Econometric Theory, **25**, 2009, 1–42. Printed in the United States of America.
doi:10.1017/S0266466608090014

NONPARAMETRIC ESTIMATION OF REGRESSION FUNCTIONS WITH DISCRETE REGRESSORS

DESHENG OUYANG
Shanghai University of Finance and Economics

QI LI
Texas A&M University
and
Tsinghua University

JEFFREY S. RACINE
McMaster University

We consider the problem of estimating a nonparametric regression model containing categorical regressors only. We investigate the theoretical properties of least squares cross-validated smoothing parameter selection, establish the rate of convergence (to zero) of the smoothing parameters for relevant regressors, and show that there is a high probability that the smoothing parameters for irrelevant regressors converge to their upper bound values, thereby automatically smoothing out the irrelevant regressors. A small-scale simulation study shows that the proposed cross-validation-based estimator performs well in finite-sample settings.

1. INTRODUCTION

Nonparametric and semiparametric methods have attracted much attention among econometricians and statisticians in the last two decades. These methods have been successfully applied to a range of problem domains, including the estimation of treatment effects (see Hahn, 1998; Hirano, Imbens, and Ridder, 2003) and the analysis of auctions (see Guerre, Perrigne, and Vuong, 2000; Li, Perrigne, and Vuong, 2002, 2003) and are also being employed in the financial econometrics literature (see Hong and Lee, 2003, 2005).

The seminal work of Aitchison and Aitken (1976) has spawned a rich literature on the kernel smoothing of discrete (categorical) variables. This literature was

We gratefully acknowledge comments from a co-editor and from two anonymous referees that led to a much improved version of this paper. All errors remain, naturally, our own. Li's research is partially supported by the Private Enterprise Research Center, Texas A&M University. Racine gratefully acknowledges support from the Social Sciences and Humanities Research Council of Canada (SSHRC: www.sshrc.ca) and the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca). Address correspondence to Jeffrey S. Racine, Department of Economics, McMaster University, Graduate Program in Statistics, McMaster University, Hamilton, ON L8S 4M4, Canada; e-mail: racinej@mcmaster.ca.

 0266-4666/09 $15.00

motivated mainly by the need to deal with the "small cell" problem frequently encountered in the analysis of multivariate discrete data. Examples of this literature include the work of Titterington (1980), Hall (1981), Wang and van Ryzin (1981), Bierens (1983), Bowman, Hall, and Titterington (1984), Hall and Wand (1988), and Grund and Hall (1993), to mention only a few (see also the monographs by Scott, 1992; Fahrmeir and Tutz, 1994; Simonoff, 1996). A brief survey of this literature leads one rather quickly to the realization that concern lies almost exclusively with the estimation of the (conditional) probability distribution of the discrete variables. Though it is not uncommon to encounter situations in which regressors are exclusively discrete (e.g., survey data, medical data, etc.), much less effort has been devoted to the regression framework when dealing with discrete regressors.

In this paper we study the theoretical properties of a data-driven least squares cross-validation (CV) method for selecting the smoothing parameters in a regression model composed solely of discrete regressors. We point out from the outset that we are in no way "interpolating" between realizations of discrete random variables; rather, we are simply smoothing a multivariate mean vector, or, alternatively, one could think of the approach as shrinking a multivariate mean toward a global mean in the Bayesian sense.

Our analysis is necessarily more complex than that underlying the probability distribution framework because of the existence of a random denominator in the nonparametric kernel estimator of a multivariate mean. We consider a general nonparametric regression model in which we allow for the possibility that some of the discrete regressors have a natural ordering, e.g., preferences (dislike, indifference, like), health (excellent, good, poor), etc. We shall distinguish between "relevant" and "irrelevant" regressors, and we derive the convergence rate of the CV smoothing parameters associated with the relevant regressors. We also demonstrate theoretically that, when irrelevant regressors are present, the associated smoothing parameters do not converge to zero; rather, with high probability they converge to their upper bound values, thereby smoothing out irrelevant regressors. A small-scale simulation study shows that the proposed CV-based estimator performs well in finite-sample settings.

Recently, Li and Racine (2004, 2007), Hall, Racine, and Li (2004), Racine and Li (2004), and Hall, Li, and Racine (2007) have considered nonparametric estimation of regression functions, conditional density, and distribution functions, and quantile functions containing a mix of discrete and continuous regressors. Theoretical results developed in the current paper highlight the fact that regression models having only discrete regressors are qualitatively different from those containing a mix of discrete and continuous regressors, both in theory and in practice. In particular, in the mixed regressor case with at least one relevant continuous regressor, irrelevant regressors can be smoothed out with probability approaching one as the sample size increases. However, in the discrete regressor only case, although the irrelevant regressors can be smoothed out with a high probability, the probability is *strictly less* than one even as the sample size goes to infinity. Also,

for the discrete regressor only case considered herein, the smoothing parameters associated with the relevant regressors converge to zero at the rate of n^{-1} or $n^{-1/2}$, where n is the sample size, depending on whether or not there exist irrelevant regressors. This result again differs markedly from the mixed regressor case, where the presence of irrelevant regressors does not affect the rate of convergence of the smoothing parameters associated with the relevant regressors. Therefore, the discrete regressor only model necessarily warrants a separate treatment because the results simply cannot be obtained as a special case of those derived for mixed discrete and continuous regressor models. We hope that the novel results developed in this paper will prove to be of interest to practitioners and theoreticians alike.

2. KERNEL REGRESSION WITH DISCRETE REGRESSORS: THE RELEVANT REGRESSOR CASE

Consider a nonparametric regression model given by

$$Y_i = g(X_i) + u_i, \tag{1}$$

where $g(\cdot)$ is an unknown function, X_i is an r-dimensional vector of discrete regressors, and u_i is an error term satisfying $\mathrm{E}(u_i|X_i) = 0$.

We use x_s to denote the sth component of x, we assume that x_s takes c_s different values in $D_s \stackrel{def}{=} \{0, 1, \ldots, c_s - 1\}$, $s = 1, \ldots, r$, and let $c_s \geq 2$ be a finite positive constant. For expositional simplicity we will mainly focus on the case in which the components of x are unordered discrete regressors,[1] and we postpone the treatment of ordered discrete regressors until Section 3.1.

For an unordered regressor, we suggest using a variant of the Aitchison and Aitken (1976) kernel function defined as

$$l(X_{is}, x_s, \lambda_s) = \begin{cases} 1, & \text{when } X_{is} = x_s, \\ \lambda_s, & \text{otherwise.} \end{cases} \tag{2}$$

Let $\mathbf{1}(A)$ denote the usual indicator function, which assumes the value one if A holds true, zero otherwise. Using (2), we can construct a product kernel function given by

$$L(X_i, x, \lambda) = \prod_{s=1}^{r} l(X_{is}, x_s, \lambda_s) = \prod_{s=1}^{r} \lambda_s^{\mathbf{1}(X_{is} \neq x_s)}. \tag{3}$$

Note that $\lambda_s = 0$ along with the convention $0^0 = 1$ leads to an indicator function, whereas $\lambda_s = 1$ leads to a uniform weight function. Therefore, the range of λ_s is $[0, 1]$ for all $s = 1, \ldots, r$.

We use $\mathcal{D}$ to denote the range assumed by X_i. For $x \in \mathcal{D}$, we estimate the probability function $p(x)$ by

$$\hat{p}(x) = \frac{1}{n} \sum_{j=1}^{n} L(X_j, x, \lambda), \tag{4}$$

and we estimate $g(x)$ by

$$\hat{g}(x) = \frac{n^{-1}\sum_{j=1}^{n} Y_j L(X_j, x, \lambda)}{\hat{p}(x)}. \tag{5}$$

Observe that the kernel weight function we use here does not add up to one when $\lambda_s \neq 0$; however, this does not affect the nonparametric estimator $\hat{g}(x)$ defined in (5) as the kernel function appears in both the numerator and the denominator of (5) and thus the kernel function can be multiplied by any nonzero constant, leaving the definition of $\hat{g}(x)$ intact.

Note that when $\lambda_s = 0$ for all $s = 1, \ldots, r$, our estimator reverts to the conventional approach whereby one uses a frequency estimator to deal with the discrete regressors, whereas if $\lambda_s = 1$ for some s, then $\hat{g}(x)$ becomes unrelated to x_s. That is, x_s is smoothed out from the regression model when $\lambda_s = 1$ (it is deemed to be an "irrelevant" regressor).

We choose $\lambda \stackrel{def}{=} (\lambda_1, \ldots, \lambda_r)$ to minimize[2]

$$CV(\lambda) = \sum_{i=1}^{n} [Y_i - \hat{g}_{-i}(X_i)]^2, \tag{6}$$

where $\lambda = (\lambda_1, \ldots, \lambda_r)$,

$$\hat{g}_{-i}(X_i) = \frac{\frac{1}{n-1}\sum_{j=1, j\neq i}^{n} Y_j L(X_i, X_j, \lambda)}{\hat{p}_{-i}(X_i)} \tag{7}$$

is the leave-one-out kernel estimator of $g(X_i)$, and

$$\hat{p}_{-i}(X_i) = \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} L(X_i, X_j, \lambda) \tag{8}$$

is the leave-one-out estimator of $p(X_i)$. We will use $\hat{\lambda} \stackrel{def}{=} (\hat{\lambda}_1, \ldots, \hat{\lambda}_r)$ to denote the CV choice of λ that minimizes (6).

First we introduce some notation and provide a definition. Let x_{-s} denote x with x_s excluded; i.e., $x_{-s} = (x_1, \ldots, x_{s-1}, x_{s+1}, \ldots, x_r)$. Let $\mathcal{D}_{-s}$ and $\mathcal{D}_s$ denote the supports of x_{-s} and x_s, respectively. We say that $g(x)$ is a *constant function* with respect to x_s if

$$g(x_s, x_{-s}) = g(z_s, x_{-s}) \quad \text{for all } x_s, z_s \in \mathcal{D}_s \quad \text{and all } x_{-s} \in \mathcal{D}_{-s}.$$

That is, $g(\cdot)$ does not vary as x_s changes. In this case, ideally one should remove x_s from the regression model. In this section we assume that all x_s's are relevant regressors. That is, $g(x)$ is *not* a constant function with respect to x_s for all $s = 1, \ldots, r$. We shall make the following assumptions.

Assumption 1.

(a) $\{X_i, Y_i\}_{i=1}^{n}$ are independent and identically distributed (i.i.d.) as (X, Y).

(b) $\mathrm{E}(Y_i^2 | X_i = x)$ is bounded for all $x \in \mathcal{D}$.

Assumption 2. For all $x \in \mathcal{D}$, the only values of $(\lambda_1, \ldots, \lambda_r)$ that make $\left\{ \sum_{z \in \mathcal{D}} p(z)[g(x) - g(z)]L(x, z, \lambda) \right\}^2 = 0$ are $\lambda_s = 0$ for all $s = 1, \ldots, r$.

Assumption 1 is quite standard. Assumption 2 implies that $g(x)$ is not a constant function with respect to any component $x_s \in \mathcal{D}_s$. This is needed to prove that $\hat{\lambda} = o_p(1)$, which in turn is needed to establish the rate of convergence of $\hat{\lambda}_s$ $(s = 1, \ldots, r)$.

Considering the case for which $r = 1$, then Assumption 2 becomes, for all $x \in \mathcal{D}$ (x is a scalar because $r = 1$), $\sum_{z \in \mathcal{D}} \{[g(x) - g(z)][\mathbf{1}(x = z) + \lambda \mathbf{1}(x \neq z)]\}^2 = 0$, which is equivalent to $\lambda^2 \sum_{z \in \mathcal{D}} [g(x) - g(z)]^2 \mathbf{1}(x \neq z) = 0$ because $[g(x) - g(z)]$ $\mathbf{1}(x = z) \equiv 0$. However, because $g(x)$ is not a constant function, we know that $\sum_{z \in \mathcal{D}} [g(x) - g(z)]^2 \mathbf{1}(x \neq z) > 0$. Hence, we must have $\lambda = 0$. Thus, Assumption 2 holds true if and only if $g(x)$ is not a constant function when x is a scalar.

The next theorem provides the rate of convergence of the CV selected smoothing parameters.

THEOREM 2.1. *Under Assumptions 1 and 2, we have*

$$\hat{\lambda}_s = O_p(n^{-1}) \quad \textit{for } s = 1, \ldots, r.$$

The proof of Theorem 2.1 is given in Appendix A.

Theorem 2.1 shows that, when all of the regressors are relevant, the CV selected smoothing parameters converge to zero at a fast rate of n^{-1}. It is interesting to note that the preceding n^{-1} rate of convergence is faster from the rate obtained by Hall et al. (2007) when the regression function also contains continuous regressors.[3] Therefore, the discrete-regressor-only case must be treated separately because one cannot obtain the preceding result as a corollary from the mixed discrete and continuous regressor model case.

From Theorem 2.1 one can easily obtain the following result.

THEOREM 2.2. *Under the same conditions as those given in Theorem 2.1, then*

$$\sqrt{n}(\hat{g}(x) - g(x))/\sqrt{\hat{\Omega}(x)} \to N(0, 1) \quad \textit{in distribution},$$

where $\hat{\Omega}(x) = \hat{\sigma}^2(x)/\hat{p}(x)$ *and where* $\hat{\sigma}^2(x) = n^{-1} \sum_i [Y_i - \hat{g}(X_i)]^2 L(X_i, x, \hat{\lambda})/\hat{p}(x)$ *is a consistent estimator of* $\sigma^2(x) = \mathrm{E}(u_i^2 | X_i = x)$.

The proof of Theorem 2.2 is given in Appendix A.

In the next section we discuss the case for which some of the regressors are irrelevant.

3. KERNEL REGRESSION WITH DISCRETE REGRESSORS: THE IRRELEVANT REGRESSOR CASE

In this section we allow for the possibility that some of the regressors are in fact irrelevant in the sense that they are independent of the dependent variable. Without loss of generality we assume that the first r_1 $(1 \leq r_1 < r)$ components of X_i are relevant, whereas the remaining $r_2 = r - r_1$ components of X_i are irrelevant. Let $\overline{X}_i$ denote the r_1-dimensional vector of relevant components of X_i and let $\widetilde{X}_i$ denote the r_2-dimensional vector of irrelevant components. Similar to the approach taken in Hall et al. (2007) we shall assume that

$$(Y, \overline{X}) \quad \text{and} \quad \widetilde{X} \quad \text{are independent of each other.} \tag{9}$$

Assumption (9) is quite strong as it requires independence not only between $\widetilde{X}$ and Y but also between $\widetilde{X}$ and $\overline{X}$. A weaker assumption would be to require that, conditional on $\overline{X}$, Y is independent of $\widetilde{X}$ (so that $\widetilde{X}$ and $\overline{X}$ can be correlated). Though simulation results reported in Section 4 support this conjecture, at this time we are unable to relax Assumption (9) theoretically. Therefore, we will impose this condition to prove results stated in Theorems 3.1 and 3.2 that are given later in this section.

The kernel estimator of $g(x)$ and the definition of the CV objective function are the same as those given in Section 2. We still use $(\hat{\lambda}_1, \ldots, \hat{\lambda}_r)$ to denote the CV selected smoothing parameters. In Theorem 3.1 we show that (i) the smoothing parameters associated with the relevant regressors converge to zero at the rate of $n^{-1/2}$, which differs from the n^{-1} rate of convergence when there do not exist any irrelevant components, and (ii) the smoothing parameters associated with the irrelevant components will not converge to zero; rather they have a high probability of converging to their upper extreme values so that these irrelevant regressors are smoothed out with high probability. This stands in stark contrast to the results for the mixed discrete and continuous regressor case considered in Hall et al. (2007). Furthermore, there is also a positive probability that these smoothing parameters do not converge to their upper extreme values even as $n \to \infty$.

Mirroring the notation used for $\bar{x}$ and $\tilde{x}$, we shall use $L(\bar{x}, \bar{z}, \bar{\lambda}) = \prod_{s=1}^{r_1} l(\bar{x}_s, \bar{z}_s, \lambda_s)$ and $L(\tilde{x}, \tilde{z}, \tilde{\lambda}) = \prod_{s=r_1+1}^{r} l(\tilde{x}_s, \tilde{z}_s, \lambda_s)$ to denote the kernel functions associated with the relevant and the irrelevant covariates, respectively. Also, using $\bar{p}(\cdot)$ and $\tilde{p}(\cdot)$ to denote the marginal probability functions of $\overline{X}$ and $\widetilde{X}$, respectively, then by independence $p(x) = \bar{p}(\bar{x})\tilde{p}(\tilde{x})$, and using $\overline{D}$ to denote the support of $\bar{x}$, Assumption 2 is modified as follows.

Assumption 3. For all $x \in \overline{D}$, the only values of $\lambda_1, \ldots, \lambda_{r_1}$ that make

$$\left\{ \sum_{\bar{z} \in \overline{D}} \bar{p}(\bar{z})[g(\bar{x}) - g(\bar{z})]L(\bar{x}, \bar{z}, \bar{\lambda}) \right\}^2 = 0$$

are $\lambda_s = 0$ for all $s = 1, \ldots, r_1$.

Assumption 3 ensures that the CV selected smoothing parameters associated with the relevant regressors will converge to zero, whereas we do not impose any assumption on the smoothing parameters associated with the irrelevant regressors except that they assume values in the unit interval [0, 1].

The next theorem provides the asymptotic behavior of the CV selected smoothing parameters.

THEOREM 3.1. *Assume that* $r_1 \geq 1$ *and* $r_2 \geq 1$ *(with* $r = r_1 + r_2 \geq 2$*). Then under Assumptions 1 and 3 and (9), we have*

$$\hat{\lambda}_s = O_p(n^{-1/2}), \quad \text{for } s = 1, \ldots, r_1, \quad \text{and}$$

$$\lim_{n\to\infty} \Pr(\hat{\lambda}_{r_1+1} = 1, \ldots, \hat{\lambda}_r = 1) \geq \alpha \quad \text{for some } \alpha \in (0, 1).$$

The proof of Theorem 3.1 is given in Appendix B. We emphasize here that the rate of $\hat{\lambda}_s = O_p(n^{-1/2})$ $(s = 1, \ldots, r_1)$ is sharp. That is, $\hat{\lambda}_s$ goes to zero at exactly the rate of $O_p(n^{-1/2})$, and it cannot be faster than this rate. In particular, it cannot reach the $O_p(n^{-1})$ rate that occurs in the case where irrelevant variables are not present; see the arguments following equation (B.6) in Appendix B for the underpinnings of this result.

Theorem 3.1 states that the smoothing parameters associated with the relevant regressors all converge to zero at the rate of $n^{-1/2}$, whereas the smoothing parameters for the irrelevant regressors have a positive probability of assuming their upper bound value of one; that is, there is a positive probability that the irrelevant regressors will be smoothed out. It is difficult to determine the exact value of α for the general case because the exact value of α depends on the unknown functions $g(\cdot)$, $p(\cdot)$, and $\sigma^2(\cdot)$. However, when u_i is symmetrically distributed around zero and is independent of X_i, then it is expected that $\alpha > 0.5$. Our simulation shows that there is usually about a 60% chance that $\hat{\lambda}_s$ takes the upper extreme value one and hence a 40% chance that $\hat{\lambda}_s$ takes values between zero and one, for $s = r_1 + 1, \ldots, r$.

For $\bar{v}, \bar{x} \in \overline{D}$, define $\mathbf{1}_s(\bar{v}, \bar{x}) = \mathbf{1}(\bar{v}_s \neq \bar{x}_s) \prod_{t=1, t\neq s}^{r_1} \mathbf{1}(\bar{v}_s = \bar{x}_s)$. Note that $\mathbf{1}_s(\bar{v}, \bar{x})$ is an indicator function equal to one if $\bar{v}$ and $\bar{x}$ only differ in their sth component, zero otherwise. Also, define $v_l(\tilde{x}) = \mathrm{E}[(L(\widetilde{X}_i, \widetilde{X}_j, \tilde{\lambda}))^l | \widetilde{X}_i = \tilde{x}]$ $(l = 1, 2)$.

The asymptotic distribution of $\hat{g}(x)$ is given by the following theorem.

THEOREM 3.2. *Under the same conditions as in Theorem 3.2, we have*

$$\sqrt{n}\left(\hat{g}(x) - g(\bar{x}) - \sum_{s=1}^{r_1} \hat{\lambda}_s B_s(\bar{x})\right) / \sqrt{\Sigma(x)} \to N(0, 1) \quad \text{in distribution},$$

where $B_s(\bar{x}) = \bar{p}(\bar{x})^{-1} \sum_{\bar{v}\in\overline{D}} \mathbf{1}_s(\bar{x}, \bar{v})(g(\bar{v}) - g(\bar{x}))$ *and* $\Sigma(x) = \sigma^2(\bar{x})v_2(\tilde{x})/[\bar{p}(\bar{x})v_1(\tilde{x})^2]$. *Moreover,* $B_s(\bar{x})$ *can be consistently estimated by* $\hat{B}_s(\bar{x}) = \hat{p}(\bar{x})^{-1} \sum_{\bar{v}\in\overline{D}} \hat{p}(\bar{v})\mathbf{1}_s(\bar{v}, \bar{x})(\hat{g}(\bar{v}) - \hat{g}(\bar{x}))$, *and* $\Sigma(x)$ *can be consistently estimated by*

$\hat{\Sigma}(x) = [n^{-1}\sum_i \hat{u}_i^2 L_{ij,\hat{\lambda}}^2]/[n^{-1}\sum_i L_{ij,\hat{\lambda}}]^2$, $\hat{u}_i = Y_i - \hat{g}(X_i)$ *and* $L_{ij,\hat{\lambda}} = L(X_j, X_i, \hat{\lambda})$.

The proof of Theorem 3.2 is given in Appendix B.

Note that in computing $\hat{\Sigma}(x)$ one does not need to know which variables are relevant or irrelevant. However, to compute $\hat{B}_s(\bar{x})$, one needs to know the set of relevant variables (ex post). If the CV method selects $\hat{\lambda}_s = 1$, then one knows that the corresponding x_s is an irrelevant variable, and if $\hat{\lambda}_s$ is very small, say, $\hat{\lambda}_s = 0.01$, then it is highly likely that x_s is a relevant variable. For $\hat{\lambda}_s$ values in the middle of the interval $[0, 1]$, it is less clear whether x_s should be treated as a relevant or an irrelevant regressor. In this case one can use the bootstrap testing procedure proposed by Racine, Hart, and Li (2006) to formally test whether x_s is an irrelevant regressor or not.

Also note that the leading bias term $B_s(\bar{x})$ does not depend on the irrelevant regressors $\tilde{x}$, whereas the leading variance term $\Sigma(x)$ depends on $\tilde{x}$. By Hölder's inequality we know that $\nu_2(\tilde{x})/\nu_1(\tilde{x}) \geq 1$. It equals one if and only if $\lambda_s = 1$ for all $s = r_1 + 1, \ldots, r$. However, by Theorem 3.1 we know that the probability that all irrelevant variables can be smoothed out is strictly less than one. Hence, there is a positive probability that the asymptotic variance $\Sigma(x)$ is larger than for the case where the irrelevant regressors are removed. That is, there exists a loss in efficiency as compared to the case where all variables are relevant (the efficiency loss arises from the presence of the irrelevant variables).

3.1. Ordered Discrete Regressors

We now discuss the case for which some of the discrete regressors have a natural ordering. For an ordered discrete regressor $X_{i,s}$ taking c_s different values, Aitchison and Aitken (1976) suggest using $l(X_{is}, x_s, \lambda_s) = \binom{c_s-1}{t}(1-\lambda_s)^{c_s-1-t}\lambda_s^t$ when $X_{is} - x_s = t$, where $\binom{c_s-1}{t} = (c_s-1)!/[t!(c_s-1-t)!]$ $(t = 0, \ldots, c_s-1)$. However, when $c_s \geq 3$, this kernel function suffers from the defect that there does not exist a value of λ_s such that $l(X_{is}, x_s, \lambda_s)$ equals a constant function. Hence, even when x_s is an irrelevant regressor, one cannot smooth it out. In this paper we suggest a simple alternative, and for an ordered regressor we suggest using the following kernel:

$$l(X_{is}, x_s, \lambda_s) = \begin{cases} 1, & \text{if } X_{is} = x_s, \\ \lambda_s^{|X_{is}-x_s|}, & \text{if } X_{is} \neq x_s. \end{cases} \tag{10}$$

When $\lambda_s = 0$, we get an indicator function, and when $\lambda_s = 1$, we get a uniform weight function. Therefore, the range of λ_s is $[0, 1]$. When λ_s takes the upper bound value of one, x_s becomes an irrelevant regressor (i.e., it is completely smoothed out). When some of the regressors are ordered discrete regressors, we use the kernel function defined in (10) and modify the definition of $\mathbf{1}_s(\bar{v}, \bar{x})$ so

that $\mathbf{1}_s(\bar{v}, \bar{x}) = \mathbf{1}(|\bar{v}_s - \bar{x}_s| = 1) \prod_{t=1, t\neq s}^{r_1} \mathbf{1}(\bar{v}_s = \bar{x}_s)$ when $\bar{x}_s$ is an ordered discrete variable. Then it can then be shown that the conclusions of Theorems 2.1, 2.2, 3.1, and 3.2 remain unchanged. That is, $\hat{\lambda}_s = O_p(n^{-1})$ when there do not exist irrelevant regressors, and $\hat{\lambda}_s = O_p(n^{-1/2})$ when there exist some irrelevant regressors but x_s is a relevant regressor, whereas $\hat{\lambda}_s$ has a positive probability of taking the upper extreme value of one when x_s is irrelevant.

We now briefly discuss how our approach handles the empty cell problem. Suppose that Y is medical expenditure and X takes four values $\{0, 1, 2, 3\}$ corresponding to poor, ordinary, good, and excellent health status for a person.[4] Now suppose that there are no sample realizations for persons having ordinary health status (i.e., $x = 1$ is an empty cell). Letting λ be the smoothing parameter, then our estimator $\hat{g}(1)$ is given by

$$\hat{g}(1) = \frac{\sum_i Y_i L(X_i, 1, \lambda)}{\sum_i L(X_i, 1, \lambda)} = \frac{\sum_{|X_i - 1|=1} Y_i \lambda + \sum_{|X_i - 1|=2} Y_i \lambda^2}{\sum_{|X_i - 1|=1} \lambda + \sum_{|X_i - 1|=2} \lambda^2}.$$

Clearly this estimate is a weighted average of the Y_i's with the weight depending on the distance $|X_i - 1|$. Individuals with poor or good health receive weight λ (they are close to "ordinary" health), whereas those with "excellent" health receive the smaller weight of λ^2. Thus, our smoothing estimator uses data on the outcomes lying in the "nearby" nonempty cells to impute a value of the unknown conditional mean $g(1)$ for empty cells, namely, $\hat{g}(1)$.

We examine the finite-sample performance of the proposed estimator in the next section.

4. MONTE CARLO SIMULATIONS

In this section we consider three simulation experiments that highlight the behavior of the proposed method in finite-sample settings. The first simulation experiment examines properties of the proposed approach, the second examines the effects of combining an irrelevant and a relevant variable into a single variable, and the third focuses on the performance of the proposed approach relative to the conventional frequency estimator and the popular linear parametric model.

4.1. Finite-Sample Performance

For this experiment we simulate data from

$$Y_i = m(X_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i, \tag{11}$$

where $\beta = (\beta_0, \beta_1, \beta_2, \beta_3) = (1, 1, 0.25, 0)$ and $u \sim N(0, 1)$. Each $X_{ji} \in \{0, 1, 2\}$ with unequal probabilities; hence there exist 27 "cells" in this application. Clearly X_{1i} is relevant ($\beta_1 = 1$), X_{3i} is irrelevant ($\beta_3 = 0$), and although X_{2i} is relevant, it is less important than X_{1i} because its coefficient is much smaller than that for X_{1i}.

We also vary the degree of correlation among X_1, X_2, and X_3 to determine whether or not our independence assumption can be weakened, letting the degree of correlation $\rho = \rho_{x_1,x_2} = \rho_{x_1,x_3} = \rho_{x_2,x_3}$ equal $(0.00, 0.25, 0.50, 0.75)$.

We draw 1,000 Monte Carlo replications, and for each replication we generate the CV bandwidths via multivariate numerical minimization and then compute the mean square error (MSE) for three models defined as $\text{MSE} = n^{-1}\sum_{i=1}^{n}(m(X_i) - \hat{m}(X_i))^2$ where $m(X_i) = \beta_0 + \beta_1 X_{1i} + \beta X_{2i} + \beta_3 X_{3i}$ and where $\hat{m}(X_i)$ represents a model's fitted value. We report the median MSEs for the 1,000 replications in Table 1. The three models are the kernel estimator with CV bandwidths (Kernel), the parametric estimator that relies on knowledge of the underlying model (Param), and the frequency estimator (Freq). We conduct a range of simulations that vary with the sample size, n. For the kernel estimator, bandwidth distributions are summarized in Figure 1 for $\rho = 0$.[5]

Table 1 reveals that, on MSE grounds, the proposed approach dominates the frequency approach in finite-sample settings as expected, whereas a *correctly specified* parametric model performs the best as expected because it exploits knowledge of the true data generating process (DGP). Of course, in practice one's parametric model may in fact be misspecified, and we shall consider this case in Section 4.3. Finally, it appears that the method performs well even when the degree of correlation among the regressors is large.

TABLE 1. Median MSE summary for the proposed method (Kernel), the frequency method (Freq), and the correctly specified parametric model (Param)

n	Kernel	Param	Freq
$\rho = 0.00$			
100	0.0603	0.0321	0.2267
500	0.0173	0.0068	0.0518
1,000	0.0092	0.0035	0.0267
$\rho = 0.25$			
100	0.0603	0.0321	0.2267
500	0.0167	0.0068	0.0532
1,000	0.0087	0.0033	0.0263
$\rho = 0.50$			
100	0.0599	0.0335	0.2314
500	0.0156	0.0068	0.0530
1,000	0.0083	0.0034	0.0264
$\rho = 0.75$			
100	0.0603	0.0321	0.2267
500	0.0146	0.0069	0.0482
1,000	0.0076	0.0035	0.0258

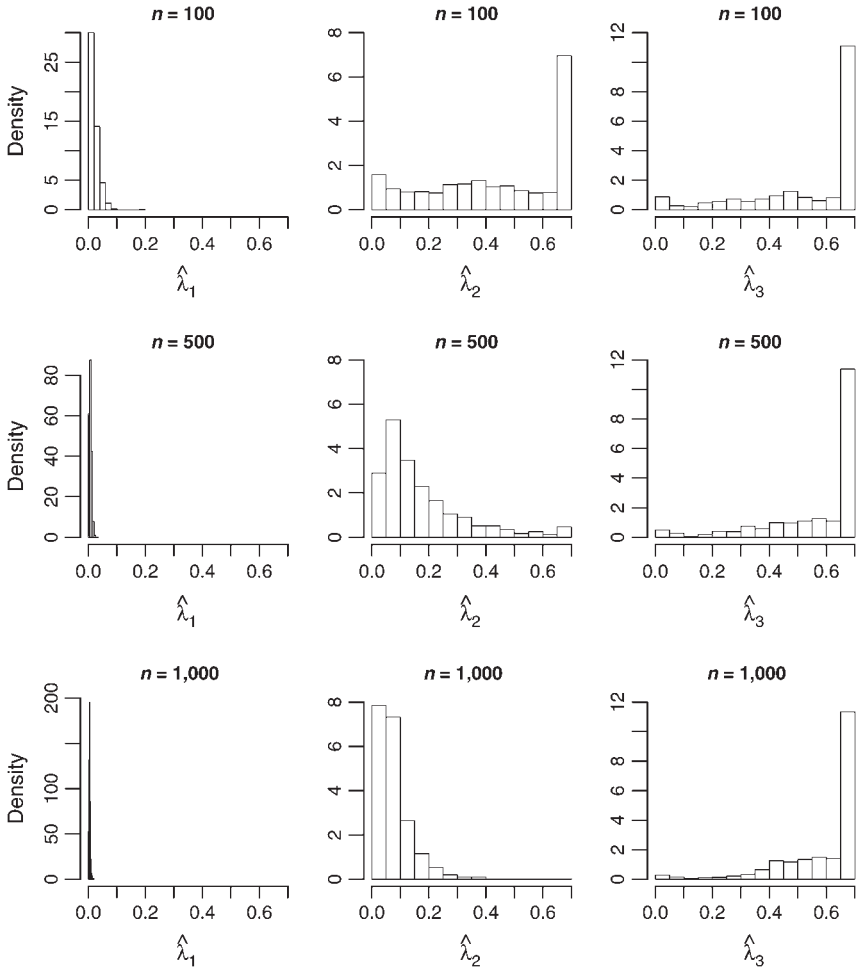

FIGURE 1. Histograms of the distribution of $\hat{\lambda}_1$, $\hat{\lambda}_2$, and $\hat{\lambda}_3$ for $n = 100, 500$, and 1,000, $\rho = 0$.

Turning to the behavior of the CV bandwidths, Figure 1 graphs the histograms of $\hat{\lambda}_j$ $(j = 1, 2, 3)$ for n from 100 to 1,000 when $\rho = 0$. A close examination of Figure 1 reveals that both $\hat{\lambda}_1$ and $\hat{\lambda}_2$ converge to zero as $n \to \infty$. In finite-sample settings $\hat{\lambda}_1$ tends to take smaller values than $\hat{\lambda}_2$ because X_{1i} is more important than X_{2i} in the sense that X_{i1} has a larger effect on Y_i than X_{2i}, whereas the distribution of $\hat{\lambda}_3$ is stable for large values of n and there is a positive probability that it will not assume its upper bound value, even as n goes to infinity, each behaving exactly as our theory predicts. We emphasize that the CV method can not only detect irrelevant regressors by oversmoothing, it can also distinguish

among more important and less important regressors by applying differential amounts of smoothing in finite-sample applications. Of course, asymptotically, $\hat{\lambda}_s \to 0$ for all relevant regressors.

4.2. Finite-Sample Performance Combining Relevant and Irrelevant Regressors

For this experiment we consider a setting with two discrete binary regressors, one of which is irrelevant. We compare the proposed approach with that obtained by creating a single binary regressor from the unique values of the two discrete binary regressors. We consider two approaches, namely, (i) conduct estimation on the two binary regressors, one of which is irrelevant, and (ii) conduct estimation on one regressor formed from the two binary regressors having 2^2 unique values. Note that the maximum value for (λ_1, λ_2) and for (λ) is one.

We consider a range of sample sizes, $n = 25, 50, 75, 100$. The regressors x_1 and x_2 are binomially distributed with $\Pr[x_j = 1] = 0.5$, $j = 1, 2$. The disturbance is $N(0, 1)$, and the DGP is given by $Y_i = X_{i1} + u_i$, $i = 1, \ldots, n$.

The median values of λ_1, λ_2, and λ are summarized in Table 2, and the median MSE values are summarized in Table 3.

It can be seen from examining Table 2 that the median value for the irrelevant bandwidth equals its maximum value and is thereby totally smoothed out, whereas that for the single regressor falls in the middle for all sample sizes considered. Table 3 reveals that the method that appropriately includes two regressors, one relevant and one irrelevant, dominates that which creates a single regressor from the relevant and irrelevant regressor on MSE grounds for all sample sizes considered.

4.3. Relative Performance of the Proposed Method versus Parametric Methods

For this experiment we consider two DGPs,

DGP 1: $Y_i = X_{i1} + X_{i2} + X_{i3} + X_{i1}X_{i2} + X_{i1}X_{i3} + X_{i2}X_{i3} + u_i$,

DGP 2: $Y_i = X_{i1} + X_{i2} + X_{i1}X_{i2} + u_i$,

TABLE 2. Median bandwidth summary for the method using two regressors (λ_1, λ_2) and that using one regressor (λ)

n	λ_1	λ_2	λ
25	0.076	1.000	0.117
50	0.039	1.000	0.058
75	0.026	1.000	0.040
100	0.020	1.000	0.030

TABLE 3. Median MSE summary for the method using two regressors (MSE(λ_1, λ_2)) and that using one regressor (MSE(λ))

n	MSE(λ_1, λ_2)	MSE(λ)
25	0.0916	0.1378
50	0.0404	0.0675
75	0.0250	0.0455
100	0.0195	0.0362

where $X_1, X_2, X_3 \in \{0, 1\}$, $\Pr[X_j = 1] = 0.5$, $j = 1, 2, 3$, and $u \sim N(0, 1)$. Note that, for both DGPs, we have only eight discrete cells, whereas for DGP 2 X_{i3} is irrelevant.

For each DGP we construct the proposed nonparametric estimator, the frequency estimator, and the following parametric models:

$$\text{Model 1: } Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1} X_{i2} + \beta_5 X_{i1} X_{i3} + \beta_6 X_{i2} X_{i3} + u_i,$$

$$\text{Model 2: } Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + u_i.$$

Note that parametric Model 1 is a correct specification for both DGPs,[6] whereas parametric Model 2 is incorrect for both DGPs in that the interaction terms are missing.

We draw 1,000 Monte Carlo replications, and for each replication we compute each model's MSE. We report the median MSEs in Table 4.

From Table 4 we observe that for DGP 1, the correctly specified parametric estimator (Model 1) performs best as expected, followed by the kernel smoothing and the frequency estimators. Not surprisingly, all three estimators dominate the misspecified parametric estimator (Model 2) because they are all consistent estimators, whereas the misspecified parametric model leads to inconsistent estimates.

Finally, the bottom block of Table 4 is based on DGP 2, where X_{i3} is an irrelevant regressor. However, no estimation method takes into account this known prior information, and all of them estimate a regression model using all three regressors, X_{i1}, X_{i2}, X_{i3}. First we observe that the nonparametric CV-based method has a smaller MSE than that obtained from the nonparametric frequency method (a 45% reduction in MSE for $n = 100$). This is because for DGP2, X_{i3} is an irrelevant regressor and our CV-based estimator has a high probability of smoothing out the irrelevant regressor; hence it leads to more efficient estimates (in finite samples) than the frequency estimator. It is interesting to observe that our nonparametric CV-based estimator also has a smaller MSE than the correctly specified parametric estimator (a 36% reduction in MSE for $n = 100$). By Theorem 3.2 we know that $\hat{g}(x)$ has the same $\sqrt{n}$ rate of convergence as the parametric estimator. Note that the parametric model estimates six parameters (the β's)

TABLE 4. Median MSE summary for the proposed nonparametric (Kernel), nonparametric frequency (Freq), and parametric models (Model 1, Model 2)

n	Kernel	Freq	Model 1	Model 2
DGP 1				
100	0.073	0.072	0.063	0.218
200	0.037	0.037	0.032	0.204
400	0.018	0.018	0.016	0.195
DGP 2				
100	0.041	0.074	0.064	0.093
200	0.020	0.036	0.031	0.078
400	0.010	0.019	0.016	0.070

and the frequency estimator estimates eight different cell means (recall that there exist eight cells). Our nonparametric estimator $\hat{g}(x)$ also estimates eight different cell means; however, if $\hat{\lambda}_3 = 1$, $\hat{g}(x)$ will smooth out the irrelevant regressor X_{i3} and only estimate four different cell means (corresponding to the discrete cells arising from the two relevant variables). Our simulations reveal that the probability that this happens is about 60%. This explains why $\hat{g}(x)$ can have a smaller MSE than the estimator based on a correctly (and over) specified linear model.

NOTES

1. Examples of unordered discrete regressors would include different regions, blood types, etc.

2. For related work that uses least squares CV for selecting smoothing parameters in a nonparametric regression model with continuous regressors, see Härdle and Marron (1985) and Härdle, Hall, and Marron (1988, 1992).

3. In a regression model with mixed continuous and discrete regressors, Hall et al. (2007) show that $\hat{\lambda} = O_p(n^{-2/(4+q)})$, where q is the dimension of the continuous regressors.

4. A more realistic example would include a group of other discrete variables such as gender, race, etc. Here for simplicity of exposition and to conserve space, we only consider a univariate discrete variable, namely, the health status of a person.

5. As results were qualitatively similar for $\rho \neq 0$ we do not report those results for space considerations.

6. Here we view an overspecified model as a correct model specification as it leads to consistent estimation of the conditional mean function. Also note that when the parametric model is overspecified, so is the nonparametric model.

REFERENCES

Aitchison, J. & Aitken, C.G.G. (1976) Multivariate binary discrimination by the kernel method. *Biometrika* 63, 413–420.

Bierens, H. (1983) Uniform consistency of kernel estimators of a regression function under generalized conditions. *Journal of the American Statistical Association* 78, 699–707.

Bowman, A.W., P. Hall & T.D.M. Titterington (1984) Cross-validation in nonparametric estimation of probabilities and probability densities. *Biometrika* 71, 341–351.

Fahrmeir, L. & G. Tutz (1994) *Multivariate Statistical Modeling Based on Generalized Linear Models.* Springer-Verlag.

Grund, B. & P. Hall (1993) On the performance of kernel estimators for high-dimensional sparse binary data. *Journal of Multivariate Analysis* 44, 321–344.

Guerre, E., I. Perrigne, & Q. Vuong (2000) Optimal nonparametric estimation of first price auction. *Econometrica* 68, 525–574.

Hahn, J. (1998) On the role of propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 66, 315–331.

Hall, P. (1981) On nonparametric multivariate binary discrimination. *Biometrika* 68, 287–294.

Hall, P., Q. Li, & J.S. Racine (2007) Nonparametric estimation of regression functions in the presence of irrelevant regressors. *Review of Economics and Statistics* 89, 784–789.

Hall, P., J. Racine, & Q. Li (2004) Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association* 99, 1015–1026.

Hall, P. & M. Wand (1988) On nonparametric discrimination using density differences. *Biometrika* 75, 541–547.

Härdle, W., P. Hall, & J.S. Marron (1988) How far are automatically chosen regression smoothing parameters from their optimum? *Journal of the American Statistical Association* 83, 86–99.

Härdle, W., P. Hall, & J.S. Marron (1992) Regression smoothing parameters that are not far from their optimum. *Journal of the American Statistical Association* 87, 227–233.

Härdle, W. & J.S. Marron (1985) Optimal bandwidth selection in nonparametric regression function estimation. *Annals of Statistics* 13, 1465–1481.

Hirano, K., G.W. Imbens, & G. Ridder (2003) Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71, 1161–1189.

Hong, Y. & T.H. Lee (2003) Inference on predictability of exchange rates via generalized spectrum and nonlinear time series models. *Review of Economics and Statistics* 85, 1048–1062.

Hong, Y. & T.H. Li (2005) Nonparametric specification testing for continuous-time models with applications to term structures of interest rates. *Review of Financial Studies* 18, 37–84.

Lee, J. (1990) U-*Statistics: Theory and Practice*. Marcel Dekker.

Li, Q. & J. Racine (2004) Cross-validated local linear nonparametric regression. *Statistica Sinica* 14, 485–512.

Li, Q. & J.S. Racine (2007) Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *Journal of Business & Economic Statistics*, *forthcoming*.

Li, T., I. Perrigne, & Q. Vuong (2002) Structural estimation of the affiliated private value auction model. *RAND Journal of Economics* 33, 171–193.

Li, T., I. Perrigne, & Q. Vuong (2003) Semiparametric estimation of the optimal reserve prices in first-price auctions. *Journal of Business & Economics Statistics* 21, 53–64.

Masry, E. (1996) Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *Journal of Time Series Analysis* 17, 571–599.

Racine, J.S., J. Hart, & Q. Li (2006) Testing the significance of categorical predictor variables in nonparametric regression models. *Econometric Reviews* 25, 523–544.

Racine, J.S. & Q. Li (2004) Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics* 119, 99–130.

Rosenthal, H.P. (1970) On the subspace of L^p $(p \geq 1)$ spanned by sequences of independent random variables. *Israel Journal of Mathematics* 8, 273–303.

Scott, D. (1992) *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley.

Simonoff, J.S. (1996) *Smoothing Methods in Statistics*. Springer-Verlag.

Titterington, D.M. (1980) A comparative study of kernel-based density estimates for categorical data. *Technometrics* 22, 259–268.

Wang, M.C. & J. van Ryzin (1981) A class of smooth estimators for discrete distributions. *Biometrika* 68, 301–309.

APPENDIX A: Proofs of Theorems 2.1 and 2.2—The Relevant Regressor Case

A.1. Preliminaries. The proof of Theorem 2.1 is quite tedious. Therefore, it is necessary to introduce some shorthand notation and preliminary manipulations to simplify the derivations that follow. For the reader's convenience we list most of the notation used in Appendix A here.

1. We will use g_i to denote $g(x_i)$ and $\hat{g}_i$ to denote $\hat{g}_{-i}(x_i)$ defined in (7). Similarly, we let $p_i = p(x_i)$ and $\hat{p}_i = \hat{p}_{-i}(x_i)$.
2. We define $\sum_i = \sum_{i=1}^n$, $\sum\sum_{j\neq i} = \sum_{i=1}^n \sum_{j=1, j\neq i}^n$, $\sum\sum\sum_{l\neq j\neq i} = \sum_{i=1}^n \sum_{j=1, j\neq i}^n \sum_{l=1, l\neq i, l\neq j}^n$.
3. We use $\mathbf{1}_s(x_i, x_j)$ to denote an indicator function that equals one if and only if x_i and x_j differs only in the sth component, and zero otherwise. That is, $\mathbf{1}_s(x_i, x_j) = \mathbf{1}(x_{is} \neq x_{js}) \prod_{t\neq s} \mathbf{1}(x_{it} = x_{jt})$.
4. Define $p_{i,0} = 1/(n-1)\sum_{j\neq i} \mathbf{1}(x_j = x_i)$ and $p_{i,1s} = 1/(n-1)\sum_{j\neq i} \mathbf{1}_s(x_i, x_j)$. Note that $p_{i,0}$ is the usual (leave-one-out) frequency estimator of p_i.
5. We write $A_n = B_n + (s.o.)$ to denote the fact that B_n is the leading term of A_n, where $(s.o.)$ denotes terms that have orders smaller than B_n. Here $A_i = B_i + (s.o.)$ always means that $n^{-1}\sum_i A_i = n^{-1}\sum_i B_i + (s.o.)$, and $A_{ij} = B_{ij} + (s.o.)$ means that $n^{-2}\sum_i \sum_j A_{ij} = n^{-2}\sum_i \sum_j B_{ij} + (s.o.)$. Also, we write $A_n \sim B_n$ to mean that A_n and B_n have the same order of magnitude in probability.
6. For notational simplicity we often ignore the difference between n^{-1} and $(n-1)^{-1}$ simply because this will have no effect on the asymptotic analysis.

We also give some results that will be used in our proofs.

Rosenthal's Inequality. Let $p \geq 2$ be a positive constant and let $X_1, \ldots, X_n$ denote i.i.d. random variables for which $\mathrm{E}(X_i) = 0$ and $\mathrm{E}(|X_i|^p) < \infty$. Then there exists a positive constant (which may depend on p) $C(p)$ such that

$$\mathrm{E}\left(\left|\sum_{i=1}^n X_i\right|^p\right) \leq C(p)\left\{\sum_{i=1}^n \mathrm{E}\left(|X_i|^p\right) + \left[\sum_{s=1}^n \mathrm{E}(X_i^2)\right]^{p/2}\right\}. \tag{A.1}$$

Equation (A.1) is widely known as Rosenthal's inequality (see Rosenthal, 1970).

The ***H***-Decomposition for ***U***-Statistics. Let $(n,k) = n!/[k!(n-k!)]$ denote the number of combinations obtained by choosing k items from n (distinct) items. Then a general kth-order U-statistic $U_{(k)}$ is defined by

$$U_{(k)} = \frac{1}{(n,k)} \sum_{1\leq i_1 < \cdots < i_k} H_n(X_{i_1}, \ldots, X_{i_k}),$$

where $H_n(X_{i_1}, \ldots, X_{i_k})$ is symmetric in its arguments and $\mathrm{E}[H_n^2(X_{i_1}, \ldots, X_{i_k})] < \infty$. In our proofs we will often use the following H-decomposition for a second-order U-statistic:

$$U_{(2)} = \theta + \frac{1}{n}\sum_i (H_{ni} - \theta) + \frac{2}{n(n-1)} \sum_i \sum_{j>i} [H_{n,ij} - H_{ni} - H_{nj} + \theta], \tag{A.2}$$

where $H_{n,ij} = H_n(X_i, x_j)$, $H_{ni} = \mathrm{E}[H_{n,ij}|X_i]$, and $\theta = \mathrm{E}[H_{nij}]$. We will also make use of the H-decomposition for a third-order U-statistic,

$$U_{(3)} = \theta + \frac{1}{n}\sum_i (H_{ni} - \theta) + \frac{2}{n(n-1)}\sum\sum_{j>i}(H_{n,ij} - H_{ni} - H_{nj} + \theta)$$

$$+ \frac{6}{n(n-1)(n-2)}\sum\sum\sum_{l>j>i}$$

$$\times (H_{n,ijl} - H_{n,ij} - H_{n,jl} - H_{n,li} + H_{ni} + H_{nj} + H_{nl} - \theta), \tag{A.3}$$

where $H_{n,ijl} = H_n(X_i, X_j, X_l)$, $H_{n,ij} = \mathrm{E}[H_{n,ijl}|X_i, X_j]$, $H_{ni} = \mathrm{E}[H_{n,ij}|X_i]$, and $\theta = \mathrm{E}[H_{n,ijl}]$. For a derivation of the preceding formulas and also an H-decomposition for a general kth-order U-statistic, see Lee (1990, p. 26).

Before we begin proving Theorem 2.1, we first provide some intermediate steps. Using (6) and $Y_i = g_i + u_i$, we have

$$CV(\lambda) = n^{-1}\sum_i (Y_i - \hat{g}_i)^2 = n^{-1}\sum_i (g_i - \hat{g}_i)^2 + 2n^{-1}\sum_i u_i(g_i - \hat{g}_i) + n^{-1}\sum_i u_i^2$$

$$= n^{-1}\sum_i (g_i - \hat{g}_i)^2 \hat{p}_i^2/\hat{p}_i^2 + 2n^{-1}\sum_i u_i(g_i - \hat{g}_i)\hat{p}_i/\hat{p}_i + n^{-1}\sum_i u_i^2. \tag{A.4}$$

In what follows, we obtain the leading terms of $CV(\lambda)$. We use $CV_0(\lambda)$ to denote the first two terms on the right-hand side of (A.4). Minimizing $CV(\lambda)$ over λ is equivalent to minimizing $CV_0(\lambda)$ as $n^{-1}\sum_i u_i^2$ does not depend on λ, where

$$CV_0(\lambda) = n^{-1}\sum_i (g_i - \hat{g}_i)^2 \hat{p}_i^2/\hat{p}_i^2 + 2n^{-1}\sum_i u_i(g_i - \hat{g}_i)\hat{p}_i/\hat{p}_i. \tag{A.5}$$

Using (7), $Y_j = g_j + u_j$, and $L_{ij} = L(x_i, x_j, \lambda)$, we have

$$CV_0(\lambda) = n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq i}(g_i - Y_j)(g_i - Y_l)L_{ij}L_{il}/\hat{p}_i^2$$

$$+ 2n^{-2}\sum_i\sum_{j\neq i} u_i(g_i - Y_j)L_{ij}/\hat{p}_i$$

$$= \left\{\frac{1}{n^3}\sum_i\sum_{j\neq i}\sum_{l\neq i}(g_i - g_j)(g_i - g_l)L_{ij}L_{il}/\hat{p}_i^2\right\}$$

$$+ \left\{\frac{1}{n^3}\sum_i\sum_{j\neq i}\sum_{l\neq i} u_j u_l L_{ij}L_{il}/\hat{p}_i^2 - \frac{2}{n^2}\sum_i\sum_{j\neq i} u_i u_j L_{ij}/\hat{p}_i\right\}$$

$$+ 2\left\{n^{-2}\sum_i\sum_{j\neq i} u_i(g_i - g_j)L_{ij}/\hat{p}_i - n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq i}(g_i - g_j)u_l L_{ij}L_{il}/\hat{p}_i^2\right\}$$

$$\equiv S_1 + S_2 + 2S_3, \tag{A.6}$$

where the definitions of S_j $(j = 1, 2, 3)$ should be apparent. Our proofs of Theorem 2.1 will be based on (A.6).

A.2. Proofs.

Proof of Theorem 2.1. In Lemma A.1 we show that $\hat{\lambda}_s = o_p(1)$ for all $s = 1, \ldots, r$. Given this result, in the proofs of Lemma A.2 and Lemma A.4 later in this section we will only consider the case in which $\lambda_s \in [0, \eta_n]$, where η_n is a positive sequence that converges to zero as $n \to \infty$.

Lemmas A.2–A.4 hold uniformly in $\lambda \in \Lambda_n = [0, \eta_n]^r$. The proof of the uniform rate of convergence is relatively lengthy. To conserve space, in this paper we only explicitly prove the uniform rate result for Lemma A.2, and for all the remaining lemmas we omit the uniform rate arguments because the detailed proofs follow along the same lines.

In Lemma A.2 we prove that

$$S_1 = \sum_{x \in \mathcal{D}} \left[\sum_{s=1}^{r} \lambda_s \left(\sum_{z \in \mathcal{D}} p(z) \mathbf{1}_s(x,z)(g(x) - g(z)) \right) \right]^2 p(x)^{-1} + o_p(|\lambda|^2), \tag{A.7}$$

where $\mathbf{1}_s(x,z)$ is an indicator function that equals one if x and z differ only in the sth component, and zero otherwise, whereas $|\lambda|^2 = \sum_{s=1}^{r} \lambda_s^2$. Furthermore, Lemma A.3 shows that

$$S_2 = -\frac{1}{n} \sum_{s=1}^{r} \lambda_s [A_s + Z_{1n,s}] + o_p(|\lambda|^2 + n^{-1}|\lambda|) + \quad \text{terms unrelated to } \lambda, \tag{A.8}$$

where A_s is a positive constant and $Z_{1n,s}$ is a zero mean $O_p(1)$ random variable. Finally, Lemma A.4 shows that

$$S_3 = \frac{1}{n} \sum_{s=1}^{r} \lambda_s Z_{2n,s} + o_p(n^{-1}|\lambda| + |\lambda|^2) + \quad \text{terms unrelated to } \lambda, \tag{A.9}$$

where $Z_{2n,s}$ is a zero mean $O_p(1)$ random variable.

Therefore, (A.7)–(A.9) lead to

$$\begin{aligned} CV_0(\lambda) = & \sum_{x \in \mathcal{D}} \left[\sum_{s=1}^{r} \lambda_s \sum_{z \in \mathcal{D}} \mathbf{1}_s(x,z) p(z) (g(x) - g(z)) \right]^2 p(x)^{-1} \\ & - \frac{1}{n} \sum_{s=1}^{r} \lambda_s \left(A_s + Z_{n,s} \right) + o_p(n^{-1}|\lambda| + |\lambda|^2) + \quad \text{terms unrelated to } \lambda, \end{aligned} \tag{A.10}$$

where $Z_{n,s} = Z_{1n,s} - 2Z_{2n,s}$ is a zero mean $O_p(1)$ random variable.

Note that (A.7) can be written as $S_1 = \lambda_{(r)}' \Omega \lambda_{(r)} + o_p(|\lambda|^2)$, where $\lambda_{(r)} = (\lambda_1, \ldots, \lambda_r)'$ and Ω is an $r \times r$ matrix with its (s,t)th element given by

$$\Omega_{st} = \sum_{x \in \mathcal{D}} \sum_{z' \in \mathcal{D}} \sum_{z \in \mathcal{D}} \mathbf{1}_s(x,z) \mathbf{1}_t(x,z') p(z) p(z') \left(g(x) - g(z) \right)(g(x) - g(z')) \, p(x)^{-1}.$$

Assumption 2 is equivalent to the assumption that Ω is a positive definite matrix.

Thus, from (A.10) we have

$$\frac{\partial CV_0(\lambda)}{\lambda_{(r)}} = 2\Omega\lambda_{(r)} - n^{-1}[A + Z_n] + (s.o.) \overset{set}{=} 0, \tag{A.11}$$

where $A = (A_1, \ldots, A_r)'$ and $Z_n = (Z_{n,1}, \ldots, Z_{n,r})'$. Equation (A.11) leads to $\hat{\lambda}_{(r)} = O_p(n^{-1})$. ■

Proof of Theorem 2.2. We write $\hat{g}(x) - g(x) = [\hat{g}(x) - g(x)]\hat{p}(x)/\hat{p}(x) = \hat{m}(x)/\hat{p}(x)$, where $\hat{m}(x) = [\hat{g}(x) - g(x)]\hat{p}(x)$. The expansion of $L(x_i, x, \hat{\lambda})$ that follows is also used frequently in the proof, namely,

$$L(x_i, x, \hat{\lambda}) = \mathbf{1}(x_i = x) + \sum_{s=1}^{r} \hat{\lambda}_s \mathbf{1}_s(x_i, x) + O_p\left(\sum_{s=1}^{r} \hat{\lambda}_s^2\right). \tag{A.12}$$

Using this expansion it is easy to see that

$$\hat{m}(x) = \tilde{m}(x) + \sum_{s=1}^{r} \hat{\lambda}_s \left[n^{-1} \sum_i Y_i \mathbf{1}_s(x_i, x)\right] + O_p(n^{-2}) = \tilde{m}(x) + O_p(n^{-1}),$$

because $\hat{\lambda}_s = O_p(n^{-1})$ by Theorem 2.1, where $\tilde{m}(x) = n^{-1} \sum_i (Y_i - g(x))\mathbf{1}(x_i = x)$.

Similarly, one can show that $\hat{p}(x) = n^{-1} \sum_i \mathbf{1}(x_i = x) + O_p(n^{-1})$. Hence we have

$$\hat{g}(x) = \tilde{g}(x) + O_p(n^{-1}), \tag{A.13}$$

where $\tilde{g}(x)$ is the frequency estimator of $g(x)$, i.e., $\tilde{g}(x)$ is obtained from $\hat{g}(x)$ by replacing λ_s by 0 for all $s = 1, \ldots, r$. It is established that $\sqrt{n}(\tilde{g}(x) - g(x)) \to N(0, \Omega(x))$ in distribution, where $\Omega(x) = \sigma^2(x)/p(x)$. Also, it is straightforward to show that $\hat{\sigma}^2(x) = \sigma^2(x) + o_p(1)$ and $\hat{p}(x) = p(x) + o_p(1)$ (the proofs for these are omitted here). Hence, $\hat{\Omega}(x) = \Omega(x) + o_p(1)$. These results, together with (A.13), prove Theorem 2.2. ■

Before we prove Lemmas A.2–A.4 that are used in the proof of Theorem 2.1, we need to resolve a technical difficulty in handling S_l $(l = 1, 2, 3)$ that arises from the presence of the random denominator $\hat{p}_i = \hat{p}_{-i}(x_i)$. We will use the following identity to handle the random denominator:

$$\frac{1}{\hat{p}_i} = \frac{1}{p_i} + \frac{(p_i - \hat{p}_i)}{p_i^2} + \frac{(p_i - \hat{p}_i)^2}{p_i^2 \hat{p}_i}. \tag{A.14}$$

Recalling that $p_{i,0} = (n-1)^{-1} \sum_{j \neq i} \mathbf{1}(x_j = x_i)$ and $p_{i,1s} = (n-1)^{-1} \sum_{j \neq i} \mathbf{1}_s(x_j, x_i)$, we have uniformly in $1 \leq i \leq n$

$$\begin{aligned} p_i - \hat{p}_i &= p_i - \frac{1}{n-1} \sum_{j \neq i} L_{ij} = p_i - \frac{1}{n-1} \\ &\quad \times \sum_{j \neq i} \left[\mathbf{1}(x_j = x_i) + \sum_{s=1}^{r} \lambda_s \mathbf{1}_s(x_j, x_i) + O(|\lambda|^2)\right] \\ &= (p_i - p_{i,0}) - \sum_{s=1}^{r} \lambda_s p_{i,1s} + O_p(|\lambda|^2) = O_p(n^{-1/2}) + O_p(|\lambda|), \end{aligned} \tag{A.15}$$

the last equality following because $\max_{1 \leq i \leq n} |p_i - p_{i,0}| \leq \sup_{x \in \mathcal{D}} |p(x) - n^{-1} \sum_i \mathbf{1}(x_i = x)| + O(n^{-1}) = O_p(n^{-1/2})$ (because $\mathcal{D}$ is a finite set) and $\max_{1 \leq i \leq n} |p_{i,1s}| = O_p(1)$.

Substituting (A.15) into (A.14), we get uniformly in $1 \le i \le n$

$$\frac{1}{\hat{p}_i} = \frac{1}{p_i} + \frac{(p_i - p_{i,0})}{p_i^2} - \sum_{s=1}^{r} \lambda_s \frac{p_{i,1s}}{p_i^2} + O_p(n^{-1}) + O_p(n^{-1/2}|\lambda| + |\lambda|^2). \tag{A.16}$$

Note that in (A.16), the $O_p(n^{-1})$ term comes from $(p_i - \tilde{p}_{i,0})^2/p_i^3$, which is unrelated to λ.

From (A.16), we also obtain uniformly in $1 \le i \le n$,

$$\frac{1}{\hat{p}_i^2} = \frac{1}{p_i^2} + 2\frac{(p_i - p_{i,0})}{p_i^3} - 2\sum_{s=1}^{r} \lambda_s \frac{p_{i,1s}}{p_i^3} + O_p(n^{-1}) + O_p(n^{-1/2}|\lambda| + |\lambda|^2), \tag{A.17}$$

where again the $O_p(n^{-1})$ is unrelated to λ.

Both (A.16) and (A.17) will be used to handle the random denominator in the proofs that follow.

The next lemma shows that the CV selected smoothing parameters all converge to zero in probability.

LEMMA A.1. *$\hat{\lambda}_s = o_p(1)$ for all $s = 1, \ldots, r$.*

Proof. When we choose $\lambda_s = 0$ for all $s = 1, \ldots, r$, it can be shown that $CV_0(0) = o_p(1)$. Because $\hat{\lambda}$ minimizes $CV_0(\lambda)$, it can be shown that $CV_0(\hat{\lambda}) \le CV_0(0)$. From the expression of $CV_0(\lambda)$ given in (A.6), we know that $CV_0(\lambda) = S_1(\lambda) + o_p(1)$ uniformly in $\lambda \in [0, 1]^r$ because both $S_2(\lambda)$ and $S_3(\lambda)$ contain u_i, which has zero mean, which necessarily makes $S_2(\lambda) = o_p(1)$ and $S_3(\lambda) = o_p(1)$, both uniformly in $\lambda \in [0, 1]^r$. Thus, we have $CV_0(\hat{\lambda}) = S_1(\hat{\lambda}) + o_p(1) \le CV_0(0) = o_p(1)$. Also, because $S_1(\hat{\lambda}) \ge 0$, we know that it must be true that

$$S_1(\hat{\lambda}) = o_p(1). \tag{A.18}$$

In what follows we consider a generic $\lambda \in [0, 1]^r$. We expand $S_1(\lambda)$ as

$$\begin{aligned} S_1(\lambda) &= n^{-3} \sum_i \sum_{j \ne i} \sum_{l \ne i} (g_i - g_j)(g_i - g_l) L_{ij} L_{il} / \hat{p}_i^2 \\ &= n^{-3} \sum \sum \sum_{l \ne j \ne i} (g_i - g_j)(g_i - g_l) L_{ij} L_{il} / \hat{p}_i^2 + n^{-3} \sum \sum_{j \ne i} (g_i - g_j)^2 L_{ij}^2 / \hat{p}_i^2 \\ &= n^{-3} \sum \sum \sum_{l \ne j \ne i} (g_i - g_j)(g_i - g_l) L_{ij} L_{il} / \hat{p}_i^2 + o_p(1) \\ &\equiv A_n(\lambda) + o_p(1) \quad \text{uniformly in } \lambda \in [0, 1]^r, \end{aligned} \tag{A.19}$$

where $A_n(\lambda) = n^{-3} \sum \sum \sum_{l \ne j \ne i} (g_i - g_j)(g_i - g_l) L_{ij} L_{il} / \hat{p}_i^2$. It can be shown that $\max_{1 \le i \le n} |\hat{p}_i - p_{i\lambda}| = o_p(1)$ uniformly in $\lambda \in [0, 1]^r$, where $p_{i\lambda} = \mathrm{E}[\hat{p}_i | x_i]$. Hence, the leading term of $A_n(\lambda)$ is $A_{1n}(\lambda)$, where $A_{1n}(\lambda)$ is obtained from $A_n(\lambda)$ by replacing $1/\hat{p}_i^2$ by $1/p_{i\lambda}^2$, i.e., $A_{1n}(\lambda) = n^{-3} \sum \sum \sum_{l \ne j \ne i} (g_i - g_j)(g_i - g_l) L_{ij} L_{il} / p_{i\lambda}^2$. Obviously, A_{1n}

can be written as a third-order U-statistic, and by the U-statistic H-decomposition (eqn. (A.3)) we know that

$$A_{1n}(\lambda) = \mathrm{E}(A_{1n}(\lambda)) + o_p(1) \qquad \text{uniformly in } \lambda \in [0,1]^r. \tag{A.20}$$

Now,

$$\begin{aligned}\mathrm{E}(A_{1n}(\lambda)) &= \mathrm{E}\{[\mathrm{E}[(g_i - g_j)L_{ij}/p_{i\lambda}|x_i]]^2\} + o(1) \\ &= \sum_{x\in\mathcal{D}} p(x)\left\{\sum_{z\in\mathcal{D}} p(z)[g(x)-g(z)]L(x,z,\lambda)/p(x,\lambda)\right\}^2 + o(1)\end{aligned} \tag{A.21}$$

uniformly in $\lambda \in [0,1]^r$, where $p(x,\lambda) = \mathrm{E}[\hat{p}_i|x_i = x]$.

Equations (A.19) and (A.20) and $A_n(\lambda) = A_{1n}(\lambda) + o_p(1)$ imply that

$$S_1(\lambda) = \mathrm{E}[A_{1n}(\lambda)] + o_p(1) \tag{A.22}$$

uniformly in $\lambda \in [0,1]^r$. Replacing λ by $\hat{\lambda}$ in (A.22) and also using (A.18) and (A.21) we obtain

$$S_1(\hat{\lambda}) = \sum_{x\in\mathcal{D}} p(x)\left\{\sum_{z\in\mathcal{D}} p(z)[g(x)-g(z)]L(x,z,\hat{\lambda})/p(x,\hat{\lambda})\right\}^2 + o_p(1) = o_p(1). \tag{A.23}$$

Equation (A.23) implies that

$$\sum_{x\in\mathcal{D}} p(x)\left\{\sum_{z\in\mathcal{D}} p(z)[g(x)-g(z)]L(x,z,\hat{\lambda})/p(x,\hat{\lambda})\right\}^2 = o_p(1). \tag{A.24}$$

Because $p(x)/p(x,\hat{\lambda})^2$ is bounded from both above and below by some positive constants, then (A.24) is equivalent to

$$\left\{\sum_{z\in\mathcal{D}} p(z)[g(x)-g(z)]L(x,z,\hat{\lambda})\right\}^2 = o_p(1) \qquad \text{for all } x \in \mathcal{D}. \tag{A.25}$$

Equation (A.25) and Assumption 2 imply that $\hat{\lambda}_s = o_p(1)$ for all $s = 1,\ldots,r$. Recall that Assumption 2 states that, for all $x \in \mathcal{D}$, $\left\{\sum_{z\in\mathcal{D}} p(z)[g(x)-g(z)]L(x,z,\lambda)\right\}^2 = 0$ if and only if $\lambda_s = 0$ for all $s = 1,\ldots,r$. It can be shown that if one (or some) of the $\hat{\lambda}_s$ does not converge to zero in probability, then the left-hand side of (A.25) is not $o_p(1)$, which contradicts (A.25). Hence, we must have $\hat{\lambda}_s = o_p(1)$ for all $s = 1,\ldots,r$ for (A.25) to hold. Thus, the CV selected smoothing parameters must all converge to zero in probability. ■

We are now ready to state and prove Lemmas A.2–A.4. Given the result of Lemma A.1, in the proofs of Lemmas A.2–A.4 we will give S_l $(l = 1,2,3)$ expansions in terms of powers of λ_s $(s = 1,\ldots,r)$.

LEMMA A.2. $S_1 = \sum_{x\in\mathcal{D}}[\sum_{s=1}^r \lambda_s(\sum_{z\in\mathcal{D}} \mathbf{1}_s(x,z)p(z)(g(x)-g(z)))]^2 p(x)^{-1} + o_p(|\lambda|^2)$.

Proof. Define S_1^0 the same way as S_1 except that $\hat{p}_i^{-2}$ is replaced by p_i^{-2}. That is,

$$S_1^0 \stackrel{def}{=} \frac{1}{n^3}\sum\sum_{i\neq j}(g_i-g_j)^2 L_{ij}^2/p_i^2 + \frac{1}{n^3}\sum\sum\sum_{i\neq j\neq l}(g_i-g_j)(g_i-g_l)L_{ij}L_{il}/p_i^2 = S_{1a}+S_{1b}.$$

Here S_{1b} can be written as a third-order U-statistic, and $S_{1b} = n^{-3}\sum\sum\sum_{i\neq j\neq l} Q_{ijl}$, where $Q(x_i,x_j,x_l)$ is a symmetrized version of $(g_i-g_j)(g_i-g_l)L_{ij}L_{il}/p_i^2$. Also, define $Q_{ij} \equiv \mathrm{E}\left(Q_{ijl}|x_i,x_j\right)$ and $Q_i \equiv \mathrm{E}\left(Q_{ijl}|x_i\right)$. Then by the U-statistic H-decomposition we have

$$\begin{aligned} S_{1b} &= \mathrm{E}Q_i + \frac{1}{n}\sum_i (Q_i - \mathrm{E}Q_i) + \frac{2}{n(n-1)}\sum\sum_{j>i}(Q_{ij}-Q_i-Q_j+\mathrm{E}Q_i) \\ &\quad + \frac{6}{n(n-1)(n-2)}\sum\sum\sum_{l>j>i} \\ &\quad \times (Q_{ijl}-Q_{ij}-Q_{jl}-Q_{li}+Q_i+Q_j+Q_l-\mathrm{E}Q_i) \\ &\equiv J_0+J_1+J_2+J_3, \end{aligned} \tag{A.26}$$

where $J_0 = \mathrm{E}Q_i$ and the definition of J_l $(l=1,2,3)$ should be apparent.

Using (A.12) and noting that $(g_i-g_j)\mathbf{1}(x_j=x_i)=0$, we obtain

$$\mathrm{E}[(g_i-g_j)L_{ji}|x_j=x] = \sum_{s=1}^r \lambda_s \sum_{z\in\mathcal{D}} \mathbf{1}_s(x,z)p(z)(g(x)-g(z)) + O(|\lambda|^2). \tag{A.27}$$

Hence, we have

$$\begin{aligned} J_0 = \mathrm{E}Q_i = \mathrm{E}(Q_{ijk}) &= \mathrm{E}[\{\mathrm{E}[(g_i-g_j)L_{ji}|x_i]\}^2/p_i^2] \\ &= \sum_{x\in\mathcal{D}} p(x)\left[\sum_{s=1}^r \lambda_s \sum_{z\in\mathcal{D}} \mathbf{1}_s(x,z)p(z)(g(x)-g(z))\right]^2 \Big/ p(x)^2 + o(|\lambda|^2) \\ &= \sum_{x\in\mathcal{D}}\left[\sum_{s=1}^r \lambda_s \sum_{z\in\mathcal{D}} \mathbf{1}_s(x,z)p(z)(g(x)-g(z))\right]^2 p(x)^{-1} + o(|\lambda|^2) \end{aligned} \tag{A.28}$$

uniformly in $x\in\mathcal{D}$ and λ.

Next, we consider J_1. The term $Q_i \equiv \mathrm{E}\left(Q_{ijl}|x_i\right)$ has the same order as $|\lambda|^2$ because

$$\mathrm{E}[(g_i-g_j)(g_i-g_l)L_{ji}L_{li}/p_i^2|x_i] = \{\mathrm{E}[(g_i-g_j)L_{ji}|x_i]\}^2/p_i^2 = O(|\lambda|^2).$$

Hence $\mathrm{E}(Q_i^k) = O(|\lambda|^{2k})$, and by Rosenthal's inequality of (A.1) we know that

$$\mathrm{E}|J_1|^{2k} \le n^{-2k}C_k(n^k|\lambda|^{4k}+n|\lambda|^{4k}) = O(n^{-k}|\lambda|^{4k}).$$

Therefore, by Markov's inequality, we have for $0<\delta<\frac{1}{2}$ and for all $k>0$,

$$\Pr(|J_1| > n^{-\delta}|\lambda|^2) \le \frac{\mathrm{E}[|J_1|^{2k}]}{n^{-2\delta k}|\lambda|^{4k}} = O(n^{-(1-2\delta)k}) = O(n^{-C}), \tag{A.29}$$

where $C \geq (1 - 2\delta k)$ and C can be arbitrarily large because k is allowed to be arbitrarily large. Then the same result holds uniformly in any set of λ with size no larger than a polynomial in n (i.e., size no larger than $O(n^a)$ for some $a > 0$). If $M_n \in \Lambda_n = [0, \eta)n]^r$ is any such set of values of λ (the size of M_n is no larger than $O(n^a)$), it follows that

$$\Pr\left(\max_{\lambda \in M_n} |J_1| > n^{-\delta}|\lambda|^2\right) \leq O(n^a) \max_{\lambda \in M_n} P(|J_1| > n^{-\delta}|\lambda|^2) = O(n^{-(C-a)}) = O(n^{-b}) \tag{A.30}$$

for some $b > 0$ ($b = C - a$, C can be arbitrarily large). Furthermore, the function $L(.,.,\lambda)$ is a polynomial function in λ and hence is Hölder continuous in λ. Therefore, taking a polynomially fine mesh of λ over $[0, \eta_n]^r$, following the standard arguments as used in Masry (1996) for establishing the uniform consistency of nonparametric estimators, we deduce that (A.30) continues to hold if M_n is replaced by $\Lambda_n = [0, \eta_n]^r$, i.e.,

$$\Pr\left(\sup_{\lambda \in \Lambda_n} |J_1| > n^{-\delta}|\lambda|^2\right) = O(n^{-b}) \tag{A.31}$$

for some $b > 0$, which implies that

$$J_1 = o_p(|\lambda|^2) \quad \text{uniformly in } \lambda \in [0, \eta_n]^r. \tag{A.32}$$

Next, we consider J_2. Note that Q_{ij} has the same order as

$$\mathrm{E}[(g_i - g_j)(g_i - g_l)L_{ji}L_{li}/p_i^2 | x_i, x_j] \sim O(|\lambda|)(g_i - g_j)L_{ji}.$$

Hence, $\mathrm{E}(Q_{ij}^k) = O(|\lambda|^{2k})$. We need to evaluate $\mathrm{E}(J_2^{2k})$. Note that $J_2 = 2n^{-2}\sum_i \sum_{j>i} \mathcal{H}_{ij}$ contains two summations, where $\mathcal{H}_{ij} = Q_{ij} - Q_i - Q_j + EQ_i$. Thus, $\mathrm{E}(J_2^{2k}) = 2^{2k}n^{-4k}\sum_{i_1}\sum_{j_1>i_1}\cdots\sum_{i_{2k}}\sum_{j_{2k}>i_{2k}} \mathrm{E}[\mathcal{H}_{i_1 j_1}\cdots\mathcal{H}_{i_{2k} j_{2k}}]$, which contains $4k$ summations. Also, because $\mathrm{E}[\mathcal{H}_{ij}|x_i] = 0$, the nonzero terms in $\mathrm{E}(J_2^{2k})$ must have the property that each summation index is equal to at least another summation index. Thus, the nonzero terms can at most contain $2k$ summations (each summation index is paired with and only with another one), the next nonzero term contains $2k - 1$ summations, etc., whereas the last nonzero term (having the smallest number of summations) has two summations.

More specifically, let us consider the case of $k = 2$. In this case $\mathrm{E}(J_2^{2k}) = \mathrm{E}(J_2^4) = 16n^{-8}\sum_{i_1}\sum_{j_1>i_1}\cdots\sum_{i_4}\sum_{j_4>i_4}\mathrm{E}[\mathcal{H}_{i_1 j_1}\ldots\mathcal{H}_{i_4 j_4}]$ contains 8 summations. However, if one of the indexes, say, i_1, differs from all other indexes, i.e., $i_1 \neq i_l$ for all $l \neq 1$, and $i_1 \neq j_l$ for $l = 1, \ldots, 4$, then $\mathrm{E}[H_{i_1 j_1}|x_{j_1}, x_{i_2}, x_{j_2}, \ldots, x_{j_4}] = 0$, which leads to $\mathrm{E}(J_2^4) = 0$ (for this case) by the law of iterated expectations. Therefore, for $\mathrm{E}(J_2^4)$ to be nonzero, all subscript indexes must pair with at least another index. Then we have the following case, situation (a), where the eight indexes take four different values, i.e., each index pairs with one and only one other index, one such case being $i_1 = i_3$, $i_2 = i_4$, $j_1 = j_3$, and $j_2 = j_4$, which gives $\mathrm{E}[\mathcal{H}_{i_1 j_1}^2 \mathcal{H}_{i_2 j_2}^2]$. Obviously, all cases in situation (a) have the same order. If we let $C_{2,1}$ (which also includes the factor 16) denote the number of cases in situation (a), then we have $\mathrm{E}(J_2^4)_{(a)} \sim C_{2,1}n^{-8}\sum_{i_1}\sum_{j_1>i_1}\sum_{i_2}\sum_{j_2>i_2}\mathrm{E}[\mathcal{H}_{i_1 j_1}^2 \mathcal{H}_{i_2 j_2}^2]$. On the other hand we could encounter situation (b), where the eight indexes take three different values, one such case being $i_1 = i_2 = i_3 = i_4$, $j_3 = j_1$, $j_4 = j_2$, which corresponds to $\mathrm{E}[\mathcal{H}_{i_1 j_1}^2 \mathcal{H}_{i_1 j_2}^2]$.

We use $C_{2,2}$ to denote other cases for situation (b), all such cases having the same order so that we have $\mathrm{E}(J_2^4)_{(b)} \sim C_{2,2} n^{-8} \sum_{i_1} \sum_{j_1 > i_1} \sum_{j_2 > i_1} \mathrm{E}[\mathcal{H}_{i_1 j_1}^2 \mathcal{H}_{i_1 j_2}^2]$. Or, we could encounter situation (c), where the eight indexes take two different values. There is only one such case, i.e., $i_1 = i_2 = i_3 = i_4$ and $j_1 = j_2 = j_3 = j_4$, which leads to $\mathrm{E}[\mathcal{H}_{i_1 j_1}^4]$. Finally, by noting that $\mathrm{E}(\mathcal{H}_{ij}^k) \sim \mathrm{E}(Q_{ij}^k) = O(|\lambda|^{2k})$, we have

$$\begin{aligned}\mathrm{E}(J_2^4) &= \mathrm{E}(J_2^4)_{(a)} + \mathrm{E}(J_2^4)_{(a)} + \mathrm{E}(J_2^4)_{(a)} \sim n^{-8}\{C_{2,1} n^4 |\lambda|^8 + C_{2,2} n^3 |\lambda|^8 + n^2 |\lambda|^8\} \\ &= O(n^{-4}|\lambda|^8).\end{aligned}$$

Therefore, for a general positive integer $k \geq 2$, using $\mathrm{E}(Q_{ij}^k) = O(|\lambda|^{2k})$, similar to the case for which $k = 2$, one can show that

$$\mathrm{E}\,|J_2|^{2k} \leq n^{-4k}\{C_{k,1} n^{2k} |\lambda|^{4k} + C_{k,2} n^{(2k-1)} |\lambda|^{4k} + \cdots + n^2 |\lambda|^{4k}\} = O(n^{-2k}|\lambda|^{4k}). \tag{A.33}$$

Hence, by Markov's inequality, we know that this leads to $\Pr(J_2 > n^{-\delta}|\lambda|^2) = O(n^{-(2-2\delta)k}) = O(n^{-C})$ (for some $\delta \in (0,1)$ and for all $C > 0$). Then by using the same arguments as those leading to (A.31), we get

$$\Pr\left(\sup_{\lambda \in \Lambda_n} |J_2| > n^{-\delta}|\lambda|^2\right) = O(n^{-b}), \tag{A.34}$$

for some $b > 0$, which implies that

$$J_2 = o_p(|\lambda|^2) \quad \text{uniformly in } \lambda. \tag{A.35}$$

The term J_3 is a third-order U-statistic. The kth-order moment of Q_{ijl} has the same order as the kth moment of the unsymmetrized quantity $(g_i - g_j)L_{ij}(g_i - g_l)L_{il}/p_i^2$. Thus,

$$\mathrm{E}(Q_{ijl}^k) \sim \mathrm{E}[(g_i - g_j)(g_i - g_l)L_{ji}L_{li}/p_i^2]^k = O(|\lambda|^{2k}).$$

The term J_3^{2k} contains $6k$ summations. However, for $\mathrm{E}(J_3^{2k})$ to be nonzero, each summation index must be equal to at least another summation index. Hence, we have

$$\mathrm{E}\left|J_3^{2k}\right| \leq n^{-6k} C_k (n^{3k}|\lambda|^{4k} + \cdots + n^3|\lambda|^{4k}) = O(n^{-3k}|\lambda|^{4k}). \tag{A.36}$$

By comparing (A.36) with (A.33), we know that J_3 has an order smaller than that of J_2. Hence,

$$J_3 = o_p(|\lambda|^2) \quad \text{uniformly in } \lambda \in \Lambda_n. \tag{A.37}$$

Summarizing (A.28) to (A.37), we have shown that

$$S_{1b} = \sum_{x \in \mathcal{D}} \left[\sum_{s=1}^{r} \lambda_s \sum_{z \in \mathcal{D}} \mathbf{1}_s(x,z) p(z)(g(x) - g(z))\right]^2 p(x)^{-1} + o(|\lambda|^2) \tag{A.38}$$

uniformly in $\lambda \in \Lambda_n$.

Now we consider S_{1a}, where

$$S_{1a} = \frac{2}{n(n-1)^2} \sum_{i}^{n} \sum_{j>i}^{n} (g_i - g_j)^2 L_{ji}^2 \left[\frac{1}{p_i^2} + \frac{1}{p_j^2} \right]. \tag{A.39}$$

By using the U-statistic H-decomposition, it is easy to see that the leading term of S_{1a} is $\mathrm{E}[S_{1a}] = n^{-1}\mathrm{E}[(g_i - g_j)^2 L_{ji}^2 / p_i^2] = O(n^{-1}|\lambda|^2) = o(|\lambda|^2)$. Hence, $S_{1a} = o(|\lambda|^2)$ uniformly in λ.

By (A.38) and (A.39) we have

$$S_1^0 = \sum_{x \in \mathcal{D}} \left[\sum_{s=1}^{r} \lambda_s \sum_{z} \mathbf{1}_s(x,z) p(x) (g(x) - g(z)) \right]^2 p(x)^{-1} + o(|\lambda|^2) \tag{A.40}$$

uniformly in λ.

Define $\Delta_{p,i}(x_i) \equiv \hat{p}_i - p_i$ and $\hat{m}_{1,i}(x) = 1/(n-1) \sum_{j \neq i} (g(x) - g(x_j)) L(x_j, x, \lambda)$. Then

$$\left| S_1 - S_1^0 \right| = \left| \frac{1}{n} \sum_{i=1}^{n} \hat{m}_{1,i}^2(x_i) \left(\frac{1}{\hat{p}_i^2} - \frac{1}{p_i^2} \right) \right| \leq C \sup_{1 \leq i \leq n} \hat{m}_{1,i}^2(x_i) \sup_{1 \leq i \leq n} \left| \Delta_{p,i}(x_i) \right|. \tag{A.41}$$

Now,

$$\mathrm{E}(\hat{m}_{1,i}(x)) = \mathrm{E}((g(x) - g(x_j)) L(x_j, x, \lambda)) \sim O(|\lambda|), \quad \text{uniformly in } x \in \mathcal{D} \text{ and } \lambda. \tag{A.42}$$

We write

$$\begin{aligned} \hat{m}_{1,i}(x) - \mathrm{E}(\hat{m}_{1,i}(x)) &= \frac{1}{n-1} \sum_{j \neq i} \big[(g(x) - g(x_j)) L(x_j, x) \\ &\qquad - \mathrm{E}((g(x) - g(x_j)) L(x_j, x)) \big] \\ &\equiv \frac{1}{n-1} \sum_{j \neq i} \big[V_j - \mathrm{E}(V_j) \big], \end{aligned} \tag{A.43}$$

where $V_j \equiv (g(x) - g(x_j)) L(x_j, x)$. It is easy to show that $\mathrm{E}(V_j^k) \sim O(|\lambda|^k)$. By Rosenthal's inequality, we have

$$\mathrm{E} \left| \hat{m}_{1,i}(x) - \mathrm{E}(\hat{m}_{1,i}(x)) \right|^{2k} \leq C_k n^{-2k} (n^k |\lambda|^{2k} + n|\lambda|^{2k}) \sim O(n^{-k}|\lambda|^{2k}).$$

By Markov's inequality, we have

$$\Pr \left(\left| \hat{m}_{1,i}(x) - \mathrm{E}(\hat{m}_{1,i}(x)) \right| > n^{-\delta} |\lambda| \right) = O(n^{-(1-2\delta)k}), \quad \text{for some } 0 < \delta < 1/2.$$

Using the same arguments as those leading to (A.34), we have for some $\delta \in (0, 1/2)$ and any $C > 0$,

$$\Pr \left(\sup_{\lambda} \left| \hat{m}_{1,i}(x) - \mathrm{E}(\hat{m}_{1,i}(x)) \right| > n^{-\delta} |\lambda| \right) \leq const. n^{-(1-2\delta)k} = O(n^{-C}),$$

which implies that

$$\sup_{\lambda} \left| \hat{m}_{1,i}(x) - \mathrm{E}(\hat{m}_{1,i}(x)) \right| = o_p(|\lambda|) \quad \text{uniformly in } x \in \mathcal{D}. \tag{A.44}$$

By (A.42) and (A.44), we have

$$\hat{m}_{1,i}(x) = O_p(|\lambda|) \quad \text{uniformly in } x \in \mathrm{D} \quad \text{and} \quad \lambda \in \Lambda_n. \tag{A.45}$$

Recalling that $\Delta_{p,i}(x_i) = \hat{p}_i - p_i$, then by the same arguments that lead to (A.45), we have that

$$\begin{aligned} \sup_{i,\lambda,x_i} \left| \Delta_{p,i}(x_i) \right| &\le \max_{\lambda, x \in \mathcal{D}} |\hat{p}(x) - \mathrm{E}\hat{p}(x)| + \max_{\lambda, x \in \mathcal{D}} |\mathrm{E}[\hat{p}(x)] - p(x)| \\ &= O_p(n^{-1/2}) + O_p(|\lambda|) = o_p(1). \end{aligned} \tag{A.46}$$

By (A.45) and (A.46) we have

$$\left| S_1 - S_1^0 \right| \le C \sup_{1 \le i \le n} \hat{m}_{1,i}^2(x_i) \sup_{1 \le i \le n} \left| \Delta_{p,i}(x_i) \right| = o_p(|\lambda|^2). \tag{A.47}$$

By (A.40) and (A.47) we have

$$S_1 = \sum_{x \in \mathcal{D}} \left[\sum_{s=1}^{r} \lambda_s \sum_{z \in \mathcal{D}} \mathbf{1}_s(x,z) p(z)(g(x) - g(z)) \right]^2 p(x)^{-1} + o_p(|\lambda|^2) \tag{A.48}$$

uniformly in λ. ■

LEMMA A.3. *$S_2 = -n^{-1} \sum_{s=1}^{r} \lambda_s A_s + n^{-1} \sum_{s=1}^{r} \lambda_s Z_{1n,s} + o_p(|\lambda|^2 + |\lambda| n^{-1}) +$ terms unrelated to λ, where $A_s > 0$ is a positive constant and $Z_{1n,s}$ is a zero mean $O_p(1)$ random variable defined in the proof that follows.*

Proof.

$$\begin{aligned} S_2 &= n^{-3} \sum \sum_{j \ne i} u_j^2 L_{ij}^2 / \hat{p}_i^2 + n^{-3} \sum \sum \sum_{i \ne j \ne l} u_j u_l L_{ij} L_{il} / \hat{p}_i^2 - 2n^{-2} \sum \sum_{j \ne i} u_i u_j L_{ij} / \hat{p}_i \\ &\equiv S_{2a} + S_{2b} - 2S_{2c}. \end{aligned}$$

Using (A.17) and noting that $L_{ij,\lambda}^2 = O(|\lambda|^2)$ if $x_j \ne x_i$, we have

$$\begin{aligned} S_{2a} &= n^{-3} \sum \sum_{j \ne i} \mathbf{1}(x_j = x_i) u_j^2 / \hat{p}_i^2 + O_p(n^{-1}|\lambda|^2) \\ &= n^{-3} \sum \sum_{j \ne i} \mathbf{1}(x_j = x_i) u_j^2 \left[1/p_i^2 + 2(p_i - p_{i,0})/p_i^3 - 2 \sum_{s=1}^{r} \lambda_s p_{i,1s}/p_i^3 \right] \\ &\quad + O(n^{-3/2}|\lambda| + n^{-1/2}|\lambda|^2) \\ &= -\frac{1}{n} \sum_{s=1}^{r} \lambda_s 2n^{-2} \sum \sum_{j \ne i} \mathbf{1}(x_j = x_i) u_j^2 p_{i,1s}/p_i^3 \\ &\quad + O_p(n^{-3/2}|\lambda| + n^{-1/2}|\lambda|^2) + \quad \text{terms unrelated to } \lambda, \end{aligned}$$

$$\equiv -n^{-1}\sum_{s=1}^{r}\lambda_s A_s + O(n^{-3/2}|\lambda| + n^{-1/2}|\lambda|^2) + \quad \text{terms unrelated to } \lambda,$$

where $A_s = 2\mathrm{E}[\mathbf{1}(x_j = x_i)u_j^2 p_{i,1s}/p_i^2]$ is a positive constant and we have used the fact that

$$2n^{-2}\sum\sum_{j\neq i}\mathbf{1}(x_j = x_i)u_j^2 p_{i,1s}/p_i^3 = A_s + O_p(n^{-1/2}). \tag{A.49}$$

Equation (A.49) follows from the U-statistic H-decomposition because $2n^{-2}\sum\sum_{j\neq i}\mathbf{1}(x_j = x_i)u_j^2 p_{i,1s}/p_i^3$ can be written as a second-order U-statistic.

Using (A.17) and (A.12) we have

$$\begin{aligned}
S_{2b} &= n^{-3}\sum\sum\sum_{i\neq j\neq l} u_j u_l \left[\mathbf{1}(x_j = x_i) + \sum_{s=1}^{r}\lambda_s \mathbf{1}_s(x_j, x_i)\right] \\
&\quad \times \left[\mathbf{1}(x_l = x_i) + \sum_{t=1}^{r}\lambda_t \mathbf{1}_t(x_l, x_i)\right]\left[1/p_i^2 + 2(p_i - p_{i,0})/p_i^3 - 2\sum_{s=1}^{r}\lambda_s p_{i,1s}/p_i^3\right] \\
&\quad + O_p(n^{-1}|\lambda|^2) \\
&= n^{-1}\sum_{s=1}^{r}\lambda_s 2n^{-2}\sum\sum\sum_{i\neq j\neq l} u_j u_l \Big[\mathbf{1}(x_j = x_i)\mathbf{1}_s(x_l, x_i) + \mathbf{1}(x_l = x_i)\mathbf{1}_s(x_j, x_i) \\
&\qquad\qquad - 2\times\mathbf{1}(x_j = x_i)\mathbf{1}(x_l = x_i)p_{i,1s}p_i^{-1}\Big]\Big/ p_i^2 \\
&\quad + O_p(n^{-3/2}|\lambda| + n^{-1/2}|\lambda|^2) + \quad \text{terms unrelated to } \lambda \\
&= n^{-1}\sum_{s=1}^{r}\lambda_s Z_{3n,s} + O_p(n^{-3/2}|\lambda| + n^{-1/2}|\lambda|^2) + \quad \text{terms unrelated to } \lambda,
\end{aligned}$$

where $Z_{3n,s}$ equals

$$\frac{2}{n^2}\sum\sum\sum_{i\neq j\neq l}\frac{u_j u_l}{p_i^2}\left[\mathbf{1}(x_j = x_i)\mathbf{1}_s(x_l, x_i) + \mathbf{1}(x_l = x_i)\mathbf{1}_s(x_j, x_i) - 2\times\mathbf{1}(x_j = x_i)\mathbf{1}(x_l = x_i)\times\frac{p_{i,1s}}{p_i}\right].$$

It is easy to see that $Z_{3n,s}$ is a zero mean $O_p(1)$ random variable by showing that $\mathrm{E}(Z_{3n,s}^2) = O(1)$ (this follows from the fact that $\mathrm{E}(u_{i_1}u_{i_2}u_{i_3}u_{i_4}|x_{i_1}, x_{i_2}, x_{i_3}, x_{i_4}) = 0$ unless i_1, i_2, i_3, i_4 take on no more than two different values). In the preceding derivation we have also used the fact that $p_i - p_{i,0} = O_p(n^{-1/2})$.

Again, using (A.16) and (A.12) and letting $\zeta_n = n^{-1}|\lambda|^2 + n^{-3/2}|\lambda| +$ terms unrelated to λ, we have

$$S_{2c} = n^{-2}\sum\sum_{j\neq i} u_i u_j \left[\mathbf{1}(x_j = x_i) + \sum_{s=1}^{r}\lambda_s \mathbf{1}_s(x_j, x_i)\right]$$

$$\times \left[1/p_i + (p_i - p_{i,0})/p_i^2 - \sum_{s=1}^{r} \lambda_s p_{i,1s}/p_i^2\right] + O_p(\zeta_n)$$

$$= n^{-1} \sum_{s=1}^{r} \lambda_s n^{-1} \sum\sum_{j\neq i} u_i u_j [\mathbf{1}_s(x_j, x_i)/p_i - \mathbf{1}(x_j = x_i) p_{i,1s}/p_i^2] + O_p(\zeta_n)$$

$$= n^{-1} \sum_{s=1}^{r} \lambda_s Z_{4n,s} + O_p(n^{-1}|\lambda|^2 + n^{-3/2}|\lambda|) + \quad \text{terms unrelated to } \lambda,$$

where $Z_{4n,s} = n^{-1} \sum\sum_{j\neq i} u_i u_j [\mathbf{1}_s(x_j, x_i)/p_i - \mathbf{1}(x_j = x_i) p_{i,1s}/p_i^2]$. It is easy to see that $Z_{4n,s}$ is a zero mean $O_p(1)$ random variable. Note that the term associated with $p_i - p_{i,0}$ is of order $O_p(n^{-3/2}|\lambda|)$ because $\max_{1\leq i\leq n} |p_i - p_{i,0}| = O_p(n^{-1/2})$.

Summarizing the preceding discussion we have shown that

$$S_2 = S_{2a} + S_{2b} - 2S_{2c}$$

$$= -n^{-1} \sum_{s=1}^{r} \lambda_s A_s + n^{-1} \sum_{s=1}^{r} \lambda_s Z_{1n,s}$$

$$+ O_p(n^{-1}|\lambda|^2 + n^{-3/2}|\lambda|) + \quad \text{terms unrelated to } \lambda, \tag{A.50}$$

where $Z_{1n,s} = Z_{3n,s} - 2Z_{4n,s}$ is a zero mean $O_p(1)$ random variable. ■

LEMMA A.4.

$$S_3 = n^{-1} \sum_{s=1}^{r} \lambda_s Z_{2n,s} + o_p(|\lambda|^2 + n^{-1}|\lambda|),$$

where $Z_{2n,s}$ is a zero mean $O_p(1)$ random variable defined in the proof that follows.

Proof.

$$S_3 = n^{-2} \sum_i \sum_{j\neq i} u_i (g_i - g_j) L_{ij}/\hat{p}_i - n^{-3} \sum\sum\sum_{i\neq j\neq l} (g_i - g_j) u_l L_{ij} L_{il}/\hat{p}_i^2$$

$$- n^{-3} \sum_i \sum_{j\neq i} (g_i - g_j) u_j L_{ij}^2/\hat{p}_i^2 \equiv S_{3a} - S_{3b} - S_{3c}.$$

As we will see subsequently, there are some subtle cancellations between S_{3a} and S_{3b} that are the key to establishing Lemma A.4.

Letting $\zeta_n = |\lambda|^2 n^{-1/2} + |\lambda| n^{-3/2} +$ terms unrelated to λ, then using (A.16) and noting that $(g_i - g_j)\mathbf{1}(x_j = x_i) = 0$, we have

$$S_{3a} = n^{-2} \sum\sum_{j\neq i} u_i (g_i - g_j) \left[0 + \sum_{s=1}^{r} \lambda_s \mathbf{1}_s(x_j, x_i)\right] [1/p_i + (p_i - p_{i,0})/p_i^2] + O_p(\zeta_n)$$

$$= \sum_{s=1}^{r} \lambda_s n^{-2} \sum\sum_{j\neq i} \mathbf{1}_s(x_j, x_i) u_i (g_i - g_j)/p_i$$

$$+\sum_{s=1}^{r}\lambda_s n^{-2}\sum\sum_{j\neq i}\mathbf{1}_s(x_j,x_i)u_i(g_i-g_j)(p_i-p_{i,0})/p_i^2+O_p(\zeta_n)$$

$$=S_{3a,1}+S_{3a,2}+O_p(\zeta_n), \tag{A.51}$$

where the definitions of $S_{3a,1}$ and $S_{3a,2}$ should be apparent. The zero term appearing in the first equality comes from the fact that $(g_i-g_j)\mathbf{1}(x_j=x_i)=0$.

Next, we consider S_{3b}. Again noting that $(g_i-g_j)\mathbf{1}(x_j=x_i)=0$ and using (A.17), we have

$$S_{3b}=n^{-3}\sum\sum\sum_{i\neq j\neq l}u_l(g_i-g_j)\left[0+\sum_{s=1}^{r}\lambda_s\mathbf{1}_s(x_j,x_i)\right]$$

$$\times[\mathbf{1}(x_l=x_i)][1/p_i^2+2(p_i-p_{i,0})/p_i^3]+O_p(n^{-1/2}|\lambda|^2+n^{-3/2}|\lambda|)$$

$$=\sum_{s=1}^{r}\lambda_s n^{-3}\sum\sum\sum_{l\neq j\neq i}\mathbf{1}_s(x_j,x_i)\mathbf{1}(x_l=x_i)u_l(g_i-g_j)/p_i^2$$

$$+2\sum_{s=1}^{r}\lambda_s n^{-3}\sum\sum\sum_{l\neq j\neq i}\mathbf{1}_s(x_j,x_i)\mathbf{1}(x_l=x_i)u_l(g_i-g_j)(p_i-p_{i,0})/p_i^3$$

$$+O_p(n^{-1/2}|\lambda|^2+n^{-3/2}|\lambda|)$$

$$\equiv S_{3b,1}+2S_{3b,2}+O_p(n^{-1/2}|\lambda|^2+n^{-3/2}|\lambda|), \tag{A.52}$$

where the definitions of $S_{3b,1}$ and $S_{3b,2}$ should be apparent.

Note that $\max_{1\le i\le n}|p_i-p_{i,0}|\le\max_{x\in\mathcal{D}}|p(x)-n^{-1}\sum_{j=1}^{n}\mathbf{1}(x_j=1)|+O(n^{-1})=O_p(n^{-1/2})$ because the support $\mathcal{D}$ only contains finitely many x. Using this result it is easy to see that both $S_{3a,2}$ and $S_{3b,2}$ are of order $O_p(|\lambda|n^{-1})$. Although $S_{3a,1}$ and $S_{3b,1}$ are both of order $O_p(|\lambda|n^{-1/2})$, we will show subsequently that $S_{3a,1}-S_{3b,1}=O_p(|\lambda|n^{-1})$. To show this, we need to rewrite $S_{3b,1}$ in a form similar to $S_{3a,1}$,

$$S_{3b,1}=\sum_{s=1}^{r}\lambda_s n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq i,l\neq j}\mathbf{1}_s(x_j,x_i)\mathbf{1}(x_l=x_i)u_l(g_i-g_j)/p_i^2$$

$$=\sum_{s=1}^{r}\lambda_s n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq i,l\neq j}\mathbf{1}_s(x_j,x_l)\mathbf{1}(x_l=x_i)u_l(g_l-g_j)/p_l^2\qquad\text{(because } x_l=x_i)$$

$$=\sum_{s=1}^{r}\lambda_r n^{-2}\sum_j\sum_{l\neq j}\left[\mathbf{1}_s(x_j,x_l)u_l(g_l-g_j)/p_l^2\right]\left[n^{-1}\sum_{i\neq j,i\neq l}\mathbf{1}(x_l=x_i)\right]$$

$$=\sum_{s=1}^{r}\lambda_s n^{-2}\sum_j\sum_{l\neq j}\mathbf{1}_s(x_j,x_l)u_l(g_l-g_j)p_{l,0}^*/p_l^2$$

$$(p^*_{l,0} \overset{def}{=} n^{-1}\textstyle\sum_{i=1,i\neq j,i\neq l}^{n} \mathbf{1}(x_l = x_i))$$

$$= \sum_{s=1}^{r} \lambda_s n^{-2} \sum_j \sum_{i\neq j} \mathbf{1}_s(x_j, x_i) u_i (g_i - g_j) p_{i,0}/p_i^2 + O_p(|\lambda| n^{-3/2}), \tag{A.53}$$

where the second equality is the key step. There we used $g_l = g_i$, $p_l = p_i$ because $x_l = x_i$ as a result of the restriction $\mathbf{1}(x_l = x_i)$. The third equality simply reorders the summations. The fourth equality follows from the definition of $p^*_{l,0}$, and in the last equality we used $\max_{1\le l\le n} |p_{l,0} - p^*_{l,0}| = O_p(n^{-1})$ ($p_{l,0} = n^{-1}\sum_{i=1,i\neq l} \mathbf{1}(x_l = x_i)$) and we changed summation indexes from (j,l) to (j,i).

Note that both $S_{3a,1}$ and $S_{3b,1}$ are of order $O_p(|\lambda| n^{-1/2})$, but we have (using (A.53))

$$S_{3a,1} - S_{3b,1} = n^{-1} \sum_{s=1}^{r} \lambda_s \left\{ n^{-1} \sum \sum_{j\neq i} \mathbf{1}_s(x_j, x_i) u_i (g_i - g_j)(p_i - p_{i,0})/p_i^2 \right\}$$

$$+ O_p(n^{-3/2}|\lambda|)$$

$$\equiv S_{3a,2} + O_p(n^{-3/2}|\lambda|), \tag{A.54}$$

which is of the order of $O_p(|\lambda| n^{-1})$ because it is easy to show that $\mathrm{E}[S^2_{3a,2}] = O(|\lambda|^2 n^{-2})$, which follows from the facts that u_i has zero mean and that $\max_{1\le i\le n} |p_{i,0} - p_i| = O_p(n^{-1/2})$.

Finally, noting that $(g_i - g_j)\mathbf{1}(x_j = x_i) = 0$, and that $\mathbf{1}(x_j \neq x_i) L_{ij}^2 = O(|\lambda|^2)$, we have

$$S_{3c} = n^{-3} \sum \sum_{j\neq i} \mathbf{1}(x_j \neq x_i)(g_i - g_j) u_j L_{ij}^2 / \hat{p}_i^2 = O(n^{-1}|\lambda|^2). \tag{A.55}$$

Now, combining (A.51), (A.52), (A.54), and (A.55), we obtain

$$S_3 = S_{3a} - S_{3b} - S_{3c} = (S_{3a,1} - S_{3b,1}) + S_{3a,2} - 2S_{3b,2} + O_p(|\lambda|^2 n^{-1/2} + |\lambda| n^{-3/2})$$

$$= 2S_{3a,2} - 2S_{3b,2} + O_p(|\lambda|^2 n^{-1/2} + |\lambda| n^{-3/2})$$

$$= n^{-1} \sum_{s=1}^{r} \lambda_s \left\{ \frac{2}{n} \sum \sum_{j\neq i} \mathbf{1}_s(x_j, x_i)(g_i - g_j)(p_i - p_{i,0})/p_i^2 \right.$$

$$\left. \times \left[u_i - \frac{1}{n} \sum_{l\neq i, l\neq j} \mathbf{1}(x_l = x_i) u_l / p_i \right] \right\} + O_p(\zeta_n)$$

$$\equiv n^{-1} \sum_{s=1}^{r} \lambda_s Z_{2n,s} + O_p(|\lambda|^2 n^{-1/2} + |\lambda| n^{-3/2}), \tag{A.56}$$

where $Z_{2n,s} = \frac{2}{n}\sum\sum_{j\neq i} \mathbf{1}_s(x_j, x_i)(g_i - g_j)(p_i - p_{i,0})/p_i^2 [u_i - (1/n)\sum_{l\neq i, l\neq j} \mathbf{1}(x_l = x_i) u_l/p_i]$ and $\zeta_n = n^{-1/2}|\lambda|^2 + n^{-3/2}|\lambda|$. Using $\max_{1\le i\le n} |p_{i,0} - p_i| = O(n^{-1/2})$, it is easy to see that $Z_{2n,s}$ is a zero mean $O_p(1)$ random variable. ■

APPENDIX B: Proofs of Theorems 3.1 and 3.2—The Irrelevant Regressor Case

Proof of Theorem 3.1. As outlined in Appendix A, we have $CV(\lambda) = CV_0(\lambda) +$ a term unrelated to λ, where $CV_0(\lambda) = S_1 + S_2 + 2S_3$, with the definitions of the S_j's given in (A.6). Now, we assume that $x = (\bar{x}, \tilde{x})$, where $\bar{x}$ contains the first relevant r_1 components of x and $\tilde{x}$ contains the last $r_2 = r - r_1$ irrelevant components of x; $g(x) = g(\bar{x})$ so that $\tilde{x}$ are irrelevant regressors.

In Lemma B.1 we show that $\hat{\lambda}_s = o_p(1)$ for $s = 1, \ldots, r_1$. Given this result, in the proofs of Lemmas B.2 and B.4 in this Appendix we will only consider the case in which $\bar{\lambda} = (\lambda_1, \ldots, \lambda_{r_1}) \in [0, \eta_n]^{r_1}$, where η_n is a positive sequence that converges to zero as $n \to \infty$. Lemmas B.2–B.4 hold uniformly in $\bar{\lambda} \in \bar{\Lambda}_n = [0, \eta_n]^{r_1}$.

Letting $|\bar{\lambda}| = \sqrt{\sum_{s=1}^{r_1} \lambda_s^2}$, then by Lemmas B.2–B.4, we have

$$S_1 = \sum_{x \in \mathcal{D}} \left[\sum_{s=1}^{r_1} \bar{\lambda}_s \sum_{\bar{z} \in \bar{D}} \mathbf{1}_s(\bar{x}, \bar{z}) \bar{p}(\bar{z})(g(\bar{x}) - g(\bar{z})) \right]^2 \tilde{p}(\tilde{x}) \bar{p}(\bar{x})^{-1} + o_p(|\bar{\lambda}|^2), \tag{B.1}$$

$$S_2 = n^{-1} B(\tilde{\lambda}) + n^{-1} \mathcal{Z}_{1n}(\tilde{\lambda}) + o_p(n^{-1/2}|\bar{\lambda}| + |\bar{\lambda}|^2), \tag{B.2}$$

where $B(\tilde{\lambda})$ equals a positive constant times $\mathrm{E}\{\widetilde{L}_{ij}^2 / [\mathrm{E}(\widetilde{L}_{ij}|\tilde{x}_i)]^2\}$, where $\widetilde{L}_{ij} = L(\tilde{x}_i, \tilde{x}_j, \tilde{\lambda})$. Hence, $B(\tilde{\lambda})$ is positive and finite for all values of $\tilde{\lambda}$. The expression $\mathcal{Z}_{1n}(\tilde{\lambda}) = n^{-1} \sum\sum_{j \neq i} [u_i u_j \mathbf{1}(\bar{x}_j = \bar{x}_i)/\bar{p}_i^2]\, \theta_{ij}(\tilde{\lambda})$ is a zero mean $O_p(1)$ random variable, with $\theta_{ij}(\tilde{\lambda}) = \mathrm{E}\left\{ \frac{\widetilde{L}_{il}\widetilde{L}_{jl}}{[\mathrm{E}(\widetilde{L}_{il}|\tilde{x}_l)]^2} | \tilde{x}_i, \tilde{x}_j \right\}$. Furthermore,

$$S_3 = n^{-1/2} \sum_{s=1}^{r_1} \bar{\lambda}_s \mathcal{Z}_{2n,s}(\tilde{\lambda}) + O_p(n^{-1/2}|\bar{\lambda}|^2 + n^{-1}|\bar{\lambda}| + n^{-3/2}), \tag{B.3}$$

where $\mathcal{Z}_{2n,s}(\tilde{\lambda})$ is a zero mean $O_p(1)$ random variable defined in Lemma B.4.

Similar to the proof of Theorem 2.1, one can write (B.1) as $S_1 = \bar{\lambda}'_{(r_1)} \bar{\Omega} \bar{\lambda}_{(r_1)}$, where $\bar{\lambda}_{(r_1)} = (\bar{\lambda}_1, \ldots, \bar{\lambda}_{r_1})'$ and $\bar{\Omega}$ is an $r_1 \times r_1$ positive definite matrix defined similarly to Ω in Appendix A but with r replaced by r_1. Then combining (B.1)–(B.3), we obtain

$$\begin{aligned} CV_0(\lambda) &= S_1 + S_2 + 2S_3 \\ &= \bar{\lambda}'_{(r_1)} \bar{\Omega} \bar{\lambda}_{(r_1)} + n^{-1} B(\tilde{\lambda}) + n^{-1} \mathcal{Z}_{1n}(\tilde{\lambda}) + 2n^{-1/2} \sum_{s=1}^{r_1} \bar{\lambda}_s \mathcal{Z}_{2n,s}(\tilde{\lambda}) \\ &\quad + o_p(|\bar{\lambda}|^2 + n^{-1}|\bar{\lambda}| + n^{-3/2}) + \quad \text{terms unrelated to } \lambda. \end{aligned} \tag{B.4}$$

Taking the derivative of $CV_0(\lambda)$ with respect to $\bar{\lambda}_{(r_1)}$ gives

$$\frac{\partial CV_0(\lambda)}{\partial \bar{\lambda}_{(r_1)}} = 2\Omega \bar{r}_{(r_1)} + n^{-1/2} \mathcal{Z}_{2n} + (s.o.) \stackrel{set}{=} 0, \tag{B.5}$$

where $\mathcal{Z}_n = (\mathcal{Z}_{2n,1}, \ldots, \mathcal{Z}_{2n,r_1})'$.

By noting that, for any values of $(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{r_2})$, $\mathcal{Z}_{2n,s}$ is an $O_p(1)$ random variable, (B.5) leads to

$$\bar{\lambda}_{(r_1)} = n^{-1/2}[2\bar{\Omega}]^{-1}\mathcal{Z}_{2n} + (s.o.) = O_p(n^{-1/2}). \tag{B.6}$$

From Lemma B.4 we know that (where $\tilde{v}_i = \mathrm{E}(\widetilde{L}_{ij}|\tilde{x}_i)$)

$$\mathcal{Z}_{2n,s}(\tilde{\lambda}) = n^{-3/2}\sum\sum_{j\neq i} u_i(g_i - g_j)\mathbf{1}_s(\bar{x}_j, \bar{x}_i)\frac{1}{\bar{p}_i} \times \left[\frac{\widetilde{L}_{ij}}{\tilde{v}_i} - \frac{1}{\bar{p}_i n}\sum_{l\neq j, l\neq i}\mathbf{1}(\bar{x}_l = \bar{x}_i)\widetilde{L}_{li}\widetilde{L}_{lj}/\tilde{v}_l^2\right],$$

which has zero mean and an asymptotic finite positive variance. This, together with the fact that $\bar{\Omega}^{-1}$ is positive definite and finite, implies that $\bar{\lambda}_{(r_1)}$ converges to zero in probability at an exact rate of $O_p(n^{-1/2})$ (i.e., it cannot go to zero at a rate faster than $O_p(n^{-1/2})$). (Note that when there do not exist irrelevant variables, $\tilde{v}_i = \mathrm{E}(\widetilde{L}_{ij}|\tilde{x}_i)$ reduces to 1 (we replace $\widetilde{L}_{ij}$ by 1 if there are no irrelevant variables). In this case $1 - (1/\bar{p}_i n)\sum_{l\neq j, l\neq i}\mathbf{1}(\bar{x}_l = \bar{x}_i) = O_p(n^{-1/2})$. Hence, $\mathcal{Z}_{2n,s} = O_p(n^{-1/2})$ and $\bar{\lambda}$ becomes $O_p(n^{-1})$ as stated in Theorem 2.1.)

Substituting (B.6) back into $CV_0(\lambda)$ yields a concentrated objective function, which we denote as $\widetilde{CV}_0(\tilde{\lambda})$. It is easy to see that $\widetilde{CV}_0(\tilde{\lambda})$ takes the following form:

$$\widetilde{CV}_0(\tilde{\lambda}) = n^{-1}\{B(\tilde{\lambda}) + \mathcal{X}_n(\tilde{\lambda})\} + \quad \text{terms unrelated to } \tilde{\lambda}, \tag{B.7}$$

where $\mathcal{X}_n(\tilde{\lambda}) = \mathcal{Z}_{1n}(\tilde{\lambda}) + 2\sqrt{n}\sum_{s=1}^{r_1}\bar{\lambda}_s\mathcal{Z}_{2n,s}(\tilde{\lambda})$ is a zero mean $O_p(1)$ random variable. In fact, $\mathcal{X}_n(\tilde{\lambda})$ is a zero mean U-process indexed by $\tilde{\lambda}$. By Lemma B.5 we know that the first term on the right-hand side of (B.7) has a unique minimization point at $\tilde{\lambda}_s = 1$ for all $s = 1, \ldots, r_2$. The second term is a zero mean U-process given by $n^{-1}\sum\sum_{j\neq i} u_i u_j A_{ij} B_{ij}(\tilde{\lambda})$, where $A_{ij} = A(\bar{x}_i, \bar{x}_j)$ is a bounded function of $(\bar{x}_i, \bar{x}_j)$ and where $B_{ij}(\tilde{\lambda})$ is a bounded function of $(\tilde{x}_i, \tilde{x}_j)$ and $\tilde{\lambda}$. Then it is easy to see that, asymptotically, this second term is minimized at $\tilde{\lambda}_s = 1$ for all $s = 1, \ldots, r_2$ with positive probability, say, $\delta \in (0, 1)$. The value of δ will depend on the distribution of x_i and u_i and can be difficult to compute exactly. Because the first term on the right-hand side of (B.7) is minimized at $\tilde{\lambda}_s = 1$ (with probability one) for all $s = 1, \ldots, r_1$, then $\widetilde{\ }_0(\tilde{\lambda})$ is minimized at $\tilde{\lambda}_s = 1$ (for $s = 1, \ldots, r_1$) with probability $\alpha \in (0, 1)$ with $\alpha > \delta$.

Hence, we have

$$\lim_{n\to\infty} \Pr(\tilde{\lambda}_1 = 1, \ldots, \tilde{\lambda}_{r_2} = 1) \geq \alpha \tag{B.8}$$

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

It does not seem to be possible to determine the value of α exactly for the general case. However, because the first term on the right-hand side of (B.7) is uniquely minimized at $\tilde{\lambda}_s = 1$ for $s = 1, \ldots, r_2$ and the other terms have zero means, this suggests that it is likely that $\alpha \in (1/2, 1)$. Indeed our simulations show that α is between 0.6 and 0.65 for a variety of DGPs. ■

Proof of Theorem 3.2. We write $\hat{g}(x) - g(\bar{x}) = (\hat{g}(x) - g(\bar{x}))\hat{p}(x)/\hat{p}(x) \equiv \hat{m}(x)/\hat{p}(x)$, where

$$\begin{aligned}\hat{m}(x) &= (\hat{g}(x) - g(\bar{x}))\hat{p}(x) = n^{-1}\sum_i [Y_i - g(\bar{x})]L_{x_i,x,\lambda} \\ &= n^{-1}\sum_i [g(\bar{x}_i) - g(\bar{x})]L_{x_i,x,\lambda} n^{-1} + \sum_i u_i L_{x_i,x,\lambda} \\ &\equiv \hat{m}_1(x) + \hat{m}_2(x)\end{aligned}$$

and where $L_{x_i,x,\lambda} \equiv L(x_i, x, \lambda)$ and the definitions of $\hat{m}_1(x)$ and $\hat{m}_2(x)$ should be apparent.

We have the following decomposition for $L_{x_i,x,\lambda}$:

$$L(x_i, x, \hat{\lambda}) = \left[\mathbf{1}(x_i = x) + \sum_{s=1}^{r_1} \hat{\lambda}_s \mathbf{1}_s(\bar{x}_i, \bar{x}) + O_p(|\bar{\lambda}|^2)\right] L_{\tilde{x}_i,\tilde{x},\tilde{\lambda}}, \tag{B.9}$$

where $L_{\tilde{x}_i,\tilde{x},\tilde{\lambda}} = \prod_{s=r_1+1}^{r} l(x_{is}, x_s, \lambda_s)$ and $|\bar{\lambda}|^2 = \sum_{s=1}^{r_1} \hat{\lambda}_s^2 = O_p(n^{-1})$.

Using (B.9) we can write $\hat{m}_1(x)$ as

$$\begin{aligned}\hat{m}_1(x) &= n^{-1}\sum_i \left[g(\bar{x}_i) - g(\bar{x})\right]\left[\mathbf{1}(\bar{x}_i = \bar{x}) + \sum_{s=1}^{r_1} \hat{\lambda}_s \mathbf{1}_s(\bar{x}_i, \bar{x}) + O_p(|\bar{\lambda}|^2)\right] L_{\tilde{x}_i,\tilde{x},\tilde{\lambda}} \\ &= \sum_{s=1}^{r_1} \hat{\lambda}_s G_s(x) + O_p(n^{-1}),\end{aligned} \tag{B.10}$$

because $[g(\bar{x}_i) - g(\bar{x})]\mathbf{1}(\bar{x}_i = \bar{x}) \equiv 0$ where $G_s(x) = n^{-1}\sum_i [g(\bar{x}_i) - g(\bar{x})]\mathbf{1}_s(\bar{x}_{is}, \bar{x}) L_{\tilde{x}_i,\tilde{x},\tilde{\lambda}}$.

It is easy to show that $\mathrm{E}[G_s(x)] = \mathrm{E}\{[g(\bar{x}_i) - g(\bar{x})]\mathbf{1}_s(\bar{x}_{is}, \bar{x})\}\mathrm{E}[L_{\tilde{x}_i,\tilde{x},\tilde{\lambda}}] = \sum_{\bar{v}\in\bar{\mathcal{D}}} \bar{p}(\bar{v})[g(\bar{v}) - g(\bar{x})]\mathbf{1}_s(\bar{v}, \bar{x})\nu_1(\tilde{x})$. Also, $\mathrm{var}(G_s(x)) = O(n^{-1})$. Hence,

$$G_s(x) = \mathrm{E}[G_s(x)] + O_p(n^{-1/2}) = B_s(\bar{x})\bar{p}(\bar{x})\nu_1(\tilde{x}) + O_p(n^{-1/2}),$$

where $B_s(\bar{x}) = \bar{p}(\bar{x})^{-1}\sum_{\bar{v}\in\bar{D}} \bar{p}(\bar{v})[g(\bar{v}) - g(\bar{x})]\mathbf{1}_s(\bar{v}, \bar{x})$. This result, together with $\hat{\lambda}_s = O_p(n^{-1/2})$ for $s = 1, \ldots, r_1$, yields

$$\hat{m}_1(x) = \bar{p}(\bar{x})\nu_1(\tilde{x}) \sum_{s=1}^{r_1} \hat{\lambda}_s B_s(\bar{x}) + O_p(n^{-1}). \tag{B.11}$$

Next, we consider $\hat{m}_2(x)$. Using (B.9) we can write $\hat{m}_2(x)$ as $\hat{m}_2(x) = n^{-1}\sum_i u_i \mathbf{1}(\bar{x}_i = \bar{x})\tilde{L}_{\tilde{x}_i,\tilde{x},\tilde{\lambda}} + O_p(n^{-1}) \equiv \hat{m}_{2,0}(x) + O_p(n^{-1})$. Obviously $\mathrm{E}[\hat{m}_{2,0}(x)] = 0$, and its variance is given by

$\mathrm{var}(\hat{m}_{2,0}(x)) = n^{-1}\mathrm{E}[u_i^2 \mathbf{1}(\bar{x}_i = \bar{x})]\mathrm{E}[\widetilde{L}_{ij}^2] = n^{-1}\sigma^2(\bar{x})\bar{p}(\bar{x})\nu_2(\tilde{x})$. Hence, by the Lindeberg central limit theorem, we have

$$\sqrt{n}\hat{m}_2(x) = \sqrt{n}\hat{m}_{2,0}(x) + O_p(n^{-1/2}) \xrightarrow{d} N(0, \sigma^2(\bar{x})\bar{p}(\bar{x})\nu_2(\tilde{x})). \tag{B.12}$$

Combining (B.11) and (B.12) we have

$$\sqrt{n}[\hat{m}(x) - \sum_{s=1}^{r_1} \hat{\lambda}_s B_s(\bar{x})\bar{p}(\bar{x})\nu_1(\tilde{x})] \xrightarrow{d} N(0, \sigma^2(\bar{x})\bar{p}(\bar{x})\nu_2(\tilde{x})). \tag{B.13}$$

Equations (B.16) and (B.13) lead to

$$\sqrt{n}\left[\hat{g}(x) - g(x) - \sum_{s=1}^{r_1} \hat{\lambda}_s B_s(\bar{x})\right] = \frac{\sqrt{n}\,[\hat{m}(x) - \sum_{s=1}^{r_1} \hat{\lambda}_s B_s(\bar{x})\bar{p}(\bar{x})\nu_1(\tilde{x})]}{\bar{p}(\bar{x})\nu_1(\tilde{x})} + O_p(n^{-1/2})$$

$$\xrightarrow{d} \quad N\left(0, \frac{\sigma^2(\bar{x})\nu_2(\tilde{x})}{\bar{p}(\bar{x})\nu_1(\tilde{x})^2}\right).$$

■

We are now ready to state and prove Lemmas B.1–B.4. We first provide some results that are needed to handle the random denominator, $\hat{p}_i$.

Define $\mu(x) = \mathrm{E}[\hat{p}(x_i)|x_i = x]$ and recall that $\nu_l(\tilde{x}) = \mathrm{E}[(\widetilde{L}_{ij})^l|\tilde{x}_i = \tilde{x}]$ (for $l = 1, 2$). Then it is obvious that $\mathrm{E}[\hat{p}(x) - \mu(x)] = 0$ and that $\mathrm{var}(\hat{p}(x) - \mu(x)) = O(n^{-1})$. Hence

$$\hat{p}(x) - \mu(x) = O_p(n^{-1/2}). \tag{B.14}$$

Also, it is easy to show that, for all $x \in \mathcal{D}$,

$$\mu(x) = \mathrm{E}[\overline{L}_{ij}|\bar{x}_i = \bar{x}]\,\mathrm{E}[\widetilde{L}_{ij}|\tilde{x}_i = \tilde{x}] = [\bar{p}(x) + O(|\bar{\lambda}|)]\nu_1(\tilde{x}), \tag{B.15}$$

where $\overline{L}_{ij} = L(\bar{x}_i, \bar{x}_j, \bar{\lambda})$ and $\widetilde{L}_{ij} = L(\tilde{x}_i, \tilde{x}_j, \tilde{\lambda})$.

Combining (B.14) and (B.15) we know that

$$\hat{p}(x) = \bar{p}(\bar{x})\nu_1(\tilde{x}) + O_p(|\bar{\lambda}|) + O_p(n^{-1/2}). \tag{B.16}$$

From (B.16) we have, uniformly in x_i and λ,

$$\frac{1}{\hat{p}_i} = \frac{1}{\bar{p}_i\tilde{\nu}_i} + O_p(n^{-1/2} + |\bar{\lambda}|), \tag{B.17}$$

where $\bar{p}_i = \bar{p}(\bar{x}_i)$ and $\tilde{\nu}_i = \nu(\tilde{x}_i)$, and

$$\frac{1}{\hat{p}_i^2} = \frac{1}{\bar{p}_i^2\tilde{\nu}_i^2} + O_p(n^{-1/2} + |\bar{\lambda}|). \tag{B.18}$$

Similar to the proof of Lemma A.1 one can show that $\hat{\lambda}_s = o_p(1)$ for $s = 1, \ldots, r_1$, which is stated in Lemma B.1, which follows.

LEMMA B.1. *$\hat{\lambda}_s = o_p(1)$ for all $s = 1, \ldots, r_1$.*

Proof. The arguments are similar to the ones used in the proof of Lemma A.1. We use $\mathbf{0}_m$ to denote a row vector of zeros of dimension $1 \times m$. When we choose $\lambda_s = 0$ for all $s = 1, \ldots, r_1$, we know that $CV_0(\mathbf{0}_{r_1}, \tilde{\lambda}) = o_p(1)$ for any value of $\tilde{\lambda} = (\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{r_2})$ because $g_i - g_j = g(\bar{x}_i) - g(\bar{x}_j)$, which does not depend on $\tilde{x}_i$ and $\tilde{x}_j$. Because $\hat{\lambda}$ minimizes

$CV_0(\lambda)$, and from $CV_0(\hat{\lambda}) \leq CV_0(\mathbf{0}_r)$, $CV_0(\lambda) = S_1 + o_p(1)$ (S_2 and S_3 are both $o_p(1)$ because they contain u_i) and $S_1 \geq 0$, we know that it must be true that

$$S_1(\hat{\lambda}) = o_p(1). \tag{B.19}$$

Now consider a generic $\lambda \in [0,1]^r$. It can be seen that

$$\begin{aligned} S_1(\lambda) &= n^{-3} \sum\sum \sum_{l \neq j \neq i} (g_i - g_j)(g_i - g_l) L_{ij} L_{il} / \hat{p}_i^2 \\ &\quad + n^{-3} \sum\sum \sum_{j \neq i} (g_i - g_j)^2 L_{ij}^2 / \hat{p}_i^2 \\ &= n^{-3} \sum\sum \sum_{l \neq j \neq i} (g_i - g_j)(g_i - g_l) L_{ij} L_{il} / \hat{p}_i^2 + O_p(n^{-1}) \equiv A_n(\lambda) + o_p(1), \end{aligned} \tag{B.20}$$

uniformly in $\lambda \in [0,1]^r$, where

$$A_n(\lambda) = n^{-3} \sum\sum \sum_{l \neq j \neq i} (g_i - g_j)(g_i - g_l) L_{ij} L_{il} / \hat{p}_i^2.$$

Define $p_{i\lambda} = \mathrm{E}(\hat{p}_i | x_i)$ and define $A_{1n}(\lambda)$ by replacing $1/\hat{p}_i^2$ by $1/p_{i\lambda}^2$ in $A_n(\lambda)$; i.e., $A_{1n}(\lambda) = n^{-3} \sum\sum\sum_{l \neq j \neq i} (g_i - g_j)(g_i - g_l) L_{ij} L_{il} / p_{i\lambda}^2$. Then it can be shown that $A_n(\lambda) = A_{1n}(\lambda) + o_p(1)$ uniformly in $\lambda \in [0,1]^r$. The term A_{1n} can be written as a third-order U-statistic, and by the H-decomposition of U-statistics we know that, uniformly in $\lambda \in [0,1]^r$,

$$A_{1n}(\lambda) = \mathrm{E}(A_{1n}(\lambda)) + o_p(1). \tag{B.21}$$

It can also be seen that

$$\begin{aligned} \mathrm{E}(A_{1n}(\lambda)) &= \mathrm{E}\left\{ \left[\mathrm{E}\left[(g_i - g_j) \overline{L}_{ij} | \bar{x}_i \right] \right]^2 [\mathrm{E}(\widetilde{L}_{ij} | \tilde{x}_i)]^2 / p_{i\lambda}^2 \right\} + o(1) \\ &= \sum_{x \in \mathcal{D}} p(x) \left\{ \sum_{\bar{z} \in \bar{D}} \bar{p}(\bar{z}) [g(\bar{x}) - g(\bar{z})] L(\bar{x}, \bar{z}, \bar{\lambda}) \right\}^2 / \bar{p}(\bar{x}, \bar{\lambda})^2 + o(1), \end{aligned} \tag{B.22}$$

where $\bar{p}(\bar{x}, \bar{\lambda}) = \mathrm{E}[\bar{L}_{ij} | \bar{x}_i = \bar{x}]$ and we used the fact that $p_{i\lambda} = \mathrm{E}[\bar{L}_{ij} | \bar{x}_i] \mathrm{E}[\widetilde{L}_{ij} | \tilde{x}_i]$ so that $[\mathrm{E}(\widetilde{L}_{ij} | \tilde{x}_i)]^2 / p_{i\lambda}^2 = 1/\bar{p}(\bar{x}, \bar{\lambda})^2$. Similar to the arguments used in the proof of Lemma A.1, the preceding results imply that

$$\sum_{x \in \mathcal{D}} p(x) \left\{ \sum_{\bar{z} \in \bar{D}} \bar{p}(\bar{z}) [g(\bar{x}) - g(\bar{z})] L(\bar{x}, \bar{z}, \bar{\lambda}) \right\}^2 / \bar{p}(\bar{x}, \bar{\lambda})^2 = o_p(1), \tag{B.23}$$

with $\bar{\lambda}$ being selected by the CV method. Equation (B.23) is equivalent to, for all $\bar{x} \in \bar{D}$,

$$\left\{ \sum_{\bar{z} \in \bar{D}} \bar{p}(\bar{z}) [g(\bar{x}) - g(\bar{z})] L(\bar{x}, \bar{z}, \bar{\lambda}) \right\}^2 = o_p(1) \quad \text{with } \bar{\lambda} \text{ selected by the CV method.} \tag{B.24}$$

Equation (B.24) and Assumption 3 imply that the CV selected smoothing parameters for the relevant variables all converge to zero in probability, i.e., $\hat{\lambda}_s = o_p(1)$ for $s = 1, \ldots, r_1$. ■

We are now ready to prove Lemmas B.2–B.4.

LEMMA B.2.

$$S_1 = \sum_{x \in \mathcal{D}} \left[\sum_{s=1}^{r_1} \bar{\lambda}_s \sum_{\bar{z} \in \bar{D}} \mathbf{1}_s(\bar{x}, \bar{z}) \bar{p}(\bar{z}) (g(\bar{x}) - g(\bar{z})) \right]^2 \tilde{p}(\bar{x}) \bar{p}(\bar{x})^{-1} + o_p(\bar{\lambda}^2).$$

Proof. Define S_1^0 the same way as S_1 but with $\hat{p}_i^{-2}$ replaced by $(\bar{p}_i \tilde{\nu}_i)^{-2}$. Then

$$\begin{aligned} S_1^0 &= \frac{1}{n(n-1)^2} \sum_i \sum_{j \neq i} (g_i - g_j)^2 L_{ji}^2 / (\bar{p}_i \tilde{\nu}_i)^2 \\ &\quad + \frac{1}{n(n-1)(n-2)} \sum_i \sum_{j \neq i} \sum_{l \neq i, l \neq j} (g_i - g_j)(g_i - g_l) L_{ji} L_{li} / (\bar{p}_i \tilde{\nu}_i)^2 \\ &\equiv S_{1a} + S_{1b}. \end{aligned} \tag{B.25}$$

The proof is similar to the proof of Lemma A.2. Here S_{1b} can be written as a third-order U-statistic, and one can show that the leading term of S_{1b} is $\mathrm{E}[S_{1b}]$. It can be seen that

$$\begin{aligned} \mathrm{E}[S_{1b}] &= \mathrm{E}\left\{ \left[\mathrm{E}\left[(g_i - g_j) \overline{L}_{ij} | \bar{x}_i \right] / \bar{p}_i \right]^2 \right\} \mathrm{E}\left\{ \left[\mathrm{E}(\widetilde{L}_{ij} | \tilde{x}_i) / \tilde{\nu}_i \right]^2 \right\} \\ &= \mathrm{E}\left\{ \left[\left[\mathrm{E}((g_i - g_j) \overline{L}_{ij} | \bar{x}_i) \right] / \bar{p}_i \right]^2 \right\} \\ &= \sum_{x \in \mathcal{D}} \left[\sum_{s=1}^{r_1} \bar{\lambda}_s \sum_{\bar{z} \in \bar{D}} \mathbf{1}_s(\bar{x}, \bar{z}) \bar{p}(\bar{z}) (g(\bar{x}) - g(\bar{z})) \right]^2 \tilde{p}(\bar{x}) \bar{p}(\bar{x})^{-1} + o_p(|\bar{\lambda}|^2), \end{aligned} \tag{B.26}$$

where the second equality follows from the fact that $\mathrm{E}[\widetilde{L}_{ij} | \tilde{x}_i] = \tilde{\nu}_i$ (so that it cancels $1/\tilde{\nu}_i$) and the third equality follows exactly the same derivation as in the proof of Lemma A.2 as it is unrelated to $\tilde{\lambda}$.

Thus, we have shown that the leading term of S_{1b} is unrelated to $\tilde{\lambda}$. Hence,

$$S_{1b} = \sum_{x \in D} \left[\sum_{s=1}^{r_1} \bar{\lambda}_s \sum_{\bar{z} \in \bar{D}} \mathbf{1}_s(\bar{x}, \bar{z}) \bar{p}(\bar{z}) (g(\bar{x}) - g(\bar{z})) \right]^2 \tilde{p}\,(\bar{x}) \bar{p}(\bar{x})^{-1} + o_p(|\bar{\lambda}|^2). \tag{B.27}$$

Now consider S_{1a}, where

$$S_{1a} = \frac{1}{n(n-1)^2} \sum_{i=1}^{n} \sum_{j \neq i}^{n} (g_i - g_j)^2 L_{ji}^2 / (\bar{p}_i \tilde{\nu}_i)^2. \tag{B.28}$$

It is easy to see that the leading term of S_{1a} is $\mathrm{E}[S_{1a}] = O(n^{-1} |\bar{\lambda}|^2)$ uniformly in $\bar{\lambda}$. Thus, we have $S_{1a} = o_p(|\bar{\lambda}|^2)$.

Summarizing (B.25), (B.27), and (B.28) we have shown that

$$S_1^0 = S_{1a} + S_{1b} = \sum_{x\in\mathcal{D}} \left[\sum_{s=1}^{r_1} \bar{\lambda}_s \left(\sum_{\bar{z}\in\bar{D}} \mathbf{1}_s(\bar{z},\bar{x}) \bar{p}(\bar{z}) (g(\bar{x}) - g(\bar{z})) \right) \right]^2 \tilde{p}(\tilde{x}) \bar{p}(\bar{x})^{-1} + o(\bar{\lambda}^2) \quad \text{uniformly in } \bar{\lambda}. \tag{B.29}$$

Define $\hat{m}_{1,i} = (n-1)^{-1} \sum_{j\neq i} [g(x_j) - g(x)] L(x_i, x, \lambda)$ and $\Delta_i(x_i) \equiv \hat{p}_i - p_i$. Then similar to the proof of Lemma A.2, one can show that $\hat{m}_{1,i}(x) = O_p(|\bar{\lambda}|^2)$ and $\Delta_i(x) = o_p(1)$, where both are uniform in x and $\bar{\lambda}$.

Using (B.18) and by arguments similar to those used in the proof of Lemma A.2, we have

$$\left| S_1 - S_1^0 \right| \leq C \sup_{1\leq i\leq n} \hat{m}_{1,i}^2(x_i) \sup_{1\leq i\leq n} |\Delta_i(x_i)| = o_p(\bar{\lambda}^2) \tag{B.30}$$

uniformly in $\bar{\lambda}$.

By (B.29) and (B.30), we obtain

$$S_1 = \sum_{x\in\mathcal{D}} \left[\sum_{s=1}^{r_1} \lambda_s \sum_{\bar{z}\in\bar{\mathcal{D}}} \mathbf{1}_s(\bar{z},\bar{x}) \bar{p}(\bar{z}) (g(\bar{x}) - g(\bar{z})) \right]^2 \tilde{p}(\tilde{x}) \bar{p}(\bar{x})^{-1} + o_p(|\bar{\lambda}|^2). \tag{B.31}$$

■

LEMMA B.3. *$S_2 = n^{-1} B(\lambda) + n^{-1} \mathcal{Z}_{1n}(\tilde{\lambda}) + o_p(|\bar{\lambda}|^2 + |\bar{\lambda}| n^{-1/2}) +$ terms unrelated to λ, where $B(\tilde{\lambda}) = \mathrm{E}\{u_j^2 \mathbf{1}(\bar{x}_j = \bar{x}_i) \widetilde{L}_{ij}^2 / [\bar{p}_i^2 \tilde{v}_i^2]\}$ is positive and finite and where $\mathcal{Z}_{1n}(\tilde{\lambda}) = n^{-1} \sum\sum_{j\neq i} [u_i u_j \mathbf{1}(\bar{x}_j = \bar{x}_i) / \bar{p}_i^2] \theta_{ij}(\tilde{\lambda})$ is a zero mean $O_p(1)$ random variable with $\theta_{ij}(\tilde{\lambda}) = \mathrm{E}[\widetilde{L}_{il} \widetilde{L}_{jl} / \tilde{v}_l^2 | \tilde{x}_i, \tilde{x}_j]$.*

Proof. Similar to the proof of Lemma A.3, we have

$$S_2 = n^{-3} \sum\sum_{j\neq i} u_j^2 L_{ij}^2 / \hat{p}_i^2 + n^{-3} \sum\sum\sum_{i\neq j\neq l} u_j u_l L_{ij} L_{il} / \hat{p}_i^2 - 2n^{-2} \sum\sum_{j\neq i} u_i u_j L_{ij} / \hat{p}_i \equiv S_{2a} + S_{2b} - S_{2c}.$$

By (B.18) and noting that $\overline{L}_{ij} = \mathbf{1}(\bar{x}_j = \bar{x}_i) + O(|\bar{\lambda}|)$, we have

$$\begin{aligned} S_{2a} &= n^{-3} \sum\sum_{j\neq i} u_j^2 \mathbf{1}(\bar{x}_j = \bar{x}_i) \widetilde{L}_{ij}^2 / \hat{p}_i^2 + O_p(n^{-1}|\bar{\lambda}|) \\ &= n^{-3} \sum\sum_{j\neq i} u_j^2 \mathbf{1}(\bar{x}_j = \bar{x}_i) \widetilde{L}_{ij}^2 / [\bar{p}_i^2 \tilde{v}_i^2] + O_p(n^{-1}|\bar{\lambda}|) \\ &= n^{-1} B(\tilde{\lambda}) + O_p(n^{-3/2} + n^{-1}|\bar{\lambda}|), \end{aligned} \tag{B.32}$$

where $B(\tilde{\lambda}) = \mathrm{E}\{u_j^2 \mathbf{1}(\bar{x}_j = \bar{x}_i) \widetilde{L}_{ij}^2 / [\bar{p}_i^2 \tilde{v}_i^2]\}$ is positive and finite and in the last equality we have used the fact that $n^{-2} \sum\sum_{j\neq i} u_j^2 \mathbf{1}(\bar{x}_j = \bar{x}_i) \widetilde{L}_{ij} / [\bar{p}_i^2 \tilde{v}_i^2] = B(\tilde{\lambda}) + O_p(n^{-1/2})$,

which follows from the U-statistic H-decomposition. Note that we can also write $B(\lambda) = B_0 B_1(\tilde{\lambda})$, where $B_0 = \mathrm{E}[\sigma^2(\bar{x}_i)\mathbf{1}(x_j = x_i)/\bar{p}_i^2]$ is a positive constant and $B_1(\tilde{\lambda}) = \mathrm{E}\{\widetilde{L}_{ij}^2/\tilde{v}_i^2\}$.

Next, we consider S_{2b}. Note that $S_{2b} = O_p(n^{-1})$ because u_j (u_l) has zero mean. Letting S_{2b}^0 denote S_{2b} but with $\hat{p}_i^{-2}$ replaced by $(\bar{p}_i\tilde{v}_i)^{-2}$, by (B.18) we know that

$$S_{2b} = S_{2b}^0 + O_p(n^{-3/2} + n^{-1}|\lambda|), \tag{B.33}$$

where

$$\begin{aligned} S_{2b}^0 &= n^{-3}\sum\sum\sum_{i\neq j\neq l} u_j u_l \mathbf{1}(x_j = x_i)\mathbf{1}(x_l = x_i)\widetilde{L}_{ij}\widetilde{L}_{il}/[\bar{p}_i^2\tilde{v}_i^2] + O_p(n^{-1}|\lambda|) \\ &\equiv S_{2b,1}^0 + O_p(n^{-1}|\lambda|), \end{aligned}$$

where $S_{2b,1}^0 = n^{-3}\sum\sum\sum_{i\neq j\neq l} u_j u_l \mathbf{1}(x_j = x_i)\mathbf{1}(x_l = x_i)\widetilde{L}_{ij}\widetilde{L}_{il}/[\bar{p}_i^2\tilde{v}_i^2]$.

Now, $S_{2b,1}^0$ can be written as a third-order U-statistic. Let $\mathcal{Q}_{ijl}$ denote the symmetrized version of $u_j u_l \mathbf{1}(x_j = x_i)\mathbf{1}(x_l = x_i)/(\bar{p}_i^2\tilde{v}_i^2)$ and define $\mathcal{Q}_{ij} = \mathrm{E}[\mathcal{Q}_{ijl}|x_i, x_j]$. Then it is easy to show that $\mathcal{Q}_{ij} = (1/3)u_i u_j \mathrm{E}[\mathbf{1}(x_l = x_i)\mathbf{1}(x_l = x_j)/\bar{p}_l^2|\tilde{x}_i, \tilde{x}_j]\mathrm{E}[\widetilde{L}_{il}\widetilde{L}_{jl}/\tilde{v}_l^2|\tilde{x}_i, \tilde{x}_j] = (1/3)u_i u_j[\mathbf{1}(x_j = x_i)/\bar{p}_i]\mathrm{E}[\widetilde{L}_{il}\widetilde{L}_{jl}/\tilde{v}_l^2|\tilde{x}_i, \tilde{x}_j]$. By the U-statistic H-decomposition we have

$$\begin{aligned} S_{2b,1}^0 &= \frac{2}{n(n-1)}\sum\sum_{j>i}\mathcal{Q}_{ij} + \frac{6}{n(n-1)(n-2)}\sum\sum\sum_{l>j>i}[\mathcal{Q}_{ijl} - \mathcal{Q}_{ij} - \mathcal{Q}_{il} - \mathcal{Q}_{jl}] \\ &= \frac{2}{n(n-1)}\sum\sum_{j>i}\mathcal{Q}_{ij} + O_p(n^{-2}) \\ &= (2/3)n^{-2}\sum\sum_{j>i} u_i u_j[\mathbf{1}(\bar{x}_j = \bar{x}_i)/\bar{p}_i]\mathrm{E}[\widetilde{L}_{il}\widetilde{L}_{jl}/\tilde{v}_l^2|\tilde{x}_i, \tilde{x}_j] + O_p(n^{-1}|\bar{\lambda}| + n^{-2}) \\ &= n^{-1}\mathcal{Z}_{1n}(\tilde{\lambda}) + O_p(n^{-1}|\bar{\lambda}| + n^{-2}), \end{aligned} \tag{B.34}$$

where $\mathcal{Z}_{1n}(\tilde{\lambda}) = (2/3)n^{-1}\sum\sum_{j>i} u_i u_j[\mathbf{1}(\bar{x}_j = \bar{x}_i)/\bar{p}_i]\mathrm{E}[\widetilde{L}_{il}\widetilde{L}_{jl}/\tilde{v}_l^2|\tilde{x}_i, \tilde{x}_j]$. It is easy to see that $\mathcal{Z}_{1n}$ is a zero mean $O_p(1)$ random variable.

Finally, it is easy to see that $S_{2c} = n^{-3}\sum\sum_{j>i} u_i u_j L_{ij}/\hat{p}_i = O_p(n^{-2})$. Summarizing the preceding results, we have shown that

$$S_2 = n^{-1}B(\tilde{\lambda}) + n^{-1}\mathcal{Z}_{1n}(\tilde{\lambda}) + O_p(n^{-1}|\bar{\lambda}| + n^{-2}). \tag{B.35}$$

■

LEMMA B.4.

$$S_3 = n^{-1/2}\sum_{s=1}^{r_1}\bar{\lambda}_s\mathcal{Z}_{2n,s}(\tilde{\lambda}) + o_p(|\bar{\lambda}|^2 + n^{-1/2}|\bar{\lambda}|),$$

where $\mathcal{Z}_{2n,s}(\tilde{\lambda}) = n^{-3/2}\sum\sum_{j\neq i}[u_i(g_i - g_j)\mathbf{1}_s(\bar{x}_j, \bar{x}_i)/\bar{p}_i][(\widetilde{L}_{ij}/v_i) - (1/\bar{p}_i n)\sum_{l\neq j, l\neq i}\mathbf{1}(\bar{x}_l = \bar{x}_i)\widetilde{L}_{li}\widetilde{L}_{lj}/\tilde{v}_l^2]$ *is a zero mean* $O_p(1)$ *random variable.*

Proof.

$$S_3 = n^{-2}\sum_i\sum_{j\neq i} u_i(g_i-g_j)L_{ij}/\hat{p}_i - n^{-3}\sum\sum\sum_{i\neq j\neq l}(g_i-g_j)u_l L_{ij}L_{il}/\hat{p}_i^2$$

$$-n^{-3}\sum_i\sum_{j\neq i}(g_i-g_j)u_j L_{ij}^2/\hat{p}_i^2 \equiv S_{3a}-S_{3b}-S_{3c}.$$

Noting that $(g_i-g_j)\mathbf{1}(\bar{x}_j=\bar{x}_i)\equiv 0$ (because $g_i-g_j=0$ if $\bar{x}_j=\bar{x}_i$) and using (B.17), we have

$$S_{3a} = n^{-2}\sum\sum_{j\neq i} u_i(g_i-g_j)\left[0+\sum_{s=1}^{r_1}\bar{\lambda}_s\mathbf{1}_s(\bar{x}_j,\bar{x}_i)\right]\widetilde{L}_{ij}/\hat{p}_i + O_p(n^{-1/2}|\bar{\lambda}|^2)$$

$$= n^{-1/2}\sum_{s=1}^{r_1}\bar{\lambda}_s n^{-3/2}\sum_i\sum_{j\neq i}\mathbf{1}_s(\bar{x}_j,\bar{x}_i)u_i(g_i-g_j)\widetilde{L}_{ij}/[\bar{p}_i\tilde{v}_i] + O_p(|\bar{\lambda}|^2 n^{-1/2})$$

$$+ O_p(|\bar{\lambda}|n^{-1/2}+n^{-1}|\bar{\lambda}|)$$

$$= n^{-1/2}\sum_{s=1}^{r_1}\bar{\lambda}_s\mathcal{Z}_{3n,s}(\tilde{\lambda}) + O_p(|\bar{\lambda}|^2 n^{-1/2}+|\bar{\lambda}|n^{-1}), \tag{B.36}$$

where the zero in the first equality comes from $(g_i-g_j)\mathbf{1}(\bar{x}_j=\bar{x}_i)=0$, the second equality uses (B.17), and $\mathcal{Z}_{3n,s}(\tilde{\lambda}) = n^{-3/2}\sum_i\sum_{j\neq i}\mathbf{1}_s(\bar{x}_j,\bar{x}_i)u_i(g_i-g_j)\widetilde{L}_{ij}/[\bar{p}_i\tilde{v}_i]$, which is obviously a zero mean $O_p(1)$ random variable.

Next, we consider S_{3b}, and again $||\bar{x}_j-\bar{x}_i||\geq 1$, for otherwise $S_{3b}=0$. Using (B.18), we have

$$S_{3b} = n^{-3}\sum_{s=1}^{r_1}\bar{\lambda}_s\sum\sum\sum_{l\neq j\neq i}\mathbf{1}_s(\bar{x}_j,\bar{x}_i)\mathbf{1}(\bar{x}_l=\bar{x}_i)u_l(g_i-g_j)\widetilde{L}_{ij}\widetilde{L}_{il}/\hat{p}_i^2$$

$$+ O_p(n^{-1/2}|\bar{\lambda}|^2)$$

$$= n^{-3}\sum_{s=1}^{r_1}\bar{\lambda}_s\sum\sum\sum_{l\neq j\neq i}\mathbf{1}_s(\bar{x}_j,\bar{x}_i)\mathbf{1}(\bar{x}_l=\bar{x}_i)u_l(g_i-g_j)\widetilde{L}_{ij}\widetilde{L}_{il}/[\bar{p}_i^2\tilde{v}_i^2]$$

$$+ O_p(n^{-1/2}|\bar{\lambda}|^2+n^{-1}|\bar{\lambda}|)$$

$$= n^{-1/2}\sum_{s=1}^{r_1}\bar{\lambda}_s\mathcal{Z}_{4n,s}(\tilde{\lambda}) + O_p(|\bar{\lambda}|^2 n^{-1/2}) + |\bar{\lambda}|n^{-1}), \tag{B.37}$$

where $\mathcal{Z}_{4n,s} = n^{-5/2}\sum\sum\sum_{l\neq j\neq i}\mathbf{1}_s(\bar{x}_j,\bar{x}_i)\mathbf{1}(\bar{x}_l=\bar{x}_i)u_l(g_i-g_j)/[\bar{p}_i^2\tilde{v}_i^2]$ is a zero mean $O_p(1)$ random variable.

Note that unlike the case for which all regressors are relevant, here there are no cancellations. Therefore, the leading term is of order $n^{-1/2}|\bar{\lambda}|$ rather than $n^{-1}|\bar{\lambda}|$. To elaborate further on this point, we use $S_{3b,L}$ to denote the leading term of S_{3b}, i.e.,

$$S_{3b,L} = n^{-3} \sum_{s=1}^{r_1} \bar{\lambda}_s \sum\sum \sum_{l \neq j \neq i} \mathbf{1}_s(\bar{x}_j, \bar{x}_i)\mathbf{1}(\bar{x}_l = \bar{x}_i) u_l (g_i - g_j) \widetilde{L}_{ij} \tilde{L}_{il} / [\bar{p}_i^2 \tilde{v}_i^2]. \quad \textbf{(B.38)}$$

Subsequently we will show that the leading term of S_{3b} will not cancel the leading term of S_{3a} when there exist irrelevant regressors. First note that

$$\mathbf{1}(\bar{x}_l = \bar{x}_i)\mathbf{1}_s(\bar{x}_j, \bar{x}_i)(g_i - g_j)/\bar{p}_i^2 = \mathbf{1}(\bar{x}_l - \bar{x}_i)\mathbf{1}_s(\bar{x}_j, \bar{x}_l)(g_l - g_j)/\bar{p}_l^2, \quad \textbf{(B.39)}$$

because $\bar{x}_i = \bar{x}_l$ (because of $\mathbf{1}(\bar{x}_l = \bar{x}_i)$).

Using (B.39) $S_{3b,L}$ can be written as

$$\begin{aligned}
S_{3b,L} &= n^{-2} \sum_{s=1}^{r_1} \bar{\lambda}_s \sum_l \sum_{j \neq l} [u_l (g_l - g_j)\mathbf{1}_s(\bar{x}_j, \bar{x}_l)/\bar{p}_l^2][n^{-1} \sum_{i \neq j, l} \mathbf{1}(\bar{x}_l = \bar{x}_i)\widetilde{L}_{ij}\widetilde{L}_{il}/\tilde{v}_i^2] \\
&= n^{-2} \sum_{s=1}^{r_1} \bar{\lambda}_s \sum_l \sum_{j \neq l} [u_l (g_l - g_j)\mathbf{1}_s(\bar{x}_j, \bar{x}_l)/\bar{p}_l^2] \mathrm{E}\left[\mathbf{1}(\bar{x}_l = \bar{x}_i)|\bar{x}_l\right] \\
&\quad \times \mathrm{E}[\widetilde{L}_{ij}\widetilde{L}_{il}/\tilde{v}_i^2 | \tilde{x}_j, \tilde{x}_l] + O_p(n^{-1}|\lambda|) \\
&= n^{-2} \sum_{s=1}^{r_1} \bar{\lambda}_s \sum_l \sum_{j \neq l} [u_l (g_l - g_j)\mathbf{1}_s(\bar{x}_j, \bar{x}_l)/\bar{p}_l] \mathrm{E}[\widetilde{L}_{ij}\widetilde{L}_{il}/\tilde{v}_i^2 | \tilde{x}_j, \tilde{x}_l] + O_p(n^{-1}|\lambda|) \\
&= n^{-2} \sum_{s=1}^{r_1} \bar{\lambda}_s \sum_i \sum_{j \neq i} [u_i (g_i - g_j)\mathbf{1}_s(\bar{x}_j, \bar{x}_i)/\bar{p}_i] \mathrm{E}[\widetilde{L}_{lj}\widetilde{L}_{li}/\tilde{v}_l^2 | \tilde{x}_j, \tilde{x}_i] + O_p(n^{-1}|\lambda|) \\
&\equiv S_{3b,1} + O_p(n^{-1}|\lambda|), \quad \textbf{(B.40)}
\end{aligned}$$

where the definition of $S_{3b,1}$ should be apparent. In the third equality we used $\mathrm{E}[\mathbf{1}(\bar{x}_l = \bar{x}_i)|\bar{x}_l] = \bar{p}_l$, and the fourth equality follows by interchanging index i with l.

From (B.36) we know that the leading term of S_{3a} is

$$S_{3a,1} \overset{def}{=} n^{-2} \sum_{s=1}^{r_1} \bar{\lambda}_s \sum_i \sum_{j \neq i} \mathbf{1}_s(\bar{x}_i, \bar{x}_j) u_i (g_i - g_j) \widetilde{L}_{ij}/(\bar{p}_i \tilde{v}_i). \quad \textbf{(B.41)}$$

Now, (B.40) and (B.41) lead to

$$\begin{aligned}
S_{3a,1} - S_{3b,1} = n^{-2} \sum_{s=1}^{r_1} \bar{\lambda}_s \sum_i \sum_{j \neq i} &\left[\mathbf{1}_s(\bar{x}_i, \bar{x}_j) u_i (g_i - g_j)/\bar{p}_i\right] \\
&\times \left[\frac{\widetilde{L}_{ij}}{\tilde{v}_i} - \mathrm{E}\left(\frac{\widetilde{L}_{il}\widetilde{L}_{jl}}{\tilde{v}_l^2} | \tilde{x}_i, \tilde{x}_j\right)\right]. \quad \textbf{(B.42)}
\end{aligned}$$

Obviously, in the absence of the irrelevant regressors, we will have $\widetilde{L}_{ij} = 1$ and $\tilde{v}_i = 1$; hence (B.42) vanishes as we have seen previously in Appendix A. However, now with the existence of the irrelevant regressors, it is easy to show that $S_{3a,1} - S_{3b,1}$ does not vanish in general. To see this, consider the simple case where $\tilde{x}_i$ is a binary variable taking values in

$\{0,1\}$. Let $C_n \overset{def}{=} (\widetilde{L}_{ij}/\tilde{v}_i) - \mathrm{E}((\widetilde{L}_{il}\widetilde{L}_{jl}/\tilde{v}_l^2)|\tilde{x}_i, \tilde{x}_j)$. Without the irrelevant variable $\tilde{x}$, we have $\widetilde{L}_{ij} = 1$ and $\tilde{v}_i = 1$ so that $C_n = 0$. However, when there exists an irrelevant variable $\tilde{x}$, C_n does not vanish. To see this, it is straightforward to show that

$$\mathrm{E}(C_n|\tilde{x}_i = 0) = \mathrm{E}(\widetilde{L}_{ij}/\tilde{v}_i|\tilde{x}_i = 0) - \mathrm{E}\left(\frac{\widetilde{L}_{il}}{\tilde{v}_l}|\tilde{x}_i = 0\right) = 1 - \sum_{\tilde{x}_l \in \{0,1\}} \tilde{p}(\tilde{x}_l)\left[\frac{\widetilde{L}_{il}}{\tilde{v}_l}\right]$$

$$= 1 - \frac{\tilde{p}(0)}{\tilde{p}(0) + \tilde{\lambda}\tilde{p}(1)} - \frac{\tilde{p}(1)\tilde{\lambda}}{\tilde{p}(1) + \tilde{\lambda}\tilde{p}(0)} = \frac{\tilde{\lambda}(\tilde{\lambda} - 1)\tilde{p}(1)^2}{[\tilde{p}(0) + \tilde{\lambda}\tilde{p}(1)][\tilde{p}(1) + \tilde{\lambda}\tilde{p}(0)]}, \tag{B.43}$$

where $\tilde{p}(0) = \tilde{p}(\tilde{x}_l = 0)$, $\tilde{p}(1) = \tilde{p}(\tilde{x}_l = 1)$, and we used $\tilde{v}(\tilde{x}_i = 0) = \tilde{p}(0) + \tilde{\lambda}\tilde{p}(1)$ and $\tilde{v}(\tilde{x}_i = 1) = \tilde{p}(1) + \tilde{\lambda}\tilde{p}(0)$. Equation (B.43) implies that $S_{3a,1} - S_{3b,1} = O_p(n^{-1/2}|\bar{\lambda}|)$, which is the same order as $S_{3a,1}$ or $S_{3b,1}$. In particular, $S_{3a,1} - S_{3b,1} \neq O_p(n^{-1}|\bar{\lambda}|)$. Hence, the leading terms from S_{3a} and S_{3b} no longer cancel when there exist irrelevant regressors.

Finally from $(g_i - g_j)\mathbf{1}(\bar{x}_j = \bar{x}_i) = 0$ and $\mathbf{1}(\bar{x}_j \neq \bar{x}_i)\bar{L}_{ij} = O(|\bar{\lambda}|^2)$, it is easy to see that

$$S_{3c} = O_p(n^{-3/2}|\bar{\lambda}|^2). \tag{B.44}$$

Now, combining (B.36), (B.37), and (B.44), we obtain

$$S_3 = n^{-1/2}\sum_{s=1}^{r_1} \bar{\lambda}_s \mathcal{Z}_{2n,s}(\tilde{\lambda}) + O_p(|\bar{\lambda}|^2 n^{-1/2} + |\bar{\lambda}|n^{-1}), \tag{B.45}$$

where $\mathcal{Z}_{2n,s}(\tilde{\lambda}) = \mathcal{Z}_{3n,s}(\tilde{\lambda}) - \mathcal{Z}_{4n,s}(\tilde{\lambda})$ is a zero mean $O_p(1)$ random variable. ■

Note that the preceding result differs from the case where all regressors are relevant. Because of the existence of irrelevant regressors, the leading terms in $\mathcal{Z}_{3n,s}$ and $\mathcal{Z}_{4n,s}$ do not cancel. Therefore, the leading term of S_3 is of order $O_p(n^{-1/2}|\bar{\lambda}|)$ rather than $O_p(n^{-1}|\lambda|)$. This is the key reason why $\hat{\lambda}_s = O_p(n^{-1/2})$ (for $s = 1, \ldots, r_1$) rather than $O_p(n^{-1})$ when there exist irrelevant regressors.

LEMMA B.5. *$B(\tilde{\lambda})$ has a unique minimization point at $\tilde{\lambda}_s = 1$ for all $s = 1, \ldots, r_2$.*

Proof. Because $B(\tilde{\lambda})$ equals a positive constant times $B_1(\tilde{\lambda})$, it suffices to prove the result for $B_1(\tilde{\lambda}) = \mathrm{E}\{\mathrm{E}(\widetilde{L}_{ij}^2|\tilde{x}_i)/[\mathrm{E}(\widetilde{L}_{ij}|\tilde{x}_i)]^2\}$. We know that $B_1(\tilde{\lambda}) \geq 1$ by the Cauchy inequality. When $\tilde{\lambda}_s = 1$ for all $s = 1, \ldots, r_2$, we have $\tilde{L}_{ij} \equiv 1$, and hence $B_1(\cdot) = 1$, reaching its minimum value. To see that this is the unique minimization value of $B_1(\tilde{\lambda})$, we show that if $B_1(\tilde{\lambda}) = 1$, then one must have that $\tilde{\lambda}_s = 1$ for all $s = 1, \ldots, r_2$. Note that $B_1(\tilde{\lambda}) = 1$ is equivalent to $\mathrm{E}[\widetilde{L}_{ij}^2|\tilde{x}_i] = [\mathrm{E}(\widetilde{L}_{ij}|\tilde{x}_i)]^2$, or

$$\mathrm{var}(\widetilde{L}_{ij}|\tilde{x}_i) = \mathrm{E}\left\{\left[\widetilde{L}_{ij} - \mathrm{E}(\widetilde{L}_{ij}|\tilde{x}_i)\right]^2|\tilde{x}_i\right\} = 0,$$

which implies that

$$\widetilde{L}_{ij} - \mathrm{E}(\widetilde{L}_{ij}|\tilde{x}_i) \equiv 0. \tag{B.46}$$

Equation (B.46) implies that, for any given $\tilde{x}_i$, $\widetilde{L}_{ij}$ does not depend on $\tilde{x}_j$. Now,

$$\begin{aligned}\widetilde{L}_{ij} &= \prod_{s=1}^{r_2}[\mathbf{1}(\tilde{x}_{is}=\tilde{x}_{js})+\tilde{\lambda}_s\mathbf{1}(\tilde{x}_{is}\neq\tilde{x}_{js})]\\ &= \prod_{s=1}^{r_2}\{\mathbf{1}(\tilde{x}_{is}=\tilde{x}_{js})+\tilde{\lambda}_s[1-\mathbf{1}(\tilde{x}_{is}=\tilde{x}_{js})]\}\\ &= \prod_{s=1}^{r_2}[(1-\tilde{\lambda}_s)\mathbf{1}(\tilde{x}_{is}=\tilde{x}_{js})+\tilde{\lambda}_s]. \qquad \textbf{(B.47)}\end{aligned}$$

From (B.47) we know that if $\widetilde{L}_{ij}$ does not depend on $\tilde{x}_{js}$, then we must have $\tilde{\lambda}_s = 1$ $(s = 1, \ldots, r_2)$. ■