ADAPTIVE NONPARAMETRIC INSTRUMENTAL VARIABLES ESTIMATION:
EMPIRICAL CHOICE OF THE REGULARIZATION PARAMETER

by

Joel L. Horowitz
Department of Economics
Northwestern University
Evanston, IL 60208
USA

July 2010

ABSTRACT
In nonparametric instrumental variables estimation, the mapping that identifies the function of interest, \( g \) say, is discontinuous and must be regularized (that is, modified) to make consistent estimation possible. The amount of modification is controlled by a regularization parameter. The optimal value of this parameter depends on unknown population characteristics and cannot be calculated in applications. Theoretically justified methods for choosing the regularization parameter empirically in applications are not yet available. This paper presents, apparently for the first time, such a method for use in series estimation, where the regularization parameter is the number of terms in a series approximation to \( g \). The method does not require knowledge of the smoothness of \( g \) or of other unknown functions. It adapts to their unknown smoothness. The estimator of \( g \) based on the empirically selected regularization parameter converges in probability at a rate that is at least as fast as the asymptotically optimal rate multiplied by \( (\log n)^{1/2} \), where \( n \) is the sample size. The asymptotic integrated mean-square error (AIMSE) of the estimator is within a specified factor of the optimal AIMSE.

Key words: Ill-posed inverse problem, regularization, sieve estimation, series estimation, nonparametric estimation

JEL Classification: C13, C14, C21

I thank Xiaohong Chen, Joachim Freyberger, Sokbae Lee, and Vladimir Spokoiny for helpful discussions. This research was supported in part by NSF grant SES-0817552.
ADAPTIVE NONPARAMETRIC INSTRUMENTAL VARIABLES ESTIMATION: EMPIRICAL CHOICE OF THE REGULARIZATION PARAMETER

1. INTRODUCTION

This paper is about estimating the unknown function $g$ in the model

\begin{align}
Y &= g(X) + U; \quad E(U \mid W = w) = 0 \\
\end{align}

for almost every $w$ or, equivalently,

\begin{align}
E[Y - g(X) \mid W = w] = 0
\end{align}

for almost every $w$. In this model, $g$ is a function that satisfies regularity conditions but is otherwise unknown, $Y$ is a scalar dependent variable, $X$ is a continuously distributed explanatory variable that may be correlated with $U$ (that is, $X$ may be endogenous), $W$ is a continuously distributed instrument for $X$, and $U$ is an unobserved random variable. The data are an independent random sample of $(Y, X, W)$. The paper presents, apparently for the first time, a theoretically justified, empirical method for choosing the regularization parameter that is needed for estimation of $g$.

Existing nonparametric estimators of $g$ in (1.1)-(1.2) can be divided into two main classes: sieve (or series) estimators and kernel estimators. Sieve estimators have been developed by Newey and Powell (2003); Blundell, Chen, and Kristensen (2007); and Horowitz (2009). Kernel estimators have been developed by Hall and Horowitz (2005) and Darolles, Florens, and Renault (2006). Florens and Simoni (2010) describe a quasi-Bayesian estimator based on kernels. Hall and Horowitz (2005) and Chen and Reiss (2007) found the optimal rate of convergence of an estimator of $g$. Horowitz (2007) gave conditions for asymptotic normality of the estimator of Hall and Horowitz (2005). Horowitz and Lee (2010) showed how to use the sieve estimator of Horowitz (2009) to construct uniform confidence bands for $g$. Newey, Powell, and Vella (1999) present a control function approach to estimating $g$ in a model that is different from (1.1)-(1.2) but allows endogeneity of $X$ and achieves identification through an instrument. The control function model is non-nested with (1.1)-(1.2) and is not discussed further in this paper. Chernozhukov, Imbens, and Newey (2007); Horowitz and Lee (2007); and Chernozhukov, Gagliardini, and Scaillet (2008) have developed methods for estimating a quantile-regression version of model (1.1)-(1.2). Chen and Pouzo (2008, 2009) developed a method for estimating a large class of nonparametric and semiparametric conditional moment models with possibly non-smooth moments. This class includes the quantile-regression version of (1.1)-(1.2).
As is explained further in Section 2 of this paper, the relation that identifies $g$ in (1.1)-(1.2) creates an ill-posed inverse problem. That is, the mapping from the population distribution of $(Y, X, W)$ to $g$ is discontinuous. Consequently, $g$ cannot be estimated consistently by replacing unknown population quantities in the identifying relation with consistent estimators. To achieve a consistent estimator it is necessary to regularize (or modify) the mapping that identifies $g$. The amount of modification is controlled by a parameter called the regularization parameter. The optimal value of the regularization parameter depends on unknown population characteristics and, therefore, cannot be calculated in applications. Although there have been proposals of informal rules-of-thumb for choosing the regularization parameter in applications, theoretically justified empirical methods are not yet available.

This paper presents an empirical method for choosing the regularization parameter in sieve or series estimation, where the regularization parameter is the number of terms in the series approximation to $g$. The method does not require a priori knowledge of the smoothness of $g$ or of other unknown functions. It adapts to their unknown smoothness. The series estimator of $g$ based on the empirically selected regularization parameter also adapts to unknown smoothness. It converges in probability at a rate that is at least as fast as the asymptotically optimal rate multiplied by $(\log n)^{1/2}$, where $n$ is the sample size. Moreover, its asymptotic integrated mean-square error (AIMSE) is within a specified factor of the optimal AIMSE. The paper does not address question of whether the factor of $(\log n)^{1/2}$ can be removed or is an unavoidable price that must be paid for adaptation. This question is left for future research.

Section 2 of the paper provides background on the estimation problem and the series estimator that is used with the adaptive estimation procedure. This section also reviews the statistics literature on selecting the regularization parameter in inverse problems. The problems treated in the statistics literature are simpler than (1.1)-(1.2). Section 3 describes the proposed method for selecting the regularization parameter. Section 4 presents the results of Monte Carlo experiments that explore the finite-sample performance of the method. Section 5 presents concluding comments. All proofs are in the appendix.

2. BACKGROUND

This section explains the estimation problem and the need for regularization, outlines the sieve estimator that is used with the adaptive estimation procedure, and reviews the relevant statistics literature on selecting the regularization parameter.
2.1 The Estimation Problem and the Need for Regularization

Let $X$ and $W$ be continuously distributed random variables. Assume that the supports of $X$ and $W$ are $[0, 1]$. This assumption does not entail a loss of generality, because it can be satisfied by, if necessary, carrying out monotone increasing transformations of $X$ and $W$. Let $f_{XW}$ and $f_W$, respectively, denote the probability density functions of $(X, W)$ and $W$. Define 

$$m(w) = E(Y | W = w) f_W(w).$$

Let $L_2[0,1]$ be the space of real-valued, square-integrable functions on $[0, 1]$. Define the operator $A$ from $L_2[0,1] \rightarrow L_2[0,1]$ by 

$$(Av)(w) = \int_{[0,1]} v(x) f_{XW}(x, w) dx,$$

where $v$ is any function in $L_2[0,1]$. Then $g$ in (1.1)-(1.2) satisfies 

$$Ag = m .$$

Assume that $A$ is one-to-one, which is a necessary condition for identification of $g$. Then

$$g = A^{-1} m .$$

If $f_{XW}^2$ is integrable on $[0,1]^2$, then zero is a limit point (and the only limit point) of the singular values of $A$ (and eigenvalues of $A^*A$, where the operator $A^*$ is the adjoint of $A$). Consequently, the singular values of $A^{-1}$ are unbounded, and $A^{-1}$ is a discontinuous operator. This is the ill-posed inverse problem. Because of this problem, $g$ could not be estimated consistently by replacing $m$ in (2.1) with a consistent estimator, even if $A$ were known. To estimate $g$ consistently, it is necessary to regularize (or modify) $A$ so as to remove the discontinuity of $A^{-1}$. A variety of regularization methods have been developed. See, for example, Engl, Hanke, and Neubauer (1996); Kress (1999); and Carrasco, Florens, and Renault (2007), among many others. The method used in this paper is series truncation, which is a modification of the Petrov-Galerkin regularization method that is well-known in the theory of integral equations. See, for example, Kress (1999, pp. 240-245). It amounts to approximating $A$ with a non-singular, finite-dimensional matrix. The singular values of this matrix are bounded away from zero, so the inverse of the approximating matrix is a continuous operator. The details of the method are described further in Section 2.2.
2.2 Sieve Estimation and Regularization by Series Truncation

The adaptive estimation procedure uses a modified version of Horowitz’s (2009) sieve estimator of \( g \). This section describes Horowitz’s (2009) estimator and the modification.

The estimator of \( g \) is defined in terms of series expansions of \( g \), \( m \), and \( A \). Let \( \{\psi_j : j = 1, 2, \ldots\} \) be a complete, orthonormal basis for \( L_2[0,1] \). The expansions are

\[
g(x) = \sum_{j=1}^{\infty} b_j \psi_j(x),
\]

\[
m(w) = \sum_{k=1}^{\infty} m_k \psi_k(w),
\]

and

\[
f_{XW}(x,w) = \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} c_{jk} \psi_j(x) \psi_k(w),
\]

where

\[
b_j = \int_{[0,1]} g(x) \psi_j(x) dx,
\]

\[
m_k = \int_{[0,1]} m(w) \psi_k(w) dw,
\]

and

\[
c_{jk} = \int_{[0,1]} f_{XW}(x,w) \psi_j(x) \psi_k(w) dxdw.
\]

To estimate \( g \), we need estimators of \( m_k \), \( c_{jk} \), \( m \), and \( f_{XW} \). Denote the data by \( \{Y_i, X_i, W_i : i = 1, \ldots, n\} \), where \( n \) is the sample size. The estimators of \( m_k \) and \( c_{jk} \), respectively, are

\[
\hat{m}_k = n^{-1} \sum_{i=1}^{n} Y_i \psi_k(W_i)
\]

and

\[
\hat{c}_{jk} = n^{-1} \sum_{i=1}^{n} \psi_j(X_i) \psi_k(W_i).
\]

The estimators of \( m \) and \( f_{XW} \), respectively, are

\[
\hat{m}(w) = \sum_{k=1}^{J} \hat{m}_k \psi_k(w)
\]

and
\[ f_{XW}(x, w) = \sum_{j=1}^{J_n} \sum_{k=1}^{J_n} \hat{c}_{jk} \psi_j(x) \psi_k(w), \]

where \( J_n \) is a nonstochastic series truncation point. It is assumed that as \( n \to \infty, J_n \to \infty \) at a rate that is specified in Section 3.1. Define the operator \( \hat{A} \) that estimates \( A \) by
\[
(\hat{A}v)(w) = \int_{[0,1]} v(x)f_{XW}(x, w)dx.
\]

For any function \( v : [0,1] \to \mathbb{R} \), define \( D_j v(x) = d^j v(x)/dx^j \) whenever the derivative exists. Define \( D_0 v = v \). Define the norm
\[
\|v\|_2^2 = \sum_{0 \leq j \leq s} \int_{[0,1]} [D_j v(x)]^2 dx.
\]

Define the function space
\[
\mathcal{H}_s = \{ v : [0,1] \to \mathbb{R} : \|v\|_{L^2} \leq C_0 \},
\]

where \( C_0 > 0 \) is a finite constant. For any integer \( J > 0 \), define the subset of \( \mathcal{H}_s \)
\[
\mathcal{H}_{J_s} = \left\{ v = \sum_{j=1}^{J} v_j \psi_j : \|v\|_s \leq C_0 \right\},
\]

where \( \{v_j : j = 1, \ldots, J\} \) are finite constants. Horowitz’s (2009) sieve estimator of \( g \) is defined as
\[
(2.2) \quad \hat{g} = \arg\min_{v \in \mathcal{H}_{J_s}} \|\hat{A}v - \hat{m}\|,
\]

where \( \|\| \) is the norm on \( L_2[0,1] \). Under the assumptions of Section 3, \( P(\hat{A} \hat{g} = \hat{m}) \to 1 \) as \( n \to \infty \). Therefore,
\[
\hat{g} = \hat{A}^{-1} \hat{m}
\]

with probability approaching 1 as \( n \to \infty \).

To obtain the modified estimator that is used with this paper’s adaptive estimation procedure, let \( \langle \cdot, \cdot \rangle \) denote the inner product in \( L_2[0,1] \). That is
\[
\langle v_1, v_2 \rangle = \int_{[0,1]} v_1(x)v_2(x)dx.
\]

For \( j = 1, \ldots, J_n \), define \( \bar{b}_j = \langle \hat{g}, \psi_j \rangle \). The \( \bar{b}_j \)'s are the generalized Fourier coefficients of \( \hat{g} \) with the basis functions \( \{\psi_j\} \). Let \( J \leq J_n \) be a positive integer. The modified estimator of \( g \) is
\[
(2.3) \quad \hat{g}_J = \sum_{j=1}^{J} \bar{b}_j \psi_j.
\]
The adaptive estimation procedure consists of choosing $J_n$ to be “large” in a sense that is defined in Section 3.1. Then $J$ in (2.3) is selected using the empirical procedure that is explained in Section 3.1. Let $\hat{J}$ denote the resulting value of $J$. The adaptive estimator of $g$ is $\hat{g}_{\hat{J}}$. Under the regularity conditions of Section 3.2, $\|\hat{g}_{\hat{J}} - g\|$ converges in probability to 0 at a rate that is at least as fast as the asymptotically optimal rate times $(\log n)^{1/2}$. Moreover, for any $\varepsilon > 0$, the AIMSE of $\hat{g}_{\hat{J}}$ is within a factor of $2 + (4 + \varepsilon)\log n$ of the optimal AIMSE. Achieving these results does not require knowledge of $s$, the smoothness of $f_X$, or the rate of convergence of the singular values of $A$.

An alternative to the estimator (2.3) consists of using the estimator (2.2) with $\hat{J}$ in place of $J_n$. However, replacing $J_n$ with $\hat{J}$ causes the lengths of the series in $\hat{A}$ and $\hat{m}$ to be random variables. The methods of proof of this paper do not apply when the lengths of these series are random. The asymptotic properties of $\hat{g}$ with $\hat{J}$ in place of $J_n$ are unknown.

### 2.3 Review of Related Mathematics and Statistics Literature

Ill-posed inverse problems arising in models that are similar to but simpler than (1.1)-(1.2) have long been studied in mathematics and statistics. This section reviews the mathematics and statistics literature on choosing regularization parameters for such inverse problems. An important distinguishing characteristic of (1.1)-(1.2) is that the operator $A$ is unknown and must be estimated from the data. Another characteristic is that the distribution of the reduced-form residual, $V = Y - E[g(X)|W]$, is unknown. The mathematics and statistics literature contains no methods for choosing the regularization parameter in (1.1)-(1.2) when $A$ must be estimated from the data and the distribution of the reduced form residual is unknown.

A variety of ways to choose regularization parameters are known in mathematics and numerical analysis. Engl, Hanke, and Neubauer (1996), Mathé and Pereverzev (2003), Bauer and Hohage (2005), Wahba (1977), and Lukas (1993, 1998) describe many. Most of these methods assume that $A$ is known. Most also assume that the “data” are deterministic or that $\text{Var}(Y | X = x)$ is known and independent of $x$. Such methods are not suitable for econometric applications.

Spokoiny and Vial (2009) describe a method for choosing the regularization parameter in an estimation problem similar to (1.1)-(1.2) in which the operator $A$ is known and $V$ is normally distributed. Loubes and Ludeña (2008) also consider a setting in which $A$ is known.
Efromovich and Koltchinskii (2001), Cavalier and Hengartner (2005), Hoffmann and Reiss (2008), and Marteau (2006, 2009) consider settings in which \( A \) is known up to a random perturbation and, possibly, a truncation error but is not estimated from the data. Loubes and Marteau (2009) consider estimation of \( g \) in (1.1)-(1.2) when the eigenfunctions of \( A^*A \) are known but the eigenvalues must be estimated from data. Among the settings in the mathematics and statistics literature, this one is the closest to the setting considered here. Loubes and Marteau obtain non-asymptotic oracle inequalities for the risk of their estimator. They also obtain an adaptive estimator whose rate of convergence is within a factor of \((\log n)^p\) of the asymptotically optimal rate for a suitable \( p > 4s \). However, the eigenfunctions of \( A^*A \) are not known in econometric applications. Section 3 describes a method for selecting \( J \) empirically when neither the eigenvalues nor eigenfunctions of \( A^*A \) are known. In contrast to Loubes and Marteau (2009), the results of this paper are asymptotic. However, Monte Carlo experiments that are reported in Section 4 indicate that the adaptive procedure works well with samples of practical size. Some parts of the proofs in this paper are similar to parts of the proofs of Loubes and Marteau (2009).

3. MAIN RESULTS

This section begins with an informal description of the method for choosing \( J \). Section 3.2 presents the formal results.

3.1 Description of the Method for Selecting \( J \)

Define the asymptotically optimal \( J \) as the value that minimizes the asymptotic integrated mean-square error (AIMSE) of \( \hat{g}_J \) as an estimator of \( g \). The AIMSE is \( E_A \| \hat{g}_J - g \|^2 \), where \( E_A(\cdot) \) denotes the expectation of the leading term of the asymptotic expansion of \( \cdot \). Denote the asymptotically optimal value of \( J \) by \( J_{opt} \). Under the regularity conditions of Section 3.2, \( E_A \| \hat{g}_{J_{opt}} - g \|^2 \) converges to zero at the fastest possible rate (Chen and Reiss 2007). However, \( E_A \| \hat{g}_J - g \|^2 \) depends on unknown population parameters, so it cannot be minimized in applications. We replace the unknown parameters with sample analogs, thereby obtaining a feasible estimator of \( E_A \| \hat{g}_J - g \|^2 \). Let \( \hat{J} \) denote the value of \( J \) that minimizes the estimate of \( E_A \| \hat{g}_J - g \|^2 \). Note that \( \hat{J} \) is a random variable. The adaptive estimator of \( g \) is \( \hat{g}_{\hat{J}} \).
The AIMSE of the adaptive estimator is $E_A \| \hat{g}_J - g \|^2$. We show that under the regularity conditions of Section 3.2,

$$E_A \| \hat{g}_J - g \|^2 \leq [2 + (4 + \epsilon)\log(n)] E_A \| \hat{g}_{opt} - g \|^2. \tag{3.1}$$

Thus, the rate of convergence in probability of $\| \hat{g}_J - g \|$ is within a factor of $(\log n)^{1/2}$ of the asymptotically optimal rate. Moreover, the AIMSE of $\hat{g}_J$ is within a factor of $2 + (4 + \epsilon)\log n$ of the optimal AIMSE.

The following notation is used in addition to that already defined. For any positive integer $J$, let $A_J$ be the operator on $L_2[0,1]$ that is defined by

$$(A_J v)(w) = \int_{[0,1]} v(x)a_J(x,w)dx,$$

where

$$a_J(x,w) = \sum_{j=1}^{J} \sum_{k=1}^{J} c_{jk} \psi_j(x)\psi_k(w).$$

Let $J_n$ be the series truncation point defined in Section 2.2. For any $x \in [0,1]$, define

$$\delta_n(x,Y,X,W) = [Y - g_{J_n}(X)] \sum_{k=1}^{J} \psi_k(W)(A_{J_n}^{-1}\psi_k)(x),$$

$$S_n(x) = n^{-1} \sum_{i=1}^{n} \delta_n(x,Y_i,X_i,W_i),$$

and

$$g_J = \sum_{j=1}^{J} b_j \psi_j.$$

The following proposition is proved in the appendix.

**Proposition 1**: Let assumptions 1-6 of Section 3.2 hold. Then, as $n \to \infty$,

$$\hat{g}_J - g_J = \sum_{j=1}^{J} \left\langle S_n, \psi_j \right\rangle \psi_j + r_n,$$

where

$$\|r_n\| = o_p \left( \left\| \sum_{j=1}^{J} \left\langle S_n, \psi_j \right\rangle \psi_j \right\| \right)$$

uniformly over $J \leq J_n$. ■

Now for any $J \leq J_n$, ...
\[ E_A \| \hat{g}_J - g \|^2 = E_A \| \hat{g}_J - g_J \|^2 + \| g - g_J \|^2 = E_A \| \hat{g}_J - g_J \|^2 + \| g \|^2 - \| g_J \|^2. \]

Therefore, it follows from Proposition 1 that
\[ E_A \| \hat{g}_J - g \|^2 = E_A \sum_{j=1}^J \{ S_n \cdot \psi_J \}^2 + \| g \|^2 - \| g_J \|^2. \]

Define
\[ T_n(J) = E_A \sum_{j=1}^J \{ S_n \cdot \psi_J \}^2 - \| g_J \|^2. \]

Assume that \( J_n \geq J_{opt} \). Then because \( \| g \|^2 \) does not depend on \( J \),
\[ J_{opt} = \arg \min_{J > 0} T_n(J). \]

We now put \( T_n(J) \) into an equivalent form that is more convenient for the analysis that follows. Observe that
\[ (A_{J_n}^{-1} \psi_k)(x) = \sum_{j=1}^J c^{jk} \psi_j(x), \]
where \( c^{jk} \) is the \((j,k)\) element of the inverse of the \( J_n \times J_n \) matrix \([c_{jk}]\). Therefore,
\[ \sum_{k=1}^{J_n} \psi_k(W)(A_{J_n}^{-1} \psi_k)(x) = \sum_{j=1}^J \sum_{k=1}^J c^{jk} \psi_k(W) \psi_j(x) = \sum_{j=1}^J \psi_j(x)[(A_{J_n}^{-1})^* \psi_j](W), \]
where \(^*\) denotes the adjoint operator. It follows that
\[ \delta_n(x, Y, X, W) = [Y - g_{J_n}(X)] \sum_{j=1}^J \psi_j(x)[(A_{J_n}^{-1})^* \psi_j](W) \]
and
\[ \{ \delta_n(\cdot, Y, X, W), \psi_j \} = [Y - g_{J_n}(X)][(A_{J_n}^{-1})^* \psi_j](W). \]

Therefore,
\[ T_n(J) = E_A \sum_{j=1}^J \left\{ n^{-1} \sum_{i=1}^n [Y_i - g_{J_n}(X_i)][(A_{J_n}^{-1})^* \psi_j](W_i) \right\}^2 - \| g_J \|^2. \]

It follows from lemma 3 of the appendix and the assumptions of Section 3.2 that
\[
E_A \sum_{j=1}^{J} \left\{ n^{-1} \sum_{i=1}^{n} [Y_i - g_{J_n}(X_i)]((A_{j_n}^{-1})^* \psi_j)(W_i) \right\}^2
= n^{-1} E_A \sum_{j=1}^{J} [Y - g_{J_n}(X)]^2 \{(A_{j_n}^{-1})^* \psi_j\}(W)^2.
\]

Therefore,
\[
T_n(J) = n^{-1} E_A \sum_{j=1}^{J} [Y - g_{J_n}(X)]^2 \{(A_{j_n}^{-1})^* \psi_j\}(W)^2 - \|g_J\|^2.
\]

This is the desired form of \( T_n(J) \).

\( T_n(J) \) depends on the unknown parameters \( g_{J_n} \) and \( A_{J_n} \) and on the operator \( E_A \).

Therefore, \( T_n(J) \) must be replaced by an estimator for use in applications. One possibility is to replace \( g_{J_n}, A_{J_n} \), and \( E_A \) with \( \hat{g}, \hat{A}, \hat{g}_J \), and the empirical expectation, respectively.

This gives the estimator
\[
\hat{T}_n(J) \equiv n^{-2} \sum_{i=1}^{n} \left\{ [Y_i - \hat{g}(X_i)]^2 \sum_{j=1}^{J} \{(\hat{A}^{-1})^* \psi_j\}(W_i) \right\} - \|\hat{g}_J\|^2.
\]

However, \( \hat{T}_n \) is unsatisfactory for two reasons. First, it does not account for the effect on \( E_A \|\hat{g}_J - g\|^2 \) of the randomness of \( \hat{J} \). This randomness is the source of the factor of \( \log n \) on the right-hand side of (3.1). Second, some algebra shows that
\[
\|\hat{g}_J\|^2 - \|g_J\|^2 = \|\hat{g}_J - g_J\|^2 + 2\langle g_J, \hat{g}_J - g_J \rangle.
\]

The right-hand side of (3.2) is asymptotically non-negligible, so the estimator of \( T_n \) must compensate for its effect.

It is shown in the appendix that these problems can be overcome by using the estimator
\[
(3.3) \quad \hat{T}_n(J) = (2 + \varepsilon / 2)(\log n)n^{-2} \sum_{i=1}^{n} \left\{ [Y_i - \hat{g}(X_i)]^2 \sum_{j=1}^{J} \{(\hat{A}^{-1})^* \psi_j\}(W_i) \right\} - \|\hat{g}_J\|^2
\]

for any \( \varepsilon > 0 \). We obtain \( \hat{J} \) by solving the problem
\[
(3.4) \quad \minimize \quad \hat{T}_n(J), \quad \text{where } j_n = c_1 / \log(n) \text{ for some finite constant } c_1 > 0, \quad J_n \text{ satisfies } \rho_{J_n} (J_n^3 / n)^{1/2} \to 0 \text{ and } \rho_{J_n} (J_n^4 / n)^{1/2} \to \infty \text{ as } n \to \infty, \quad \text{and } \rho_{J_n} \text{ is defined as}
\]
Section 3.2 gives conditions under which $\hat{g}_j$ satisfies inequality (3.1). The problem of choosing $J_n$ in applications is discussed at the end of Section 3.2.

### 3.2 Formal Results

This section begins with the assumptions under which $\hat{g}_j$ is shown to satisfy (3.1). A theorem that states the result formally follows the presentation of the assumptions.

Let $112(, )EXw x w$ denote the Euclidean distance between $(x_1, w_1)$ and $(x_2, w_2)$. Let $D_jf_{XW}$ denote any $j$th partial or mixed partial derivative of $f_{XW}$. Let $D_0f_{XW}(x, w) = f_{XW}(x, w)$. Let $A^*$ denote the adjoint operator of $A$. Define $U = Y - g(X)$ and

$$
\rho_J = \sup_{v \in H_m} \frac{\|v\|}{\left\|A^*A\right\|^{1/2} \|v\|}.
$$

Blundell, Chen, and Kristensen (2007) call $\rho_J$ the sieve measure of ill-posedness and discuss its relation to the eigenvalues of $A^*A$. Finally, let $\bar{f}_{XW}(t_1, t_2)$ denote the characteristic function of $f_{XW}$. Define $\|x\|_E = (t_1^2 + t_2^2)^{1/2}$.

The assumptions are as follows.

**Assumption 1:** (i) The supports of $X$ and $W$ are $[0,1]$. (ii) $(X, W)$ has a probability density function $f_{XW}$ with respect to Lebesgue measure. The probability density function of $W$, $f_w$, is non-zero almost everywhere on $[0,1]$. (iii) There are an integer $r \geq 2$ and a constant $C_f < \infty$ such that $|D_jf_{XW}(x, w)| \leq C_f$ for all $(x, w) \in [0,1]^2$ and $j = 0, 1, ..., r$. (iv) $|D_rf_{XW}(x_1, w_1) - D_rf_{XW}(x_2, w_2)| \leq C_f \|x_1 - x_2\|_E$ for any order $r$ derivative and any $(x_1, w_1)$ and $(x_2, w_2)$ in $[0,1]^2$. (v) If $r = \infty$, then $\bar{f}_{XW}(t_1, t_2) = O[\exp(-\kappa \|x\|_E^\beta)]$ for finite constants $\kappa > 0$ and $\beta > 1$.

**Assumption 2:** (i) There is a finite constant $C_Y$ such that $E(Y^2 | W = w) \leq C_Y$ for each $w \in [0,1]$. (ii) There are finite constants $C_U > 0$, $c_{U_1} > 0$ and $c_{U_2} > 0$ such that $E(\|U\|_E | W = w) \leq C_U^{-2} j! E(U^2 | W = w)$ and $c_{U_1} \leq E(U^2 | W = w) \leq c_{U_2}$ for each $w \in [0,1]$. 

(3.5) \[ \rho_{J_n} = \sup_{v \in H_m} \frac{\|v\|}{\left\|A^*A\right\|^{1/2} \|v\|}. \]
Assumption 3: (i) (1.1) has a solution \( g \in \mathcal{H}_s \) with \( \|g\|_s < C_0 \) and \( s \geq 2 \). (ii) The estimators \( \hat{g} \) and \( \hat{g}_J \) are as defined in (2.2)-(2.3). (iii) The function \( m \) has \( r+s \) square-integrable derivatives.

Assumption 4: (i) The basis functions \{\psi_j\} are orthonormal, complete on \( L_2[0,1] \), and bounded uniformly over \( j \). (ii) \( \|A_j - A\| = O(J^{-r}) \) if \( r < \infty \) and \( O(e^{-cJ}) \) for some \( c > 0 \) if \( r = \infty \). (iii) For any \( \nu \in L_2[0,1] \) with \( \ell \) square integrable derivatives, there are coefficients \( \nu_j \) \( (j = 1, 2, \ldots) \) such that

\[
\|\nu - \sum_{j=1}^J \nu_j \psi_j\| \leq C_J J^{-\ell},
\]

where \( C_J < \infty \) is a constant that does not depend on \( \nu \).

Assumption 5: (i) The operator \( A \) is one-to-one. (ii)

\[
\rho_J \sup_{\nu \in \mathcal{H}_s} \frac{\|A_j - A\|}{\|\nu\|} \leq C_A J^{-s}
\]

for some constant \( C_A < \infty \) that does not depend on \( J \).

Assumption 6: As \( n \to \infty \), \( \rho_j_n (J_n^3 / n)^{1/2} \to 0 \) and \( \rho_j_n (J_n^4 / n)^{1/2} \to \infty \).

Assumptions 1 and 2 are smoothness and boundedness conditions. At the cost of slower convergence of \( \|\hat{g}_j - g\| \), assumption 2(ii) can be relaxed to allow \( U \) to have only finitely many moments. Assumption 3 defines the estimator of \( g \) and ensures that the function \( m \) is sufficiently smooth. The assumption requires \( \|g\|_s < C_0 \) (strict inequality) to avoid complications that arise when \( g \) is on the boundary of \( \mathcal{H}_s \). Assumption 4 is satisfied by trigonometric bases and B-splines that have been orthogonalized by, say, the Gram-Schmidt procedure. Orthogonal polynomials do not satisfy the boundedness requirement. However, this does not prevent the use of orthogonal polynomials in applications because, for any fixed integer \( J \), the basis can consist of the first \( J \) orthogonal polynomials plus a rotation of B-splines or trigonometric functions that is orthogonal to the polynomials. Assumption 5(i) is required for identification of \( g \). Assumption 5(ii) ensures that \( A_n \) is a “sufficiently accurate” approximation to \( A \) on \( \mathcal{H}_{ns} \). This assumption complements 4(ii), which specifies the accuracy of \( A_n \) as an approximation to \( A \) on the larger set \( \mathcal{H}_s \). Assumption 5(ii) can be interpreted as a smoothness restriction on \( f_{XW} \). For example, 5(ii) is satisfied if assumptions 4 and 6 hold and \( A \) maps \( \mathcal{H}_s \) to \( \mathcal{H}_{r+s} \). Assumption
5(ii) also can be interpreted as a restriction on the sizes of the values of $c_{jk}$ for $j \neq k$. Hall and Horowitz (2005) used a similar diagonality restriction. Assumption 6 specifies the rate of increase of $J_n$ as $n \to \infty$.

For sequences of positive numbers $\{a_n\}$ and $\{b_n\}$, define $a_n \asymp b_n$ if $a_n / b_n$ is bounded away from 0 and $\infty$ as $n \to \infty$. The problem of estimating $g$ is said to be mildly ill-posed if $\rho_J \asymp J^\beta$ for some finite $\beta > 0$. In this case, $J_{opt} \propto n^{1/(2r+2s+1)}$ and 
$$\|\hat{g}_{J_{opt}} - g\| = O_p[n^{-\beta/(2r+2s+1)}]$$ (Horowitz 2009). The estimation problem is severely ill-posed if $\rho_J \asymp e^{\beta J}$ for some finite $\beta > 0$. In this case, $J_{opt} = O(\log n)$ and $\|\hat{g}_{J_{opt}} - g\| = O_p[(\log n)^{-s}]$.

The results of this section hold in both the mildly and severely ill-posed cases.

The main result of this paper is given by the following theorem.

**Theorem 3.1**: Let assumptions 1-6 hold. Then $\hat{g}_J$ satisfies inequality (3.1). ■

Because $J_n$ in problem (3.4) depends on $\rho_J$, it is necessary, in principle, to estimate $\rho_J$ as a function of $J$ to choose $J_n$. This can be done by replacing $A$ with $\hat{A}$ in the definition of $\rho_J$ and solving the resulting version of (3.5) for several different values of $J$. Denote the resulting estimates by $\hat{\rho}_J$. Plotting $\log \hat{\rho}_J$ against $J$ and $\log J$ will usually reveal whether the problem of estimating $g$ is mildly or severely ill-posed. The exponent $r$ in the mildly ill-posed case or $c$ in the severely ill-posed case can then be estimated by ordinary least squares. In practice, however, this procedure is unnecessary. $J$ and $\hat{g}_J$ are not sensitive to the precise value of $J_n$. In applications, it matters only that $J_n$ is large enough to make the bias of $\hat{g}_{J_n}$ small compared to its variance but not so large as to make $\hat{g}_{J_n}$ excessively noisy. With samples of the sizes in most applications, the estimated Fourier coefficients $\tilde{b}_j$ are well within random sampling error of zero when $j$ exceeds 5 or 10. Therefore, it typically suffices to solve (3.4) with $J_n = 10$.

4. **MONTE CARLO EXPERIMENTS**

This section describes the results of a Monte Carlo study of the finite-sample performance of $\hat{g}_J$. The experiments use a sample size of 1000, and there are 1000 Monte Carlo replications in each experiment. The basis functions $\{\psi_j\}$ are Legendre polynomials that are centered and scaled to be orthonormal on $[0,1]$. 

---

13
There were 4 Monte Carlo experiments. In each experiment, realizations of \((X, W)\) were generated from the model

\[
f_{X,W}(x, w) = 1 + 2 \sum_{j=1}^{\infty} c_j \cos(j \pi x) \cos(j \pi w),
\]

where \(c_j = 0.7^{-1}\) in experiment 1, \(c_j = 0.6^{-2}\) in experiment 2, \(c_j = 0.52^{-4}\) in experiment 3, and \(c_j = 1.3 \exp(-0.5j)\) in experiment 4. In all experiments, the marginal distributions of \(X\) and \(W\) are \(U[0,1]\), and the conditional distributions are unimodal with an arch-like shape. The estimation problem is mildly ill-posed in experiments 1-3 and severely ill-posed in experiment 4.

The function \(g\) is

\[
g(x) = b_0 + \sqrt{2} \sum_{j=1}^{\infty} b_j \cos(j \pi x),
\]

where \(b_0 = 0.5\) and \(b_j = j^{-4}\) for \(j \geq 1\). This function is plotted in Figure 1. The series in (4.1) and (4.2) were truncated at \(j = 100\) for computational purposes. Realizations of \(Y\) were generated from

\[
Y = E[g(x) \mid W] + V,
\]

where \(V \sim N(0,0.01)\). The value of \(\varepsilon\) in (3.3) is 0.01.

The results of the experiments are displayed in Table 1, which shows the empirical means of \(\|\hat{g}_{J_{opt}} - g\|\) and \(\|\hat{g}_j - g\|\) for \(J_{n} = 10\). The results are similar with \(5 \leq J_{n} < 9\). The differences between the empirical means of \(\|\hat{g}_{J_{opt}} - g\|\) and \(\|\hat{g}_j - g\|\) are very small. The integrated mean-square error (IMSE) of \(\hat{g}_j\) is very close the IMSE that would be obtained if \(J_{opt}\) were known.

5. CONCLUSIONS

This paper has presented, for what appears to be the first time, a theoretically justified, empirical method for choosing the regularization parameter in nonparametric instrumental variables estimation. The method does not require a priori knowledge of smoothness or other unknown population parameters. The method and the resulting estimator of the unknown function \(g\) adapt to the unknown smoothness of \(g\) and the density of \((X, W)\). The results of Monte Carlo experiments indicate that the method performs well with samples of practical size.
It is likely that the ideas in this paper can be applied to the multivariate model 
\[ Y = g(X,Z) + U, \quad E(U \mid W,Z) = 0, \]  
where \( Z \) is a continuously distributed, exogenous explanatory variable or vector and \( W \) is an instrument for the endogenous variable \( X \). This model is more difficult than (1.1)-(1.2), because it requires selecting at least two regularization parameters, one for \( X \) and one or more for the components of \( Z \). The multivariate model will be addressed in future research.

**APPENDIX: PROOFS OF PROPOSITION 1 AND THEOREM 3.1**

Assumptions 1-6 hold throughout this appendix. Define \( J = \{ J : j \leq J \leq J_n \} \). For \( J \in J \), define

\[ S_n(J) = n^{-1} E[ Y - g_{J_n}(X)]^2 \sum_{j=1}^{J} \{ (A_j^{-1})^* \psi_j \}(W)^2, \]

\[ \hat{S}_n(J) = n^{-2} \sum_{i=1}^{n} \left[ Y_i - g_{J_n}(X_i) \right]^2 \sum_{j=1}^{J} \{ (A_j^{-1})^* \psi_j \}(W_j)^2, \]

and

\[ \hat{S}_n(J) = n^{-2} \sum_{i=1}^{n} \left[ Y_i - \tilde{g}(X_i) \right]^2 \sum_{j=1}^{J} \{ (\hat{A}_j^{-1})^* \psi_j \}(W_j)^2. \]

Define \( U_i = Y_i - g(X_i) \) \((i = 1, \ldots, n)\).

We begin with three lemmas that are used in the proof of Proposition 1. Then Proposition 1 is proved. Four additional lemmas that are used in the proof of Theorem 3.1 are presented after the proof of Proposition 1. Finally, Theorem 3.1 is proved.

**Lemma 1:** Let \( J \leq J_n \). Then

\[ \sup_{\nu \in \mathcal{H}_J, \| \nu \|^2 = 1} \| A_j^{-1} \nu \| = \rho_J[1 + O(J^{-3})] \]

and

\[ \sup_{\nu \in \mathcal{H}_J, \| \nu \|^2 = 1} \| (A_j^{-1})^* \nu \| = \rho_J[1 + O(J^{-3})]. \]

**Proof:** Only (A.1) is proved. The proof of (A.2) is similar. Let \( I \) denote the identity operator. For \( \nu \in \mathcal{H}_J \),

\[ (A_j^{-1} - I)\nu = -(I + A_j^{-1}(A_j^{-1} - A))^{-1}A_j^{-1}(A_j^{-1} - A)A_j^{-1}\nu. \]
But $A_{J_n}^{-1}(A_{J_n} - A_J)A_{J_n}^{-1}v = 0$ for $v \in \mathcal{H}_{J_n}$. Therefore, $\|A_{J_n}^{-1}v\| = \|A_{J_n}^{-1}v\|$. The lemma now follows from lemma A.1 of Horowitz and Lee (2010). Q.E.D.

**Lemma 2:** As $n \to \infty$,

$$\sup_{v \in \mathcal{H}_{J_n}, \|v\| \leq 1} \|\hat{A} - A_{J_n}\|v\| = O_p[(J_n/n)^{1/2}]$$

and

$$\sup_{v \in \mathcal{H}_{J_n}, \|v\| \leq 1} \|\hat{A} - A_{J_n}\|^*v\| = O_p[(J_n/n)^{1/2}] .$$

**Proof:** Only (A.3) is proved. The proof of (A.4) is similar. For any $v \in \mathcal{H}_{J_n},$ 

$$\|\hat{A} - A_{J_n}\|v\|^2 = \sum_{k=1}^{J} \sum_{j=1}^{J} (\hat{c}_{jk} - c_{jk})v_j^2,$$

where $v_j = \langle \psi_j, v \rangle$. Now $\sum_{j=1}^{J} |v_j| \alpha / J$ is bounded uniformly over $v \in \mathcal{H}_{J_n}$. Moreover,

$$\sum_{j=1}^{J} \sum_{j=1}^{J} \hat{c}_{jk} = \sum_{j=1}^{J} v_j n^{-1} \sum_{i=1}^{n} \psi_j(X_i)\psi_k(W_i).$$

The $\psi_j$’s are bounded by assumption 4. Therefore, it follows from Hoeffding’s inequality that

$$\sum_{j=1}^{J} v_j (\hat{c}_{jk} - c_{jk}) = O_p(n^{-1/2})$$

uniformly over $v \in \mathcal{H}_{J_n}$. The lemma follows by combining (A.5) and (A.6). Q.E.D.

**Lemma 3:** As $n \to \infty$,

$$\sum_{j=1}^{J} \{S_n, \psi_j\}^2 = \sum_{j=1}^{J} \left[ n^{-1} \sum_{i=1}^{n} U_i(A_{J_n}^{-1})^* \psi_j(W_i) \right]^2 + o_p(\rho_j^2 J/n),$$

and

$$E \sum_{j=1}^{J} \left[ n^{-1} \sum_{i=1}^{n} U_i[(A_{J_n}^{-1})^* \psi_j(W_i)] \right]^2 \leq \rho_j^2 J/n \text{ uniformly over } J \in \mathcal{J}.$$

**Proof:** We have
\[ S_n(x) = n^{-1} \sum_{i=1}^{n} \delta_n(x, Y_i, X_i, W_i) = \sum_{j=1}^{J} \psi_j(x) \left\{ n^{-1} \sum_{i=1}^{n} [Y_i - g_{J_\alpha}(X_i)][(A_{J_\alpha}^{-1})^* \psi_j(W_i)] \right\}. \]

Therefore,
\[ \sum_{j=1}^{J} \langle S_n, \psi_j \rangle^2 = \sum_{j=1}^{J} \left\{ n^{-1} \sum_{i=1}^{n} [Y_i - g_{J_\alpha}(X_i)][(A_{J_\alpha}^{-1})^* \psi_j(W_i)] \right\}^2. \]

But
\[ n^{-1} \sum_{i=1}^{n} [Y_i - g_{J_\alpha}(X_i)][(A_{J_\alpha}^{-1})^* \psi_j(W_i)] = R_{nj1} + R_{nj2}, \]

where
\[ R_{nj1} = n^{-1} \sum_{i=1}^{n} U_i[(A_{J_\alpha}^{-1})^* \psi_j(W_i)] \]

and
\[ R_{nj2} = -n^{-1} \sum_{i=1}^{n} [g_{J_\alpha}(X_i) - g(X_i)][(A_{J_\alpha}^{-1})^* \psi_j(W_i)]. \]

\( E(R_{nj1}) = 0, \) and \( Var(R_{nj1}) \asymp \rho_j^2 / n \) for every \( j \leq J \) by lemma 1 and assumption 2. Therefore,

\( (A.7) \quad E R_{nj1}^2 \asymp \rho_j^2 / n \)

for every \( j \leq J. \) In addition,
\[ E(R_{nj2}) = -\int_{[0,1]^2} f_{XW}(x, w)[g_{J_\alpha}(x) - g(x)][(A_{J_\alpha}^{-1})^* \psi_j](w) dx dw \]

\[ = \left\langle A(g_{J_\alpha} - g), (A_{J_\alpha}^{-1})^* \psi_j \right\rangle \]

\[ = \left\langle (A - A_{J_\alpha})(g_{J_\alpha} - g), (A_{J_\alpha}^{-1})^* \psi_j \right\rangle. \]

Therefore, the Cauchy-Schwarz inequality gives
\[ \left| E(R_{nj2}) \right| \leq \left\| (A - A_{J_\alpha})(g_{J_\alpha} - g) \right\| \left\| (A_{J_\alpha}^{-1})^* \psi_j \right\| \]

\[ \leq \left\| (A - A_{J_\alpha}) \right\| \left\| g_{J_\alpha} - g \right\| \left\| (A_{J_\alpha}^{-1})^* \psi_j \right\|. \]
\[ \|A - A_{n}\| = O(\rho_{n}^{-1}) \quad \text{and} \quad \|g_{J_{n}} - \hat{g}\| = O(J_{n}^{-s}) \quad \text{by assumption 4, and} \quad \|\left(A_{J_{n}}^{-1}\right)^{\psi_{j}}\| \leq O(\rho_{J}) \quad \text{by lemma 1.} \]

Therefore, \[ |E(R_{nj/2})| = O(\rho_{J_{n}}^{-1}\rho_{J}J_{n}^{-s}) = o(\rho_{J}/n^{1/2}) \quad \text{for every} \quad j \leq J \quad \text{by the definition of} \quad J_{n}. \]

In addition, \[ \text{Var}(R_{nj/2}) = O(J_{n}^{-2s}\rho_{J}^{2}/n) = o(\rho_{J}^{2}/n). \]

It follows that

(A.8) \[ R_{nj/2} = o_{p}(\rho_{J}/n^{1/2}) \]

for every \( j \leq J \). Combining (A.7) and (A.8) gives

\[
n^{-1}\sum_{i=1}^{n} [Y_{i} - g_{J_{n}}(X_{i})][(A_{J_{n}}^{-1})^{\psi_{j}}](W_{i}) = R_{nj/1} + o_{p}(\rho_{J}/n^{1/2})
\]

for every \( j \leq J \). It also follows that,

\[
\sum_{j=1}^{J} \langle S_{n}, \psi_{j} \rangle^{2} = \sum_{j=1}^{J} R_{nj/1}^{2} + o_{p}(\rho_{J}^{2}J/n).
\]

uniformly over \( J \in \mathcal{J} \). This establishes the first conclusion of the lemma. The second conclusion follows from (A.7). Q.E.D.

**Proof of Proposition 1:** Define \( \hat{h} = \hat{A}^{-1}\hat{m} \). Because \( \hat{g} = \hat{h} \) with probability approaching 1 as \( n \to \infty \), it suffices to prove the proposition with \( \hat{h} \) in place of \( \hat{g} \). We have

\[ A_{J_{n}}\hat{h} + (\hat{A} - A_{J_{n}})\hat{h} = \hat{m}. \]

Therefore

\[
\hat{h} = A_{J_{n}}^{-1}\hat{m} - A_{J_{n}}^{-1}(\hat{A} - A_{J_{n}})\hat{h}
\]

\[
= A_{J_{n}}^{-1}\hat{m} - A_{J_{n}}^{-1}(\hat{A} - A_{J_{n}})g_{J_{n}} - A_{J_{n}}^{-1}(\hat{A} - A_{J_{n}})(\hat{h} - g_{J_{n}}).
\]

It follows that

\[
\hat{h} - g_{J_{n}} = A_{J_{n}}^{-1}\hat{m} - A_{J_{n}}^{-1}\hat{A}g_{J_{n}} - R_{n},
\]

where \( R_{n} = A_{J_{n}}^{-1}(\hat{A} - A_{J_{n}})(\hat{h} - g_{J_{n}}) \). Some algebra shows that \( A_{J_{n}}^{-1}\hat{m} - A_{J_{n}}^{-1}\hat{A}g_{J_{n}} = S_{n} \). Therefore,

\[
\hat{g}_{J} - g_{J} = \sum_{j=1}^{J} \langle S_{n}, \psi_{j} \rangle\psi_{j} - \sum_{j=1}^{J} \langle R_{n}, \psi_{j} \rangle\psi_{j},
\]

and

\[
r_{n} = -\sum_{j=1}^{J} \langle R_{n}, \psi_{j} \rangle\psi_{j}
\]

Now
\[ \langle \psi_j, R_n \rangle = \langle \psi_j, A_{J_n}^{-1}(\hat{A} - A_{J_n})(\hat{h} - g_{J_n}) \rangle \]

\[ = \langle (A_{J_n}^{-1})^* \psi_j, (\hat{A} - A_{J_n})(\hat{h} - g_{J_n}) \rangle. \]

The Cauchy-Schwarz inequality gives
\[ \left\| \langle \psi_j, R_n \rangle \right\| \leq \left\| (A_{J_n}^{-1})^* \psi_j \right\| \left\| (\hat{A} - A_{J_n})(\hat{h} - g_{J_n}) \right\|. \]

Therefore,
\[ \left\| \langle \psi_j, R_n \rangle \right\| = O_p[\rho_J(J_n / n)^{1/2}] \left\| \hat{h} - g_{J_n} \right\| \]

by lemmas 1 and 2. It follows that
\[ \left\| r_n \right\| = \left( \sum_{j=1}^{J_n} \langle R_n, \psi_j \rangle^2 \right)^{1/2} \]

\[ = O_p[\rho_J(J_n / n)^{1/2} J_n^{1/2}] \left\| \hat{h} - g_{J_n} \right\| \]

uniformly over \( J \in J \). The definition of \( J_n \) combined with Theorem 4.1 of Horowitz (2009) give the result that
\[ (A.9) \quad J_n^{1/2} \left\| \hat{h} - g_{J_n} \right\| = o_p(1) \]

as \( n \to \infty \). Therefore, the proposition follows by combining (A.9) with lemma 3. Q.E.D.

**Lemma 4:** As \( n \to \infty \),
\[ \max_{J \in J} \left| \frac{\tilde{S}_n(J) - S_n(J)}{S_n(J)} \right| = o_p(1). \]

**Proof:** Define
\[ \tilde{S}_{n1}(J) = n^{-2} \sum_{i=1}^{n} U_i^2 \sum_{j=1}^{J_n} \left\{ \langle (A_{J_n}^{-1})^* \psi_j \rangle [W_i] \right\}^2, \]
\[ \tilde{S}_{n2}(J) = -2n^{-2} \sum_{i=1}^{n} U_i \left[ g_{J_n}(X_i) - g(X_i) \right] \sum_{j=1}^{J_n} \left\{ \langle (A_{J_n}^{-1})^* \psi_j \rangle [W_i] \right\}^2, \]

and
\[ \tilde{S}_{n3}(J) = n^{-2} \sum_{i=1}^{n} \left[ g_{J_n}(X_i) - g(X_i) \right]^2 \sum_{j=1}^{J_n} \left\{ \langle (A_{J_n}^{-1})^* \psi_j \rangle [W_i] \right\}^2. \]

Then
\[ \tilde{S}_n(J) = \tilde{S}_{n1}(J) + \tilde{S}_{n2}(J) + \tilde{S}_{n3}(J). \]
Consider, first, convergence of $\tilde{S}_{n1}(J)/S_n(J)$. By Lemma 1,
\[
\left\| (A_{J_n}^{-1})^* \psi_{J_n} \right\|^2 \leq c_1 \rho_j^2
\]
for every $J \in \mathcal{J}$, some constant $c_1 < \infty$, and all sufficiently large $n$. This result together with boundedness of the $\psi_{J_n}$'s implies that $\{(A_{J_n}^{-1})^* \psi_{J_n}(w)\}^2 \leq c_2 J \rho_j^2$ for some constant $c_2 < \infty$, every $J \in \mathcal{J}$, and all sufficiently large $n$. Define
\[
K_{J_n}(w) = \{(A_{J_n}^{-1})^* \psi_{J_n}(w)\}^2.
\]
Then
\[
\tilde{S}_{n1}(J) = n^{-2} \sum_{i=1}^{n} U_i^2 \sum_{j=1}^{J_n} K_{J_n}(W_i)
\]
with probability 1 uniformly over $J \in \mathcal{J}$. Moreover,
\[
\tilde{K}_{nJ}(w) = \left( \frac{\rho_j^2 J}{n} \right)^{-1} n^{-1} \sum_{j=1}^{J} K_{J_n}(w) \leq c_2 J
\]
for all sufficiently large $n$. Now let $a_n = n^d$ for some constant $d > 0$ such $n^{1-2d}/J_n^2 \to \infty$ as $n \to \infty$. Let $B_n$ denote the event $\max_{1 \leq i \leq n} U_i^2 \leq a_n$. Let $\bar{B}_n$ denote the complement of $B_n$. It follows from Markov’s inequality that $P(\bar{B}_n) \to 0$ as $n \to \infty$. We have
\[
\left( \frac{\rho_j^2 J}{n} \right)^{-1} \tilde{S}_{n1}(J) = n^{-1} \sum_{i=1}^{n} U_i^2 \bar{K}_{nJ}(W_i) I(B_n) + n^{-1} \sum_{i=1}^{n} U_i^2 \bar{K}_{nJ}(W_i) I(\bar{B}_n),
\]
where $I$ is the indicator function. Define
\[
\tilde{S}_{n1a}(J) = n^{-1} \sum_{i=1}^{n} U_i^2 \bar{K}_{nJ}(W_i) I(B_n)
\]
and
\[
\tilde{S}_{n1b}(J) = n^{-1} \sum_{i=1}^{n} U_i^2 \bar{K}_{nJ}(W_i) I(\bar{B}_n).
\]
For any $\varepsilon > 0$,
\[ P \left[ \sup_{J \in \mathcal{J}} \left( \frac{\rho_J^2}{n} \right)^{-1} | \tilde{S}_{n1}(J) - E\tilde{S}_{n1}(J) | > 2\varepsilon \right] \leq P \left[ \sup_{J \in \mathcal{J}} | \tilde{S}_{n1a}(J) - E\tilde{S}_{n1a}(J) | > \varepsilon \right] \]

\[ + P \left[ \sup_{J \in \mathcal{J}} | \tilde{S}_{n1b}(J) - E\tilde{S}_{n1b}(J) | > \varepsilon \right]. \]

Now

(A.10) \[ P \left[ \sup_{J \in \mathcal{J}} | \tilde{S}_{n1b}(J) - E\tilde{S}_{n1b}(J) | > \varepsilon \right] \rightarrow 0 \]

as \( n \rightarrow \infty \) because \( P(\bar{B}_n) \rightarrow 0 \). Now consider \( \tilde{S}_{n1a} \). By Hoeffding’s inequality

\[ P[| \tilde{S}_{n1a}(J) - E(\tilde{S}_{n1a} | B_n) | > \varepsilon | B_n] \leq 2 \exp\left[-2\varepsilon^2 n/(c_2^2 J^2 a_n^2)\right] \]

\[ \leq 2 \exp\left[-2\varepsilon^2 n^{-2d} / (c_2^2 J_n^2)\right] \]

for every \( J \in \mathcal{J} \) and all sufficiently large \( n \). Therefore,

\[ P \left[ \max_{J \in \mathcal{J}} | \tilde{S}_{n1a}(J) - E(\tilde{S}_{n1a} | B_n) | > \varepsilon | B_n \right] \leq 2 J_n \exp\left[-2\varepsilon^2 n^{-2d} / (c_2^2 J_n^2)\right]. \]

In the mildly ill-posed case, \( J_n = o\left[n^{\log(2d+2)}\right] \). In the severely ill-posed case, \( J_n = o(\log n) \).

Therefore,

(A.11) \[ P \left[ \max_{J \in \mathcal{J}} | \tilde{S}_{n1a}(J) - E(\tilde{S}_{n1a} | B_n) | > \varepsilon | B_n \right] \rightarrow 0 \]

as \( n \rightarrow \infty \). Because \( P(B_n) \rightarrow 1 \), it follows from (A.10) and (A.11) that

\[ P \left[ \max_{J \in \mathcal{J}} \left( \frac{\rho_J^2}{n} \right)^{-1} | \tilde{S}_{n1}(J) - E(\tilde{S}_{n1}) | > \varepsilon \right] \rightarrow 0. \]

But \( E\tilde{S}_{n1}(J) = S_n(J) \) and \( S_n(J) \sim O(\rho_J^2 J / n) \) uniformly over \( J \in \mathcal{J} \). Therefore,

(A.12) \[ \frac{\tilde{S}_{n1}(J) - S_n(J)}{S_n(J)} = o_p(1) \]

as \( n \rightarrow \infty \) uniformly over \( J \in \mathcal{J} \).

Now consider \( \tilde{S}_{n2}(J)/S_n(J) \). Smoothness of \( g \) implies that

\[ |g_{J_n}(x) - g(x)| \leq (C/2)J_n^{-s} \]

uniformly over \( x \in [0,1] \) for some finite constant \( C \). Therefore,

\[ |\tilde{S}_{n2}(J)| \leq CJ_n^{-s} n^{-2} \sum_{i=1}^{n} |U_i| \sum_{j=1}^{J} \left\{ \left[ (A_{J_n}^{-1})^* \psi_j \right](W_j) \right\}^2. \]
Now \( \{(A_{j'}^{-1} \psi_j)(w)\}^2 \leq c_2^2 J \rho_J^2 \) for almost every \( w \in [0,1] \) every \( j \leq J \). Therefore,
\[
|\tilde{S}_{n2}(J)| \leq c_2^2 CJ_J^{-s} J^2 \rho_J^2 n^{-2} \sum_{i=1}^{n} |U_i|
\]
with probability 1 uniformly over \( J \in \mathcal{J} \). It follows from the strong law of large numbers that
\[
|\tilde{S}_{n2}(J)| \leq c_2^2 CJ_J^{-s} J^2 \rho_J^2 n^{-1}[E(|U|) + o_p(1)]
\]
uniformly over \( J \in \mathcal{J} \). Therefore,
\[
(A.13) \quad \frac{|\tilde{S}_{n2}(J)|}{S_n(J)} = O_p(J^{-s}) = o_p(1)
\]
uniformly over \( J \in \mathcal{J} \). A similar argument gives
\[
|\tilde{S}_{n3}(J)| \leq CJ_J^{-2s} J^2 \rho_J^2 n^{-1}
\]
for some constant \( C < \infty \) with probability 1 uniformly over \( J \in \mathcal{J} \), so
\[
(A.14) \quad \frac{|\tilde{S}_{n3}(J)|}{S_n(J)} = O_p(J^{-2s}) = o_p(1)
\]
uniformly over \( J \in \mathcal{J} \). The lemma follows by combining (A.12)-(A.14). Q.E.D.

Lemma 5: As \( n \to \infty \),
\[
\max_{J \in \mathcal{J}} \left| \frac{\tilde{S}_{n}(J) - \hat{S}_{n}(J)}{S_n(J)} \right| = o_p(1).
\]

Proof: Define
\[
K_j(w) = \{(A_{j'}^{-1} \psi_j)(w)\}^2,
\]
\[
\hat{K}_j(w) = \{(\hat{A}^{-1})^* \psi_j)(w)\}^2,
\]
\[
\Delta g(x) = \tilde{g}(x) - g_{J_n}(x),
\]
\[
\Delta K_j(w) = \hat{K}_j(w) - K_j(w),
\]
\[
\Delta S_{n1}(J) = -2n^{-2} \sum_{i=1}^{n} [Y_i - g_{J_n}(X_i)] \Delta g(X_i) \sum_{j=1}^{J} K_j(W_i),
\]
\[
\Delta S_{n2}(J) = n^{-2} \sum_{i=1}^{n} [\Delta g(X_i)]^2 \sum_{j=1}^{J} K_j(W_i),
\]
\[
\Delta S_{n3}(J) = n^{-2} \sum_{i=1}^{n} [Y_i - g_{J_n}(X_i)]^2 \sum_{j=1}^{J} \Delta K_j(W_i),
\]
\[ \Delta S_{n4}(J) = -2n \sum_{i=1}^{n} [Y_i - g_{J_n}(X_i)] \Delta g(X_i) \sum_{j=1}^{J} \Delta K_j(W_i), \]

and

\[ \Delta S_{n5}(J) = n^{-2} \sum_{i=1}^{n} [\Delta g(X_i)]^2 \sum_{j=1}^{J} \Delta K_j(W_i). \]

Then

\[ \hat{S}_n(J) - \tilde{S}_n(J) = \sum_{k=1}^{\bar{s}} \Delta S_{nk}(J). \]

Because \( S_n(J) \asymp \rho^2 J / n \), it suffices to prove that

\[ \max_{J \in J} (\rho^2 J / n)^{-1} | \hat{S}_n(J) - \tilde{S}_n(J) | = o_p(1). \]

Consider \( \Delta S_{n1} \). By Theorem 4.1 of Horowitz (2009), \( \| \tilde{g} - g_{J_n} \| = O_p[\rho_{J_n} (J_n / n)^{1/2}] \). In addition, \( K_j(w) = O(J \rho^2 J) \) uniformly over \( J \in J \) and \( w \in [0,1] \) by Lemma 1 and boundedness of the \( \psi_j \)'s. Therefore,

\[ |\Delta S_{n1}(J)| = O_p[\rho_{J_n} (J_n / n)^{1/2}] n^{-2} \sum_{i=1}^{n} |Y_i - g_{J_n}(X_i)| \sum_{j=1}^{J} K_j(W_i) \]

\[ = O_p[\rho_{J_n} (J_n / n)^{1/2}] n^{-2} \sum_{i=1}^{n} |U_i| \sum_{j=1}^{J} K_j(W_i) \]

\[ = O_p[\rho_{J_n} (J_n / n)^{1/2}] \left\{ n^{-2} \sum_{i=1}^{n} |U_i| \sum_{j=1}^{J} K_j(W_i) \right\} + O_p(J_n^{-\delta}) n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{J} K_j(W_i) \]

\[ = o_p(\rho^2 J / n) \left[ \sum_{i=1}^{n} |U_i| + o_p(J_n^{-\delta}) \right] \]

uniformly over \( J \in J \). It follows that

(A.15) \[ |\Delta S_{n1}(J)| = o_p(\rho J / n) \]

uniformly over \( J \in J \). A similar argument shows that

(A.16) \[ |\Delta S_{n2}(J)| = o_p(\rho J / n) \]

uniformly over \( J \in J \).

Now consider \( \Delta S_{n3}(J) \). We have
\[
\Delta S_{n3}(J) = n^{-2} \sum_{i=1}^{n} \{U_i - [g_{J_i}(X_i) - g(X_i)]\}^2 \sum_{j=1}^{J} \Delta K_j(W_i)
\]

\[
= n^{-2} \sum_{i=1}^{n} U_i^2 \sum_{j=1}^{J} \Delta K_j(W_i) - 2n^{-2} \sum_{i=1}^{n} U_i[g_{J_i}(X_i) - g(X_i)] \sum_{j=1}^{J} \Delta K_j(W_i)
\]

\[
+ n^{-2} \sum_{i=1}^{n} [g_{J_i}(X_i) - g(X_i)]^2 \sum_{j=1}^{J} \Delta K_j(W_i)
\]

\[
\equiv \Delta S_{n3a}(J) + \Delta S_{n3b}(J) + \Delta S_{n3c}(J).
\]

Some algebra shows that

\[
\Delta K_j = 2[(A_{J_i}^{-1})^* \Psi_j] \{[(A_{J_i}^{-1})^* - (A_{J_i}^{-1})^* \Psi_j]^* + [(A_{J_i}^{-1})^* - (A_{J_i}^{-1})^* \Psi_j]^2
\]

Moreover,

\[
(A_{J_i}^{-1})^* - (A_{J_i}^{-1})^* = -[I + (A_{J_i}^{-1})^* \Delta A^*]^{-1} (A_{J_i}^{-1})^* (\Delta A^*) (A_{J_i}^{-1})^*,
\]

where \(I\) is the identity operator and \(\Delta A^* = \hat{A}^* - A_{J_i}^*\). Now

\[
\|\Delta A^*\| = \|\hat{A} - A_{J_i}\|
\]

\[
= O_p[(J_n/n)^{1/2}]
\]

uniformly over \(J \in \mathcal{J}\) by lemma 2. Therefore, it follows from Lemma 1 that

\[
\left\|[(A_{J_i}^{-1})^* - (A_{J_i}^{-1})^* \Psi_j]\right\| = O_p[(\rho_j^3(J_n/n)^{1/2})],
\]

\[
\|\Delta K_j\| = O_p[(\rho_j^3(J_n/n)^{1/2})],
\]

\[
n^{-1} \sum_{j=1}^{J} \Delta K_j = O_p(\rho_j^3 J^2 J_n^{1/2} n^{-3/2}),
\]

and

\[
\Delta S_{n3a}(J) = O_p(\rho_j^3 J^2 J_n^{1/2} n^{-3/2})
\]

uniformly over \(J \in \mathcal{J}\). It follows that \(\Delta S_{n3a}(J) = o_p(\rho_j^2 J/n)\). Similar arguments apply to \(\Delta S_{n3b}, \Delta S_{n3c}, \Delta S_{n4}, \text{ and } \Delta S_{n5}\). The lemma follows by combining these results with (A.15) and (A.16). Q.E.D.

Lemma 6: For any \(\kappa > 0\), the following inequality holds uniformly over \(J \in \mathcal{J}\) as \(n \to \infty\).
\[ \|\hat{g}_J - g_J\|^2 \leq (4 + \varepsilon)(\log n)S_n(J)[1 + o_p(1)]. \]

Proof: Let \( s_{nJ} \) denote the leading term of the asymptotic expansion of \( \|\hat{g}_J - g_J\|^2 \). By Proposition 1 and lemma 3,

\begin{equation}
(A.17) \quad s_{nJ} = \sum_{j=1}^{J} \left\{ n^{-1} \sum_{i=1}^{n} U_i [(A^{-1}_j)^* \psi_j(W_j)]^2 \right\}. \label{eq:asymptotic_expansion}
\end{equation}

Define

\[ \sigma_j^2 = E[U_i [(A^{-1}_j)^* \psi_j(W_i)]^2] \]

and

\[ V_{ij} = \sigma_j^{-1} U_i [(A^{-1}_j)^* \psi_j(W_i)] \cdot \]

Then

\[ s_{nJ} = \sum_{j=1}^{J} \left( \sigma_j n^{-1} \sum_{i=1}^{n} V_{ij} \right)^2. \]

Define \( \xi_n = [(4 + \varepsilon)n^{-1}\log n]^{1/2} \),

\[ R_{nj} = n^{-1} \sum_{i=1}^{n} V_{ij}, \]

\[ R_{nj_1} = R_{nj} I(|R_{nj}| \leq \xi_n), \]

and

\[ R_{nj_2} = R_{nj} I(|R_{nj}| > \xi_n). \]

Then \( R_{nj} = R_{nj_1} + R_{nj_2} \). We now prove that \( \max_{1 \leq j \leq J_n} n^{1/2} R_{nj_2} = o_p(1) \) and, therefore, that

\[ \max_{1 \leq j \leq J_n} \left| \frac{R_{nj_1}}{R_{nj}} - 1 \right| = o_p(1). \]

Given any \( \delta > 0 \),

\[ P(n^{1/2} |R_{nj_2}| > \delta) \leq P(|R_{nj}| > \xi_n). \]

By Bernstein’s inequality

\[ P(|R_{nj}| > \xi_n) \leq 2 \exp \left( -\frac{n\xi^2_n}{4 + 2cJ\xi_n} \right) \]

for some finite constant \( c > 0 \) that does not depend on \( j \). Therefore,
\[ P(| R_{nj} | > \xi_n) \leq 2 \exp \left( -\frac{n \xi_n^2}{2 + \epsilon} \right) \]

\[ = 2n^{-1} \]

for all \( j \) whenever \( 2cJ \xi_n < \epsilon \). Now

\[ P(\max_{1 \leq j \leq J_n} n^{1/2} | R_{nj} | > \delta) \leq P(\max_{1 \leq j \leq J_n} | R_{nj} | > \xi_n) \]

\[ \leq \sum_{j=1}^{J_n} P( | R_{nj} | > \xi_n ) \]

\[ \leq 2J_n / n \to 0 \]

as \( n \to \infty \). It follows that \( R_{nj} = R_{nj}[1 + o_p(1)] \) uniformly over \( j \leq J_n \). Combining this result with (A.17) gives

\[ s_{nj} = \sum_{j=1}^{J_n} \sigma_j^2 \{ R_{nj}[1 + o_p(1)] \}^2 \]

\[ \leq \frac{\epsilon^2}{2n} \sum_{j=1}^{J_n} \sigma_j^2 [1 + o_p(1)] \]

\[ = (4 + \epsilon)(\log n) S_n(J)[1 + o_p(1)] \]

uniformly over \( J \in \mathcal{J} \). Q.E.D.

**Lemma 7**: The following inequality holds.

\[ \| \hat{\beta}_j \|^2 - \| g_j \|^2 \leq 3 \| \hat{\beta}_j - g_j \|^2 + 0.5 \| g_j - g \|^2 + 0.5 \| g_{j_{opt}} - g \|^2 \]

\[ + 2 \| \hat{\beta}_{j_{opt}} - g_{j_{opt}} \|^2 + 2 \langle g_{j_{opt}}, \hat{\beta}_{j_{opt}} - g_{j_{opt}} \rangle. \]

**Proof**: The proof of this lemma is similar to the proof of lemma 3.4(ii) of Loubes and Marteau (2009). We have

\[ \| \hat{\beta}_j \|^2 = \| (\hat{\beta}_j - g_j) + g_j \|^2 \]

\[ = \| \hat{\beta}_j - g_j \|^2 + 2 \langle g_j, \hat{\beta}_j - g_j \rangle + \| g_j \|^2. \]

Therefore,
\[ \| \hat{g}_j \|^2 - \| g_j \|^2 = \| \hat{g}_j - g_j \|^2 + 2 \langle g_j, \hat{g}_j - g_j \rangle. \]

Define \( \Sigma_{J_{opt}, J_{opt}} = \sum_{j=J_{opt}+1}^{j_{\hat{J}}} J_{opt} \), if \( \hat{J} > J_{opt} \), \( -\sum_{j=J_{opt}+1}^{j_{\hat{J}}} J_{opt} \) if \( \hat{J} < J_{opt} \), and 0 if \( \hat{J} = J_{opt} \). Then,

\[
2 \langle g_j, \hat{g}_j - g_j \rangle = 2 \sum_{j=1}^{j_{\hat{J}}} b_j (\tilde{b}_j - b_j)
\]

\[
= 2 \sum_{j=1}^{J_{opt}} b_j (\tilde{b}_j - b_j) + 2 \sum_{J_{opt}+1}^{j_{\hat{J}}} b_j (\tilde{b}_j - b_j)
\]

\[
= 2 \langle g_{J_{opt}}, \hat{g}_{J_{opt}} - g_{J_{opt}} \rangle + 2 \sum_{J_{opt}+1}^{j_{\hat{J}}} b_j (\tilde{b}_j - b_j)
\]

and

\[
(A.18) \quad \| \hat{g}_j \|^2 - \| g_j \|^2 = \| \hat{g}_j - g_j \|^2 + 2 \sum_{J_{opt}+1}^{j_{\hat{J}}} b_j (\tilde{b}_j - b_j) + 2 \langle g_{J_{opt}}, \hat{g}_{J_{opt}} - g_{J_{opt}} \rangle.
\]

Define

\[ R_n = 2 \sum_{J_{opt}+1}^{j_{\hat{J}}} b_j (\tilde{b}_j - b_j). \]

Then

\[ | R_n | \leq 2 \sum_{j=1}^{\infty} | I(j \leq \hat{J}) - I(j \leq J_{opt}) | | b_j (\tilde{b}_j - b_j) | . \]

But

\[ | I(j \leq \hat{J}) - I(j \leq J_{opt}) | = [ I(j \leq \hat{J}) + I(j \leq J_{opt}) ] | I(j \leq \hat{J}) - I(j \leq J_{opt}) | \]

\[ \leq I(j \leq \hat{J}) I(j > J_{opt}) + I(j \leq J_{opt}) I(j > \hat{J}). \]

Therefore,

\[ | R_n | \leq 2 \sum_{j=1}^{\infty} I(j \leq J_{opt}) I(j > \hat{J}) | b_j (\tilde{b}_j - b_j) | + 2 \sum_{j=1}^{\infty} I(j \leq \hat{J}) I(j > J_{opt}) | b_j (\tilde{b}_j - b_j) | . \]

By the Cauchy-Schwarz inequality,

\[ | R_n | \leq 2 \left( \sum_{j=J_{opt}+1}^{\infty} b_j \right)^{1/2} \left( \sum_{j=J_{opt}+1}^{\infty} (\tilde{b}_j - b_j)^2 \right)^{1/2} + 2 \left( \sum_{j=J_{opt}}^{\infty} b_j^2 \right)^{1/2} \left( \sum_{j=J_{opt}+1}^{\infty} (\tilde{b}_j - b_j)^2 \right)^{1/2} . \]

In addition, \( 2ab \leq a^2 / 2 + 2b^2 \) for any real numbers \( a \) and \( b \). Therefore,
\[ |R_n| \leq 0.5 \sum_{j=J_{\text{op}}}^{\infty} b_j^2 + 0.5 \sum_{j=J_1}^{J_{\text{op}}} b_j^2 + 2 \sum_{j=J_1}^{J_{\text{op}}} (\tilde{b}_j - b_j)^2 + 2 \sum_{j=1}^{J_1} (\tilde{b}_j - b_j)^2 \]

(A.19) \[ = 0.5 \left\| g_{J_{\text{op}}} - g \right\|^2 + 0.5 \left\| g_j - g \right\|^2 + 2 \left\| \hat{g}_{J_{\text{op}}} - g_{J_{\text{op}}} \right\|^2 + 2 \left\| \hat{g}_j - g_j \right\|^2. \]

The lemma follows by substituting (A.19) into (A.18). Q.E.D.

**Proof of Theorem 3.1:** Define \( a_n = (2 + \epsilon / 2) \log(n) \) and
\[
\hat{Q}_n(J) = \hat{T}_n(J) + \left\| g \right\|^2
\]
\[ = a_n \hat{S}_n(J) + \left\| g \right\|^2 - \left\| \hat{g}_j \right\|^2. \]

Then \( \hat{J} \) minimizes \( \hat{Q}_n(J) \) over \( J \in \mathcal{J} \). By lemmas 4 and 5,
\[
\hat{Q}_n(J) = a_n S_n(J)[1 + o_p(1)] + \left\| g \right\|^2 - \left\| \hat{g}_j \right\|^2
\]
uniformly over \( J \in \mathcal{J} \), so
\[
\hat{Q}_n(J) = a_n S_n(J)[1 + o_p(1)] + \left\| g \right\|^2 - \left\| \hat{g}_j \right\|^2.
\]

It follows that
\[
a_n S_n(\hat{J})[1 + o_p(1)] + \left\| g \right\|^2 - \left\| g_j \right\|^2 = \hat{Q}_n(\hat{J}) + \left\| \hat{g}_j \right\|^2 - \left\| g_j \right\|^2.
\]

An application of lemma 7 gives
\[
a_n S_n(\hat{J})[1 + o_p(1)] + \left\| g \right\|^2 - \left\| g_j \right\|^2 \leq \hat{Q}_n(\hat{J}) + 3 \left\| \hat{g}_j - g_j \right\|^2 + 0.5 \left( \left\| g \right\|^2 - \left\| g_j \right\|^2 \right)
\]
\[ + 0.5 \left\| g_{J_{\text{op}}} - g \right\|^2 + 2 \left\| \hat{g}_{J_{\text{op}}} - g_{J_{\text{op}}} \right\|^2 + 2 \left\langle g_{J_{\text{op}}}, \hat{g}_{J_{\text{op}}} - g_{J_{\text{op}}} \right\rangle.
\]

Therefore,
\[
a_n S_n(\hat{J})[1 + o_p(1)] + 0.5 \left( \left\| g \right\|^2 - \left\| g_j \right\|^2 \right) - 3 \left\| \hat{g}_j - g_j \right\|^2
\]
\[ \leq \hat{Q}_n(\hat{J}) + 0.5 \left\| g_{J_{\text{op}}} - g \right\|^2 + 2 \left\| \hat{g}_{J_{\text{op}}} - g_{J_{\text{op}}} \right\|^2 + 2 \left\langle g_{J_{\text{op}}}, \hat{g}_{J_{\text{op}}} - g_{J_{\text{op}}} \right\rangle.
\]

By lemma 6,
\[
a_n [(4 + \epsilon) \log n]^{-1} \left\| \hat{g}_j - g_j \right\|^2 \leq a_n S_n(\hat{J})[1 + o_p(1)].
\]

Therefore,
\[
0.5\left\| \hat{g}_{J} - g \right\|^2 [1 + o_p(1)] \leq \hat{Q}_n(\hat{J}) + 0.5\left\| g_{J_{opt}} - g \right\|^2
\]
\[
+ 2\left\| \hat{g}_{J_{opt}} - g_{J_{opt}} \right\|^2 + 2\left\langle g_{J_{opt}}, \hat{g}_{J_{opt}} - g_{J_{opt}} \right\rangle.
\]
But \( \hat{Q}_n(\hat{J}) \leq \hat{Q}_n(J_{opt}) \), so
\[
0.5\left\| \hat{g}_{J} - g \right\|^2 [1 + o_p(1)] \leq \hat{Q}_n(J_{opt}) + 0.5\left\| g_{J_{opt}} - g \right\|^2
\]
\[
+ 2\left\| \hat{g}_{J_{opt}} - g_{J_{opt}} \right\|^2 + 2\left\langle g_{J_{opt}}, \hat{g}_{J_{opt}} - g_{J_{opt}} \right\rangle.
\]
In addition,
\[
\hat{Q}_n(J_{opt}) = a_n S_n(J_{opt}) + \left\| s \right\|^2 - \left\| \hat{g}_{J_{opt}} \right\|^2.
\]
Therefore, by lemmas 4 and 5,
\[
\hat{Q}_n(J_{opt}) = a_n S_n(J_{opt})[1 + o_p(1)] + \left\| s \right\|^2 - \left\| \hat{g}_{J_{opt}} \right\|^2.
\]
But
\[
\left\| \hat{g}_{J_{opt}} \right\|^2 = \left\| \hat{g}_{J_{opt}} - g_{J_{opt}} \right\|^2 + \left\| g_{J_{opt}} \right\|^2 + 2\left\langle g_{J_{opt}}, \hat{g}_{J_{opt}} - g_{J_{opt}} \right\rangle,
\]
so
\[
\hat{Q}_n(J_{opt})
\]
\[
= a_n S_n(J_{opt})[1 + o_p(1)] + \left\| g \right\|^2 - \left\| \hat{g}_{J_{opt}} - g_{J_{opt}} \right\|^2 - \left\| g_{J_{opt}} \right\|^2 - 2\left\langle g_{J_{opt}}, \hat{g}_{J_{opt}} - g_{J_{opt}} \right\rangle
\]
and
\[
0.5\left\| \hat{g}_{J} - g \right\|^2 [1 + o_p(1)] \leq a_n S_n(J_{opt})[1 + o_p(1)] + \left\| \hat{g}_{J_{opt}} - g_{J_{opt}} \right\|^2 + 1.5\left\| g_{J_{opt}} - g \right\|^2.
\]
Now, \( E_A \left\| \hat{g}_{J_{opt}} - g_{J_{opt}} \right\|^2 = S_n(J_{opt}) \). Therefore, for any \( n > 1 \),
\[
0.5E_A \left\| \hat{g}_{J} - g \right\|^2 \leq (a_n + 1) S_n(J_{opt}) + 1.5\left\| g_{J_{opt}} - g \right\|^2
\]
\[
\leq (a_n + 1) E_A \left\| \hat{g}_{J_{opt}} - g \right\|^2,
\]
and
\[
E_A \left\| \hat{g}_{J} - g \right\|^2 \leq 2(a_n + 1) E_A \left\| \hat{g}_{J_{opt}} - g \right\|^2. \quad \text{Q.E.D.}
\]
Figure 1: Graph of $g(x)$
## TABLE 1: RESULTS OF MONTE CARLO EXPERIMENTS

<table>
<thead>
<tr>
<th>Exp’t No.</th>
<th>Empirical mean of $\parallel \hat{g}<em>{j</em>{opt}} - g \parallel^2$</th>
<th>Empirical mean of $\parallel \hat{g}_j - g \parallel^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.119</td>
<td>0.121</td>
</tr>
<tr>
<td>2</td>
<td>0.163</td>
<td>0.165</td>
</tr>
<tr>
<td>3</td>
<td>0.187</td>
<td>0.189</td>
</tr>
<tr>
<td>4</td>
<td>0.127</td>
<td>0.127</td>
</tr>
</tbody>
</table>
REFERENCES


Chen, X. and D. Pouzo (2008). Estimation of nonparametric conditional moment models with possibly nonsmooth moments, working paper, Department of Economics, Yale University, New Haven, CT.


Florens, J.-P. and A. Simoni (2010). Nonparametric estimation of an instrumental regression: a quasi-Bayesian approach based on regularized posterior, working paper, Department of Decision Sciences, Bocconi University, Milan, Italy.


Horowitz, J.L. and S. Lee (2010). Uniform confidence bands for functions estimated nonparametrically with instrumental variables, Cemmap working paper cwp1809, Department of Economics, University College London.


