B.5}$$

where $\varrho_{g,J,n}(t)$ and $\varrho_{\text{IV},\Delta,n}(t)$ are defined in (A.107) and (A.108), respectively.

Proof of Proposition B.3. We apply the scalar Bernstein inequality (Massart, 2007, Proposition 2.9) to the sieve scores. As established in the proof of (A.25), it follows that $E\left[p^J(X_i)\left(Y_i - \Pi_J g_0(X_i)\right)\right] = 0$. Moreover, by (A.93), for every integer $k \geq 2$ and every unit vector $a \in \mathbb{R}^J$,

$$E\left|a' G_X^{-1/2} p^J(X_i)\left(Y_i - \Pi_J g_0(X_i)\right)\right|^k \lesssim \frac{k!}{2}(C\zeta_{X,J})^{k-2},$$

where we use $\|G_X^{-1/2} p^J(X_i)\| \leq \zeta_{X,J}$ and $E[(a' G_X^{-1/2} p^J(X_i))^2] = 1$.

Thus, the scalar summands satisfy the moment conditions (2.17)–(2.18) of Massart (2007, Proposition 2.9). Applying this result gives that

$$\left\| G_X^{-1/2} \frac{1}{n} \sum_{i=1}^n p^J(X_i)\left(Y_i - \Pi_J g_0(X_i)\right) \right\| \lesssim \sqrt{\frac{J+t}{n}} + \zeta_{X,J}\frac{J+t}{n} \tag{B.6}$$

with probability at least $1 - C\exp(-t)$.

Next, by the nestedness of the sieve spaces, $\zeta_{X,J}\sqrt{J/n} = O(1)$ uniformly over $J \leq J_{\Delta,\max}$. Consequently, $\zeta_{X,J} J/n \lesssim \sqrt{J/n}$, so the preceding bound (B.6) reduces to

$$\left\| G_X^{-1/2} \frac{1}{n} \sum_{i=1}^n p^J(X_i)\left(Y_i - \Pi_J g_0(X_i)\right) \right\| \lesssim \sqrt{\frac{J+t}{n}} + \zeta_{X,J}\frac{t}{n} = \varrho_{g,J,n}(t),$$

which establishes (B.4).

For the second inequality, recall the decomposition (A.32)–(A.33). We first control its centered component (A.32). By (A.94), the same Bernstein moment argument gives that

$$E\left|a' G_{Z,\Delta}^{-1/2} q^{K_{\Delta,\max}}(Z_i)\left(Y_i - \Xi_{J_{\Delta,\max}} h_0(X_i)\right)\right|^k \lesssim \frac{k!}{2}(C\zeta_\Delta)^{k-2}.$$

Applying Bernstein's inequality yields

$$\left\| G_{Z,\Delta}^{-1/2} \frac{1}{n} \sum_{i=1}^n \left\{ q^{K_{\Delta,\max}}(Z_i)\left(Y_i - \Xi_{J_{\Delta,\max}} h_0(X_i)\right) - E\left[q^{K_{\Delta,\max}}(Z_i)\left(Y_i - \Xi_{J_{\Delta,\max}} h_0(X_i)\right)\right] \right\} \right\|$$

$$\lesssim \sqrt{\frac{K_{\Delta,\max}+t}{n}} + \zeta_\Delta \frac{K_{\Delta,\max}+t}{n} \tag{B.7}$$

with probability at least $1 - C\exp(-t)$.

By the pilot sieve restriction (A.16), $\zeta_\Delta\sqrt{K_{\Delta,\max}/n} = O(1)$. Hence, the preceding bound (B.7) reduces to

$$\sqrt{\frac{K_{\Delta,\max}+t}{n}} + \zeta_\Delta \frac{t}{n}. \tag{B.8}$$

It remains to bound the population component (A.33). As established following (A.33), the IV

restriction and Assumption 4.4 (ii) imply that

$$\left\| G_{Z,\Delta}^{-1/2} E\left[q^{K_{\Delta,\max}}(Z_i)\left(Y_i - \Xi_{J_{\Delta,\max}} h_0(X_i)\right)\right] \right\| = \left\| \Pi_{K_{\Delta,\max}} T\left(h_0 - \Xi_{J_{\Delta,\max}} h_0\right)\right\|_Z \lesssim \sigma_\Delta J_{\Delta,\max}^{-s/d_x}.$$

Combining this deterministic bound with the concentration bound Equations (B.7) and (B.8) establishes (B.5). ■

Proposition B.4. *Suppose that Assumptions A.1, A.2, 4.1 to 4.4 and 5.1 hold. In addition, suppose that the corresponding smoothness restriction in (B.11) or (B.12) holds. For every $J \in \mathcal{J}_n$ with $J \le J_{\Delta,\max}$ and every $1 \le t \le C \log n$, uniformly over $F_n \in \mathcal{F}(\rho_n)$, the following inequalities hold simultaneously with probability at least $1 - C\exp(-t)$:*

$$\left\| \frac{1}{n}\sum_{i=1}^n \xi_{\mathrm{IV},J_{\Delta,\max}}(\mathcal{D}_i) \right\|_X \lesssim \sqrt{\frac{t}{n}\sum_{j=1}^{J_{\Delta,\max}} \nu_j^{-2}} + \sigma_\Delta^{-1}\zeta_\Delta \frac{t}{n}, \tag{B.9}$$

$$\left\| \left(\widehat{\Pi}_J - \Pi_J\right) \Xi_{J_{\Delta,\max}} h_0 \right\|_X \lesssim \zeta_{X,J} J^{-s/d_x}\sqrt{\frac{t}{n}} + \zeta_{X,J}\frac{t}{n}. \tag{B.10}$$

Proof of Proposition B.4. We first establish (B.9). Recall from (A.27) that $\xi_{\mathrm{IV},J_{\Delta,\max}}(\mathcal{D}_i)$ is centered. By (A.80) with $J = J_{\Delta,\max}$, together with $\sigma_\Delta \asymp \nu_{J_{\Delta,\max}}$, we have

$$\left\| p^{J_{\Delta,\max}}(\cdot)' G_{X,\Delta}^{-1/2} A_\Delta q^{K_{\Delta,\max}}(Z_i) \right\|_X \lesssim \sigma_\Delta^{-1}\zeta_\Delta.$$

Moreover, (A.43) implies that

$$E\left\| p^{J_{\Delta,\max}}(\cdot)' G_{X,\Delta}^{-1/2} A_\Delta q^{K_{\Delta,\max}}(Z_i) \right\|_X^2 \asymp \sum_{j=1}^{J_{\Delta,\max}} \nu_j^{-2}.$$

Combining these bounds with (A.94) gives, for every integer $k \ge 2$,

$$E\left\| \xi_{\mathrm{IV},J_{\Delta,\max}}(\mathcal{D}_i) \right\|_X^k \lesssim \frac{k!}{2}\left(C\sigma_\Delta^{-1}\zeta_\Delta\right)^{k-2} \sum_{j=1}^{J_{\Delta,\max}} \nu_j^{-2}.$$

Thus, the pilot influence score satisfies the Bernstein moment condition in Lemma A.1. Applying that lemma with $W_i = n^{-1}\xi_{\mathrm{IV},J_{\Delta,\max}}(\mathcal{D}_i)$ and $\mathbb{H} = \mathcal{H}_{J_{\Delta,\max}}$, using $B^2 \asymp \frac{1}{n}\sum_{j=1}^{J_{\Delta,\max}} \nu_j^{-2}, \Gamma \asymp \sigma_\Delta^{-1}\zeta_\Delta \frac{1}{n}$, yields that

$$\left\| \frac{1}{n}\sum_{i=1}^n \xi_{\mathrm{IV},J_{\Delta,\max}}(\mathcal{D}_i) \right\|_X \lesssim \sqrt{\frac{t}{n}\sum_{j=1}^{J_{\Delta,\max}} \nu_j^{-2}} + \sigma_\Delta^{-1}\zeta_\Delta \frac{t}{n}$$

with probability at least $1 - C\exp(-t)$. This establishes (B.9).

We next establish (B.10). By the empirical projection representation (A.62),

$$\left(\widehat{\Pi}_J - \Pi_J\right) \Xi_{J_{\Delta,\max}} h_0 = p^J(\cdot)' \widehat{G}_X^{-1} \frac{1}{n} \sum_{i=1}^n p^J(X_i) \left[\Xi_{J_{\Delta,\max}} h_0(X_i) - \Pi_J \Xi_{J_{\Delta,\max}} h_0(X_i)\right].$$

Moreover, (A.61) implies that

$$\left\|\Xi_{J_{\Delta,\max}} h_0 - \Pi_J \Xi_{J_{\Delta,\max}} h_0\right\|_X \lesssim J^{-s/d_x}.$$

We also require a uniform envelope bound for the approximation residual. By (4.1) and the stated smoothness restrictions,

$$\sup_{J \geq 1} \left\|\Pi_J h_0\right\|_\infty \leq L_h \left(\sum_{j=1}^\infty j^{-2s/d_x} \|\psi_j\|_\infty^2\right)^{1/2} \lesssim 1.$$

Furthermore, (A.60) and the normalized basis-envelope condition give

$$\left\|\Xi_{J_{\Delta,\max}} h_0 - \Pi_{J_{\Delta,\max}} h_0\right\|_\infty \lesssim \zeta_\Delta J_{\Delta,\max}^{-s/d_x} = O(1).$$

Since the sieve spaces are nested, these bounds imply

$$\max_{\substack{J \in \mathcal{J}_n \\ J \leq J_{\Delta,\max}}} \left\|\Xi_{J_{\Delta,\max}} h_0 - \Pi_J \Xi_{J_{\Delta,\max}} h_0\right\|_\infty \lesssim 1.$$

Consequently, for every integer $k \geq 2$,

$$\begin{aligned}
&E\left\|G_X^{-1/2} p^J(X_i) \left[\Xi_{J_{\Delta,\max}} h_0(X_i) - \Pi_J \Xi_{J_{\Delta,\max}} h_0(X_i)\right]\right\|^k \\
&\leq \left(\zeta_{X,J} \left\|\Xi_{J_{\Delta,\max}} h_0 - \Pi_J \Xi_{J_{\Delta,\max}} h_0\right\|_\infty\right)^{k-2} \times \zeta_{X,J}^2 \left\|\Xi_{J_{\Delta,\max}} h_0 - \Pi_J \Xi_{J_{\Delta,\max}} h_0\right\|_X^2 \\
&\lesssim (C\zeta_{X,J})^{k-2} \zeta_{X,J}^2 J^{-2s/d_x} \leq \frac{k!}{2} (C\zeta_{X,J})^{k-2} \zeta_{X,J}^2 J^{-2s/d_x}.
\end{aligned}$$

Applying Lemma A.1 with $\mathbb{H} = \mathbb{R}^J$ yields

$$\left\|G_X^{-1/2} \frac{1}{n} \sum_{i=1}^n p^J(X_i) \left[\Xi_{J_{\Delta,\max}} h_0(X_i) - \Pi_J \Xi_{J_{\Delta,\max}} h_0(X_i)\right]\right\| \lesssim \zeta_{X,J} J^{-s/d_x} \sqrt{\frac{t}{n}} + \zeta_{X,J} \frac{t}{n}$$

with probability at least $1 - C\exp(-t)$.

Finally, by Proposition B.2, $\left\|G_X^{1/2} \widehat{G}_X^{-1} G_X^{1/2}\right\|_{\text{op}} \lesssim 1$ on the Gram-matrix concentration event.

Combining this bound with (A.62) and the preceding Bernstein inequality gives

$$\left\| \left(\widehat{\Pi}_J - \Pi_J \right) \Xi_{J_{\Delta,\max}} h_0 \right\|_X \lesssim \zeta_{X,J} J^{-s/d_x} \sqrt{\frac{t}{n}} + \zeta_{X,J} \frac{t}{n},$$

which establishes (B.10). ■

Proposition B.5. *Suppose that Assumptions 4.1 and 4.2 and Assumption 4.4(iii)–(iv) hold, and that $J \leq K \leq C_K J$. In the mildly ill-posed case, $\nu_j \asymp j^{-\varsigma/d_x}$, assume that*

$$s > \begin{cases} d_x, & \zeta_J = O(\sqrt{J}), \\ \frac{3}{2} d_x, & \zeta_J = O(J). \end{cases} \tag{B.11}$$

In the severely ill-posed case, $\nu_j \asymp \exp(-j^{\varsigma/d_x}/2)$, assume that

$$s > \begin{cases} d_x + \frac{1}{2}\min\{\varsigma, d_x\}, & \zeta_J = O(\sqrt{J}), \\ \frac{3}{2} d_x + \frac{1}{2}\min\{\varsigma, d_x\}, & \zeta_J = O(J). \end{cases} \tag{B.12}$$

Then the deterministic pilot defined in (A.14) satisfies Assumption A.1.

Proof of Proposition B.5. First, we check the first condition in (A.16). By the definition of $J_{\Delta,\max}$,

$$C_\Delta \, \zeta_\Delta \sqrt{\frac{J_{\Delta,\max} K_{\Delta,\max} \log J_{\Delta,\max}}{n}} < \sigma_\Delta^2 \sqrt{\operatorname{tr}\left(A_\Delta G_{Z,\Delta} A_\Delta' \right)}.$$

Dividing both sides by $\sigma_\Delta^2 \sqrt{J_{\Delta,\max}}$ gives

$$\frac{\sigma_\Delta^{-2} \zeta_\Delta \sqrt{K_{\Delta,\max} \log J_{\Delta,\max}/n}}{\left(J_{\Delta,\max}^{-1} \sum_{j=1}^{J_{\Delta,\max}} \nu_j^{-2} \right)^{1/2}} < \frac{1}{C_\Delta} \left[\frac{\operatorname{tr}\left(A_\Delta G_{Z,\Delta} A_\Delta' \right)}{\sum_{j=1}^{J_{\Delta,\max}} \nu_j^{-2}} \right]^{1/2} = O(1),$$

where the last step follows from $\operatorname{tr}(A G_Z A') \asymp \sum_{j=1}^{J} \nu_j^{-2}$. This establishes the first rate condition.

We verify the second condition in (A.16). First consider the mildly ill-posed case $\nu_j \asymp j^{-\varsigma/d_x}$. Then

$$\sum_{j=1}^{J} \nu_j^{-2} \asymp J^{1+2\varsigma/d_x}.$$

and hence, using $\sigma_\Delta \asymp \nu_{J_{\Delta,\max}}$,

$$\frac{\sigma_\Delta^{-1} \zeta_\Delta J_{\Delta,\max}^{-s/d_x} \sqrt{\log J_{\Delta,\max}}}{\left(J_{\Delta,\max}^{-1} \sum_{j=1}^{J_{\Delta,\max}} \nu_j^{-2} \right)^{1/2}} \lesssim \zeta_\Delta J_{\Delta,\max}^{-s/d_x} \sqrt{\log J_{\Delta,\max}}.$$

If $\zeta_J = O(\sqrt{J})$, the right-hand side is

$$O\left(J_{\Delta,\max}^{1/2-s/d_x}\sqrt{\log J_{\Delta,\max}}\right) = o(1)$$

under (B.11). If $\zeta_J = O(J)$, it is

$$O\left(J_{\Delta,\max}^{1-s/d_x}\sqrt{\log J_{\Delta,\max}}\right) = o(1).$$

Thus the second condition in (A.16) holds in the mildly ill-posed case.

Now consider the severely ill-posed case $\nu_j \asymp \exp(-j^{\varsigma/d_x}/2)$. A sum–integral comparison when $\varsigma < d_x$, and domination by the final term when $\varsigma \geq d_x$, give

$$\sum_{j=1}^{J}\nu_j^{-2} \asymp \exp\left(J^{\varsigma/d_x}\right)J^{(1-\varsigma/d_x)_+}. \tag{B.13}$$

It follows that

$$\left(J^{-1}\sum_{j=1}^{J}\nu_j^{-2}\right)^{1/2} \asymp \nu_J^{-1}J^{-\frac{1}{2}\min\{\varsigma/d_x,1\}}.$$

Using again $\sigma_\Delta \asymp \nu_{J_{\Delta,\max}}$, we obtain

$$\frac{\sigma_\Delta^{-1}\zeta_\Delta J_{\Delta,\max}^{-s/d_x}\sqrt{\log J_{\Delta,\max}}}{\left(J_{\Delta,\max}^{-1}\sum_{j=1}^{J_{\Delta,\max}}\nu_j^{-2}\right)^{1/2}} \lesssim \zeta_\Delta J_{\Delta,\max}^{-s/d_x+\frac{1}{2}\min\{\varsigma/d_x,1\}}\sqrt{\log J_{\Delta,\max}}.$$

Under (B.12), this expression is $o(1)$ for either basis-envelope condition. The second condition in (A.16) therefore holds in the severely ill-posed case as well.

It remains to verify (A.17). Maximality of $J_{\Delta,\max}$ is crucial for this step. Since $J_{\Delta,\max}+1$ does not belong to the set in (A.14),

$$\sigma_{JK}^2\sqrt{\mathrm{tr}(AG_ZA')} \leq C_\Delta\,\zeta_J\sqrt{\frac{JK\log J}{n}} \qquad \text{at } J = J_{\Delta,\max}+1.$$

Consider first the mildly ill-posed case. Since

$$\nu_J\left(\sum_{j=1}^{J}\nu_j^{-2}\right)^{1/2} \asymp J^{1/2},$$

together with $\sigma_{JK} \asymp \nu_J$, $\mathrm{tr}(AG_ZA') \asymp \sum_{j=1}^{J}\nu_j^{-2}$, and $K \asymp J$, implies

$$\sqrt{n}\,\nu_{J_{\Delta,\max}+1} \lesssim \zeta_{J_{\Delta,\max}+1}J_{\Delta,\max}^{1/2}\sqrt{\log J_{\Delta,\max}}.$$

Therefore,

$$\max_{\substack{J\in\mathcal{J}_n\\ J\le J_{\Delta,\max}}} \frac{\sqrt{n}\,\nu_{J_{\Delta,\max}+1}J_{\Delta,\max}^{-s/d_x}}{\nu_J\left(\sum_{j=1}^J \nu_j^{-2}\right)^{1/2}} \lesssim \zeta_{J_{\Delta,\max}+1}J_{\Delta,\max}^{1/2-s/d_x}\sqrt{\frac{\log J_{\Delta,\max}}{\underline{J}}}. \tag{B.14}$$

If $\zeta_J = O(\sqrt{J})$, this is

$$O\left(J_{\Delta,\max}^{1-s/d_x}\sqrt{\frac{\log J_{\Delta,\max}}{\underline{J}}}\right) = o(1)$$

when $s > d_x$. If $\zeta_J = O(J)$, it is

$$O\left(J_{\Delta,\max}^{3/2-s/d_x}\sqrt{\frac{\log J_{\Delta,\max}}{\underline{J}}}\right) = o(1)$$

when $s > 3d_x/2$. This proves (A.17) in the mildly ill-posed case.

In the severely ill-posed case, (B.13) implies

$$\nu_J\left(\sum_{j=1}^J \nu_j^{-2}\right)^{1/2} \asymp J^{(1-\varsigma/d_x)_+/2}.$$

The maximality inequality at $J = J_{\Delta,\max}+1$, together with $K \asymp J$, therefore gives

$$\sqrt{n}\,\nu_{J_{\Delta,\max}+1} \lesssim \zeta_{J_{\Delta,\max}+1}J_{\Delta,\max}^{\frac{1}{2}+\frac{1}{2}\min\{\varsigma/d_x,1\}}\sqrt{\log J_{\Delta,\max}}.$$

Hence,

$$\max_{\substack{J\in\mathcal{J}_n\\ J\le J_{\Delta,\max}}} \frac{\sqrt{n}\,\nu_{J_{\Delta,\max}+1}J_{\Delta,\max}^{-s/d_x}}{\nu_J\left(\sum_{j=1}^J \nu_j^{-2}\right)^{1/2}} \lesssim \frac{\zeta_{J_{\Delta,\max}+1}J_{\Delta,\max}^{\frac{1}{2}+\frac{1}{2}\min\{\varsigma/d_x,1\}-s/d_x}\sqrt{\log J_{\Delta,\max}}}{\underline{J}^{(1-\varsigma/d_x)_+/2}}.$$

If $\zeta_J = O(\sqrt{J})$, the numerator is of order

$$J_{\Delta,\max}^{1+\frac{1}{2}\min\{\varsigma/d_x,1\}-s/d_x}\sqrt{\log J_{\Delta,\max}},$$

which converges to zero under (B.12). If $\zeta_J = O(J)$, it is of order

$$J_{\Delta,\max}^{\frac{3}{2}+\frac{1}{2}\min\{\varsigma/d_x,1\}-s/d_x}\sqrt{\log J_{\Delta,\max}},$$

which also converges to zero under (B.12). Hence (A.17) holds in the severely ill-posed case.

Therefore the deterministic pilot defined in (A.14) satisfies Assumption A.1. ■

References

Baraud, Y. (2000): "Model selection for regression on a fixed design," *Probability Theory and Related Fields*, 117, 467–493.

Barron, A., L. Birgé, and P. Massart (1999): "Risk bounds for model selection via penalization," *Probability theory and related fields*, 113, 301–413.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015): "Some new asymptotic theory for least squares series: Pointwise and uniform results," *Journal of Econometrics*, 186, 345–366.

Birgé, L. and P. Massart (1997): "From model selection to adaptive estimation," in *Festschrift for Lucien Le Cam: Research papers in probability and statistics*, Springer, 55–87.

Breunig, C. and X. Chen (2024): "Adaptive, Rate-Optimal Hypothesis Testing in Nonparametric IV Models," *Econometrica*, 92, 2027–2067.

Cavalier, L. and Y. Golubev (2006): "Risk Hull Method and Regularization by Projections of Ill-Posed Inverse Problems," *The Annals of Statistics*, 34, 1653–1677.

Chen, X. and T. M. Christensen (2015b): "Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions," *Journal of Econometrics*, 188, 447–465.

——— (2018): "Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression," *Quantitative Economics*, 9, 39–84.

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

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

Kreyszig, E. (1978): *Introductory Functional Analysis with Applications*, John Wiley & Sons.

Loubes, J.-M. and C. Marteau (2009): "Oracle inequality for instrumental variable regression," *arXiv preprint arXiv:0901.4321*.

Massart, P. (2007): *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*, Springer.

Pinelis, I. (1994): "Optimum bounds for the distributions of martingales in Banach spaces," *The Annals of Probability*, 1679–1706.

——— (1999): "Correction:"Optimum bounds for the distributions of martingales in Banach spaces"[Ann. Probab. 22 (1994), no. 4, 1679–1706; MR 96b: 60010]," *The Annals of Probability*, 27, 2119–2119.

Tsybakov, A. B. (2009): *Introduction to Nonparametric Estimation*, Springer Series in Statistics, New York: Springer.