

Empirical Likelihood Covariate Adjustment for Regression Discontinuity Designs*

Jun Ma[†]

Zhengfei Yu[‡]

Abstract

This paper proposes a novel approach to incorporate covariates in regression discontinuity (RD) designs. We represent the covariate balance condition as over-identifying moment restrictions. The empirical likelihood (EL) RD estimator efficiently incorporates the information from covariate balance and thus has an asymptotic variance no larger than that of the standard estimator without covariates. It achieves efficiency gain under weak conditions. We resolve the indeterminacy raised by [Calonico, Cattaneo, Farrell, and Titiunik \(2019, Page 448\)](#) regarding the asymptotic efficiency gain from incorporating covariates to RD estimator, as their estimator has the same asymptotic variance as ours. We then propose a robust corrected EL (RCEL) confidence set which achieves the fast n^{-1} coverage error decay rate even though the point estimator converges at a nonparametric rate. In addition, the coverage accuracy of the RCEL confidence set is automatically robust against slight perturbation to the covariate balance condition, which may happen in cases such as data contamination and misspecified “unaffected” outcomes used as covariates. We also show a uniform-in-bandwidth Wilks theorem, which is useful in sensitivity analysis for the proposed RCEL confidence set in the sense of [Armstrong and Kolesár \(2018\)](#). We conduct Monte Carlo simulations to assess the finite-sample performance of our method and also apply it to a real dataset.

Keywords: Covariate adjustment, coverage error, empirical likelihood, local misspecification, regression discontinuity

JEL classification: C12, C14, C31, C36

This version: May 7, 2022

*We thank Matias Cattaneo, Vadim Marmer and Taisuke Otsu for their helpful comments. All errors are ours. We acknowledge the financial support from the National Natural Science Foundation of China under grant 71903190 (Ma) and JSPS KAKENHI under Grant Number 21K01419 (Yu).

[†]School of Economics, Renmin University of China

[‡]Faculty of Humanities and Social Sciences, University of Tsukuba

1 Introduction

The RD design resembles a randomized experiment conducted near the cut-off of the score (forcing variable) and exploits the discontinuous variation in the probability of treatment to nonparametrically identify the LATE at the cut-off under mild continuity assumptions on the latent variables.¹ The transparent close-form identification (Hahn et al., 2001) of the RD LATE calls for nonparametric estimation and inference methods as they avoid functional form assumptions. See Cattaneo et al. (2019) for a recent review of RD. In practical implementations, information from pre-treatment covariates (i.e., variables that have already been determined before the assignment of the treatment) is incorporated to enhance efficiency and compensate for low accuracy of nonparametric methods. A widely-used procedure is augmented local polynomial (LP) regression where the covariates enter linearly. Calonico et al. (2019, CCFT, hereafter) formalizes this augmented regression approach and derives its (first-order) asymptotic properties. CCFT shows that augmented LP regression estimator consistently estimates the RD local average treatment effect (LATE) under the covariate balance condition, i.e., the expectations of covariates coincide at both sides of the cut-off. Apart from CCFT, covariate adjustment for RD receives much attention in recent literature. See Frölich and Huber (2019) for an alternative approach which requires smoothing over covariates but allows for potential failure of covariate balance. Arai et al. (2021) and Kreiß and Rothe (2021) extend CCFT’s approach to control for a high-dimensional covariate vector by regularization. Noack et al. (2021) extends CCFT’s linear regression adjustment to nonparametric adjustment with machine learning methods. See Cattaneo et al. (2021) for a recent review of covariate adjustment for RD.

This paper studies a novel approach to incorporate covariates in a generalized method of moments (GMM) framework with local smoothing. We formulate the close-form identification of (sharp or fuzzy) RD treatment effect as LP moment conditions. Then covariate balance is characterized by a set of over-identifying LP moment conditions and used as “side information”. The LP moment conditions are derived from a population-level minimum contrast problem (see Bickel and Doksum, 2015, Chapter 11.3 and Jiang and Doksum, 2003). CCFT treats covariate balance as a maintained assumption and our approach is not more restrictive in this regard. Our framework naturally calls for (efficient) GMM estimation. EL and generalized EL (Newey and Smith, 2004) are popular alternatives to GMM which do not require first-step estimation of the efficient weighting matrix.² We show in Theorem 1 and Remark 1 that the asymptotic variance of the EL RD estimator with the covariate-balancing-induced over-identifying moment conditions included is no larger than

¹In a recent study, Hyytinen et al. (2018) confirmed that RD produces estimates that are in line with the results from a comparable experiment if inference is implemented with the method of Calonico et al. (2014).

²See, e.g., Kitamura (2006) for a comprehensive review of EL and generalized EL. See, e.g., Chen and Qin (2000); Otsu et al. (2013, 2015); Ma et al. (2019) for EL inference in the context of non-parametric curves. It was shown that EL has favorable properties relative to GMM. See, e.g., Chen and Cui (2007); Kitamura (2001); Matsushita and Otsu (2013); Newey and Smith (2004); Otsu (2010); Ma (2017) among many others.

the standard LP regression RD estimator without covariates. In addition, we show that the EL estimator is first-order equivalent to the regression estimator of CCFT. The first contribution of this paper is that we provide new insights into CCFT’s method and resolve the indeterminacy raised on Page 448 of CCFT regarding asymptotic efficiency. We show that CCFT’s estimator weakly dominates the standard LP regression estimator without covariates and achieves efficiency gain as long as the true projection coefficients of some covariates are nonzero. See Remarks 1, 2 and 3. We explain such an asymptotic efficiency ranking from the perspective of local randomization and provide a GMM interpretation of CCFT’s estimator: its potential efficiency gain can be attributed to efficient inclusion of covariate balance as side information. Our result also provides a simple characterization of “irrelevant” covariates (Remark 3).

Then we show that inference using the EL ratio has several favorable theoretical advantages. EL inference does not require calculation of standard errors and explicit studentization. Theorem 2 shows a new uniform-in-bandwidth extension of the standard Wilks theorem (i.e., the EL ratio statistic is asymptotically χ^2). Our uniform-in-bandwidth version adjusts for specification search over multiple bandwidths known as bandwidth snooping (Armstrong and Kolesár, 2018, AK, hereafter) and takes into account the effects from data-dependent bandwidths in a robust manner (Remarks 4 and 6). It also provides a powerful tool for sensitivity analysis in the sense of AK (Remark 5). By deriving distributional expansions, Calonico et al. (2018) shows that proper studentization for Wald-type inference is crucial for having desirable coverage properties. We follow the same approach and show that implicit studentization of EL achieves favorable coverage properties. Theorem 3 characterizes the leading coverage error (i.e., the discrepancy between the nominal and finite-sample coverage probabilities, see, e.g., Calonico et al. (2020)) term in the distributional expansion of the EL ratio statistic. The coverage expansion for the EL confidence set for the RD LATE is strikingly as simple as the asymptotic mean square error (AMSE) for the point estimator. The coverage optimal bandwidth, which is defined as the minimizer of this leading coverage error in the spirit of Calonico et al. (2020), has a simple close form (Remark 9). Theorem 3 constructs a simple data-driven robust corrected EL (RCEL) confidence set with favorable robustness properties, which is the second contribution of this paper. This method does not require resampling and is thus computationally inexpensive. It complements the Wald-type inference method of CCFT and addresses common concerns in empirical applications. In particular, Theorem 3 shows that the EL confidence set admits partial Bartlett correction (Chen, 1996), i.e., rescaling the EL ratio for improving the coverage accuracy, which can also be combined with internalized bias removal (Calonico et al., 2014).³ In the context of EL inference on nonparametric curves, partial Bartlett correctability is a

³Compared with the expressions of correction factors for EL in other contexts (e.g., Chen and Cui, 2007; Matsushita and Otsu, 2013; Ma, 2017), our correction factor is very simple and thus can be estimated with good accuracy in finite samples, thanks to a special property of the moment conditions under consideration (i.e., asymptotic uncorrelatedness between the conditions and their derivatives).

stronger property than Bartlett correctability in the conventional sense (Remarks 10 and 11). We show that our RCEL confidence set achieves a coverage error decay rate of n^{-1} ($n \in \mathbb{N}$ denotes the sample size) under minimal smoothness assumptions and also the covariate balance condition. Note that n^{-1} is the coverage error decay rate of standard two-sided confidence intervals for parameters that can be estimated at the $n^{-1/2}$ parametric rate (see, e.g., Hall, 1991). The RCEL confidence set achieves the same rate even though the EL point estimator converges at a slower nonparametric rate. Therefore, our method is particularly useful when the researcher is faced with a small sample but the researcher wishes to have the coverage error under control.

Theorem 4 considers deviation from covariate balance and shows that the coverage accuracy of the RCEL confidence set for our parameter of interest is highly insensitive to mild deviation (Remark 15), which we refer to as local imbalance in this paper. Failure of the covariate balance assumption may happen in at least two realistic situations. In real applications (see, e.g., Cattaneo et al., 2019), the researcher may have access to observations on outcomes which are determined after treatment but are considered unaffected by the treatment (“unaffected” outcomes). Covariate balance should also hold for “unaffected” outcomes. However, specification of the “unaffected” outcomes is based on prior knowledge or empirical evidence, which might be mistaken. Another concern is that the balance condition holds for pre-treatment covariates in theory but our sample observations on these covariates are contaminated (possibly due to measurement errors that occur after treatment) so that they are drawn from a perturbed population (Kitamura et al., 2013) that slightly violates the balance condition. When covariate balance does not hold exactly, the coverage accuracy of our RCEL confidence set stays relatively unaffected, while other inference methods may exhibit severe undercoverage (Remark 16). To the best of our knowledge, these robustness properties are novel in the literature. We are unaware of any other inference method that has similar properties. These properties result from the intrinsic second-order properties of EL and the fact that the LP moment conditions for RD are asymptotically uncorrelated with their derivatives. Combination of RCEL and AK-type correction is straightforward (Remark 14) and provides a more accurate uniform confidence band that is useful for sensitivity analysis and robust inference.

In relation to the literature, Otsu et al. (2015) proposed EL inference for RD without covariates. Their method was based on first-order conditions from standard local linear regression. This paper focuses on covariate adjustment and uses different moment conditions. In another related paper, Ma et al. (2019) studied EL inference for the parameter of interest in the density discontinuity design (Jales and Yu, 2016). The scope of this paper is different from Ma et al. (2019) but the LP moment conditions in both papers are from population-level LP fitting (minimum contrast problem) in Bickel and Doksum (2015). Our paper uses

a similar approach to covariate adjustment as [Wu and Ying \(2011\)](#); [Zhang \(2018\)](#) who formulated covariate balance in randomized experiments as moment conditions and proposed EL-type methods. We formulate local imbalance and study the impact of it on the coverage accuracy by using standard local asymptotic analysis (e.g., the Pitman approach to local power analysis). Local imbalance can be also viewed as a special case of local misspecification of the moment conditions in the GMM framework (see, e.g., [Armstrong and Kolesár, 2021](#) and references therein). But the approach we take differs from those employed by papers in this strand of literature. Our approach follows [Bravo \(2003\)](#) and is based on second-order asymptotic expansion of the coverage probability under drifting alternative hypotheses (i.e., local imbalance). Lastly we note that our approach is potentially more flexible than the augmented regression approach of CCFT. In the literature, nonlinear estimators are proposed for RD with limited outcome variables (e.g., [Xu, 2017, 2018](#)). To the best of our knowledge, covariate adjustment to nonlinear estimation for RD has not been studied. Extension of CCFT’s linear regression approach in these contexts is not straightforward and the desired properties (consistency and potential efficiency gain) may no longer hold. Incorporating covariates by the EL probabilities (see, e.g., [Brown and Newey, 2002](#)) derived in this paper seems a simple solution and is able to deliver efficiency gain under covariate balance. Such extensions are beyond the scope of this paper.

Section 2 quickly reviews the RD design. Section 3 introduces our EL method for RD with covariates. Section 4 provides results on its first-order asymptotic properties. Section 5 is devoted to second-order properties. Sections 6 and 7 present results from simulation and empirical exercises. Section 8 concludes. Proofs of the theorems are collected in the appendix. Proofs of the lemmas are relegated to a supplement available at ruc-econ.github.io/supplement_RD.pdf.

2 Regression discontinuity designs

Let $X \in \mathbb{R}$ be a continuous score supported on $[\underline{x}, \bar{x}]$. Let f_X denote its density function. We normalize the cutoff point to zero (so that $0 \in [\underline{x}, \bar{x}]$ without loss of generality) for notational brevity. For any k -times differentiable univariate function f , let $f^{(k)}$ denote the k -th order derivative. In this paper, “ $a := b$ ” means that a is defined by b and “ $a = b$ ” means that b is defined by a . Denote $\varphi := f_X(0)$ and $\varphi^{(k)} := f_X^{(k)}(0)$ for simplicity. For a random vector (or matrix) V , denote $g_V(x) := E[V \mid X = x]$ and $m_V(x) := g_V(x) f_X(x)$. Denote $\mu_{V,-}^{(k)} := \lim_{x \uparrow 0} g_V^{(k)}(x)$ and $\psi_{V,-}^{(k)} := \lim_{x \uparrow 0} m_V^{(k)}(x)$. $(\mu_{V,+}^{(k)}, \psi_{V,+}^{(k)})$ are defined similarly with $\lim_{x \uparrow 0}$ replaced by $\lim_{x \downarrow 0}$. For simplicity, also denote $\mu_{V,s} := \mu_{V,s}^{(0)}$, $\psi_{V,s} := \psi_{V,s}^{(0)}$ ($s \in \{-, +\}$) and $\mu_{V,\pm} := \mu_{V,+} + \mu_{V,-}$. Let a^\top denote the transpose of a . For random vectors (V, U) , denote $\Sigma_{VU^\top, s} = \mu_{VU^\top, s} - \mu_{V,s} \mu_{U^\top, s}$ ($s \in \{-, +\}$) and $\Sigma_{VU^\top, \pm} := \Sigma_{VU^\top, +} + \Sigma_{VU^\top, -}$.

Let $Y \in \mathbb{R}$ denote the outcome variable, $D \in \{0, 1\}$ be the binary treatment and $Z \in \mathbb{R}^{d_z}$ be pre-treatment covariates or “unaffected” outcomes. Variables in Z can be continuous, discrete or mixed. We observe (Y, D, Z) and the score X . Let $1(\cdot)$ denote the indicator function. In an RD model, incentive is assigned if $X \geq 0$. In a sharp RD case $D = I := 1(X \geq 0)$ (i.e., perfect compliance). In the electoral RD model (see Lee, 2008; Hyytinen et al., 2018), (X, D, Y) correspond to the vote share margin in the last election, results of the last election (win or lose) and this election. Researchers almost always have access to some pre-treatment covariates. In the electoral RD case, commonly observed covariates such as candidates’ age, gender and the incumbency status are determined prior to the election considered. The more general fuzzy RD model assumes $D \neq I$ but g_D has a jump discontinuity at $x = 0$ ($\mu_{D,+} \neq \mu_{D,-}$) due to the incentive. This is known as limited compliance in the literature.

The RD model can be embedded in the potential outcome and treatment framework. Let $(Y(1), Y(0))$ be potential outcomes with or without treatment. Let (D_+, D_-) denote the potential treatments with or without incentives. The observed outcome Y and treatment D are determined by $Y = DY(1) + (1 - D)Y(0)$ and $D = ID_+ + (1 - I)D_-$ respectively. The complier group is defined to be individuals with $D_+ > D_-$ (i.e., $(D_+, D_-) = (1, 0)$).⁴ Following CCFT, we let $(Z(1), Z(0))$ denote potential covariates and then $Z = DZ(1) + (1 - D)Z(0)$. Let $\mathfrak{V}(k) := (Y(k), Z(k))$, $\forall k \in \{0, 1\}$. Denote $g_{dd'}(x) := \Pr[D_+ = d, D_- = d' | X = x]$ and $g_{\mathfrak{V}(k)|dd'}(x) := E[\mathfrak{V}(k) | D_+ = d, D_- = d', X = x]$. Similarly, let $\mathfrak{V} := (Y, Z)$ and $g_{\mathfrak{V}|dd'}(x) := E[\mathfrak{V} | D_+ = d, D_- = d', X = x]$. By the law of iterated expectations (LIE), $g_{\mathfrak{V}} = \sum_{(d,d') \in \{0,1\}^2} g_{dd'} g_{\mathfrak{V}|dd'}$. Let $\mathcal{T}_Y := E[Y(1) - Y(0) | X = 0, D_+ > D_-]$ be the RD LATE (the average treatment effect for individuals with zero score in the complier group) and similarly, $\mathcal{T}_Z := E[Z(1) - Z(0) | X = 0, D_+ > D_-]$ denotes the RD LATE on Z . The following assumption is implicit in CCFT.

Assumption 1. (a) $g_{\mathfrak{V}(k)|dd'}$ and $g_{dd'}$ are continuous at 0, $\forall (k, d, d') \in \{0, 1\}^3$; (b) $\Pr[D_- \leq D_+ | X = 0] = 1$; (c) $\Pr[D_+ > D_- | X = 0] \neq 0$; (d) $\mathcal{T}_Z = 0$.

The RD model imposes only a few weak identifying assumptions. In Assumption 1, (a), (b) and (c) are local versions of the LATE assumptions: (a) and (b) impose local continuity and monotonicity assumptions respectively and (c) imposes existence of the local complier group. These are key identifying assumptions for the RD model (see Dong, 2018). Under (a), $(\mu_{Y,+}, \mu_{D,+}, \mu_{Z,+})$ and $(\mu_{Y,-}, \mu_{D,-}, \mu_{Z,-})$ exist. (c) implies that $\mu_{D,+} > \mu_{D,-}$. These assumptions have testable implications (Arai et al., 2021). It can be shown that under these assumptions, \mathcal{T}_Y is nonparametrically identified: $\mathcal{T}_Y = \vartheta_0 := (\mu_{Y,+} - \mu_{Y,-}) / (\mu_{D,+} - \mu_{D,-})$

⁴RD can be represented by a triangular model. See Dong (2018). (Y, D) are assumed to be generated by a triangular model $Y = g(D, X, Z, \epsilon)$ and $D = 1(X \geq 0)h_+(X, Z, \eta) + 1(X < 0)h_-(X, Z, \eta)$, where (g, h_+, h_-) are unknown functions and (ϵ, η) are (potentially correlated) unobserved disturbances of unrestricted dimensionality. Then the potential outcomes and treatments are given by $Y(1) = g(1, X, Z, \epsilon)$, $Y(0) = g(0, X, Z, \epsilon)$, $D_+ = h_+(X, Z, \eta)$ and $D_- = h_-(X, Z, \eta)$.

(see [Hahn et al., 2001](#); [Dong, 2018](#)), where ϑ_0 is an observable population feature.⁵ Similarly, under (a), (b) and (c), $\mathcal{T}_Z = (\mu_{Z,+} - \mu_{Z,-}) / (\mu_{D,+} - \mu_{D,-})$. Following CCFT, we impose (d), which means that there is no RD treatment effect on Z . If Z includes only pre-treatment variables, this assumption holds by definition. Under (a), (b) and (c), (d) is equivalent to the covariate balance condition $\mu_{Z,+} = \mu_{Z,-}$, which is a testable restriction on the population of the observed variables. Indeed, it is the null hypothesis of a popular falsification or placebo test for the RD model.⁶ See, e.g., [Lee \(2008\)](#); [Canay and Kamat \(2017\)](#). $\mu_{Z,+} = \mu_{Z,-}$ is satisfied if the conditional distribution of Z given $X = x$ is continuous at $x = 0$. Evidence against $\mu_{Z,+} = \mu_{Z,-}$ in the data (so that a hypothesis test of $\mu_{Z,+} = \mu_{Z,-}$ is rejected) casts doubts on the validity of the continuity assumption (a). We can also augment the list of potential covariates to include outcomes that are determined after the assignment but unaffected by the treatment. “Unaffected” outcomes can be found in many applications. See, e.g., [Cattaneo and Titiunik \(2022, Section 4.1\)](#) for discussion. Unlike pre-treatment variables, the assumption that the “unaffected” outcomes satisfy (d) is based on our prior knowledge or evidence.

3 Covariate balance as moment restrictions

This section introduces a GMM framework that formulates the RD estimand and the covariate balance condition as a set of over-identifying moment restrictions. First, we show that the RD estimand ϑ_0 , which has causal interpretation under the identifying assumptions of the RD model, can be approximately identified by two just-identified LP moment conditions. Let K denote the kernel function and let h denote the bandwidth. Denote $K_h(t) := h^{-1}K(t/h)$. Let $M := Y - \vartheta_0 D$ and note that

$$\lim_{x \downarrow 0} \mathbb{E}[Y - \theta_0 D \mid X = x] = \lim_{x \uparrow 0} \mathbb{E}[Y - \theta_0 D \mid X = x] \text{ if and only if } \theta_0 = \vartheta_0.$$

Denote $\vartheta_1 := \mu_{M,+} = \mu_{M,-}$. Let $p \geq 1$ be the integer-valued LP order. Denote $r_p(t) := (1, t, \dots, t^p)^\top$. According to [Jiang and Doksum \(2003\)](#), p -th order LP approximation of $(\psi_{M,-}, \psi_{M,+})$ can be derived from solving the following minimum contrast problem. Let $e_{k,s}$ denote the s -th unit vector in \mathbb{R}^k . Let

$$\psi_- := e_{p+1,1}^\top \operatorname{argmin}_{z \in \mathbb{R}^{p+1}} \int_{\underline{x}}^0 \{m_M(x) - z^\top r_p(x)\}^2 K_h(x) dx \quad (1)$$

⁵In the sharp RD model ($\mu_{D,+} = 1$ and $\mu_{D,-} = 0$ in this case) or under a stronger conditional independence assumption ([Hahn et al., 2001](#)), a causal parameter that corresponds to a broader subpopulation (conditional average treatment effect) is identified by the same ratio: $\mathbb{E}[Y(1) - Y(0) \mid X = 0] = \vartheta_0$.

⁶While most empirical works conduct the balance test separately for each covariate, some researchers have noted that the problem of multiple testing may generate statistical imbalance of some covariates by chance. See, e.g., [Hyytinen et al. \(2018\)](#). In a separate paper, we propose a joint EL test for the smoothness of multiple covariates at the cut-off.

and ψ_+ be defined by the minimizer on the right hand side with the integral range $[x, 0]$ replaced by $[0, \bar{x}]$. Denote $V_{p;-} := \int_{-1}^0 r_p(t) r_p(t)^\top K(t) dt$ and $\mathcal{K}_{p;-}(t) := e_{p+1,1}^\top V_{p;-}^{-1} r_p(t) K(t)$. Let $(V_{p;+}, \mathcal{K}_{p;+})$ be defined by the same equations with the integral range $[-1, 0]$ replaced by $[0, 1]$.⁷ Let the data $\{(Y_i, D_i, X_i, Z_i)\}_{i=1}^n$ be i.i.d. copies of (Y, D, X, Z) . Then, let $W_{p;-i} := 1(X_i < 0) \mathcal{K}_{p;-}(X_i/h)$, $W_{p;+i} := 1(X_i > 0) \mathcal{K}_{p;+}(X_i/h)$ and $W_{p,i} := (W_{p;+i}, W_{p;-i})^\top$. $(W_{p;-}, W_{p;+}, W_p)$ are defined by the same formulae with X_i replaced by X . Then, by solving the first-order conditions of (1), we have $\psi_s = E[h^{-1} W_{p;s} g_M(X)] = E[h^{-1} W_{p;s} M]$. By Taylor expansion (see Jiang and Doksum, 2003), $\psi_s = \psi_{M,s} + O(h^{p+1})$ under suitable smoothness assumptions imposed on g_M . From (1) with $m_M(x)$ replaced by $f_X(x)$, we have $E[h^{-1} W_{p;s}] = \varphi + O(h^{p+1})$, $\forall s \in \{-, +\}$. Therefore,

$$E[W_{p;s}(Y - \vartheta_0 D - \vartheta_1)] = O(h^{p+2}), \forall s \in \{-, +\}, \quad (2)$$

which are the two LP moment conditions that (approximately) identify $(\vartheta_0, \vartheta_1)$.

Next, we incorporate the information from the covariates by directly formulating the covariate balance condition as over-identifying moment restrictions. This differs from CCFT where covariates are included as additional regressors in the LP regression (see (6)). Let $\vartheta_2 := \mu_{Z,+} = \mu_{Z,-}$ denote the common value. By solving (1) with m_M replaced by m_Z and f_X , we have

$$E[W_{p;+}(Z - \vartheta_2)] = O(h^{p+2}) \text{ and } E[W_{p;-}(Z - \vartheta_2)] = O(h^{p+2}). \quad (3)$$

We restrict the bandwidths on the left and the right of the cut-off to be the same. It is possible to extend all of the theorems in this paper to accommodate different bandwidths on different sides. Now combining restrictions (2) and (3) we have the following over-identified LP moment conditions:

$$E \left[W_p \otimes \begin{pmatrix} Y - \vartheta_0 D - \vartheta_1 \\ Z - \vartheta_2 \end{pmatrix} \right] = O(h^{p+2}), \quad (4)$$

where \otimes denotes the Kronecker product. Note that we have $2(1 + d_z)$ LP moment conditions that approximately identify $2 + d_z$ parameters. $\vartheta_0 = \mathcal{T}_Y$ is the parameter of interest and $(\vartheta_1, \vartheta_2)$ are nuisance parameters. Denote $\vartheta := (\vartheta_0, \vartheta_1, \vartheta_2) \in \mathbb{R}^{d_\vartheta}$ ($d_\vartheta := 2 + d_z$), $\vartheta_\dagger := (\vartheta_1, \vartheta_2) \in \mathbb{R}^{d_\dagger}$ ($d_\dagger := 1 + d_z$), $\theta := (\theta_0, \theta_1, \theta_2)$ and $\theta_\dagger := (\theta_1, \theta_2)$.

We define the EL criterion function:

$$\ell_p(\theta | h) := \min_{w_1, \dots, w_n} -2 \sum_i \log(n \cdot w_i)$$

⁷ $(\mathcal{K}_{p;+}, \mathcal{K}_{p;-})$ coincide with the “equivalent kernel” of LP regression. See, e.g., Section S2.1 of AK.

$$\text{subject to } \sum_i w_i W_{p,i} \otimes \begin{pmatrix} Y_i - \theta_0 D_i - \theta_1 \\ Z_i - \theta_2 \end{pmatrix} = 0, \sum_i w_i = 1 \text{ and } w_i \geq 0, \forall i, \quad (5)$$

where \sum_i is understood as $\sum_{i=1}^n$. $-\sum_i \log(n \cdot w_i)/n$ is the Kullback-Leibler divergence from (w_1, \dots, w_n) to the uniform weights $1/n$. Denote $U_i(\theta) := (Y_i - \theta_0 D_i - \theta_1, Z_i^\top - \theta_2^\top)^\top$, $U_i := U_i(\vartheta)$ for notational simplicity and $d_u := 1 + d_z$. $U(\theta)$ and U are defined by the same formulae with (Y_i, D_i, X_i, Z_i) replaced by (Y, D, X, Z) . The p -th order EL estimator is given by $\hat{\vartheta}_p := (\hat{\vartheta}_{p,0}, \hat{\vartheta}_{p,1}, \hat{\vartheta}_{p,2}) := \operatorname{argmin}_{\theta \in \Theta} \ell_p(\theta | h)$, where $\Theta \subseteq \mathbb{R}^{d_\vartheta}$ is a compact parameter space such that ϑ is an interior point of Θ . Also denote the constrained EL estimator: $\tilde{\vartheta}_p(\theta_0) := (\tilde{\vartheta}_{p,1}(\theta_0), \tilde{\vartheta}_{p,2}(\theta_0)) := \operatorname{argmin}_{\theta \in \Theta_\dagger} \ell_p(\theta_0, \theta | h)$, where $\Theta_\dagger \subseteq \mathbb{R}^{d_\dagger}$ is a compact constrained parameter space such that ϑ_\dagger is in the interior of Θ_\dagger and θ_0 is some hypothesized value. The EL ratio statistic is given by $LR_p(\theta_0 | h) := \ell_p(\theta_0, \tilde{\vartheta}_p(\theta_0) | h) - \ell_p(\hat{\vartheta}_p | h)$, which is a function of θ_0 . It is shown in the proof of Theorem 3 that $\ell_p(\hat{\theta}_p | h) = \inf_{\theta_2} \sup_{\lambda} 2 \sum_i \log(1 + \lambda^\top W_{p,i} \otimes (Z_i - \theta_2))$ and therefore it suffices to solve a simpler optimization problem. Let $\tau \in (0, 1)$ be the significance level. Let $F_{\chi_1^2}$ and $f_{\chi_1^2}$ denote the cumulative distribution function (CDF) and probability density function (PDF) of a χ_1^2 (χ^2 with one degree of freedom) random variable respectively. Let $c_\tau := F_{\chi_1^2}^{-1}(1 - \tau)$ be the $(1 - \tau)$ quantile of the χ_1^2 distribution. An EL confidence set for ϑ_0 with nominal coverage probability $1 - \tau$ is $CS_\tau(h) := \{\theta_0 : LR_p(\theta_0 | h) \leq c_\tau\}$. For fuzzy RD, as Noack and Rothe (2019)'s method, the EL confidence set avoids a “delta method” argument used by the Wald-type inference of CCFT. The EL probabilities (weights) $\hat{w}_1, \dots, \hat{w}_n$ are those corresponding to the minimizer of the problem (5) with $\theta = \hat{\vartheta}_p$ (Brown and Newey, 2002). These EL probabilities can be used for covariate adjustment in nonlinear estimation associated with RD (e.g., Xu, 2017, 2018 among others), for which extension of CCFT's approach is involved.⁸

4 Efficiency gain and uniform-in-bandwidth Wilks theorem

This section provides asymptotic properties of the EL estimator and EL ratio statistic. Theorem 1 shows asymptotic normality and gives the expression for the AMSE. We then compare it with the asymptotic result from CCFT. Theorem 2 provides uniform-in-bandwidth large sample approximation to the distribution of the EL ratio with $\theta_0 = \vartheta_0$. For a vector z , let $z^{(j)}$ denote its j -th coordinate. Similarly, $A^{(jk)}$ denotes the jk -th element of a matrix A . By abuse of notation, for a d_v -dimensional random

⁸It is shown in the proof of Theorem 3 that $n\hat{w}_i = \left(1 + \hat{\lambda}_p^\top W_{p,i} \otimes (Z_i - \hat{\vartheta}_{p,2})\right)^{-1}$ where $\hat{\lambda}_p$ solves $\sum_i W_{p,i} \otimes (Z_i - \hat{\vartheta}_{p,2}) / \left(1 + \hat{\lambda}_p^\top W_{p,i} \otimes (Z_i - \hat{\vartheta}_{p,2})\right) = 0$. Indeed, it is clear that the EL estimator $(\hat{\vartheta}_{p,0}, \hat{\vartheta}_{p,1})$ is numerically equivalent to the plug-in (method of moments) estimator that solves the sample analogue of the moment conditions (2) with the empirical distribution replaced by the EL probabilities: $\sum_i \hat{w}_i W_{p,i} \otimes (Y_i - \hat{\vartheta}_{p,0} D_i - \hat{\vartheta}_{p,1}) = 0$. We conjecture that the same reweighting adjustment can be extended to the nonlinear cases considered by Xu (2017, 2018).

vector V , let V^2 denote $V \otimes V$ with duplicated coordinates removed, i.e., $V^2 := \text{vech}(VV^\top)$, where $\text{vech}(A)$ denotes the half vectorization of a matrix A . Similarly, V^3 denotes the vector consisting of $V^{(1)}\text{vech}(VV^\top), V^{(2)}\text{vech}\left(\left(V^{(2)}, \dots, V^{(d_v)}\right)^\top \left(V^{(2)}, \dots, V^{(d_v)}\right)\right), \dots, \left(V^{(d_v)}\right)^3$. $\|x\|$ denotes the Euclidean norm of the vector x . We assume the following assumptions hold.

Assumption 2. (a) On a neighborhood around 0, $g_{\mathfrak{Y}(k)|dd'}$ and $g_{dd'}$ are $(p+1)$ -times continuously differentiable with Lipschitz continuous $(p+1)$ -th order derivatives, $g_{\mathfrak{Y}(k)^2|dd'}$ is Lipschitz continuous and $g_{\|\mathfrak{Y}(k)\|^{12}|dd'}$ is bounded and f_X is $(p+1)$ -times continuously differentiable with Lipschitz continuous $(p+1)$ -th order derivative; (b) $E\left[U(k)U(k)^\top \mid D_+ > D_-, X = 0\right]$ is positive definite, $\forall k \in \{0, 1\}$.

Assumption 3. (a) K is a symmetric continuous PDF supported on $[-1, 1]$; (b) $\mathcal{K}_{p;+}$ is differentiable with bounded first-order derivatives on $(-1, 0)$ and $(0, 1)$.

Assumption 2 is imposed on the latent variables in the RD model and parallels Assumption SA-5 of CCFT. Since $g_{\mathfrak{Y}} = \sum_{(d,d') \in \{0,1\}^2} g_{dd'} g_{\mathfrak{Y}|dd'}$, $\mathfrak{Y} = D_+ \mathfrak{Y}(1) + (1 - D_+) \mathfrak{Y}(0)$ if $X > 0$ and $\mathfrak{Y} = D_- \mathfrak{Y}(1) + (1 - D_-) \mathfrak{Y}(0)$ if $X < 0$, Assumption 2(a) guarantees that $g_{\mathfrak{Y}}$ and $m_{\mathfrak{Y}}$ have continuous derivatives up to $(p+1)$ -th order on the left and right neighborhoods of 0. Similarly, $g_{\mathfrak{Y}^2}$ is continuous and $g_{\|\mathfrak{Y}\|^{12}}$ is bounded on the left and right neighborhoods of 0, under Assumption 2(a). Denote $\gamma_{\text{adj}} := \Sigma_{ZZ^\top, \pm}^{-1} \Sigma_{ZM, \pm}$, $\epsilon := M - Z^\top \gamma_{\text{adj}}$ and $\sigma_{\text{adj}}^2 := \Sigma_{M^2, \pm} - \Sigma_{MZ^\top, \pm} \gamma_{\text{adj}}$. Existence of these quantities is guaranteed by Assumption 2(a). Under Assumption 1, $\mu_{\epsilon, +} = \mu_{\epsilon, -} =: \mu_\epsilon$. Assumption 2(a) guarantees that g_ϵ and m_ϵ admit continuous derivatives up to $(p+1)$ -th order on the left and right neighborhoods of 0 so that the leading bias terms can be characterized. Denote $\zeta_{p;s} := \omega_{p;s}^{p+1,1} \left(\psi_{\epsilon,s}^{(p+1)} - \mu_\epsilon \varphi^{(p+1)} \right)$, $s \in \{-, +\}$, where $\omega_{p;+}^{j,k} := \int_0^1 t^j \mathcal{K}_{p;+}(t)^k dt$ and $\omega_{p;-}^{j,k} := \int_{-1}^0 t^j \mathcal{K}_{p;-}(t)^k dt$. Assumption 2(b) guarantees that $\mu_{UU^\top, +}$ and $\mu_{UU^\top, -}$ exist and are both positive definite. Assumption 3(a) is standard and also imposed in CCFT. Assumption 3(b) is also found in AK. Assumption 3(a) implies that $\mathcal{K}_{p;+}(t) = \mathcal{K}_{p;-}(-t) \forall t \in \mathbb{R}$ and therefore (b) also holds for $\mathcal{K}_{p;-}$. The following result shows the asymptotic distribution of $\hat{\vartheta}_{p,0}$, the EL estimator of the RD LATE.

Theorem 1. Suppose that Assumptions 1, 2 and 3 hold. Assume that the bandwidth satisfies $nh^{2p+3} = O(1)$ and $\log(n)^2 / (nh) = o(1)$. Then, $\sqrt{nh} \left(\hat{\vartheta}_{p,0} - \vartheta_0 - \mathcal{B}_p^{\text{EL}} h^{p+1} \right) \rightarrow_d N(0, \mathcal{V}_p^{\text{EL}})$, where

$$\mathcal{B}_p^{\text{EL}} := \frac{\zeta_{p;+} - \zeta_{p;-}}{\varphi(\mu_{D,+} - \mu_{D,-})(p+1)!} \text{ and } \mathcal{V}_p^{\text{EL}} := \frac{\omega_{p;+}^{0,2} \sigma_{\text{adj}}^2}{\varphi(\mu_{D,+} - \mu_{D,-})^2}.$$

Remark 1. We consider the special case of sharp RD and $p = 1$. The local linear estimator $\hat{\vartheta}_0^{\text{LL}}$ can be obtained from a single localized regression. CCFT's approach augments the regression to incorporate pre-treatment covariates. CCFT's covariate adjusted estimator $\hat{\vartheta}_0^{\text{CCFT}}$ is given by the regression coefficient of

$I_i := 1 (X_i \geq 0)$ in

$$\hat{\vartheta}_0^{\text{CCFT}} := e_{4+d_z,3}^\top \underset{(a_0, b_0, a_1, b_1, d) \in \mathbb{R}^{4+d_z}}{\operatorname{argmin}} \sum_i K_h(X_i) \{Y_i - a_0 - b_0 X_i - a_1 I_i - b_1 X_i I_i - Z_i^\top d\}^2. \quad (6)$$

The covariates enter linearly and kernel smoothing over the covariates is not needed. $\hat{\vartheta}_0^{\text{CCFT}}$ converges in probability to the sharp RD estimand, under the covariate balance assumption. $\hat{\vartheta}_0^{\text{LL}} - \vartheta_0$ is approximately $N(0, \mathcal{V}^{\text{LL}}/(nh))$ distributed and $\hat{\vartheta}_0^{\text{CCFT}} - \vartheta_0$ is approximately $N(0, \mathcal{V}^{\text{CCFT}}/(nh))$, under the undersmoothing assumption $nh^5 = o(1)$. $\text{Var}_{|0}$ and $\text{Cov}_{|0}$ are understood as $\text{Var}[\cdot | X=0]$ and $\text{Cov}[\cdot | X=0]$ and \sum_k is understood as $\sum_{k \in \{0,1\}}$. We compare the asymptotic variance $\mathcal{V}_1^{\text{EL}} = \omega_{1,+}^{0,2} \sigma_{\text{adj}}^2 / \varphi$ with $\mathcal{V}^{\text{LL}} = \omega_{1,+}^{0,2} \sigma_{\text{LL}}^2 / \varphi$ and $\mathcal{V}^{\text{CCFT}} = \omega_{1,+}^{0,2} \sigma_{\text{CCFT}}^2 / \varphi$, where $\sigma_{\text{LL}}^2 := \sum_k \text{Var}_{|0}[Y(k)]$ (see, e.g., [Imbens and Kalyanaraman, 2011](#)) and $\sigma_{\text{CCFT}}^2 := \sum_k \text{Var}_{|0}[Y(k) - Z(k)^\top \gamma_{\text{CCFT}}]$ with $\gamma_{\text{CCFT}} := (\sum_k \text{Var}_{|0}[Z(k)])^{-1} (\sum_k \text{Cov}_{|0}[Z(k), Y(k)])$.⁹ It is easy to check that in the definition σ_{adj}^2 and γ_{adj} , $\Sigma_{M^2, \pm} = \sigma_{\text{LL}}^2$, $\Sigma_{MZ^\top, \pm} = \sum_k \text{Cov}_{|0}[Y(k), Z(k)^\top]$ and $\Sigma_{ZZ^\top, \pm} = \sum_k \text{Var}_{|0}[Z(k)]$. To see $\sigma_{\text{adj}}^2 \leq \sigma_{\text{LL}}^2$, observe that by definition,

$$\begin{aligned} \sigma_{\text{adj}}^2 &:= \Sigma_{M^2, \pm} - \Sigma_{MZ^\top, \pm} \gamma_{\text{adj}} \\ &= \sigma_{\text{LL}}^2 - \left(\sum_k \text{Cov}_{|0}[Y(k), Z(k)^\top] \right) \left(\sum_k \text{Var}_{|0}[Z(k)] \right)^{-1} \left(\sum_k \text{Cov}_{|0}[Z(k), Y(k)] \right) \leq \sigma_{\text{LL}}^2. \end{aligned} \quad (7)$$

Next, we show that $\sigma_{\text{adj}}^2 = \sigma_{\text{CCFT}}^2$.¹⁰ Observe that $\gamma_{\text{adj}} = \gamma_{\text{CCFT}}$ and also

$$\begin{aligned} \text{Var}_{|0}[Z(0)^\top \gamma_{\text{adj}}] + \text{Var}_{|0}[Z(1)^\top \gamma_{\text{adj}}] &= \gamma_{\text{adj}}^\top (\text{Var}_{|0}[Z(0)] + \text{Var}_{|0}[Z(1)]) \gamma_{\text{adj}} \\ &= \text{Cov}_{|0}[Y(1), Z(1)^\top \gamma_{\text{adj}}] + \text{Cov}_{|0}[Y(0), Z(0)^\top \gamma_{\text{adj}}] = \Sigma_{MZ^\top, \pm} \gamma_{\text{adj}}. \end{aligned} \quad (8)$$

Therefore,

$$\begin{aligned} \sigma_{\text{adj}}^2 &= \text{Var}_{|0}[Y(0)] + \text{Var}_{|0}[Y(1)] - \text{Cov}_{|0}[Y(1), Z(1)^\top \gamma_{\text{adj}}] \\ &\quad - \text{Cov}_{|0}[Y(0), Z(0)^\top \gamma_{\text{adj}}] = \text{Var}_{|0}[Y(0) - Z(0)^\top \gamma_{\text{adj}}] + \text{Var}_{|0}[Y(1) - Z(1)^\top \gamma_{\text{adj}}] = \sigma_{\text{CCFT}}^2, \end{aligned}$$

where the second equality follows from

$$\text{Var}_{|0}[Y(k) - Z(k)^\top \gamma_{\text{adj}}] = \text{Var}_{|0}[Y(k)] + \text{Var}_{|0}[Z(k)^\top \gamma_{\text{adj}}] - 2 \cdot \text{Cov}_{|0}[Y(k), Z(k)^\top \gamma_{\text{adj}}]$$

⁹It is easy to see that the method of moments estimator of ϑ_0 based on (2) without using (3) has the same asymptotic variance \mathcal{V}^{LL} as the standard local linear estimator.

¹⁰Indeed, it can be shown that the EL and CCFT's estimators with undersmoothing (i.e., $nh^5 = o(1)$) are first-order equivalent in a stronger sense: $\hat{\vartheta}_0^{\text{CCFT}} - \hat{\vartheta}_{1,0}^{\text{EL}} = o_p((nh)^{-1/2})$.

and (8). The equivalence result $\mathcal{V}^{\text{CCFT}} = \mathcal{V}_1^{\text{EL}}$ can be generalized to the case of arbitrary p . The equivalence also holds for fuzzy RD with an arbitrary LP order p .

Remark 2. The conclusion $\sigma_{\text{adj}}^2 = \sigma_{\text{CCFT}}^2 \leq \sigma_{\text{LL}}^2$ implies that the asymptotic variance of both EL and CCFT estimators is smaller or equal to that of the standard local linear estimator without covariates. This asymptotic efficiency ranking is achieved without additional assumptions other than covariate balance. At the first glance, such an asymptotic efficiency ranking seems surprising given that CCFT (on their Page 448) finds no definite ranking between σ_{CCFT}^2 and σ_{LL}^2 and interprets the indeterminacy as “in perfect agreement with those in the literature on analysis of experiments,..., where it is also found that incorporating covariates in randomized controlled trials using linear regression leads to efficiency gains only under particular assumptions”. As the RD design is often viewed as local randomization, let us reconcile our finding and CCFT’s comment from the perspective of randomized experiments. In RD designs, continuity of the density of the score X implies that the shares of units with X being in small neighborhoods to the left and right of the cutoff are equal (Noack et al., 2021, Section 5.2). Therefore, the RD design is analogous to a randomized experiment with equal probabilities of being in treatment and control groups. In the literature of randomized experiments, Negi and Wooldridge (2014, Theorem 5.2(iv)) show that when the assignment probability is equal to 1/2, the pooled regression adjustment (see Negi and Wooldridge, 2014 for its definition), whose algorithm is analogous to that of the CCFT estimator, always leads to a smaller or equal asymptotic variance. The assignment probability assumption is automatically fulfilled in RD designs. Theorem 1 and the first-order equivalence between the EL and CCFT estimators explains the asymptotic efficiency ranking from a different perspective: CCFT’s estimator can be interpreted as being efficiently incorporating the side information from the covariate balance condition, which will typically reduce the asymptotic variance, and in the worst scenario, will yield the same asymptotic variance if the side information is irrelevant.¹¹

Remark 3. Theorem 1 also implies that including a covariate will not change the asymptotic variance if and only if the corresponding element in γ_{adj} is zero. Note that the (true) projection coefficients γ_{adj} are the probabilistic limits of the regression coefficients of Z_i in the “long” regression (6) including all covariates. Consider the partition $Z = (Z_1^\top, Z_2^\top)^\top$ of Z and let $\gamma_{\text{adj}}^\top = (\gamma_{\text{adj},1}^\top, \gamma_{\text{adj},2}^\top)^\top$ be the conformable partition of γ_{adj} such that the dimension of $\gamma_{\text{adj},k}^\top$ coincides with that of Z_k , $k = 1, 2$. Using Theorem 1 and the representation on the right hand side of the second equality of (7), then writing $\sum_k \text{Var}_{|0}[Z(k)]$ as a block matrix and inverting it, we can easily show that $\mathcal{V}_1^{\text{EL}}$ is equal to the asymptotic variance of the CCFT (EL) covariate adjusted estimator using only Z_1 if and only if $\gamma_{\text{adj},2} = 0$. In this case, Z_2 is irrelevant in the sense that

¹¹Such an argument is analogous to that of Hirano et al. (2003), which explains the puzzling phenomenon that the inverse probability weighting estimator using the nonparametrically estimated propensity score has a smaller asymptotic variance relative to that uses the true propensity score. Hirano et al. (2003) shows that the former is equivalent to an EL estimator that incorporates the side information from knowing the true propensity score efficiently.

dropping Z_2 has no first-order impact: it neither leads to efficiency loss nor changes the leading asymptotic bias. In conclusion, if we say that an estimator achieves efficiency gain when its asymptotic variance is smaller than that of the standard estimator without covariates, then both EL and CCFT estimators achieve efficiency gain as long as the coefficients of some covariates are nonzero.¹²

The following theorem establishes uniform-in-bandwidth validity of the EL confidence set. Let $\ell^\infty(\mathfrak{S})$ denote the space of all bounded functions $f : \mathfrak{S} \rightarrow \mathbb{R}$ endowed with the sup-norm $\|f\|_{\mathfrak{S}} := \sup_{s \in \mathfrak{S}} |f(s)|$. Let $\mathbb{H} := [\underline{h}, \bar{h}]$ be a compact bandwidth set where $\underline{h} = \underline{h}_n > 0$ and $\bar{h} = \bar{h}_n > 0$ ($\underline{h} < \bar{h}$) are bandwidths that depend on the sample size. The following theorem parallels the main result of AK and is a substantial extension of the standard Wilks theorem which states that $LR_p(\vartheta_0 | h) \rightarrow_d \chi_1^2$. Our result incorporates covariates and accommodates unbounded outcomes. The proof techniques we use differ from those employed by AK.

Theorem 2. *Suppose that Assumptions 1, 2 and 3 hold. Suppose that (\underline{h}, \bar{h}) satisfy $n\bar{h}^{-2p+3} = o(1)$ and $n^{1/12}/(n\underline{h})^{1/2} + (n\underline{h})^{-1/6} = o(\log(n)^{-3})$. There exists a zero-mean Gaussian process $\{G(s) : s \in [1, \bar{h}/\underline{h}]\}$ which is a tight random element in $\ell^\infty([1, \bar{h}/\underline{h}])$ with the covariance structure given by*

$$\mathbb{E}[G(s)G(t)] = \sqrt{\frac{s}{t}} \frac{\int_0^\infty \mathcal{K}_{p;+}(z) \mathcal{K}_{p;+}((s/t)z) dz}{\int_0^\infty \mathcal{K}_{p;+}(z)^2 dz}. \quad (9)$$

Then, $\Pr[LR_p(\vartheta_0 | h) \leq z_\tau (\bar{h}/\underline{h})^2, \forall h \in \mathbb{H}] \rightarrow 1 - \tau$, as $n \uparrow \infty$, where $z_\tau (\bar{h}/\underline{h})$ denotes the $1 - \tau$ quantile of $\|G\|_{[1, \bar{h}/\underline{h}]}$.

Remark 4. Theorem 2 generalizes the standard Wilks theorem with a single bandwidth. It implies that when $h = \underline{h} = \bar{h}$, $\Pr[LR_p(\vartheta_0 | h) \leq c_\tau] = \Pr[\vartheta_0 \in CS_\tau(h)] \rightarrow 1 - \tau$. I.e., with a single bandwidth, the EL confidence set is asymptotically valid. The standard EL confidence set $CS_\tau(h)$ may undercover if the bandwidth is selected after specification search over \mathbb{H} . As an example, suppose that $\hat{h} := \arg\max_{h \in \mathbb{H}} LR_p(0 | h)$ is selected to maximize the p -value for the two-sided hypothesis test of $\vartheta_0 = 0$. AK shows that $z_\tau (\bar{h}/\underline{h})^2 > c_\tau$ when $\bar{h}/\underline{h} > 1$ but $z_\tau (\bar{h}/\underline{h})$ grows at a logarithmic speed as $\bar{h}/\underline{h} \uparrow \infty$. It is clear from Theorem 2 that $\Pr[\vartheta_0 \in CS_\tau(\hat{h})] \rightarrow 1 - \tilde{\tau}$, where $\tilde{\tau} > \tau$ solves $z_{\tilde{\tau}} (\bar{h}/\underline{h})^2 = c_\tau$ if $\vartheta_0 = 0$ and the test of $\vartheta_0 = 0$ does not have asymptotically correct size. Theorem 2 justifies a simple correction for bandwidth snooping as AK by replacing the critical value c_τ used by $CS_\tau(h)$ with $z_\tau (\bar{h}/\underline{h})^2$. Let $CS_\tau^{\text{sc}}(h | \bar{h}/\underline{h}) := \{\theta_0 : LR_p(\theta_0 | h) \leq z_\tau (\bar{h}/\underline{h})^2\}$ be the snooping corrected confidence set. Then, $CS_\tau^{\text{sc}}(h | \bar{h}/\underline{h})$ has asymptotically correct coverage no matter how h is selected from \mathbb{H} , i.e., $\liminf_{n \uparrow \infty} \Pr[\vartheta_0 \in CS_\tau^{\text{sc}}(h | \bar{h}/\underline{h})] \geq 1 - \tau \forall h \in \mathbb{H}$. The critical value $z_\tau (\bar{h}/\underline{h})$

¹²Noack et al. (2021) shows that covariate adjustment for RD in an arbitrary way may not lead to efficiency gain. Their optimal nonparametric adjustment leads to efficiency gain under an assumption that is more stringent than having nonzero coefficients.

can be easily simulated.¹³

Remark 5. Theorem 4 shows that $\{CS_\tau^{\text{sc}}(h \mid \bar{h}/\underline{h}) : h \in \mathbb{H}\}$ is an asymptotically valid confidence band for the constant function $\mathbb{H} \ni h \mapsto \vartheta_0$. By using it for inference on ϑ_0 , we take multiple bandwidth choices into account. Such an inference procedure is therefore more robust and less sensitive to bandwidth choice. The uniform confidence band can also be used for analysis of the sensitivity of the result from the pointwise confidence set to bandwidth choice. See AK for detailed discussion. AK’s argument can be extended to our case. Let h_{rf} denote a reference bandwidth and one computes $CS_\tau(h_{\text{rf}})$. In case of a statistically insignificant result (i.e., $0 \in CS_\tau(h_{\text{rf}})$), it can be argued that using a smaller (larger) bandwidth is necessary due to high bias (variance) incurred by h_{rf} . However, the specification search or multiple testing issue undermines the validity of a significant result ($CS_\tau(h) \subseteq (0, \infty)$ or $CS_\tau(h) \subseteq (-\infty, 0)$) corresponding to some $h \neq h_{\text{rf}}$. In such a case, with suitable lower and upper bounds (\underline{h}, \bar{h}) such that $\underline{h} < h_{\text{rf}} < \bar{h}$, one may follow AK’s approach and use the band $\{CS_\tau^{\text{sc}}(h \mid \bar{h}/\underline{h}) : h \in \mathbb{H}\}$. If $\exists h \in \mathbb{H}$ such that $CS_\tau^{\text{sc}}(h \mid \bar{h}/\underline{h}) \subseteq (0, \infty)$ or $CS_\tau^{\text{sc}}(h \mid \bar{h}/\underline{h}) \subseteq (-\infty, 0)$, one may conclude that the RD LATE is different from zero and validity of such a result is guaranteed by Theorem 2. On the other hand, if $0 \in CS_\tau^{\text{sc}}(h \mid \bar{h}/\underline{h}) \forall h \in \mathbb{H}$, we conclude that the insignificant result is insensitive to bandwidth choice. In case of $0 \notin CS_\tau(h_{\text{rf}})$, it is still necessary to examine the sensitivity of such a significant result to bandwidth choice (Imbens and Lemieux, 2008). As AK, with suitable (\underline{h}, \bar{h}) , one may conclude that $\vartheta_0 > 0$ in a robust sense if $\exists h \in \mathbb{H}$ such that $CS_\tau^{\text{sc}}(h \mid \bar{h}/\underline{h}) \subseteq (0, \infty)$ and $\forall h \in \mathbb{H}$, $CS_\tau^{\text{sc}}(h \mid \bar{h}/\underline{h}) \cap (0, \infty) \neq \emptyset$. Compared with AK, our confidence band incorporates information from covariates and the robust inference based on it is more powerful.

Remark 6. Theorem 2 also provides correction to obtain asymptotic validity under criterion-based data-driven bandwidth selection. In practical implementation, one may take the bandwidth to be $\hat{h} \in [\underline{h}, \bar{h}]$, where (\underline{h}, \bar{h}) are deterministic lower and upper bounds and \hat{h} is the minimizer of some data-dependent criterion function defined on $[\underline{h}, \bar{h}]$. Theorem 2 shows that snooping correction takes all noise in \hat{h} into account, by replacing the χ_1^2 quantile with $z_{1-\tau}(\bar{h}/\underline{h})^2$. By Theorem 2, asymptotic validity of the robust confidence set $CS_\tau^{\text{sc}}(\hat{h} \mid \bar{h}/\underline{h})$ is guaranteed without assuming that \hat{h} fulfills any property such as the stochastic order of $\hat{h}/h - 1$ is sufficiently small so that the noise in \hat{h} is negligible, where h is some deterministic bandwidth that \hat{h} tries to capture.

Remark 7. As the main result of AK, Theorem 2 assumes deterministic upper and lower bounds. Let

¹³See the R package `BWSnooping` from github.com/kolesarm/BWSnooping. If $\bar{h}/\underline{h} \uparrow \infty$ as $n \uparrow \infty$, then $z_\tau(\bar{h}/\underline{h})$ can be replaced by its asymptotic counterpart $z_\tau^{\text{asy}}(\bar{h}/\underline{h})$, where $z_{1-\tau}^{\text{asy}}(\bar{h}/\underline{h}) = -\log(-\log(1-\tau))/a_n + b_n$ with constants (a_n, b_n) that depend on \bar{h}/\underline{h} and the kernel function. $z_\tau^{\text{asy}}(\bar{h}/\underline{h})$ is not recommended to be used in practice since its justification is based on the asymptotic theory of suprema of stationary Gaussian processes, which converge at a slow speed. See AK for more detailed discussion on the critical values.

$(\underline{h}^*, \bar{h}^*)$ denote some deterministic bounds that some data-dependent bounds (\underline{h}, \bar{h}) capture. As argued by AK, the conclusion of Theorem 2 still holds under data-dependent bounds, if the orders of $\bar{h}/\bar{h}^* - 1$ and $\underline{h}/\underline{h}^* - 1$ are sufficiently small and $(\underline{h}^*, \bar{h}^*)$ satisfy the assumptions of Theorem 2.

5 Robust corrected empirical likelihood inference

In this section, we investigate the second-order properties of the EL inference method. Theorem 3 provides the distributional expansion of $LR_p(\vartheta_0 | h)$ and characterizes the leading term. By using this result, we drive the coverage optimal bandwidth and propose a simple and feasible correction to the EL ratio that leads to a fast coverage error decay rate. Theorem 4 provides the distributional expansion of the corrected EL ratio under local perturbation to the covariate balance condition. By this result, we show that the corrected EL confidence set enjoys a favorable property that its good coverage accuracy is maintained even if covariate balance assumption is slightly violated. We assume the following assumptions hold.

Assumption 4. *On a neighborhood around 0, $g_{\mathfrak{V}(k)^3|dd'}$ and $g_{\mathfrak{V}(k)^4|dd'}$ are both Lipschitz continuous and $g_{\|\mathfrak{V}(k)\|^{20}|dd'}$ is bounded, $\forall k \in \{0, 1\}$.*

Assumption 5. *$(1, \mathcal{K}_{p;+}, \mathcal{K}_{p;+}^2, \mathcal{K}_{p;+}^3)$ are linearly independent as elements in the vector space of continuous functions.*

Assumption 4 is a stronger condition than Assumption 2(a). Assumption 5 is a mild condition which is satisfied by all commonly-used kernels. Clearly, the same property also holds for $(1, \mathcal{K}_{p;-}, \mathcal{K}_{p;-}^2, \mathcal{K}_{p;-}^3)$. Assumptions 4 and 5 are used when establishing validity of the Edgeworth expansions in the proofs of Theorems 3 and 4. Let $\text{tr}(\mathbf{A})$ denote the trace of a square matrix \mathbf{A} . Denote $\Xi_1 := \mu_{UU^\top, \pm}^{-1}$, $\Xi_2 := (\mu_{UU^\top, +}^{-1} + \mu_{UU^\top, -}^{-1})^{-1}$, $\Psi_1^{uw} := \text{tr}(\Xi_1 \mu_{U^{(u)}U^{(w)}UU^\top, \pm})$ and $\Psi_2^{uw} := \text{tr}(\Xi_1 \mu_{U^{(u)}UU^\top, +} \Xi_1 \mu_{U^{(w)}UU^\top, +}) - 2 \cdot \text{tr}(\Xi_1 \mu_{U^{(u)}UU^\top, -} \Xi_1 \mu_{U^{(w)}UU^\top, +}) + \text{tr}(\Xi_1 \mu_{U^{(u)}UU^\top, -} \Xi_1 \mu_{U^{(w)}UU^\top, -})$. Let

$$\mathcal{V}_p^\dagger := \sum_{\mathbf{u}, \mathbf{w}=1, \dots, d_u} \left\{ \frac{1}{2} \frac{\omega_{p;+}^{0,4}}{\omega_{p;+}^{0,2}} \Xi_1^{(uw)} \Psi_1^{uw} - \frac{1}{3} \frac{(\omega_{p;+}^{0,3})^2}{(\omega_{p;+}^{0,2})^2} \Xi_1^{(uw)} \Psi_2^{uw} + \left(4\omega_{p;+}^{0,3} - 2(\omega_{p;+}^{0,2})^2 \right) \text{tr}(\Xi_1 \Xi_2) \right\} / (\omega_{p;+}^{0,2} \varphi). \quad (10)$$

Let \mathcal{V}_p^\ddagger be defined by the same formula with U replaced by $\bar{U} := Z - \vartheta_2$ and the range changed to $\mathbf{u}, \mathbf{w} = 1, \dots, d_z$ accordingly. Let $\mathcal{V}_p^{\text{LR}} := \mathcal{V}_p^\dagger - \mathcal{V}_p^\ddagger$. We write $a_n \asymp b_n$, if $a_n = O(b_n)$ and $b_n = O(a_n)$. The following result is similar to Calonico et al. (2020, Theorem 3.1) and provides an asymptotic expansion of the coverage probability $\Pr[\vartheta_0 \in CS_\tau(h)] = \Pr[LR_p(\vartheta_0 | h) \leq c_\tau]$. The proof uses the method of Calonico et al. (2022) and calculations in Matsushita and Otsu (2013); Ma (2017).

Theorem 3. Suppose that Assumptions 1 - 5 hold. Suppose that h satisfies $nh^{2p+3} = o(1)$ and $(nh^3)^{-1} = O(1)$. Then, $\Pr[LR_p(\vartheta_0 | h) \leq x] = F_{\chi_1^2}(x) - \mathcal{C}_p(n, h) x f_{\chi_1^2}(x) + o(v_n^*)$, where $v_n^* := nh^{2p+3} + h^{p+1} + (nh)^{-1}$, $\mathcal{C}_p(n, h) := nh^{2p+3} \mathcal{B}_p^{\text{LR}} + (nh)^{-1} \gamma_p^{\text{LR}}$ and $\mathcal{B}_p^{\text{LR}} := (\mathcal{B}_p^{\text{EL}})^2 / \gamma_p^{\text{EL}}$. Let the RCEL ratio be $LR_p^{\text{rc}}(\vartheta_0 | h) := LR_{p+1}(\vartheta_0 | h) / (1 + (nh)^{-1} \gamma_{p+1}^{\text{LR}})$. Then, $\Pr[LR_p^{\text{rc}}(\vartheta_0 | h) \leq x] = F_{\chi_1^2}(x) + O(n^{-1})$, if $h \asymp n^{-1/(p+2)}$.

Remark 8. Theorem 3 shows that the coverage error $\Pr[\vartheta_0 \in CS_\tau(h)] - (1 - \tau)$ is approximately equal to $-\mathcal{C}_p(n, h) c_\tau f_{\chi_1^2}(c_\tau)$. In the leading coverage error term, $nh^{2p+3} \mathcal{B}_p^{\text{LR}}$ is the “bias” term that is brought by the smoothing bias and $(nh)^{-1} \gamma_p^{\text{LR}}$ is the “variance term” that stems from the stochastic variability. Note that typically the distributional expansion corresponding to a nonparametric kernel-based Wald or likelihood ratio statistic (e.g., Calonico et al., 2020, Theorem 3.1) involves another “bias-variability” interaction term of order h^{p+1} . In our case, such a term is of order h^{p+2} and partial Bartlett correctability (Chen, 1996) crucially relies on this fact. In addition, because the moment conditions $W_p \otimes (Y - \theta_0 D - \theta_1, Z^\top - \theta_2^\top)^\top$ (with $\theta = \vartheta$) is asymptotically uncorrelated with the derivatives with respect to (θ_1, θ_2) at $\theta = \vartheta$, the coefficient γ_p^{LR} of the variance term has a simple expression given by the matrix formula (10).

Remark 9. Since $h \asymp n^{-1/(p+2)}$ gives the best coverage error decay rate, we restrict our attention to bandwidths that satisfy $h = H \cdot n^{-1/(p+2)}$ for some $H > 0$. Now the leading coverage error is proportional to $-n^{-(p+1)/(p+2)} C_1^{\text{EL}}(H)$, where $C_1^{\text{EL}}(H) := \mathcal{B}_p^{\text{LR}} H^{2p+3} + \gamma_p^{\text{LR}} H^{-1}$.¹⁴ Parallel to Calonico et al. (2018), we define the optimal constant H_{co} as the minimizer of the absolute value of the leading coverage error: $H_{\text{co}} := \operatorname{argmin}_{H>0} |C_1^{\text{EL}}(H)|$. Hence the coverage optimal bandwidth is given by $h_{\text{co}} = H_{\text{co}} n^{-1/(p+2)}$.¹⁵ Note that h_{co} in our EL approach is independent of the nominal coverage probability $1 - \tau$. This property is not shared by the coverage optimal bandwidth for the Wald-type approach.

Remark 10. It is shown in the proof that if $h \asymp n^{-1/(p+2)}$ the remainder term in the expansion of $\Pr[LR_p(\vartheta_0 | h) \leq x]$ is $O(n^{-1})$. Let $LR_p^{\text{bc}}(\vartheta_0 | h) := LR_p(\vartheta_0 | h) / (1 + \mathcal{C}_p(n, h))$ be the Bartlett corrected EL ratio. One can show that Bartlett correction removes the leading coverage error: $\Pr[LR_p^{\text{bc}}(\vartheta_0 | h) \leq x] = F_{\chi_1^2}(x) + o(v_n^*)$. The infeasible Bartlett corrected EL confidence set $\{\vartheta_0 : LR_p^{\text{bc}}(\vartheta_0 | h) \leq c_\tau\}$ has coverage accuracy with error rate $O(n^{-1})$. However, $(\mathcal{B}_p^{\text{LR}}, \gamma_p^{\text{LR}})$ depend on unknown parameters. One can replace these unknown quantities with their consistent nonparametric estimators to get the feasible Bartlett corrected

¹⁴The coverage expansion for the EL confidence set takes a much simpler form than its Wald-type counterpart. Let $WS_p(\vartheta_0 | h)$ denote a Wald-type statistic using the p -th order LP regression estimator (Calonico et al., 2020). If $h = H \cdot n^{-1/(p+2)}$, which leads to the best coverage error decay rate, the first-order approximation to the coverage error of the Wald-type confidence set is of the form $\bar{C}(H, x) n^{-(p+1)/(p+2)}$, where $\bar{C}(H, x) := (C_1(H)x + C_3(H)x^3 + C_5(H)x^5) f_{\chi_1^2}(x)$, $C_k(H) := c_{k,1} H^{2p+3} + c_{k,2} H^{p+1} + c_{k,3} H^{-1}$ and $(c_{k,1}, c_{k,2}, c_{k,3})$ are constants. The distributional expansion corresponding to the EL ratio is similar but much simpler. Its leading error term satisfies $C_3(H) = C_5(H) = c_{1,2} = 0$.

¹⁵Note that $\mathcal{B}_p^{\text{LR}} > 0$. If $\gamma_p^{\text{LR}} > 0$, $C_1^{\text{EL}}(H) > 0$ and clearly $\lim_{H \downarrow 0} C_1^{\text{EL}}(H) = \lim_{H \uparrow \infty} C_1^{\text{EL}}(H) = \infty$. The unique minimizer H_{co} satisfies the first-order condition. An explicit solution is available from solving it: $H_{\text{co}} = (\gamma_p^{\text{LR}} / ((2p+3) \mathcal{B}_p^{\text{LR}}))^{1/(2p+4)}$. If $\gamma_p^{\text{LR}} < 0$, it is easy to see that $H_{\text{co}} = (-\gamma_p^{\text{LR}} / \mathcal{B}_p^{\text{LR}})^{1/(2p+4)}$ and $C_1^{\text{EL}}(H_{\text{co}}) = 0$. In this case, the first-order coverage error vanishes at the optimal bandwidth.

EL confidence set. Note that $\mathcal{B}_p^{\text{LR}}$ involves higher-order derivatives up to the order $p+1$ while γ_p^{LR} depends only on conditional expectations. Hence the latter can be estimated by a simple plug-in estimator $\hat{\gamma}_p^{\text{LR}}$ that is based on local linear regression with standard rule-of-thumb (ROT) bandwidths (Hansen, 2021, Chapter 21.6). By standard theory, $\hat{\gamma}_p^{\text{LR}} - \gamma_p^{\text{LR}} = O_p(n^{-2/5})$. On the other hand, a fully nonparametric estimator $\hat{\mathcal{B}}_p^{\text{LR}}$ of $\mathcal{B}_p^{\text{LR}}$ is highly variable. As a result, the practical performance of the feasible Bartlett corrected EL confidence set is highly dependent on the estimation error for $\mathcal{B}_p^{\text{LR}}$ and for this reason, its coverage error decay rate can be much slower than $O(n^{-1})$.¹⁶

Remark 11. To avoid estimating $\mathcal{B}_p^{\text{LR}}$, we follow Chen (1996)’s partial Bartlett correction approach. We use the “partial” rescaling factor $1 + (nh)^{-1} \hat{\gamma}_p^{\text{LR}}$ and then choose h optimally to reduce the effects from the smoothing bias on the coverage accuracy. Since the bias-variability term is small in our case, such an approach delivers a coverage error of smaller order.¹⁷ We also use internalized bias removal by increasing the order of LP by one in the spirit of Calonico et al. (2014) and exhaust the smoothness.¹⁸ Note that Assumption 2(a) parallels the smoothness assumption of Calonico et al. (2020, Theorem 3.1). Under Assumption 2, the smoothing bias is now of order h^{p+2} . EL automatically accounts for the change in variability by implicit studentization. In the proof of Theorem 3, we show that $\Pr[LR_{p+1}(\vartheta_0 | h) \leq x]$ is equal to the sum of $F_{\chi_1^2}(x) - (nh)^{-1} \gamma_{p+1}^{\text{LR}} x f_{\chi_1^2}(x)$ and a remainder term that absorbs effects from the smoothing bias. The leading “variance” term becomes $(nh)^{-1} \gamma_{p+1}^{\text{LR}}$ and rescaling the EL ratio by $\left(1 + (nh)^{-1} \gamma_{p+1}^{\text{LR}}\right)^{-1}$ eliminates it. Essentially this approach trades bias for variance, as the latter can be estimated with good accuracy. In the proof we show that the rate optimal bandwidth ($h \asymp n^{-1/(p+2)}$) balances the terms of order nh^{2p+5} (bias) and h^{p+3} (bias-variability interaction) so that effects from the smoothing bias are made negligible. The remainder is now $O(n^{-1})$. Let the feasible RCEL ratio be $LR_p^{\text{frc}}(\theta_0 | h) := LR_{p+1}(\theta_0 | h) / \left(1 + (nh)^{-1} \hat{\gamma}_{p+1}^{\text{LR}}\right)$, where $\hat{\gamma}_{p+1}^{\text{LR}}$ is a plug-in estimator of γ_{p+1}^{LR} . Since $\hat{\gamma}_{p+1}^{\text{LR}} - \gamma_{p+1}^{\text{LR}} = O_p(n^{-2/5})$, the distributions of $LR_p^{\text{frc}}(\vartheta_0 | h)$ and $LR_p^{\text{rc}}(\vartheta_0 | h)$ differ by an error of order $o(n^{-1})$ and the second conclusion of Theorem 3 holds for $\Pr[LR_p^{\text{frc}}(\vartheta_0 | h) \leq x]$. Then it follows that the feasible RCEL confidence set $CS_\tau^{\text{frc}}(h) := \{\theta_0 : LR_p^{\text{frc}}(\theta_0 | h) \leq c_\tau\}$ has a coverage error of order $O(n^{-1})$. Following Gelman and Imbens (2019), we recommend using lower order local polynomials and setting $p = 1$ or $p = 2$. In both cases, $CS_\tau^{\text{frc}}(h)$ has coverage error decay rate $O(n^{-1})$. The rate optimal bandwidth obeys $h \asymp n^{-1/3}$ ($p = 1$) or $h \asymp n^{-1/4}$ ($p = 2$). In the former situation, we require a weaker

¹⁶It follows from standard theory that the best possible rate for $\hat{\mathcal{B}}_p^{\text{LR}}$ is given by $\hat{\mathcal{B}}_p^{\text{LR}} - \mathcal{B}_p^{\text{LR}} = O_p(n^{-1/(2p+5)})$ under our smoothness assumptions. The coverage error of the feasible Bartlett corrected confidence set is of order $n^{-(1/(2p+5) + (p+1)/(p+2))}$ if $h \asymp n^{-1/(p+2)}$.

¹⁷In case of a typical distributional expansion with leading terms of order nh^{2p+3} (bias), $(nh)^{-1}$ (variability) and h^{p+1} (bias-variability interaction), the best coverage error decay rate of the remainder stays unchanged even if the variability term is removed.

¹⁸Calonico et al. (2014, Remark 7) shows that subtracting the p -th order LP estimator by the nonparametric estimator for the leading bias term with the same bandwidth is the same as a $(p+1)$ -th order LP estimator. By increasing the order of LP by one, this approach makes the order of bias smaller but brings one more term that contributes to the stochastic variability. Calonico et al. (2014) proposed bias-correction-aware standard errors that account for the change in variability.

smoothness assumption, achieve the same fast coverage error decay rate but use a smaller effective sample of size nh . In this situation, the length of $CS_{\tau}^{\text{frc}}(h)$ is of larger order of magnitude.

Remark 12. We now compare the feasible RCEL to CCFT's inference method with $p = 1$. CCFT proposes Wald-type inference using their local linear estimator with bias correction and standard errors that take into account estimation of the bias. CCFT's bias-corrected local linear estimator with common bandwidths is equivalent to the augmented local quadratic regression estimator. It is well-expected that an extension of [Calonico et al. \(2020, Theorem 3.1\)](#) holds and CCFT's confidence interval admits a similar distributional expansion. Hence for $p = 1$, CCFT's method with a bandwidth h that obeys the optimal rate (i.e., $h \asymp n^{-1/4}$) has coverage error decay rate $n^{-3/4}$ (see [Calonico et al., 2020, Theorem 3.1\(a\)](#)). As CCFT, we use local quadratic moment conditions ($p + 1 = 2$) in (4) to reduce the smoothing bias. Meanwhile, our method analytically accounts for the effect of stochastic variability on the coverage error and then chooses the bandwidth rate optimally (i.e., setting $h \asymp n^{-1/3}$) so that a faster $O(n^{-1})$ coverage error decay rate is achieved. Note that the same smoothness assumption underlies such comparison. If more smoothness (thrice differentiability in Assumption 2(a)) is available (see [Calonico et al., 2020, Theorem 3.1\(b\)](#)), we can further increase the LP order by one (i.e., local cubic). Partial Bartlett correction takes the increase in variability into account. The length of the resulting $CS_{\tau}^{\text{frc}}(h)$ with $p = 2$ and the rate optimal bandwidth (i.e., $h \asymp n^{-1/4}$) has the same order as CCFT's confidence interval but $CS_{\tau}^{\text{frc}}(h)$ enjoys a faster $O(n^{-1})$ coverage error decay rate.

Remark 13. Like most existing results on second-order properties of kernel-based nonparametric inference, Theorem 3 and our previous discussion assume a deterministic bandwidth. In practical data-driven implementation of the corrected confidence set, one selects a deterministic bandwidth of the form $h = H \cdot n^{-1/(p+2)}$, replaces H with a consistent estimator \hat{H} and reports $CS_{\tau}^{\text{frc}}(\hat{h})$ where $\hat{h} := \hat{H} \cdot n^{-1/(p+2)}$. [Calonico et al. \(2020\)](#) (see Section 5.3 therein) proposes an approach that takes the estimated AMSE optimal bandwidth and rescales it to make it obey the coverage optimal rate (see Section IV(C) of CCFT). We can follow this approach to use the rescaled versions of CCFT's bandwidth. Alternatively, we can use a simpler ROT bandwidth proposed in [Hansen \(2021, Chapter 21.6\)](#). The partial Bartlett correction we take makes the coverage accuracy less sensitive to the choice of the constant part of the bandwidth. In simulations, we find that $CS_{\tau}^{\text{frc}}(\hat{h})$ with \hat{h} taken to be any of the aforementioned data-driven bandwidth selectors has good coverage accuracy.

Remark 14. By using the AK-type correction proposed in Theorem 2, we can also construct a confidence band that uses a continuous range of bandwidths to analyze the sensitivity of the result from the RCEL confidence set $CS_{\tau}^{\text{frc}}(\hat{h})$ to bandwidth choice. It can be easily verified that the conclusion of The-

orem 2 with p changed to $p + 1$ still holds when $LR_{p+1}(\vartheta_0 | h)$ is replaced by $LR_p^{\text{frc}}(\vartheta_0 | h)$ since they are first-order equivalent, uniformly in $h \in \mathbb{H}$. We can take the lower and upper bounds in \mathbb{H} to be proportional to any of the reference bandwidths \hat{h} discussed in the preceding remark. Let $CS_\tau^{\text{frc}}(h | \bar{h}/\underline{h}) := \left\{ \theta_0 : LR_p^{\text{frc}}(\theta_0 | h) \leq z_\tau (\bar{h}/\underline{h})^2 \right\}$ be the “doubly corrected” confidence set and we construct the RCEL confidence band $\{CS_\tau^{\text{frc}}(h | \bar{h}/\underline{h}) : h \in \mathbb{H}\}$ accordingly for sensitivity analysis. We also expect a small coverage error for the RCEL confidence band.¹⁹ Our method thus serves as a very effective tool for AK-type robust inference that explicitly takes multiple bandwidths into account.

We have shown that the RCEL confidence set has superb coverage accuracy, under the covariate balance assumption. We now consider scenarios in which covariate balance fails to hold and analyze the sensitivity of the coverage accuracy to this assumption. Cattaneo and Titiunik (2022) note that “the principle of covariate balance can be extended beyond pre-determined covariates to variables that are determined after the treatment is assigned but are known to be unaffected by the treatment...” Such extension of the scope of covariate is more than welcome in our GMM framework because LP moment conditions (4) allow for any Z with $\mathcal{T}_Z = 0$, regardless of whether Z is a pre-determined covariate or an “unaffected” outcome. While expanding the set of covariates may improve the efficiency of estimation and inference, it bears the risk that the prior belief $\mathcal{T}_Z = 0$ is actually wrong for some “unaffected” outcomes included in (4). If the falsification test in the first stage rejects the balance hypothesis for such “unaffected” outcomes, we can exclude it from covariate adjusted estimation. However, the usual falsification test sets $\mathcal{T}_Z = 0$ as the null hypothesis and may fail to reject if \mathcal{T}_Z is close to the hypothesized value 0 under the null hypothesis.²⁰ Another possibility is that Assumption 1 is indeed satisfied by the true probability law but our sample observations are subject to data contamination or measurement errors that occur after treatment (Kitamura et al., 2013). Z_1, \dots, Z_n may be drawn from some perturbed probability law which slightly violates $\mu_{Z,+} = \mu_{Z,-}$. CCFT shows that the covariate adjusted estimator is inconsistent and the confidence interval fails to have asymptotically correct coverage probability in both situations when $\mu_{Z,+} \neq \mu_{Z,-}$. When implementing covariate adjustment, the researcher may mistakenly include covariates that slightly violate the assumption $\mu_{Z,+} = \mu_{Z,-}$. Theorem 4 shows that our method is useful when our prior belief about the “unaffected” outcomes is imperfect or the

¹⁹In the proof of the asymptotic validity result $\Pr \left[\vartheta_0 \in CS_\tau^{\text{frc}}(h | \bar{h}/\underline{h}), \forall h \in \mathbb{H} \right] \rightarrow 1 - \tau$, we show that the distribution of $\sup_{h \in \mathbb{H}} LR_p^{\text{frc}}(\vartheta_0 | h)$ is approximated by the distribution of $\|G\|_{[1, \bar{h}/\underline{h}]}^2 = \sup_{h \in \mathbb{H}} \Gamma_G(h/\underline{h})^2$ with a vanishing error, where $\Gamma_G(h/\underline{h})^2$ follows the χ_1^2 distribution $\forall h \in \mathbb{H}$. We expect that the distributional approximation of $\sup_{h \in \mathbb{H}} \Gamma_G(h/\underline{h})^2$ to $\sup_{h \in \mathbb{H}} LR_p^{\text{frc}}(\vartheta_0 | h)$ inherits the good accuracy of the pointwise-in-bandwidth distributional approximation of $\Gamma_G(h/\underline{h})^2$ to $LR_p^{\text{frc}}(\vartheta_0 | h)$.

²⁰E.g., Ludwig and Miller (2007) found no discontinuity in child mortality from injuries near the cutoff that divides the treatment and control groups, where the treatment refers to high participation and funding rate for the Head Start program. Therefore, child mortality from injuries can serve as a covariate Z when the interested outcome is the effect on child mortality rate from causes affected by Head Start. However, the fact that the balance condition involving Z is not empirically rejected may be caused by lack of sufficient power. Hence, there is a risk of covariate imbalance.

data on covariates are contaminated but the incurred imbalance is slight.

By using local asymptotic analysis, we analyze the performance of our RCEL confidence set in the framework of local misspecification (see, e.g., [Armstrong and Kolesár, 2021](#)). We assume that \mathcal{T}_Z approaches the hypothesized value 0 under covariate balance at the rate of $(nh)^{-1/2}$ so that the coverage probability $\Pr[\vartheta_0 \in CS_\tau^{\text{frc}}(h)] = \Pr[LR_p^{\text{frc}}(\vartheta_0 | h) \leq c_\tau]$ has a limit in $(0, 1 - \tau)$, which captures the phenomenon that covariate imbalance results in undercoverage in finite samples. We assume that the bandwidth for our RCEL confidence set has been set to obey the optimal rate that minimizes the coverage error, i.e., $h = H \cdot n^{-1/(p+2)}$ for some constant $H > 0$, when covariate balance holds. Let $l_n := n^{-(p+1)/(2p+4)} \asymp (nh)^{-1/2}$. As the standard Pitman approach to analyzing power properties of tests of parametric hypotheses, we think of the local imbalance hypothesis $\mathcal{T}_Z = \delta l_n$ as reparametrization of values of \mathcal{T}_Z that lie in a small neighborhood around 0, where $\delta \in \mathbb{R}^{d_z}$ denotes the localizing parameter.²¹ $\mathcal{T}_Z = \delta l_n$ is equivalent to $\mu_{Z,-} = \mu_{Z,+} - (\mu_{D,+} - \mu_{D,-}) \delta l_n$ under Assumption 1 (a), (b) and (c). Then it is clear that then the moment conditions (4) are locally misspecified in the sense of [Armstrong and Kolesár \(2021\)](#) since $E[h^{-1}W_{p,-}(Z - \vartheta_2)] = O(l_n)$. Our result differs from [Armstrong and Kolesár \(2021\)](#) and focuses on the coverage performance of the RCEL confidence set when δ is close to 0.²² This is in accordance with the fact that local imbalance with a large δ can be detected with a high probability in the first-stage RD falsification test.

We now consider $\Pr[\vartheta_0 \in CS_\tau^{\text{frc}}(h)]$ as a function of δ under local imbalance. Theorem 3 shows that if $\delta = 0$, $\Pr[\vartheta_0 \in CS_\tau^{\text{frc}}(h)] = 1 - \tau + O(n^{-1})$. A measure of sensitivity of the coverage accuracy to local imbalance (i.e., how the coverage probability drops relative to that under $\delta = 0$) is given by the slope of $\Pr[\vartheta_0 \in CS_\tau^{\text{frc}}(h)]$ as a function of δ at $\delta = 0$. We extend Theorem 3 and derive a two-term asymptotic expansion for $\Pr[\vartheta_0 \in CS_\tau^{\text{frc}}(h)]$ and take the sum of the leading terms, denoted by $R(\delta)$, as approximation to $\Pr[\vartheta_0 \in CS_\tau^{\text{frc}}(h)]$ in finite samples. We show that $R(0) = 1 - \tau$ and the gradient $\nabla R(\delta) := \partial R(\delta) / \partial \delta$ of $R(\delta)$ at $\delta = 0$ is equal to 0, so that $R(\delta)$ is locally constant around $\delta = 0$. This shows that coverage accuracy of the RCEL confidence set is relatively unaffected by local imbalance in finite samples. We note that this is indeed a unique property of the RCEL confidence set (Remark 16) and any other inference method does not have the same property in general. We get the same conclusion in case of local imbalance due to data contamination, when Z_1, \dots, Z_n are drawn from a locally perturbed population that satisfies $\mu_{Z,+} - \mu_{Z,-} = \delta l_n$. Let $F(\cdot | \eta)$ denote the CDF of a $\chi_1^2(\eta)$ (non-central χ^2 with one degree of freedom

²¹Our specification follows [Gallant and White \(1988, Chapter 7\)](#)’s “fixed data-generating process (DGP), drifting hypothesis” approach. $\mathcal{T}_Z = \delta l_n$ is understood as the assumption that our hypothesized value 0 for the true RD LATE \mathcal{T}_Z is chosen in such a way that 0 lies in the $\|\delta\| l_n$ neighborhood of \mathcal{T}_Z , which is assumed to be fixed. For the alternative “fixed hypothesis, drifting DGP” approach, we fix a DGP for the latent variables and the score that satisfies Assumptions 1, 2 and 4. $\mathcal{T}_Z = \delta l_n$ is understood as the assumption that the latent variables and the score follow a joint distribution that is given by the given DGP with a location shift. By taking this alternative approach, we can show a result similar to Theorem 4 and get an identical conclusion. The proof requires more complicated arguments and suitable modification of the assumptions.

²²The approach of [Armstrong and Kolesár \(2021\)](#) specifies a set in which δ possibly lies and then adjust the critical value to take into account the maximal misspecification bias. We take a very different approach in this paper.

and non-centrality parameter $\eta \geq 0$) random variable. Let $F^{(k)}(x | \eta) := \partial^k F(x | \eta) / \partial \eta^k$ be the k -times partial derivative of $F(x | \eta)$ with respect to η . The following result gives a second-order approximation to the distribution of $LR_p^{\text{rc}}(\vartheta_0 | h)$ under $\mathcal{T}_Z = \delta l_n$. The same result holds for $LR_p^{\text{frc}}(\vartheta_0 | h)$.

Theorem 4. *Suppose that Assumptions 1 - 5 with Assumption 1(d) replaced by $\mathcal{T}_Z = \delta l_n$ hold. Suppose that h satisfies $h = H \cdot n^{-1/(p+2)}$ for some constant $H > 0$. Then, $\Pr[LR_p^{\text{rc}}(\vartheta_0 | h) \leq x] = F\left(x | H(\gamma_\Delta^\top \delta / \bar{\sigma}_{p+1})^2\right) + P(x, \delta) l_n + o(l_n)$, where $P(x, \delta) := \mathcal{P}_1(\delta) F^{(1)}\left(x | H(\gamma_\Delta^\top \delta / \bar{\sigma}_{p+1})^2\right) + \mathcal{P}_2(\delta) F^{(2)}\left(x | H(\gamma_\Delta^\top \delta / \bar{\sigma}_{p+1})^2\right)$, $\gamma_\Delta \rightarrow \gamma_{\text{adj}}$ and $\bar{\sigma}_{p+1}^2 \rightarrow \mathcal{V}_{p+1}^{\text{EL}}$ as $n \uparrow \infty$ and $(\mathcal{P}_1, \mathcal{P}_2)$ are homogeneous cubic polynomials with constant coefficients. The expressions of $(\gamma_\Delta, \bar{\sigma}_{p+1}^2, \mathcal{P}_1, \mathcal{P}_2)$ are in the appendix.*

Remark 15. In the approximation to $\Pr[\vartheta_0 \in CS_\tau^{\text{frc}}(h)] = \Pr[LR_p^{\text{frc}}(\vartheta_0 | h) \leq c_\tau]$, the first-order term $F\left(c_\tau | H(\gamma_\Delta^\top \delta / \bar{\sigma}_{p+1})^2\right)$ is an even function of δ , $\partial F\left(c_\tau | H(\gamma_\Delta^\top \delta / \bar{\sigma}_{p+1})^2\right) / \partial \delta|_{\delta=0} = 0$ and the second-order approximation $P(c_\tau, \delta)$ is an odd function of δ . Theorem 4 also implies that $\partial P(c_\tau, \delta) / \partial \delta|_{\delta=0} = 0$ and $P(c_\tau, \cdot)$ is locally constant around the origin. Let $R(\delta) := F\left(c_\tau | H(\gamma_\Delta^\top \delta / \bar{\sigma}_{p+1})^2\right) + P(c_\tau, \delta) l_n$. Then we have $\nabla R(0) = 0$ and this shows that the coverage accuracy of the RCEL confidence set is highly insensitive to local perturbation to covariate balance ($\delta = 0$). If $\|\nabla R(0)\|$ is large in magnitude, a slight perturbation would incur severe undercoverage. To see that the slope is a measure of sensitivity to local imbalance, we consider the approximate minimal coverage $\min_{\delta \in \mathbb{S}_\nu} R(\delta)$ on \mathbb{S}_ν , where ν denotes a positive constant and $\mathbb{S}_\nu := \{\delta \in \mathbb{R}^{d_z} : \|\delta\| = \nu\}$ represents perturbations with equal magnitude ν in all directions. $\delta_R^* := \text{argmin}_{\delta \in \mathbb{S}_\nu} R(\delta)$ corresponds to the direction in which the perturbation results in the most severe undercoverage. Clearly, $R(\delta_R^*) < 1 - \tau$ and we have the approximation $R(\delta_R^*) = (1 - \tau) - \|\nabla R(0)\| \nu + o(\nu)$ when ν is small.²³ Therefore, the RCEL confidence set has minimal sensitivity due to $\|\nabla R(0)\| = 0$.

Remark 16. Having a locally constant second-order approximation (as a function of δ) is a unique property. Let $\rho_\varsigma(x) := (x^{-\varsigma} - 1) / \{\varsigma(1 + \varsigma)\}$ for $\varsigma \in \mathbb{R}$. We interpret the special case $\rho_0(x) = -\log(x)$ (EL) as the limit of $\rho_\varsigma(x)$ as $\varsigma \rightarrow 0$. The nonparametric likelihood (NPL) criterion function $\ell_p^\varsigma(\theta | h)$ is defined by (5) with $\sum_i \rho_0(n \cdot w_i) = -\sum_i \log(w_i/n^{-1})$ replaced by the more general Cressie-Read divergence $\sum_i \rho_\varsigma(n \cdot w_i)$. The NPL ratio $LR_{p+1}^\varsigma(\theta_0 | h)$ and confidence set are defined analogously. Under the same assumptions as in Theorem 4, we can show that $\Pr[LR_{p+1}^\varsigma(\vartheta_0 | h) \leq x]$ admits a similar two-term asymptotic expansion.²⁴ The first-order term in the expansion for $\Pr[LR_{p+1}^\varsigma(\vartheta_0 | h) \leq x]$ is still given by $F\left(x | H(\gamma_\Delta^\top \delta / \bar{\sigma}_{p+1})^2\right)$. The second-order term is of the form $(P(x, \delta) + P_\varsigma(x, \delta)) l_n$, where $P_\varsigma(x, \delta)$ is an odd function of δ , $P_\varsigma(x, \delta) = 0$ if $\varsigma = 0$ and $\partial P_\varsigma(x, \delta) / \partial \delta|_{\delta=0} \neq 0$ in general if $\varsigma \neq 0$. In this case, there can be perturbation associated

²³By using the Lagrange multiplier method to solve the constrained minimization problem $\min_{\delta \in \mathbb{S}_\nu} R(\delta)$ and mean value expansion, $\delta_R^* = -(\nabla R(\delta_R^*) / \|\nabla R(\delta_R^*)\|) \nu$ and therefore, $R(\delta_R^*) = (1 - \tau) - \left(\nabla R(\delta_R^*)^\top \nabla R(\delta_R^*) / \|\nabla R(\delta_R^*)\|\right) \nu$, where δ_R^* is the mean value that lies between δ_R^* and 0. Clearly, $\nabla R(\delta_R^*)^\top \nabla R(\delta_R^*) / \|\nabla R(\delta_R^*)\| \rightarrow \|\nabla R(0)\| = 0$, as $\nu \downarrow 0$.

²⁴The proof of this result is omitted for brevity but available from the authors.

with a large drop in coverage probability of the NPL confidence set. Let $R_\varsigma(\delta) := F\left(c_\tau \mid H\left(\gamma_\Delta^\top \delta / \bar{\sigma}_{p+1}\right)^2\right) + (P(x, \delta) + P_\varsigma(x, \delta))l_n$ and $\delta_{R_\varsigma}^* := \operatorname{argmin}_{\delta \in \mathbb{S}_\nu} R_\varsigma(\delta)$. Then by similar arguments, $R_\varsigma(\delta_{R_\varsigma}^*) = (1 - \tau) - \|\nabla R_\varsigma(0)\| \nu + o(\nu)$, which is highly sensitive if $\varsigma \neq 0$ and $\|\nabla R_\varsigma(0)\| > 0$ is large. In contrast, the RCEL confidence set exhibits good coverage accuracy uniformly in all $\delta \in \mathbb{S}_\nu$ when ν is small.

6 Monte Carlo simulations

We conduct simulations to evaluate the finite sample performance of the proposed RCEL inference for sharp RD designs with covariates. The DGP of the outcome variable Y_i , the score X_i and the first covariate $Z_i^{(1)}$ is based on the simulation design of CCFT. Incorporation of additional covariates $Z_i^{(2)}, \dots, Z_i^{(l)}$ follows that of [Arai et al. \(2021\)](#). I.e., $Y_i = \mu_y(X_i, Z_i^{(1)}) + \sum_{j=2}^l \pi^{j-1} Z_i^{(j)} + \varepsilon_{y,i}$ and $Z_i^{(1)} = \mu_z(X_i) + \varepsilon_{z,i}$, where

$$\begin{aligned} \mu_y(x, z_1) &:= \begin{cases} 0.36 + 0.96x + 5.47x^2 + 15.28x^3 + 15.87x^4 + 5.14x^5 + 0.22z_1 & \text{if } x < 0, \\ 0.38 + 0.62x - 2.84x^2 + 8.42x^3 - 10.24x^4 + 4.31x^5 + 0.28z_1 & \text{if } x \geq 0; \end{cases} \\ \mu_z(x) &:= \begin{cases} 0.49 + 1.06x + 5.74x^2 + 17.14x^3 + 19.75x^4 + 7.47x^5 & \text{if } x < 0, \\ 0.49 + 0.61x - 0.23x^2 - 3.46x^3 + 6.43x^4 - 3.48x^5 & \text{if } x \geq 0. \end{cases} \end{aligned}$$

Error terms $(\varepsilon_{y,i}, \varepsilon_{z,i})$ are bivariate normal with mean 0, standard deviation 1 and correlation coefficient $\rho = 0.269$. Additional covariates $(Z_i^{(2)}, \dots, Z_i^{(l)})$ have a multivariate normal distribution with mean zero and covariance matrix given by $\operatorname{Cov}[Z_i^{(j)}, Z_i^{(k)}] = 0.5^{|j-k|}$, $\forall j, k \geq 2$. We take $\pi = 0.2$. We consider three scenarios with $l = 0, 2, 4$, which correspond to the total number of covariates $d_z = l + 1$ being 1, 3, 5. We take the LP order $p = 1$ or 2. The sample sizes are $n = 500, 1000, 2000$. The number of Monte Carlo replications is 2000.

Table 1 presents the empirical coverage rates of the feasible RCEL confidence set $CS_\tau^{\text{frc}}(\hat{h})$ proposed in Remark 11. The nominal coverage $1 - \tau$ is set as 0.90, 0.95, 0.99. Following Remark 13, we consider bandwidths in the form of $\hat{h} = \hat{H} \cdot n^{-1/(p+2)}$ and two choices of the constant part \hat{H} : ROT in the table corresponds to the ROT bandwidth recommended in [Hansen \(2021, Chapter 21.6\)](#) and CCFT corresponds to rescaled CCFT's bandwidth.²⁵ Both ROT and CCFT bandwidths used for RCEL obey the coverage optimal rate discussed in Remark 11. For comparison, Table 1 also includes results from CCFT's method that uses a bandwidth

²⁵As Remark 12 notes, CCFT's coverage optimal bandwidth has the rate $n^{-1/4}$ for $p = 1$. For each simulation replication, let h_{CCFT} be the CCFT bandwidth computed from R function `rdrobust` with the options `p=1, rho=1, and bwselect="cerrd"`. Then in Table 1 and Figure 1, our CCFT bandwidth used for "RCEL, $p = 1$ " is $h_{\text{CCFT}} \cdot n^{-1/12}$ (rescaled to the coverage optimal rate $n^{-1/(p+2)}$ discussed in Remark 11). The CCFT bandwidth used for "RCEL, $p = 2$ " is h_{CCFT} itself, as now the coverage optimal rate for RCEL is $n^{-1/4}$.

with the coverage optimal rate and restricts $\rho = h/b = 1$, where b stands for the pilot bandwidth used for bias estimation. For each (n, τ) combination, the number most close to the nominal coverage probability is bold-faced. We observe that both RCEL and CCFT yield coverage probabilities close to their nominal levels for all the considered scenarios. For RCEL, both ROT and (rescaled) CCFT bandwidths perform reasonably well. When the sample size is small ($n = 500$), RCEL exhibits a small advantage over CCFT, which is in line with the theoretical result that RCEL achieves a faster coverage error decay rate than CCFT. See Remark 12. We then examine how the coverage rate of RCEL confidence set changes when the covariate balance assumption is slightly violated. We consider the case with one covariate ($d_z = 1$) and modify the DGP of $Z_i^{(1)}$ in the following way that the imbalance is characterized by a perturbation parameter δ :

$$\mu_z(x | \delta) := \begin{cases} 0.49 + 1.06x + 5.74x^2 + 17.14x^3 + 19.75x^4 + 7.47x^5 & \text{if } x < 0, \\ 0.49 + \delta + 0.61x - 0.23x^2 - 3.46x^3 + 6.43x^4 - 3.48x^5 & \text{if } x \geq 0. \end{cases}$$

Figure 1 plots the simulated coverage rates of RCEL and CCFT as a function of $\delta \in [-0.4, 0.4]$ for different combinations of sample size n and nominal coverage $1 - \tau$. We observe that the coverage rate of RCEL is less sensitive to the change of δ , which parallels the theoretical finding discussed in Remark 15. Overall, our simulation results show that RCEL inference method can be a useful addition to practitioners' toolkit.

7 Empirical illustration

We apply the RCEL inference method to analyze the individual incumbent advantage in Finnish municipal elections, which was first studied by Hyytinen et al. (2018). In the RD framework, the binary outcome variable Y indicates whether the candidate is elected in the next election, and the score X is the vote share margin in the previous election. Two covariates are included: candidates' age and gender. The main results are presented in Table 2. The ROT bandwidth is computed in the same way as the simulation exercise. For $p = 1$, the CCFT bandwidth and its rescaled version correspond to h_{CCFT} and $h_{\text{CCFT}} \cdot n^{-1/12}$ described in Footnote 25. For $p = 2$, the rescaled CCFT bandwidth corresponds to h_{CCFT} (for $p = 1$) in Footnote 25, and the CCFT bandwidth is equal to the one used by CCFT for $p = 2$. The columns of Table 2 present the estimates of RD LATE ϑ_0 , p -values for testing $H_0 : \vartheta_0 = 0$, the 95% confidence intervals and the selected bandwidths. The dataset also includes 1351 candidates "for whom the (previous) electoral outcome was determined via random seat assignment due to ties in vote counts" (Hyytinen et al., 2018, Page 1020), which constitutes a experiment benchmark to evaluate the credibility of the RD treatment effect estimated from the non-experimental data (candidates with previous electoral ties are excluded from the RD sample).

As reproduced as “Experiment benchmark” in Table 2, Hyytinen et al. (2018) find zero treatment effect (see their Table 2, Column 4, the p -value is imputed by us). We notice that RCEL and CCFT methods using different bandwidths presented in Table 2 deliver non-significant inference results comparable to the experiment benchmark, except for CCFT ($p = 2$) that does not restrict $\rho := h/b$ (p -value = 0.028).²⁶ This is not due to the larger bandwidth ($h = 1.103$) used by CCFT when $p = 2$, as RCEL ($p = 2$) using the same bandwidth does not reject the null hypothesis (p -value = 0.229).

The robustness of RCEL is further confirmed by Figure 2, which conducts sensitivity analysis of the RCEL inference method with respect to the bandwidth choice. Figure 2 plots the RCEL confidence band (Remarks 5 and 14) over a continuous range of bandwidths $h \in [\underline{h}, \bar{h}]$. Here we choose the lower bound $\underline{h} = 0.12$, which would include about 3% of the sample. The upper bound $\bar{h} = 0.72$ is approximately two times the ROT bandwidth in Table 2 and includes 17% of the total sample. The snooping corrected critical value $(2.413)^2$ for the triangular kernel and bandwidth ratio $\bar{h}/\underline{h} = 6$ is calculated from R package `BWSnooping`. In Figure 2, the solid (or dotted) line corresponds to 95% uniform (or pointwise) confidence band. The vertical dashed lines indicate the first two bandwidths used for RCEL in Table 2. For small bandwidth (say, less than 0.2), the RCEL uniform confidence band is wide. However, as long as the bandwidth is not so small, the confidence band looks quite stable. Moreover, the confidence band includes zero over the entire bandwidth range we consider, which demonstrates the robustness of the finding of no incumbency advantage with respect to the bandwidth choice. Overall, our example illustrates the practicality of the RCEL inference method.

8 Conclusion

This paper proposes a novel EL approach to covariate adjustment for regression discontinuity designs. Our approach incorporates covariates through over-identifying restrictions which represent the covariate balance condition. We derive the first-order and second-order asymptotic properties of our method. We show that the widely-used regression estimator of Calonico et al. (2019) is never less efficient than the standard estimator without covariates. It achieves efficiency gain as long as the true projection coefficients of some covariates are nonzero. By establishing the first-order equivalence between our EL estimator and Calonico et al. (2019)’s regression estimator, we show that the efficiency gain can be attributed to incorporating the covariate balance condition as side information. We show a uniform-in-bandwidth Wilks theorem, which can be used for sensitivity analysis and robust inference along the lines of Armstrong and Kolesár (2018). We derive the distributional expansion for the EL ratio statistic under the covariate balance condition and show that it

²⁶This is in line Hyytinen et al. (2018, columns (2) of Table 4)’s results that the CCT (without covariates) using the MSE-optimal bandwidth (ρ not restricted) rejects the null hypothesis of zero treatment effect, while the one restricting $\rho = 1$ does not reject.

admits a simple data-driven correction that substantially improves the coverage performance. We also derive the distributional expansion for the robust corrected EL ratio statistic under the local imbalance condition. It shows that the robust corrected EL confidence set is self-guarded against undercoverage in case of slight perturbation to covariate balance.

References

- Arai, Y., Y.-c. Hsu, T. Kitagawa, I. Mourifie, and Y. Wan (2021). Testing identifying assumptions in fuzzy regression discontinuity designs. *Quantitative Economics*.
- Arai, Y., T. Otsu, and M. H. Seo (2021). Regression discontinuity design with potentially many covariates. *arXiv preprint arXiv:2109.08351*.
- Armstrong, T. B. and M. Kolesár (2018). A simple adjustment for bandwidth snooping. *The Review of Economic Studies* 85(2), 732–765.
- Armstrong, T. B. and M. Kolesár (2021). Sensitivity analysis using approximate moment condition models. *Quantitative Economics* 12(1), 77–108.
- Bickel, P. J. and K. A. Doksum (2015). *Mathematical statistics: basic ideas and selected topics*, Volume 2. CRC Press.
- Bravo, F. (2003, 12). Second order power comparisons for a class of nonparametric likelihood based tests. *Biometrika* 90(4), 881–890.
- Brown, B. W. and W. K. Newey (2002). Generalized Method of Moments, Efficient Bootstrapping, and Improved Inference. *Journal of Business & Economic Statistics* 20(4), 507–517.
- Calonico, S., M. D. Cattaneo, and M. H. Farrell (2018). On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association* 113(522), 767–779.
- Calonico, S., M. D. Cattaneo, and M. H. Farrell (2020). Optimal bandwidth choice for robust bias corrected inference in regression discontinuity designs. *Econometrics Journal*.
- Calonico, S., M. D. Cattaneo, and M. H. Farrell (2022). Coverage error optimal confidence intervals for local polynomial regression. *Bernoulli*.
- Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik (2019). Regression discontinuity designs using covariates. *Review of Economics and Statistics* 101(3), 442–451.

- Calonico, S., M. D. Cattaneo, and R. Titiunik (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica* 82(6), 2295–2326.
- Canay, I. A. and V. Kamat (2017). Approximate Permutation Tests and Induced Order Statistics in the Regression Discontinuity Design. *The Review of Economic Studies* 85(3), 1577–1608.
- Cattaneo, M., L. Keele, and R. Titiunik (2021). Covariate adjustment in regression discontinuity designs. *Handbook of Matching and Weighting in Causal Inference*.
- Cattaneo, M. and R. Titiunik (2022). Regression discontinuity designs. *Annual Review of Economics*.
- Cattaneo, M. D., N. Idrobo, and R. Titiunik (2019). A Practical Introduction to Regression Discontinuity Designs. *arXiv*.
- Chen, S. X. (1996). Empirical likelihood confidence intervals for nonparametric density estimation. *Biometrika*, 329–341.
- Chen, S. X. and H. Cui (2007). On the second-order properties of empirical likelihood with moment restrictions. *Journal of Econometrics* 141(2), 492–516.
- Chen, S. X. and Y. S. Qin (2000). Empirical likelihood confidence intervals for local linear smoothers. *Biometrika*, 946–953.
- Chen, X. and K. Kato (2020). Jackknife multiplier bootstrap: finite sample approximations to the u-process supremum with applications. *Probability Theory and Related Fields* 176(3-4), 1–67.
- Chernozhukov, V., D. Chetverikov, and K. Kato (2014a). Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics* 42(5), 1787–1818.
- Chernozhukov, V., D. Chetverikov, and K. Kato (2014b). Gaussian approximation of suprema of empirical processes. *Annals of Statistics* 42(4), 1564–1597.
- Chernozhukov, V., D. Chetverikov, and K. Kato (2016). Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related Gaussian couplings. *Stochastic Processes and their Applications* 126(12), 3632–3651.
- Cohen, J. D. (1988). Noncentral chi-square: Some observations on recurrence. *The American Statistician* 42(2), 120–122.
- Dong, Y. (2018). Alternative assumptions to identify late in fuzzy regression discontinuity designs. *Oxford Bulletin of Economics and Statistics* 80(5), 1020–1027.

- Dudley, R. M. (2002). *Real Analysis and Probability*. Cambridge University Press.
- Frölich, M. and M. Huber (2019). Including covariates in the regression discontinuity design. *Journal of Business & Economic Statistics* 37(4), 736–748.
- Gallant, A. R. and H. White (1988). *A unified theory of estimation and inference for nonlinear dynamic models*. Blackwell.
- Gelman, A. and G. Imbens (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics* 37(3), 447–456.
- Giné, E. and R. Nickl (2015). *Mathematical foundations of infinite-dimensional statistical models*, Volume 40. Cambridge University Press.
- Hahn, J., P. Todd, and W. Van der Klaauw (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* 69(1), 201–209.
- Hall, P. (1991). Edgeworth expansions for nonparametric density estimators, with applications. *Statistics* 22(2), 215–232.
- Hansen, B. (2021). *Econometrics*. Princeton University Press.
- Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica* 71(4), 1161–1189.
- Hyytinen, A., J. Meriläinen, T. Saarimaa, O. Toivanen, and J. Tukiainen (2018). When does regression discontinuity design work? evidence from random election outcomes. *Quantitative Economics* 9(2), 1019–1051.
- Imbens, G. and K. Kalyanaraman (2011). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies* 79(3), 933–959.
- Imbens, G. W. and T. Lemieux (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics* 142(2), 615–635.
- Jales, H. and Z. Yu (2016). Identification and estimation using a density discontinuity approach. *Advances in Econometrics*. forthcoming.
- Jiang, J. and K. A. Doksum (2003). Empirical plug-in curve and surface estimates. In B. Lindquist and K. Doksum (Eds.), *Mathematical and Statistical Methods in Reliability*, pp. 433–453. World Scientific.

- Kitamura, Y. (2001). Asymptotic optimality of empirical likelihood for testing moment restrictions. *Econometrica* 69(6), 1661–1672.
- Kitamura, Y. (2006). Empirical likelihood methods in econometrics: theory and practice. *Cowles Foundation Discussion Paper*.
- Kitamura, Y., T. Otsu, and K. Evdokimov (2013). Robustness, infinitesimal neighborhoods, and moment restrictions. *Econometrica* 81(3), 1185–1201.
- Kosorok, M. R. (2007). *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media.
- Kreiß, A. and C. Rothe (2021). Inference in regression discontinuity designs with high-dimensional covariates. *arXiv preprint arXiv:2110.13725*.
- Lee, D. S. (2008). Randomized experiments from non-random selection in us house elections. *Journal of Econometrics* 142(2), 675–697.
- Ludwig, J. and D. L. Miller (2007). Does head start improve children’s life chances? evidence from a regression discontinuity design. *The Quarterly journal of economics* 122(1), 159–208.
- Ma, J. (2017). Second-order refinement of empirical likelihood ratio tests of nonlinear restrictions. *The Econometrics Journal* 20(1), 139–148.
- Ma, J., H. Jales, and Z. Yu (2019). Minimum contrast empirical likelihood inference of discontinuity in density. *Journal of Business & Economic Statistics*, DOI:10.1080/07350015.2019.1617155.
- Matsushita, Y. and T. Otsu (2013). Second-order refinement of empirical likelihood for testing overidentifying restrictions. *Econometric Theory* 29(02), 324–353.
- Negi, A. and J. M. Wooldridge (2014). Revisiting Regression Adjustment in Experiments with Heterogeneous Treatment Effects. *Econometric Reviews*.
- Newey, W. K. and R. J. Smith (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* 72(1), 219–255.
- Noack, C., T. Olma, and C. Rothe (2021). Flexible Covariate Adjustments in Regression Discontinuity Designs. *arXiv*.
- Noack, C. and C. Rothe (2019). Bias-aware inference in fuzzy regression discontinuity designs. *arXiv preprint arXiv:1906.04631*.

- Otsu, T. (2010). On Bahadur efficiency of empirical likelihood. *Journal of Econometrics* 157(2), 248–256.
- Otsu, T., K.-L. Xu, and Y. Matsushita (2013). Estimation and inference of discontinuity in density. *Journal of Business & Economic Statistics* 31(4), 507–524.
- Otsu, T., K.-L. Xu, and Y. Matsushita (2015). Empirical likelihood for regression discontinuity design. *Journal of Econometrics* 186(1), 94–112.
- Skovgaard, I. M. (1981). Transformation of an edgeworth expansion by a sequence of smooth functions. *Scandinavian Journal of Statistics*, 207–217.
- Wu, X. and Z. Ying (2011). An Empirical Likelihood Approach to Nonparametric Covariate Adjustment in Randomized Clinical Trials. *arXiv*.
- Xu, K.-L. (2017). Regression discontinuity with categorical outcomes. *Journal of Econometrics* 201(1), 1–18.
- Xu, K.-L. (2018). A semi-nonparametric estimator of regression discontinuity design with discrete duration outcomes. *Journal of Econometrics* 206(1), 258–278.
- Zhang, B. (2018). Empirical likelihood inference in randomized clinical trials. *Statistical Methods in Medical Research* 27(12), 3770–3784.

Table 1: Sharp RD with covariates: robust corrected EL (RCEL) inference using ROT and CCFT bandwidths (h) and CCFT Wald-type inference using the bandwidth with coverage optimal (CO) rate and $\rho := h/b = 1$, d_z = the number of covariates, p = the polynomial order, $1 - \tau$ = nominal coverage probability, n = sample size.

d_z	Methods	h	$1 - \tau = 0.99$			$1 - \tau = 0.95$			$1 - \tau = 0.90$		
			$n = 500$	1,000	2,000	500	1,000	2,000	500	1,000	2,000
1	RCEL, $p = 1$	ROT	.9805	.9825	.9840	.9265	.9430	.9450	.8750	.8795	.8900
	RCEL, $p = 1$	CCFT	.9825	.9850	.9865	.9350	.9420	.9505	.8765	.8905	.9000
	RCEL, $p = 2$	ROT	.9880	.9865	.9885	.9400	.9500	.9475	.8870	.8960	.8960
	RCEL, $p = 2$	CCFT	.9885	.9900	.9900	.9490	.9495	.9580	.8960	.8980	.9080
	CCFT, $p = 1$	CO, $\rho = 1$.9760	.9825	.9905	.9315	.9420	.9535	.8815	.9005	.9130
	CCFT, $p = 2$	CO, $\rho = 1$.9740	.9760	.9890	.9265	.9340	.9525	.8745	.8925	.9015
3	RCEL, $p = 1$	ROT	.9780	.9680	.9745	.9155	.9090	.9250	.8560	.8470	.8780
	RCEL, $p = 1$	CCFT	.9785	.9740	.9830	.9195	.9195	.9360	.8680	.8660	.8905
	RCEL, $p = 2$	ROT	.9840	.9765	.9825	.9325	.9290	.9395	.8730	.8715	.8835
	RCEL, $p = 2$	CCFT	.9870	.9790	.9865	.9450	.9310	.9430	.8925	.8735	.8925
	CCFT, $p = 1$	CO, $\rho = 1$.9670	.9767	.9860	.9130	.9320	.9500	.8580	.8695	.9015
	CCFT, $p = 2$	CO, $\rho = 1$.9640	.9750	.9850	.9130	.9240	.9430	.8520	.8695	.8985
5	RCEL, $p = 1$	ROT	.9645	.9720	.9765	.9035	.9145	.9125	.8525	.8515	.8510
	RCEL, $p = 1$	CCFT	.9730	.9690	.9820	.9150	.9125	.9200	.8665	.8515	.8495
	RCEL, $p = 2$	ROT	.9800	.9810	.9810	.9185	.9320	.9180	.8620	.8655	.8675
	RCEL, $p = 2$	CCFT	.9775	.9815	.9815	.9245	.9320	.9270	.8710	.8745	.8630
	CCFT, $p = 1$	CO, $\rho = 1$.9590	.9740	.9795	.8960	.9245	.9345	.8420	.8820	.8805
	CCFT, $p = 2$	CO, $\rho = 1$.9630	.9690	.9760	.9060	.9185	.9295	.8475	.8745	.8745

Figure 1: Coverage rates under covariate imbalance $\delta = \text{perturbation}$, $n = \text{sample size}$.

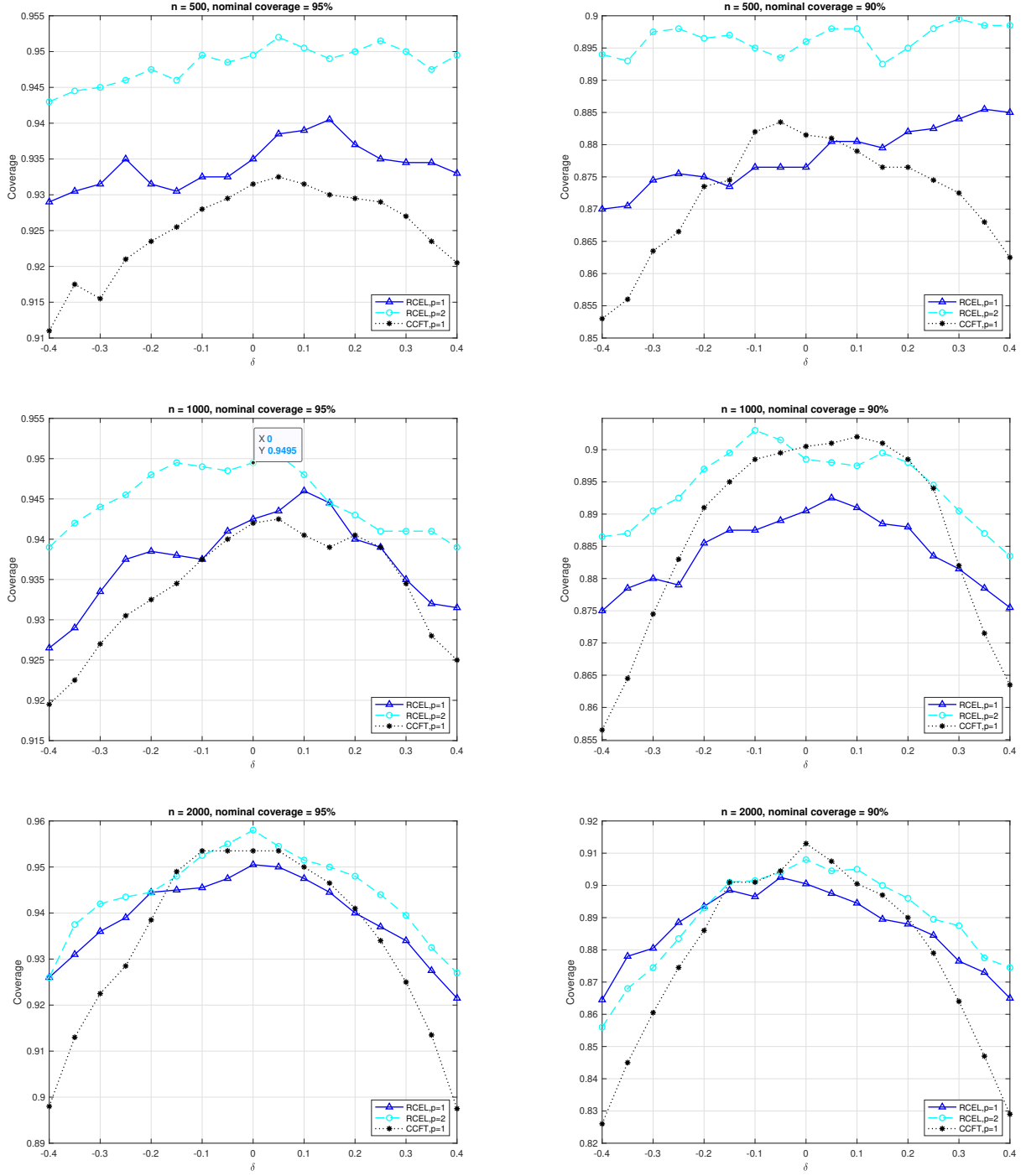
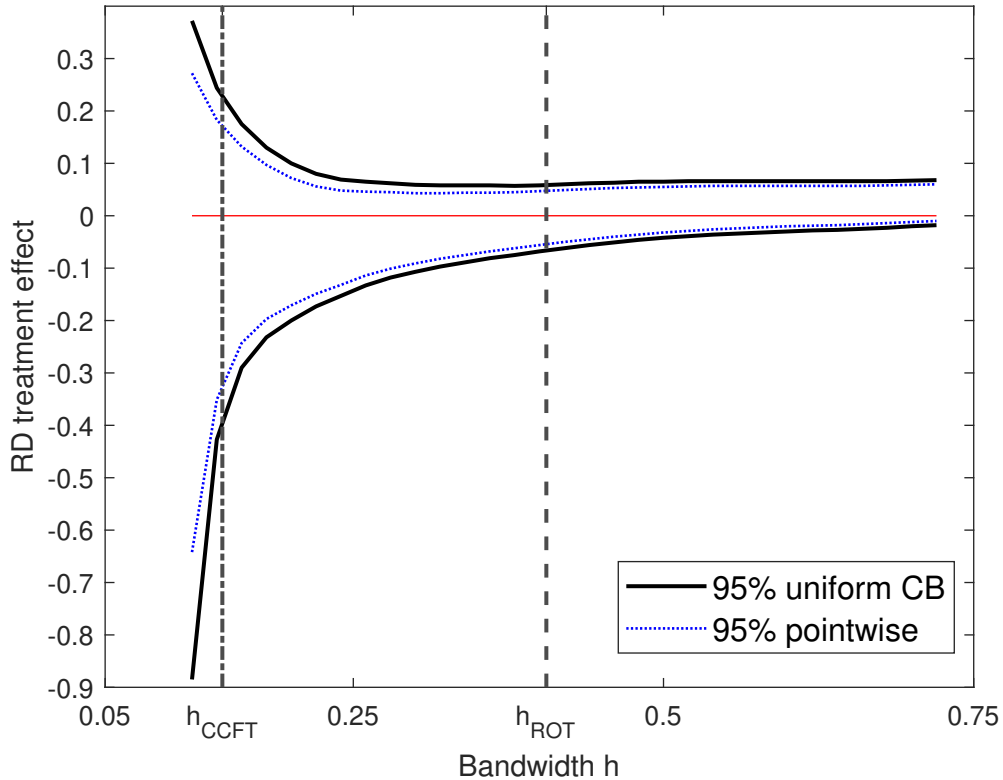


Table 2: Incumbency advantage in Finnish municipal election: $\hat{\vartheta}_0 = \text{RD LATE estimator}$, $h = \text{bandwidth}$.

	Methods	Bandwidth selector	$\hat{\vartheta}_0$	p -value	95% CI	CI length	h
RD with covariates $n = 154,543$	RCEL, $p = 1$	ROT	-.003	.896	[-.054, .047]	.101	.406
	RCEL, $p = 1$	CCFT, rescaled	-.066	.587	[-.319, .171]	.490	.144
	RCEL, $p = 1$	CCFT	-.007	.807	[-.059, .046]	.105	.391
	CCFT, $p = 1$	CO	.024	.161	[-.010, .059]	.069	.391
	CCFT, $p = 1$	CO, $\rho = 1$	-.012	.671	[-.068, .044]	.111	.391
	RCEL, $p = 2$	ROT	.023	.234	[-.014, .060]	.074	1.098
	RCEL, $p = 2$	CCFT, rescaled	-.045	.346	[-.140, .048]	.188	.391
	RCEL, $p = 2$	CCFT	.023	.229	[-.014, .060]	.074	1.103
	CCFT, $p = 2$	CO	.033	.028	[.003, .062]	.065	1.103
	CCFT, $p = 2$	CO, $\rho = 1$.017	.402	[-.023, .057]	.801	1.103
Experiment benchmark (Hyytinen et al., 2018) $n = 1,351$			-.010	.516	[-.060, .040]	.100	

Figure 2: Robust corrected EL inference applied to Finnish municipal election: 95% uniform (solid) and pointwise (dotted) confidence bands over the bandwidth range $[0.12, 0.72]$. $p = 1$. Corrected critical value $= (2.413)^2$. Vertical lines indicate two bandwidth choices in Table 2: ROT bandwidth $h_{\text{ROT}} = 0.406$ and rescaled CCFT bandwidth $h_{\text{CCFT}} = 0.144$.



Appendix A Proofs of Theorems 1 and 2

For a sequence of classes of \mathbb{R} -valued functions \mathfrak{F}_n defined on \mathcal{S} (a compact set in a finite-dimensional Euclidean space), let $\|f\|_{Q,2} := (\int f^2 dQ)^{1/2}$ and $N(\varepsilon, \mathfrak{F}_n, \|\cdot\|_{Q,2})$ denote the ε -covering number, i.e., the smallest integer m such that there are m balls of radius $\varepsilon > 0$ (with respect to $\|\cdot\|_{Q,2}$) centered at points in \mathfrak{F}_n whose union covers \mathfrak{F}_n . A function $F_{\mathfrak{F}_n} : \mathcal{S} \rightarrow \mathbb{R}_+$ is an envelope of \mathfrak{F}_n if $\sup_{f \in \mathfrak{F}_n} |f| \leq F_{\mathfrak{F}_n}$. We say that \mathfrak{F}_n is a (uniform) Vapnik–Chervonenkis-type (VC-type) class with respect to the envelope $F_{\mathfrak{F}_n}$ (see, e.g., Chernozhukov et al., 2014b, Definition 2.1) if there exist some positive constants (VC characteristics) $A_{\mathfrak{F}_n} \geq e$ and $V_{\mathfrak{F}_n} > 1$ that are independent of the sample size n such that $\sup_{Q \in \mathcal{Q}_{\mathcal{S}}^{\text{fd}}} N(\varepsilon \|F_{\mathfrak{F}_n}\|_{Q,2}, \mathfrak{F}_n, \|\cdot\|_{Q,2}) \leq (A_{\mathfrak{F}_n}/\varepsilon)^{V_{\mathfrak{F}_n}} \forall \varepsilon \in (0, 1]$ where $\mathcal{Q}_{\mathcal{S}}^{\text{fd}}$ denotes the collection of all finitely discrete probability measures on \mathcal{S} . \lesssim denotes an inequality up to a universal constant that does not depend on the sample size or the population. For a real sequence $\{a_n\}_{n=1}^\infty$, we denote $b_n \propto a_n$ if $b_n = c \cdot a_n$ for some constant $c > 0$. Proofs of all lemmas are in the supplement available at ruc-econ.github.io/supplement_RD.pdf.

Lemma 1. *Let V denote a random variable and $\{V_1, \dots, V_n\}$ are i.i.d. copies of V . Let $\mathbb{B}(0)$ denote an open neighborhood of 0. Suppose that (\underline{h}, \bar{h}) satisfy $\bar{h} = o(1)$. $\forall (s, k) \in \{-, +\} \times \mathbb{N}$, the following results hold uniformly in $h \in \mathbb{H}$: (a) if g_V is Lipschitz continuous on $\mathbb{B}(0)$, for $k \geq 2$, $\mathbb{E}[h^{-1}W_{p;s}^k V] = \psi_{V,s}\omega_{p;s}^{0,k} + O(\bar{h})$; (b) if g_V is $(p+1)$ -times continuously differentiable with Lipschitz continuous $g_V^{(p+1)}$ on $\mathbb{B}(0)$, $\mathbb{E}[h^{-1}W_{p;s} V] = \psi_{V,s} + \omega_{p;s}^{p+1,1}\psi_{V,s}^{(p+1)}h^{p+1}/(p+1)! + O(\bar{h}^{p+2})$ and $\mathbb{E}[h^{-1}W_{p+1;s} V] = \psi_{V,s} + O(\bar{h}^{p+2})$; (c) if $g_{|V|^r}$ is bounded on $\mathbb{B}(0)$ for some integer $r > 2$, $(nh)^{-1/2} \sum_i (W_{p;s,i}^k V_i - \mathbb{E}[W_{p;s}^k V]) = O_p(\sqrt{\log(n)} + \log(n) \left\{ (n\bar{h})^{1/r} / (n\bar{h})^{1/2} \right\})$.*

Let $0_{J \times K}$ denote the $J \times K$ matrix in which all elements are zeros. Let I_K denote the $K \times K$ identity matrix. Let 0_J denote the J -dimensional vector in which all elements are zeros. Let $G_i := \partial U_i(\theta) / \partial \theta^\top = \begin{bmatrix} G_{0,i} & G_{\dagger,i} \end{bmatrix}$, where $G_{0,i} := \begin{bmatrix} D_i & 0_{d_z}^\top \end{bmatrix}^\top$ and $G_{\dagger,i} := I_{d_u}$. (G, G_0, G_\dagger) are defined by the same formulae with D_i replaced by D . Denote $\mathcal{U}_i(\theta) := W_{p,i} \otimes U_i(\theta)$ and $\hat{\mathcal{U}}_i := W_{p,i} \otimes U_i(\hat{\vartheta}_p)$. Let $\mathcal{U}_i := W_{p,i} \otimes U_i$, $\mathcal{G}_i := W_{p,i} \otimes G_i$, $\mathcal{G}_{0,i} := W_{p,i} \otimes G_{0,i}$ and $\mathcal{G}_{\dagger,i} := W_{p,i} \otimes G_{\dagger,i}$. $(\mathcal{U}, \mathcal{G}, \mathcal{G}_0, \mathcal{G}_\dagger)$ are defined similarly. Let $\mathcal{D}_s := W_{p;s} D$. Denote $\Delta_s := \mathbb{E}[h^{-1}W_{p;s}]$ and $\Delta_A := \mathbb{E}[h^{-1}A]$ for a random variable/vector/matrix A . Let $\bar{\Delta}_{\mathcal{U}\mathcal{U}^\top} := (nh)^{-1} \sum_i \mathcal{U}_i \mathcal{U}_i^\top$, $\hat{\Delta}_{\mathcal{U}\mathcal{U}^\top} := (nh)^{-1} \sum_i \hat{\mathcal{U}}_i \hat{\mathcal{U}}_i^\top$ and $\bar{\Delta}_{\mathcal{G}} := (nh)^{-1} \sum_i \mathcal{G}_i$. Let $\bar{\mathcal{U}} := (nh)^{-1/2} \sum_i \mathcal{U}_i$ and $\hat{\mathcal{U}} := (nh)^{-1/2} \sum_i \hat{\mathcal{U}}_i$.

Denote $S(\lambda, \theta) := 2 \sum_i \log(1 + \lambda^\top \mathcal{U}_i(\theta))$ and $\tilde{\vartheta}_p := \tilde{\vartheta}_p(\vartheta_0)$. Note that the dual form of the EL criterion function is $\ell_p(\theta | h) = \sup_{\lambda \in \mathcal{L}(\theta)} S(\lambda, \theta)$, where $\mathcal{L}(\theta) := \{\lambda \in \mathbb{R}^{2d_u} : \lambda^\top \mathcal{U}_i(\theta) > -1, \forall i\}$. \max_i is understood as $\max_{1 \leq i \leq n}$. For square matrices A and B , $\text{diag}(A, B)$ denotes the block diagonal matrix. $\|A\|$ is understood as the spectral norm of A and $\varrho_{\min}(A)$ denotes the smallest eigenvalue of A . In the remaining proofs in

Appendix A, whenever applied to quantities that depend on h , $O_p(\cdot)$, $o_p(\cdot)$, $O(\cdot)$ and $o(\cdot)$ notations are understood as being uniform in $h \in \mathbb{H}$. For notational simplicity, denote $\bar{n} := n\bar{h}$, $\underline{n} := n\underline{h}$, $\hat{\eta}_p := \hat{\vartheta}_p - \vartheta$ and $\tilde{\eta}_p := \tilde{\vartheta}_p - \vartheta$. “With probability approaching one” is abbreviated as “wpa1”. The proof of the following lemma follows the arguments in [Newey and Smith \(2004\)](#).

Lemma 2. *Suppose that Assumptions 1, 2 and 3 hold. Suppose that (\underline{h}, \bar{h}) satisfy $n\bar{h}^{-2p+3} = O(1)$ and $\log(n) (\bar{n}^{1/12}/\underline{n}^{1/2}) = o(1)$. Then, the following results hold uniformly in $h \in \mathbb{H}$: (a) $\sqrt{nh}\hat{\eta}_p = O_p(\sqrt{\log(n)})$; (b) $\hat{\lambda}_p := \arg\max_{\lambda \in \mathcal{L}(\hat{\vartheta}_p)} S(\lambda, \hat{\vartheta}_p)$ exists wpa1 and $\sqrt{nh}\hat{\lambda}_p = O_p(\sqrt{\log(n)})$; (c) $\sqrt{nh}\tilde{\eta}_p = O_p(\sqrt{\log(n)})$; (d) $\tilde{\lambda}_p := \arg\max_{\lambda \in \mathcal{L}(\vartheta_0, \tilde{\vartheta}_p)} S(\lambda, \vartheta_0, \tilde{\vartheta}_p)$ exists wpa1 and $\sqrt{nh}\tilde{\lambda}_p = O_p(\sqrt{\log(n)})$.*

Denote $\mathbf{O} := (\Delta_{\mathcal{G}}^\top \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1} \Delta_{\mathcal{G}})^{-1}$, $\mathbf{N} := \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1} \Delta_{\mathcal{G}} \mathbf{O}$ and $\mathbf{Q} := \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1} - \mathbf{N} \Delta_{\mathcal{G}}^\top \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1}$. Let $(\mathbf{O}_\dagger, \mathbf{N}_\dagger, \mathbf{Q}_\dagger)$ be defined by the same formulae with $\Delta_{\mathcal{G}}$ replaced by $\Delta_{\mathcal{G}_\dagger}$.

Lemma 3. *Suppose that the same assumptions as Lemma 2 hold. Then, the following results hold uniformly in $h \in \mathbb{H}$: (a) $\sqrt{nh}(\hat{\lambda}_p^\top, \hat{\eta}_p^\top) = \bar{\mathbf{U}}^\top \begin{bmatrix} \mathbf{Q} & \mathbf{N} \end{bmatrix} + O_p(v_n^\dagger)$; (b) $\sqrt{nh}(\tilde{\lambda}_p^\top, \tilde{\eta}_p^\top) = \bar{\mathbf{U}}^\top \begin{bmatrix} \mathbf{Q}_\dagger & \mathbf{N}_\dagger \end{bmatrix} + O_p(v_n^\dagger)$, where $v_n^\dagger := \log(n)/\sqrt{\underline{n}} + \log(n)^{3/2}(\bar{n}^{1/6}/\underline{n})$.*

Proof of Theorem 1. Let $\mathcal{M}_s := W_{p;s}(M - \vartheta_1)$, $\mathcal{Z}_s := W_{p;s}(Z - \vartheta_2)$, $\bar{\mathcal{M}}_s := (nh)^{-1/2} \sum_i W_{p;s,i}(M_i - \vartheta_1)$, $\bar{\mathcal{Z}}_s := (nh)^{-1/2} \sum_i W_{p;s,i}(Z_i - \vartheta_2)$ and $\gamma_\Delta := (\Delta_{\mathcal{Z}_+ \mathcal{Z}_+} / \Delta_+^2 + \Delta_{\mathcal{Z}_- \mathcal{Z}_-} / \Delta_-^2)^{-1} (\Delta_{\mathcal{Z}_+ \mathcal{M}_+} / \Delta_+^2 + \Delta_{\mathcal{Z}_- \mathcal{M}_-} / \Delta_-^2)$. Also denote $\mathcal{U}_s := W_{p;s}U$, $\bar{\mathcal{U}}_s := (nh)^{-1/2} \sum_i W_{p;s,i}U_i$, $\Phi_{00} := \Delta_{\mathcal{G}_0}^\top \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1} \Delta_{\mathcal{G}_0}$, $\Phi_{0\dagger} := \Delta_{\mathcal{G}_0}^\top \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1} \Delta_{\mathcal{G}_\dagger}$, $\Phi_{\dagger 0} := \Phi_{0\dagger}^\top$, $\Phi_{\dagger\dagger} := \Delta_{\mathcal{G}_\dagger}^\top \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1} \Delta_{\mathcal{G}_\dagger}$ and $\Phi_\pm := \Delta_{\mathcal{U}_+ \mathcal{U}_+^\top} / \Delta_+^2 + \Delta_{\mathcal{U}_- \mathcal{U}_-^\top} / \Delta_-^2$. Then we have

$$\begin{aligned} \Sigma_\Delta &:= (\mathbf{e}_{d_u,1}^\top \Phi_\pm^{-1} \mathbf{e}_{d_u,1})^{-1} = \left(\Delta_{\mathcal{M}_+^2} / \Delta_+^2 + \Delta_{\mathcal{M}_-^2} / \Delta_-^2 \right) - \left(\Delta_{\mathcal{M}_+ \mathcal{Z}_+^\top} / \Delta_+^2 + \Delta_{\mathcal{M}_- \mathcal{Z}_-^\top} / \Delta_-^2 \right) \gamma_\Delta \\ &= \Delta_{(\mathcal{M}_+ - \mathcal{Z}_+^\top \gamma_\Delta)^2} / \Delta_+^2 + \Delta_{(\mathcal{M}_- - \mathcal{Z}_-^\top \gamma_\Delta)^2} / \Delta_-^2, \end{aligned} \quad (11)$$

where the second equality follows from writing Φ_\pm as a block matrix and inverting and the third equality follows from simple algebra. And similarly,

$$\mathbf{e}_{d_u,1}^\top \Phi_\pm^{-1} (\bar{\mathcal{U}}_+ / \Delta_+ - \bar{\mathcal{U}}_- / \Delta_-) = \left\{ (\bar{\mathcal{M}}_+ / \Delta_+ - \bar{\mathcal{M}}_- / \Delta_-) - (\bar{\mathcal{Z}}_+ / \Delta_+ - \bar{\mathcal{Z}}_- / \Delta_-)^\top \gamma_\Delta \right\} / \Sigma_\Delta. \quad (12)$$

By simple algebra, $(\Phi_{00} - \Phi_{0\dagger} \Phi_{\dagger\dagger}^{-1} \Phi_{\dagger 0})^{-1} = \Sigma_\Delta / (\Delta_{\mathcal{D}_+} / \Delta_+ - \Delta_{\mathcal{D}_-} / \Delta_-)^2$. Then, by this result, writing $\Delta_{\mathcal{G}}^\top \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1} \Delta_{\mathcal{G}}$ as a block matrix and inverting,

$$(\Delta_{\mathcal{G}}^\top \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1} \Delta_{\mathcal{G}})^{-1} = (\Delta_{\mathcal{D}_+} / \Delta_+ - \Delta_{\mathcal{D}_-} / \Delta_-)^{-2} \begin{bmatrix} \Sigma_\Delta & -\Sigma_\Delta \Phi_{0\dagger} \Phi_{\dagger\dagger}^{-1} \\ -\Phi_{\dagger\dagger}^{-1} \Phi_{\dagger 0} \Sigma_\Delta & \Phi_{\dagger\dagger}^{-1} + \Phi_{\dagger\dagger}^{-1} \Phi_{\dagger 0} \Sigma_\Delta \Phi_{0\dagger} \Phi_{\dagger\dagger}^{-1} \end{bmatrix}. \quad (13)$$

By straightforward algebraic calculation,

$$\left(\Delta_{\mathcal{G}_0}^\top - \Phi_{0\dagger}\Phi_{\dagger\dagger}^{-1}\Delta_{\mathcal{G}_\dagger}^\top\right)\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1}\bar{\mathcal{U}} = (\Delta_{\mathcal{D}_+}/\Delta_+ - \Delta_{\mathcal{D}_-}/\Delta_-)e_{d_u,1}^\top\Phi_\pm^{-1}(\bar{\mathcal{U}}_+/\Delta_+ - \bar{\mathcal{U}}_-/\Delta_-). \quad (14)$$

Then, by this result, (12) and (13),

$$\begin{aligned} e_{d_\vartheta,1}^\top(\Delta_{\mathcal{G}}^\top\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1}\Delta_{\mathcal{G}})^{-1}(\Delta_{\mathcal{G}}^\top\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1}\bar{\mathcal{U}}) &= (\Delta_{\mathcal{D}_+}/\Delta_+ - \Delta_{\mathcal{D}_-}/\Delta_-)^{-2}\Sigma_\Delta\left(\Delta_{\mathcal{G}_0}^\top - \Phi_{0\dagger}\Phi_{\dagger\dagger}^{-1}\Delta_{\mathcal{G}_\dagger}^\top\right)\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1}\bar{\mathcal{U}} \\ &= \frac{(\bar{\mathcal{M}}_+/\Delta_+ - \bar{\mathcal{M}}_-/\Delta_-) - (\bar{\mathcal{Z}}_+/\Delta_+ - \bar{\mathcal{Z}}_-/\Delta_-)^\top\gamma_\Delta}{\Delta_{\mathcal{D}_+}/\Delta_+ - \Delta_{\mathcal{D}_-}/\Delta_-}. \end{aligned} \quad (15)$$

It follows from Lemma 1 with $\underline{h} = \bar{h} = h$ that $\gamma_\Delta = \gamma_{\text{adj}} + O(h)$, $\Delta_{\mathcal{D}_s} = \psi_{D,s} + O(h^{p+1})$ and $\Delta_s = \varphi + O(h^{p+1})$ $\forall s$. By Lemma 1 with $\underline{h} = \bar{h} = h$ and Markov's inequality, $\bar{\mathcal{M}}_s = O_p(1)$ and $\bar{\mathcal{Z}}_s = O_p(1)$ $\forall s$. Then, it follows from Lemma 3 with $\underline{h} = \bar{h} = h$ and these results that

$$\begin{aligned} \sqrt{nh}\left(\hat{\vartheta}_{p,0} - \vartheta_0\right) &= e_{d_\vartheta,1}^\top(\Delta_{\mathcal{G}}^\top\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1}\Delta_{\mathcal{G}})^{-1}(\Delta_{\mathcal{G}}^\top\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1}\bar{\mathcal{U}}) + o_p(1) \\ &= \left\{(\bar{\mathcal{M}}_+ - \bar{\mathcal{Z}}_+\gamma_{\text{adj}}) - (\bar{\mathcal{M}}_- - \bar{\mathcal{Z}}_-\gamma_{\text{adj}})\right\} / (\psi_{D,+} - \psi_{D,-}) + o_p(1) \\ &= (nh)^{-1/2} \sum_i (W_{p;+,i} - W_{p;- ,i})(\epsilon_i - \mu_\epsilon) / (\psi_{D,+} - \psi_{D,-}) + o_p(1), \end{aligned} \quad (16)$$

where $\epsilon_i := M_i - Z_i^\top\gamma_{\text{adj}}$. Let $\mathcal{E}_i := (W_{p;+,i} - W_{p;- ,i})(\epsilon_i - \mu_\epsilon)$ and \mathcal{E} be defined similarly. Then,

$$\sqrt{nh}\left(\hat{\vartheta}_{p,0} - \vartheta_0 - \frac{\Delta_{\mathcal{E}}}{\psi_{D,+} - \psi_{D,-}}\right) = \sum_i \left(\frac{\mathcal{E}_i}{\sqrt{nh}} - \mathbb{E}\left[\frac{\mathcal{E}}{\sqrt{nh}}\right]\right) / (\psi_{D,+} - \psi_{D,-}) + o_p(1) \quad (17)$$

follows from subtracting both sides of (16) by $\sqrt{nh}\Delta_{\mathcal{E}}/(\psi_{D,+} - \psi_{D,-})$. By Lemma 1 with $\underline{h} = \bar{h} = h$, $\Delta_{\mathcal{E}} = (\zeta_{p;+} - \zeta_{p;-})h^{p+1}/(p+1)! + O(h^{p+2})$ and $\Delta_{\mathcal{E}^2} = \omega_{p;+}^{0,2} + \sum_{s \in \{-,+\}} \psi_{(\epsilon - \mu_\epsilon)^2,s} + O(h)$. It follows from simple algebraic calculations that $\sum_{s \in \{-,+\}} \psi_{(\epsilon - \mu_\epsilon)^2,s} = \sigma_{\text{adj}}^2\varphi$. Then, $\text{Var}\left[\mathcal{E}/\sqrt{h}\right] = \Delta_{\mathcal{E}^2} - h\Delta_{\mathcal{E}}^2 = \omega_{p;+}^{0,2} + \sigma_{\text{adj}}^2\varphi + O(h)$. By LIE and change of variables, $\Delta_{\mathcal{E}^4} = O(1)$. Then,

$$\sum_i \mathbb{E}\left[\left(\frac{\mathcal{E}_i/\sqrt{nh} - \mathbb{E}\left[\mathcal{E}/\sqrt{nh}\right]}{\sqrt{\text{Var}\left[\mathcal{E}/\sqrt{h}\right]}}\right)^4\right] \lesssim \frac{\Delta_{\mathcal{E}^4} + h^3\Delta_{\mathcal{E}}^4}{(nh)\left(\text{Var}\left[\mathcal{E}/\sqrt{h}\right]\right)^2} = O\left((nh)^{-1}\right), \quad (18)$$

where the inequality follows from Loève's c_r inequality and the equality follows from $\text{Var}\left[\mathcal{E}/\sqrt{h}\right] = \omega_{p;+}^{0,2} + \sigma_{\text{adj}}^2\varphi + O(h)$, $\Delta_{\mathcal{E}^4} = O(1)$ and $\Delta_{\mathcal{E}} = O(h^{p+1})$. (18) verifies Lyapunov's condition. By Lyapunov's central limit theorem, $\sum_i \left(\mathcal{E}_i/\sqrt{nh} - \mathbb{E}\left[\mathcal{E}/\sqrt{nh}\right]\right) / \sqrt{\text{Var}\left[\mathcal{E}/\sqrt{h}\right]} \rightarrow_d \text{N}(0,1)$. The conclusion follows from this result, (17), $\text{Var}\left[\mathcal{E}/\sqrt{h}\right] = \omega_{p;+}^{0,2} + \sigma_{\text{adj}}^2\varphi + O(h)$ and Slutsky's lemma. \blacksquare

The following lemma shows that $\{LR_p(\vartheta_0 | h) : h \in \mathbb{H}\}$ can be approximated by the square of an empirical process indexed by $h \in \mathbb{H}$. Denote $\mathbb{P}_n^T f := n^{-1} \sum_i f(T_i, X_i)$, $\mathbb{P}^T f := \mathbb{E}[f(T, X)]$ and $\mathbb{G}_n^T := \sqrt{n}(\mathbb{P}_n^T - \mathbb{P}^T)$, where $T_i := (Y_i, D_i, Z_i^\top)^\top$ (similarly, $T := (Y, D, Z^\top)^\top$). Denote $\|F\|_{\mathbb{P}^T, r} := (\mathbb{P}^T |F|^r)^{1/r}$. Let $\xi(x) := \mathbb{E}[(\epsilon - \mu_\epsilon)^2 | |X| = x]$ and $q(\cdot | h)$ be defined by $q(T_i, X_i | h) := h^{-1/2} \mathcal{E}_i / \sqrt{\xi(|X_i|) f_{|X|}(|X_i|) \omega_{p;+}^{0,2}}$, where $f_{|X|}$ denotes the PDF of $|X|$.

Lemma 4. *Suppose that the assumptions of Lemma 2 hold. Then, uniformly in $h \in \mathbb{H}$, $LR_p(\vartheta_0 | h) = \{\mathbb{G}_n^T q(\cdot | h)\}^2 + O_p(\log(n) \bar{h} + \log(n)^{3/2} (\bar{n}^{1/12} / \underline{n}^{1/2}))$.*

Proof of Theorem 2. Denote $Z_{\Omega_\pm} := \sup_{f \in \Omega_\pm} \mathbb{G}_n^T f = \|\mathbb{G}_n^T\|_{\Omega}$. Since F_Ω is also an envelope of $\Omega_\pm := \Omega \cup (-\Omega)$ ($-\Omega := \{-f : f \in \Omega\}$) and the covering number of Ω_\pm is at most twice that of Ω , Ω_\pm is also VC-type with respect to F_Ω . By standard calculus calculations (see, e.g., the proof of Chernozhukov et al., 2014b, Corollary 5.1) and Chernozhukov et al. (2014b, Lemma 2.1), there exists a zero-mean Gaussian process $\{G^T(f) : f \in \Omega_\pm\}$ that is a tight random element in $\ell^\infty(\Omega_\pm)$ and also satisfies $\mathbb{E}[G^T(f) G^T(g)] = \text{Cov}[f(T, X), g(T, X)]$, $\forall f, g \in \Omega_\pm$.²⁷ By Giné and Nickl (2015, Theorem 3.7.28), almost surely the sample paths $\Omega_\pm \ni f \mapsto G^T(f)$ are prelinear and therefore, almost surely, $\forall f \in \Omega$, $G^T(f) + G^T(-f) = 0$, and $\sup_{f \in \Omega_\pm} G^T(f) = \|G^T\|_{\Omega}$. Let $\bar{F}_G(h) := G^T(q(\cdot | h))$ and therefore, the zero-mean Gaussian process $\{\bar{F}_G(h) : h \in \mathbb{H}\}$ is a tight random element in $\ell^\infty(\mathbb{H})$ and has the covariance structure $\mathbb{E}[\bar{F}_G(h) \bar{F}_G(h')] = \text{Cov}[q(T, X | h), q(T, X | h')]$, $\forall (h, h') \in \mathbb{H}^2$. By definition, $\|\bar{F}_G\|_{\mathbb{H}} = \|G^T\|_{\Omega}$. By change of variables and LIE, $\sup_{f \in \Omega} \mathbb{P}^T |f|^3 = \sup_{h \in \mathbb{H}} \mathbb{E}[|q(T, X | h)|^3] \lesssim \underline{h}^{-1/2}$ and similarly $\sup_{f \in \Omega} \mathbb{P}^T |f|^4 \lesssim \underline{h}^{-1}$. Also, $\mathbb{P}^T F_\Omega^{12} \lesssim \bar{h} / \underline{h}^6$. By Chernozhukov et al. (2016, Theorem 2.1) with $B(f) = 0$, $\mathcal{F} = \Omega_\pm$, $q = 12$, $K_n = \log(n)$, $\sigma = 1$, $b \lesssim \underline{h}^{-1/2}$ and $\gamma = \log(n)^{-1}$, there exists $\tilde{Z}_{\Omega_\pm} =_d \sup_{f \in \Omega_\pm} G^T(f) = \|G^T\|_{\Omega}$ which satisfies $Z_{\Omega_\pm} - \tilde{Z}_{\Omega_\pm} = O_p(v_n^*)$, where “ $=_d$ ” is understood as being equal in distribution and $v_n^* := \left\{ \log(n) (\log(n) n)^{1/12} \right\} / \underline{n}^{1/2} + \log(n) / \underline{n}^{1/6}$. By Dudley’s entropy integral bound (Giné and Nickl, 2015, Theorem 2.3.7), Chen and Kato (2020, Lemma A.2) and standard calculus calculations (see, e.g., calculations in the proof of Chernozhukov et al., 2014b, Corollary 5.1),

$$\begin{aligned} \mathbb{E}[\|G^T\|_{\Omega}] &\lesssim \int_0^{\sigma_\Omega \vee n^{-1/2} \|F_\Omega\|_{\mathbb{P}^T, 2}} \sqrt{1 + \log\left(N\left(\varepsilon, \Omega, \|\cdot\|_{\mathbb{P}^T, 2}\right)\right)} d\varepsilon \\ &\lesssim \left(\sigma_\Omega \vee n^{-1/2} \|F_\Omega\|_{\mathbb{P}^T, 2}\right) \sqrt{\log(n)} = O\left(\sqrt{\log(n)}\right). \end{aligned} \quad (19)$$

By Lemma 4, $\sup_{h \in \mathbb{H}} LR_p(\vartheta_0 | h) = \|\mathbb{G}_n^T\|_{\Omega}^2 + O_p(\log(n) \bar{h} + \log(n)^{3/2} (\bar{n}^{1/12} / \underline{n}^{1/2}))$. By (19) and the fact

²⁷Tightness of Ω is equivalent to the condition that Ω endowed with the intrinsic pseudo metric $(f, g) \mapsto \|f - g\|_{\mathbb{P}^T, 2} := (\mathbb{P}^T (f - g)^2)^{1/2}$ is totally bounded and almost surely the sample paths $f \mapsto G^T(f)$ are uniformly continuous with respect to the intrinsic pseudo metric. By Kosorok (2007, Lemmas 7.2 and 7.4), $\{G^T(f) : f \in \Omega\}$ is also separable as a stochastic process.

that $\mathbb{E} \left[\|G_n^T\|_{\Omega} \right] = O \left(\sqrt{\log(n)} \right)$, we have $Z_{\Omega_{\pm}}^2 - \tilde{Z}_{\Omega_{\pm}}^2 = O_p \left(\sqrt{\log(n)} v_n^* \right)$. Therefore, $\sup_{h \in \mathbb{H}} LR_p(\vartheta_0 | h) = \tilde{Z}_{\Omega_{\pm}}^2 + O_p \left(\sqrt{\log(n)} v_n^* + \log(n) \bar{h} \right)$. By [Dudley \(2002, Theorem 9.2.2\)](#) and $\sup_{h \in \mathbb{H}} LR_p(\vartheta_0 | h) - \tilde{Z}_{\Omega_{\pm}}^2 = o_p \left(\log(n)^{-1} \right)$, there exists a null sequence $\varepsilon_n \downarrow 0$ such that $\Pr \left[\left| \sup_{h \in \mathbb{H}} LR_p(\vartheta_0 | h) - \tilde{Z}_{\Omega_{\pm}}^2 \right| > \varepsilon_n / \log(n) \right] \leq \varepsilon_n$ and by the fact that $(a - b)^2 \leq |a^2 - b^2| \forall a, b \geq 0$,

$$\Pr \left[\left| \sqrt{\sup_{h \in \mathbb{H}} LR_p(\vartheta_0 | h)} - \tilde{Z}_{\Omega_{\pm}} \right| > \sqrt{\varepsilon_n / \log(n)} \right] \leq \varepsilon_n. \quad (20)$$

It is easy to check that for random variables (V, W) and constants $r_1, r_2, t > 0$ such that $\Pr[|V - W| > r_1] \leq r_2$,

$$|\Pr[V \leq t] - \Pr[W \leq t]| \leq \Pr[|W - t| \leq r_1] + r_2. \quad (21)$$

Then, by (20) and (21),

$$\begin{aligned} \left| \Pr \left[\sup_{h \in \mathbb{H}} LR_p(\vartheta_0 | h) \leq z_{1-\tau} (\bar{h}/\underline{h})^2 \right] - \Pr \left[\tilde{Z}_{\Omega_{\pm}}^2 \leq z_{1-\tau} (\bar{h}/\underline{h})^2 \right] \right| \\ \leq \Pr \left[\left| \tilde{Z}_{\Omega_{\pm}} - z_{1-\tau} (\bar{h}/\underline{h}) \right| \leq \sqrt{\varepsilon_n / \log(n)} \right] + \varepsilon_n. \end{aligned} \quad (22)$$

Since $\tilde{Z}_{\Omega_{\pm}} =_d \|G^T\|_{\Omega}$ and $\{G^T(f) : f \in \Omega\}$ is a centered Gaussian process with $\mathbb{E} \left[G^T(f)^2 \right] = 1, \forall f$, by using the Gaussian anti-concentration inequality ([Chernozhukov et al., 2014a](#), Corollary 2.1) and (19),

$$\Pr \left[\left| \tilde{Z}_{\Omega_{\pm}} - z_{1-\tau} (\bar{h}/\underline{h}) \right| \leq \sqrt{\varepsilon_n / \log(n)} \right] \lesssim \sqrt{\varepsilon_n / \log(n)} \left(\mathbb{E} \left[\|G^T\|_{\Omega} \right] + 1 \right) = O(\sqrt{\varepsilon_n}). \quad (23)$$

It then follows from (22) and (23) that $\Pr \left[LR_p(\vartheta_0 | h) \leq z_{1-\tau} (\bar{h}/\underline{h})^2, \forall h \in \mathbb{H} \right] = \Pr \left[\|\bar{\Gamma}_G\|_{\mathbb{H}} \leq z_{1-\tau} (\bar{h}/\underline{h}) \right] + o(1)$. Let N be an $N(0, 1)$ random variable that is independent of $\{\bar{\Gamma}_G(h) : h \in \mathbb{H}\}$. Let $\tilde{\Gamma}_G(h) := \bar{\Gamma}_G(h) + \mathbb{E}[q(T, X | h)] \cdot N$. By change of variables, $\sup_{h \in \mathbb{H}} |\mathbb{E}[q(T, X | h)]| = O(\bar{h}^{1/2})$. $\{\tilde{\Gamma}_G(h) : h \in \mathbb{H}\}$ is a zero-mean Gaussian process which satisfies $\|\tilde{\Gamma}_G\|_{\mathbb{H}} = \|\bar{\Gamma}_G\|_{\mathbb{H}} + O_p(\bar{h}^{1/2})$ and has the covariance structure $\mathbb{E}[\tilde{\Gamma}_G(h) \tilde{\Gamma}_G(h')] = \mathbb{E}[q(T, X | h) q(T, X | h')]$, $\forall (h, h') \in \mathbb{H}^2$. By LIE and change of variables, $\mathbb{E}[q(T, X | h) q(T, X | h')] = \sqrt{h/h'} \int_0^\infty \mathcal{K}_{p;+}(z) \mathcal{K}_{p;+}((h/h')z) dz / \omega_{p;+}^{0,2}$. Let $\Gamma_G(s) := \tilde{\Gamma}_G(s \cdot \underline{h})$, $s \in [1, \bar{h}/\underline{h}]$. Then it is easy to see that the zero-mean Gaussian process $\{\Gamma_G(s) : s \in [1, \bar{h}/\underline{h}]\}$ has a covariance structure given by (9) and $\|\Gamma_G\|_{[1, \bar{h}/\underline{h}]} = \|\tilde{\Gamma}_G\|_{\mathbb{H}}$. By [Dudley \(2002, Theorem 9.2.2\)](#) and $\|\tilde{\Gamma}_G\|_{\mathbb{H}} - \|\bar{\Gamma}_G\|_{\mathbb{H}} = o_p \left(\log(n)^{-1/2} \right)$, there exists a null sequence $\tilde{\varepsilon}_n \downarrow 0$ such that $\Pr \left[\left| \|\tilde{\Gamma}_G\|_{\mathbb{H}} - \|\bar{\Gamma}_G\|_{\mathbb{H}} \right| > \tilde{\varepsilon}_n / \sqrt{\log(n)} \right] \leq \tilde{\varepsilon}_n$. By similar arguments, we have $\Pr \left[\|\tilde{\Gamma}_G\|_{\mathbb{H}} \leq z_{1-\tau} (\bar{h}/\underline{h}) \right] - \Pr \left[\|\bar{\Gamma}_G\|_{\mathbb{H}} \leq z_{1-\tau} (\bar{h}/\underline{h}) \right] = o(1)$. By the definition of $z_{1-\tau} (\bar{h}/\underline{h})$ and $\|\Gamma_G\|_{[1, \bar{h}/\underline{h}]} = \|\tilde{\Gamma}_G\|_{\mathbb{H}}$, $\Pr \left[\|\tilde{\Gamma}_G\|_{\mathbb{H}} \leq z_{1-\tau} (\bar{h}/\underline{h}) \right] = 1 - \tau$. It then follows that $\Pr \left[LR_p(\vartheta_0 | h) \leq z_{1-\tau} (\bar{h}/\underline{h})^2, \forall h \in \mathbb{H} \right] = 1 - \tau + o(1)$. \blacksquare

Appendix B Proofs of Theorems 3 and 4

We denote $\bar{n} := nh$ for notational simplicity and write $\delta = O_p^*(a_n)$ for some bounded sequence a_n if there exists some positive constants $c_1, c_2 > 0$ such that $\Pr[|\delta| > c_1 a_n] \leq c_2 (\log(n) / \bar{n}^{3/2})$. It is straightforward to check that if $\delta_1 = O_p^*(a_n)$ and $\delta_2 = O_p^*(b_n)$, then $\delta_1 \delta_2 = O_p^*(a_n b_n)$ and $\delta_1 + \delta_2 = O_p^*(a_n + b_n)$, i.e., the algebra of the O_p notations carry over to O_p^* notations. We say that an event occurs wp^* if its probability is $1 - O(\log(n) / \bar{n}^{3/2})$.

Lemma 5. *Suppose that the same assumptions as Lemma 1 hold with $\underline{h} = \bar{h} = h$. If $g_{|V|^5}$ is bounded on $\mathbb{B}(0)$, $\bar{n}^{-1/2} \sum_i (W_{p;s,i}^k V_i - \mathbb{E}[W_{p;s}^k V]) = O_p^*(\sqrt{\log(n)})$, $\forall (k, s) \in \mathbb{N} \times \{-, +\}$.*

The following result is an analogue of Lemma 2. Its proof essentially follows similar arguments.

Lemma 6. *Suppose that the same assumptions as Theorem 3 hold. (a) $\sqrt{\bar{n}} \hat{\eta}_p = O_p^*(\sqrt{\log(n)})$; (b) $\hat{\lambda}_p := \operatorname{argmax}_{\lambda \in \mathcal{L}(\hat{\vartheta}_p)} S(\lambda, \hat{\vartheta}_p)$ exists wp^* and $\sqrt{\bar{n}} \hat{\lambda}_p = O_p^*(\sqrt{\log(n)})$; (c) $\sqrt{\bar{n}} \tilde{\eta}_p = O_p^*(\sqrt{\log(n)})$; (d) $\tilde{\lambda}_p := \operatorname{argmax}_{\lambda \in \mathcal{L}(\vartheta_0, \tilde{\vartheta}_p)} S(\lambda, \vartheta_0, \tilde{\vartheta}_p)$ exists wp^* and $\sqrt{\bar{n}} \tilde{\lambda}_p = O_p^*(\sqrt{\log(n)})$.*

Consider the singular value decomposition of $\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1/2}(-\Delta_{\mathcal{G}})$: $\mathbf{S}^\top \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1/2}(-\Delta_{\mathcal{G}}) \mathbf{T} = \begin{bmatrix} \Lambda & 0_{d_\vartheta \times d_z} \end{bmatrix}^\top$, where $\mathbf{S}^\top \mathbf{S} = \mathbf{I}_{2d_u}$, $\mathbf{T}^\top \mathbf{T} = \mathbf{I}_{d_\vartheta}$ and Λ is a d_ϑ -dimensional diagonal matrix with the square roots of the eigenvalues of $\Delta_{\mathcal{G}}^\top \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1} \Delta_{\mathcal{G}}$ being on its diagonal. We follow [Chen and Cui \(2007\)](#) to rotate the moment conditions by $\Gamma := \mathbf{S}^\top \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1/2}$ so that results from [Chen and Cui \(2007\)](#); [Ma \(2017\)](#) can be applied. Let $\mathcal{V}_i(\theta) := \Gamma \mathcal{U}_i(\theta)$, $\mathcal{V}_i := \Gamma \mathcal{U}_i$, $\mathcal{H}_i := \Gamma(-\mathcal{G}_i)$, $\mathcal{H}_{\dagger,i} := \Gamma(-\mathcal{G}_{\dagger,i})$ ($\mathcal{V}, \mathcal{H}, \mathcal{H}_\dagger$ defined similarly) and $\hat{\mathcal{V}}_i := \Gamma \hat{\mathcal{U}}_i$. Denote $\bar{\Delta}_{\mathcal{V}\mathcal{V}^\top} := \bar{n}^{-1} \sum_i \mathcal{V}_i \mathcal{V}_i^\top$ and $\bar{\Delta}_{\mathcal{H}} := \bar{n}^{-1} \sum_i \mathcal{H}_i$. Note that the EL criterion function is invariant to such a rotation, i.e., $\ell_p(\theta | h) = \sup_\lambda 2 \sum_i \log(1 + \lambda^\top \mathcal{V}_i(\theta))$. For notational simplicity, we still use $\hat{\lambda}_p$ and $\tilde{\lambda}_p$ to denote the Lagrange multipliers. Clearly, (b) and (d) of Lemma 6 still hold. Let $\Pi := \Lambda \mathbf{T}^\top$ and $\Omega := \Pi^{-1}$. Then, $\Delta_{\mathcal{V}\mathcal{V}^\top} = \mathbf{I}_{2d_u}$, $\Delta_{\mathcal{H}} := \Gamma(-\Delta_{\mathcal{G}}) = \begin{bmatrix} \Pi^\top & 0_{d_\vartheta \times d_z} \end{bmatrix}^\top$ and $\Delta_{\mathcal{H}_\dagger} := \Gamma(-\Delta_{\mathcal{G}_\dagger}) = \begin{bmatrix} \Pi_\dagger^\top & 0_{d_\dagger \times d_z} \end{bmatrix}^\top$, where Π_\dagger is a $d_\vartheta \times d_\dagger$ matrix collecting the last d_\dagger columns of Π . Denote $\mathbf{J} := (\Pi_\dagger^\top \Pi_\dagger)^{-1} \Pi_\dagger^\top$, $\mathbf{P} := \Pi_\dagger \mathbf{J}$ and $\mathbf{M} := -\mathbf{I}_{d_\vartheta} + \mathbf{P}$. Then, by inverting the block matrices,

$$\begin{aligned} \begin{bmatrix} -\Delta_{\mathcal{V}\mathcal{V}^\top} & \Delta_{\mathcal{H}} \\ \Delta_{\mathcal{H}}^\top & 0_{d_\vartheta \times d_\vartheta} \end{bmatrix}^{-1} &= \begin{bmatrix} -\Delta_{\mathcal{V}\mathcal{V}^\top}^{-1} + \Delta_{\mathcal{V}\mathcal{V}^\top}^{-1} \Delta_{\mathcal{H}} (\Delta_{\mathcal{H}}^\top \Delta_{\mathcal{V}\mathcal{V}^\top}^{-1} \Delta_{\mathcal{H}})^{-1} \Delta_{\mathcal{H}}^\top \Delta_{\mathcal{V}\mathcal{V}^\top}^{-1} & \Delta_{\mathcal{V}\mathcal{V}^\top}^{-1} \Delta_{\mathcal{H}} (\Delta_{\mathcal{H}}^\top \Delta_{\mathcal{V}\mathcal{V}^\top}^{-1} \Delta_{\mathcal{H}})^{-1} \\ (\Delta_{\mathcal{H}}^\top \Delta_{\mathcal{V}\mathcal{V}^\top}^{-1} \Delta_{\mathcal{H}})^{-1} \Delta_{\mathcal{H}}^\top \Delta_{\mathcal{V}\mathcal{V}^\top}^{-1} & (\Delta_{\mathcal{H}}^\top \Delta_{\mathcal{V}\mathcal{V}^\top}^{-1} \Delta_{\mathcal{H}})^{-1} \end{bmatrix} \\ &= \begin{bmatrix} 0_{d_\vartheta \times d_\vartheta} & 0_{d_\vartheta \times d_z} & \Omega^\top \\ 0_{d_z \times d_\vartheta} & -\mathbf{I}_{d_z} & 0_{d_z \times d_\vartheta} \\ \Omega & 0_{d_\vartheta \times d_z} & \Omega \Omega^\top \end{bmatrix} = \begin{bmatrix} -(\Gamma^\top)^{-1} \mathbf{Q} \Gamma^{-1} & -(\Gamma^\top)^{-1} \mathbf{N} \\ -\mathbf{N}^\top \Gamma^{-1} & \mathbf{O} \end{bmatrix} \quad (24) \end{aligned}$$

and

$$\begin{bmatrix} -\Delta_{\mathcal{V}\mathcal{V}^\top} & \Delta_{\mathcal{H}\dagger} \\ \Delta_{\mathcal{H}\dagger}^\top & 0_{d_\dagger \times d_\dagger} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{M} & 0_{d_\vartheta \times d_z} & \mathbf{J}^\top \\ 0_{d_z \times d_\vartheta} & -\mathbf{I}_{d_z} & 0_{d_z \times d_\dagger} \\ \mathbf{J} & 0_{d_\dagger \times d_z} & \mathbf{J}\mathbf{J}^\top \end{bmatrix} = \begin{bmatrix} -(\Gamma^\top)^{-1} \mathbf{Q}_\dagger \Gamma^{-1} & -(\Gamma^\top)^{-1} \mathbf{N}_\dagger \\ -\mathbf{N}_\dagger^\top \Gamma^{-1} & \mathbf{O}_\dagger \end{bmatrix}. \quad (25)$$

By similar arguments as in the proof of Lemma 3, the first order conditions $\sum_i \hat{\mathcal{V}}_i / (1 + \hat{\lambda}_p^\top \hat{\mathcal{V}}_i) = 0$ and $\sum_i \mathcal{H}_i^\top \hat{\lambda}_p / (1 + \hat{\lambda}_p^\top \hat{\mathcal{V}}_i) = 0$ hold wp*. Expanding the left hand sides yields

$$\begin{aligned} 0 &= \sum_i \hat{\mathcal{V}}_i \left\{ 1 - \hat{\lambda}_p^\top \hat{\mathcal{V}}_i + (\hat{\lambda}_p^\top \hat{\mathcal{V}}_i)^2 - (\hat{\lambda}_p^\top \hat{\mathcal{V}}_i)^3 + \frac{(\hat{\lambda}_p^\top \hat{\mathcal{V}}_i)^4}{1 + \hat{\lambda}_p^\top \hat{\mathcal{V}}_i} \right\} \\ 0 &= \sum_i \mathcal{H}_i^\top \hat{\lambda}_p \left\{ 1 - \hat{\lambda}_p^\top \hat{\mathcal{V}}_i + (\hat{\lambda}_p^\top \hat{\mathcal{V}}_i)^2 - \frac{(\hat{\lambda}_p^\top \hat{\mathcal{V}}_i)^3}{1 + \hat{\lambda}_p^\top \hat{\mathcal{V}}_i} \right\}. \end{aligned} \quad (26)$$

By Lemma 6 and $\max_i \|\hat{\mathcal{U}}_i\| = O_p^*(\bar{n}^{1/5})$, $\max_i |\hat{\lambda}_p^\top \hat{\mathcal{V}}_i| \lesssim \|\hat{\lambda}_p\| (\max_i \|\hat{\mathcal{V}}_i\|) = O_p^*(\sqrt{\log(n)}/\bar{n}^{3/10})$. Therefore, $\max_i |\hat{\lambda}_p^\top \hat{\mathcal{V}}_i| < 1/2$ wp*. By this result, Lemma 6 and $\bar{n}^{-1} \sum_i \|\hat{\mathcal{U}}_i\|^4 = O_p^*(1)$, which follows from boundedness of Θ , Markov's inequality and $\hat{\mathcal{U}}_i = \mathcal{U}_i - \mathcal{G}_i \hat{\eta}_p$, $\sum_i (\hat{\lambda}_p^\top \hat{\mathcal{V}}_i)^4 / (1 + \hat{\lambda}_p^\top \hat{\mathcal{V}}_i) = O_p^*(\log(n)^2/\bar{n})$. Similarly, $\sum_i \mathcal{H}_i^\top \hat{\lambda}_p (\hat{\lambda}_p^\top \hat{\mathcal{V}}_i)^3 / (1 + \hat{\lambda}_p^\top \hat{\mathcal{V}}_i) = O_p^*(\log(n)^2/\bar{n})$. By $\hat{\mathcal{V}}_i = \mathcal{V}_i + \mathcal{H}_i \hat{\eta}_p$ and Lemma 6, $\sum_i \hat{\lambda}_p^\top \hat{\mathcal{V}}_i = \sum_i \hat{\lambda}_p^\top \mathcal{V}_i + \sum_i \hat{\lambda}_p^\top \mathcal{H}_i \hat{\eta}_p$, $\sum_i (\hat{\lambda}_p^\top \hat{\mathcal{V}}_i)^2 = \sum_i (\hat{\lambda}_p^\top \mathcal{V}_i)^2 + \sum_i (\hat{\lambda}_p^\top \mathcal{H}_i \hat{\eta}_p)^2 + 2 \sum_i (\hat{\lambda}_p^\top \mathcal{V}_i) (\hat{\lambda}_p^\top \mathcal{H}_i \hat{\eta}_p)$ and $\sum_i (\hat{\lambda}_p^\top \hat{\mathcal{V}}_i)^3 = \sum_i (\hat{\lambda}_p^\top \mathcal{V}_i)^3 + 3 \sum_i (\hat{\lambda}_p^\top \mathcal{V}_i)^2 (\hat{\lambda}_p^\top \mathcal{H}_i \hat{\eta}_p) + O_p^*(\log(n)^{5/2}/\bar{n}^{3/2})$. By plugging these results into the right hand side of (26),

$$\begin{aligned} -\Delta_{\mathcal{V}\mathcal{V}^\top} \hat{\lambda}_p + \Delta_{\mathcal{H}\hat{\eta}_p} &= -\frac{1}{\bar{n}} \sum_i \mathcal{V}_i + \frac{1}{\bar{n}} \sum_i \mathcal{V}_i (\hat{\lambda}_p^\top \mathcal{H}_i \hat{\eta}_p) - \frac{1}{\bar{n}} \sum_i \mathcal{V}_i (\hat{\lambda}_p^\top \mathcal{V}_i)^2 - \frac{2}{\bar{n}} \sum_i \mathcal{V}_i (\hat{\lambda}_p^\top \mathcal{V}_i) (\hat{\lambda}_p^\top \mathcal{H}_i \hat{\eta}_p) \\ &\quad + \frac{1}{\bar{n}} \sum_i \mathcal{V}_i (\hat{\lambda}_p^\top \mathcal{V}_i)^3 + \frac{1}{\bar{n}} \sum_i \mathcal{H}_i \hat{\eta}_p (\hat{\lambda}_p^\top \mathcal{V}_i) + \frac{1}{\bar{n}} \sum_i \mathcal{H}_i \hat{\eta}_p (\hat{\lambda}_p^\top \mathcal{H}_i \hat{\eta}_p) \\ &\quad - \frac{1}{\bar{n}} \sum_i \mathcal{H}_i \hat{\eta}_p (\hat{\lambda}_p^\top \mathcal{V}_i)^2 + (\bar{\Delta}_{\mathcal{V}\mathcal{V}^\top} - \Delta_{\mathcal{V}\mathcal{V}^\top}) \hat{\lambda}_p - (\bar{\Delta}_{\mathcal{H}} - \Delta_{\mathcal{H}}) \hat{\eta}_p + O_p^*((\log(n)/\bar{n})^2) \\ \Delta_{\mathcal{H}}^\top \hat{\lambda}_p &= \frac{1}{\bar{n}} \sum_i \mathcal{H}_i^\top \hat{\lambda}_p (\hat{\lambda}_p^\top \mathcal{V}_i) + \frac{1}{\bar{n}} \sum_i \mathcal{H}_i^\top \hat{\lambda}_p (\hat{\lambda}_p^\top \mathcal{H}_i \hat{\eta}_p) - \frac{1}{\bar{n}} \sum_i \mathcal{H}_i^\top \hat{\lambda}_p (\hat{\lambda}_p^\top \mathcal{V}_i)^2 \\ &\quad - (\bar{\Delta}_{\mathcal{H}}^\top - \Delta_{\mathcal{H}}^\top) \hat{\lambda}_p + O_p^*((\log(n)/\bar{n})^2). \end{aligned} \quad (27)$$

By fifth-order Taylor expansion and $\max_i |\hat{\lambda}_p^\top \hat{\mathcal{V}}_i| = O_p^*(\sqrt{\log(n)}/\bar{n}^{3/10})$, $\ell_p(\hat{\mathcal{V}}_p | h)$ can be written as the sum of $2 \sum_i \hat{\lambda}_p^\top \hat{\mathcal{V}}_i - \sum_i (\hat{\lambda}_p^\top \hat{\mathcal{V}}_i)^2 + 2 \sum_i (\hat{\lambda}_p^\top \hat{\mathcal{V}}_i)^3 / 3 - \sum_i (\hat{\lambda}_p^\top \hat{\mathcal{V}}_i)^4 / 2$ and a remainder term bounded up to a constant by $\sum_i |\hat{\lambda}_p^\top \hat{\mathcal{V}}_i|^5 \lesssim \|\hat{\lambda}_p\|^5 \sum_i \|\hat{\mathcal{V}}_i\|^5 = O_p^*(\log(n)^{5/2}/\bar{n}^{3/2})$ wp*. By $\hat{\mathcal{V}}_i = \mathcal{V}_i + \mathcal{H}_i \hat{\eta}_p$ and Lemma 6,

$\sum_i (\hat{\lambda}_p^\top \hat{\mathbf{v}}_i)^4 = \sum_i (\hat{\lambda}_p^\top \mathbf{v}_i)^4 + O_p^* \left(\log(n)^{5/2} / \bar{n}^{3/2} \right)$ and therefore,

$$\begin{aligned} \bar{n}^{-1} \ell_p \left(\hat{\vartheta}_p \mid h \right) &= \frac{2}{\bar{n}} \sum_i \hat{\lambda}_p^\top \mathbf{v}_i + \frac{2}{\bar{n}} \sum_i \hat{\lambda}_p^\top \mathcal{H}_i \hat{\eta}_p - \frac{1}{\bar{n}} \sum_i (\hat{\lambda}_p^\top \mathbf{v}_i)^2 - \frac{1}{\bar{n}} \sum_i (\hat{\lambda}_p^\top \mathcal{H}_i \hat{\eta}_p)^2 - \frac{2}{\bar{n}} \sum_i (\hat{\lambda}_p^\top \mathbf{v}_i) (\hat{\lambda}_p^\top \mathcal{H}_i \hat{\eta}_p) \\ &\quad + \frac{2}{3} \frac{1}{\bar{n}} \sum_i (\hat{\lambda}_p^\top \mathbf{v}_i)^3 + \frac{2}{\bar{n}} \sum_i (\hat{\lambda}_p^\top \mathbf{v}_i)^2 (\hat{\lambda}_p^\top \mathcal{H}_i \hat{\eta}_p) - \frac{1}{2\bar{n}} \sum_i (\hat{\lambda}_p^\top \mathbf{v}_i)^4 + O_p^* \left(\log(n)^{5/2} / \bar{n}^{5/2} \right). \end{aligned} \quad (28)$$

A stochastic expansion (e.g., [Newey and Smith, 2004](#)) is understood as an approximation that is a polynomial of centered sample averages and has an approximation error of desired order of magnitude. We use (24) to invert (27) and get higher-order approximations for $(\hat{\lambda}_p, \hat{\eta}_p)$. We then replace all sample averages except $\bar{n}^{-1} \sum_i \mathbf{v}_i$ which is approximately centered ($\|\Delta_{\mathbf{v}}\| = O(h^{p+1})$) with the sums of their population means and their centered versions. By iteratively replacing $(\hat{\lambda}_p, \hat{\eta}_p)$ on the right hand side of (27) with the approximations, using Lemmas 5 and 6 and dropping terms that are $O_p^* \left((\log(n)/\bar{n})^2 \right)$, we get cubic stochastic expansions of $(\hat{\lambda}_p, \hat{\eta}_p)$. By the same steps and plugging stochastic expansions of $(\hat{\lambda}_p, \hat{\eta}_p)$ into the right hand side of (28), we have a stochastic expansion of $\bar{n}^{-1} \ell_p \left(\hat{\vartheta}_p \mid h \right)$ so that $\bar{n}^{-1} \ell_p \left(\hat{\vartheta}_p \mid h \right) = \hat{\ell}^* + O_p^* \left(\log(n)^{5/2} / \bar{n}^{5/2} \right)$, where the leading term $\hat{\ell}^*$ is a quartic polynomial of centered sample averages. Similarly, by using Lemmas 5 and 6, the first-order conditions and (25), we get cubic stochastic expansions of $(\tilde{\lambda}_p, \tilde{\eta}_p)$ and a quartic stochastic expansion of $\ell_p \left(\vartheta_0, \tilde{\vartheta}_p \mid h \right)$ so that $\bar{n}^{-1} \ell_p \left(\vartheta_0, \tilde{\vartheta}_p \mid h \right) = \tilde{\ell}^* + O_p^* \left(\log(n)^{5/2} / \bar{n}^{5/2} \right)$. The same algebraic calculations have been done in [Chen and Cui \(2007\)](#); [Ma \(2017\)](#) so that we use them directly here. We switch to coordinate notations and apply the calculations from [Chen and Cui \(2007\)](#); [Ma \(2017\)](#). In the rest of the proofs, summation over repeated indices is taken implicitly with the “ \sum ” notation suppressed and ranges of indices fixed: $k, l, m, n, o, v, q = 1, \dots, d_\vartheta$, $\mathbf{k}, \mathbf{l}, \mathbf{m}, \mathbf{n}, \mathbf{o}, \mathbf{v} = 1, \dots, 2d_u$, $u, w = 1, \dots, d_\dagger$, $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}, \mathbf{e}, \mathbf{f} = 1, \dots, 2d_z$, $s, t, a, b, c, d, e = 1, \dots, d_z$ and $\mathbf{u}, \mathbf{w} = 1, \dots, d_u$. Let $\alpha^{\mathbf{kl}} := \Delta_{\mathbf{v}^{(\mathbf{k})} \mathbf{v}^{(\mathbf{l})}}$, $\alpha^{\mathbf{klm}} := \Delta_{\mathbf{v}^{(\mathbf{k})} \mathbf{v}^{(\mathbf{l})} \mathbf{v}^{(\mathbf{m})}}$, $\alpha^{\mathbf{klmn}} := \Delta_{\mathbf{v}^{(\mathbf{k})} \mathbf{v}^{(\mathbf{l})} \mathbf{v}^{(\mathbf{m})} \mathbf{v}^{(\mathbf{n})}}$, $\gamma^{\mathbf{k},n} := \Delta_{\mathcal{H}^{(kn)}}$, $\gamma^{\mathbf{k};\mathbf{l},n} := \Delta_{\mathbf{v}^{(\mathbf{k})} \mathcal{H}^{(\mathbf{l}n)}}$, $\gamma^{\mathbf{k};\mathbf{l};\mathbf{m},n} := \Delta_{\mathbf{v}^{(\mathbf{k})} \mathbf{v}^{(\mathbf{l})} \mathcal{H}^{(\mathbf{m}n)}}$, $\gamma^{\mathbf{k},n;\mathbf{l},o} := \Delta_{\mathcal{H}^{(kn)} \mathcal{H}^{(\mathbf{l}o)}}$, $A^{\mathbf{k}} := \bar{n}^{-1} \sum_i \mathbf{v}_i^{(\mathbf{k})}$, $A^{\mathbf{kl}} := \bar{n}^{-1} \sum_i \mathbf{v}_i^{(\mathbf{k})} \mathbf{v}_i^{(\mathbf{l})} - \alpha^{\mathbf{kl}}$, $A^{\mathbf{klm}} := \bar{n}^{-1} \sum_i \mathbf{v}_i^{(\mathbf{k})} \mathbf{v}_i^{(\mathbf{l})} \mathbf{v}_i^{(\mathbf{m})} - \alpha^{\mathbf{klm}}$, $C^{\mathbf{k},n} := \bar{n}^{-1} \sum_i \mathcal{H}_i^{(kn)} - \gamma^{\mathbf{k},n}$ and $C^{\mathbf{k};\mathbf{l},n} := \bar{n}^{-1} \sum_i \mathbf{v}_i^{(\mathbf{k})} \mathcal{H}_i^{(\mathbf{l}n)} - \gamma^{\mathbf{k};\mathbf{l},n}$. By Lemma 5, $\bar{n}^{-1} \sum_i \mathbf{v}_i^{(\mathbf{k})} \mathbf{v}_i^{(\mathbf{l})} \mathbf{v}_i^{(\mathbf{m})} \mathbf{v}_i^{(\mathbf{n})} - \alpha^{\mathbf{klmn}}$, $\bar{n}^{-1} \sum_i \mathbf{v}_i^{(\mathbf{k})} \mathbf{v}_i^{(\mathbf{l})} \mathcal{H}_i^{(\mathbf{m}n)} - \gamma^{\mathbf{k};\mathbf{l};\mathbf{m},n}$ and $\bar{n}^{-1} \sum_i \mathcal{H}_i^{(kn)} \mathcal{H}_i^{(\mathbf{l}o)} - \gamma^{\mathbf{k},n;\mathbf{l},o}$ are all $O_p^* \left(\sqrt{\log(n)/\bar{n}} \right)$. We can show that $(\hat{\lambda}_p, \hat{\eta}_p)$ and $\bar{n}^{-1} \ell_p \left(\hat{\vartheta}_p \mid h \right)$ admit stochastic expansions with leading terms that are polynomials of $(A^{\mathbf{k}}, A^{\mathbf{kl}}, A^{\mathbf{klm}}, C^{\mathbf{k},n}, C^{\mathbf{k};\mathbf{l},n})$ with coefficients given by $(\alpha^{\mathbf{kl}}, \alpha^{\mathbf{klm}}, \alpha^{\mathbf{klmn}})$, $(\gamma^{\mathbf{k},n}, \gamma^{\mathbf{k};\mathbf{l},n}, \gamma^{\mathbf{k};\mathbf{l};\mathbf{m},n}, \gamma^{\mathbf{k},n;\mathbf{l},o})$ and Ω . Formally, their expressions are the same as those given in the special case of [Chen and Cui \(2007\)](#) (see (2.6) and (2.8) therein) when the moment restrictions are linear in parameters and terms that depend on the second and third derivatives of the moment restrictions are removed. Similar stochastic expansions of $(\tilde{\lambda}_p, \tilde{\eta}_p)$ and $\bar{n}^{-1} \ell_p \left(\vartheta_0, \tilde{\vartheta}_p \mid h \right)$ that are polynomials of $(A^{\mathbf{k}}, A^{\mathbf{kl}}, A^{\mathbf{klm}}, C^{\mathbf{k},n}, C^{\mathbf{k};\mathbf{l},n})$ can also be obtained. Formally, their expressions are the same as those given in the special case of [Ma \(2017\)](#)

when the moment restrictions and the null restrictions are both linear in parameters and hence omitted. See Ma (2017, (C.4)). Let $LR^* := \bar{n} \left(\tilde{\ell}^* - \hat{\ell}^* \right)$ so that $LR_p(\vartheta_0 | h) = LR^* + O_p^* \left(\log(n)^{5/2} / \bar{n}^{3/2} \right)$.

Let $(\Upsilon^{kl}, \Upsilon^{klm}, \Upsilon^{klmn})$ be defined by the same formulae as those of $(\alpha^{kl}, \alpha^{klm}, \alpha^{klmn})$ with \mathcal{V} replaced by \mathcal{U} . Let $(\Gamma_{\dagger}^{k,u}, \Gamma_{\dagger}^{k;l,u}, \Gamma_{\dagger}^{k;l;m,u}, \Gamma_{\dagger}^{k,u;l,w})$ be defined by the formulae of $(\gamma^{k,n}, \gamma^{k;l,n}, \gamma^{k;l;m,n}, \gamma^{k,n;l,o})$ with $(\mathcal{V}, \mathcal{H})$ replaced by $(\mathcal{U}, \mathcal{G}_{\dagger})$. Denote $\bar{U}(\theta_2) := Z - \theta_2$, $\bar{U} := \bar{U}(\vartheta_2)$ and $\bar{G} := -\partial \bar{U}(\theta_2) / \partial \theta_2^{\top} = I_{d_z}$. Also let $\bar{U}(\theta_2) := W_p \otimes \bar{U}(\theta_2)$, $\bar{U} := W_p \otimes \bar{U}$, $\bar{G} := W_p \otimes \bar{G}$ and $(\bar{U}_i(\theta_2), \bar{U}_i, \bar{G}_i)$ be defined by the same formulae with (X, Z) replaced by (X_i, Z_i) . Let $(\bar{O}, \bar{N}, \bar{Q})$ be defined by the formulae of (O, N, Q) with $(\Delta_{\mathcal{U}\mathcal{U}^{\top}}, \Delta_{\mathcal{G}})$ replaced by $(\Delta_{\bar{U}\bar{U}^{\top}}, \Delta_{\bar{G}})$. Let $(\bar{\Upsilon}^{ab}, \bar{\Upsilon}^{abc}, \bar{\Upsilon}^{abcd})$ and $(\bar{\Gamma}^{a,s}, \bar{\Gamma}^{a;b,s}, \bar{\Gamma}^{a;b;c,s}, \bar{\Gamma}^{a,s;b,t})$ be defined by the formulae of $(\alpha^{kl}, \alpha^{klm}, \alpha^{klmn})$ and $(\gamma^{k,n}, \gamma^{k;l,n}, \gamma^{k;l;m,n}, \gamma^{k,n;l,o})$ with $(\mathcal{V}, \mathcal{H})$ replaced by $(\bar{\mathcal{U}}, \bar{\mathcal{G}})$. Let $\Upsilon^k := \Delta_{\mathcal{U}^{(k)}}$ and $\bar{\Upsilon}^a := \Delta_{\bar{\mathcal{U}}^{(a)}}$. Let $\mathcal{B}_p^{\dagger} := Q_{\dagger}^{(kl)} \Upsilon^k \Upsilon^l$, $\gamma_{p,1}^{\dagger} := \Upsilon^{klmn} Q_{\dagger}^{(kl)} Q_{\dagger}^{(mn)} / 2$, $\gamma_{p,2}^{\dagger} := -\Upsilon^{klm} Q_{\dagger}^{(kn)} Q_{\dagger}^{(lo)} Q_{\dagger}^{(mv)} \Upsilon^{nov} / 3$, $\gamma_{p,3}^{\dagger} := 2\Gamma_{\dagger}^{k;l;m,w} N_{\dagger}^{(kw)} Q_{\dagger}^{(lm)}$, and $\gamma_{p,4}^{\dagger} := -\Gamma_{\dagger}^{k,u;l,w} Q_{\dagger}^{(kl)} O_{\dagger}^{(uw)}$. Let $(\mathcal{B}_p^{\dagger}, \gamma_{p,1}^{\dagger}, \gamma_{p,2}^{\dagger}, \gamma_{p,3}^{\dagger}, \gamma_{p,4}^{\dagger})$ be defined by the same formulae with $(Q_{\dagger}, N_{\dagger}, O_{\dagger}, \Gamma_{\dagger}, \Upsilon)$ replaced by $(\bar{Q}, \bar{N}, \bar{O}, \bar{\Gamma}, \bar{\Upsilon})$. Let $\mathcal{C}_p^{\text{pre}}(n, h) := \bar{n} (\mathcal{B}_p^{\dagger} - \bar{\mathcal{B}}_p^{\dagger}) + \bar{n}^{-1} \sum_{j=1}^4 (\gamma_{p,j}^{\dagger} - \bar{\gamma}_{p,j}^{\dagger})$ denote the pre-asymptotic coverage error.

Lemma 7. Suppose that the same assumptions as Theorem 3 hold. Then, $\Pr[LR^* \leq x] = F_{\chi_1^2}(x) - \mathcal{C}_p^{\text{pre}}(n, h) x f_{\chi_1^2}(x) + O(v_n^{\#})$, where $v_n^{\#} := (\log(n) \|\Delta_{\mathcal{U}}\|) / \sqrt{\bar{n}} + \log(n)^{5/2} / \bar{n}^{3/2} + h \|\Delta_{\mathcal{U}}\| + n^{-1} + \bar{n}^2 \|\Delta_{\mathcal{U}}\|^4 + \bar{n} \|\Delta_{\mathcal{U}}\|^3$.

Proof of Theorem 3. By simple algebra, $Q_{\dagger} = \begin{bmatrix} Q_{\dagger 11} & Q_{\dagger 12} \\ Q_{\dagger 21} & Q_{\dagger 22} \end{bmatrix}$, where $Q_{\dagger 11} := \Delta_{+}^{-2} \Phi_{\pm}^{-1}$, $Q_{\dagger 22} := \Delta_{-}^{-2} \Phi_{\pm}^{-1}$, $Q_{\dagger 12} = Q_{\dagger 21} := -\Delta_{+}^{-1} \Delta_{-}^{-1} \Phi_{\pm}^{-1}$, $O_{\dagger} = \left(\Delta_{+}^2 \Delta_{\mathcal{U}_{+}\mathcal{U}_{+}^{\top}}^{-1} + \Delta_{-}^2 \Delta_{\mathcal{U}_{-}\mathcal{U}_{-}^{\top}}^{-1} \right)^{-1}$ and $N_{\dagger} = \begin{bmatrix} N_{\dagger 1}^{\top} & N_{\dagger 2}^{\top} \end{bmatrix}^{\top}$, where $N_{\dagger 1} := \Delta_{+} \Delta_{\mathcal{U}_{+}\mathcal{U}_{+}^{\top}}^{-1} O_{\dagger}$ and $N_{\dagger 2} := \Delta_{-} \Delta_{\mathcal{U}_{-}\mathcal{U}_{-}^{\top}}^{-1} O_{\dagger}$. For simplicity, denote $\Pi_s^u := \Delta_{\mathcal{U}_s^{(u)} \mathcal{U}_s \mathcal{U}_s^{\top}}$ and $\Pi_s^{uw} := \Delta_{\mathcal{U}_s^{(u)} \mathcal{U}_s^{(w)} \mathcal{U}_s \mathcal{U}_s^{\top}}$, $s \in \{-, +\}$. First, write $\Upsilon^{klmn} Q_{\dagger}^{(kl)} Q_{\dagger}^{(mn)} = Q_{\dagger}^{(kl)} \text{tr}(Q_{\dagger} \Delta_{\mathcal{U}^{(k)} \mathcal{U}^{(l)} \mathcal{U} \mathcal{U}^{\top}})$. Then it is easy to check that

$$\begin{aligned} \Upsilon^{klmn} Q_{\dagger}^{(kl)} Q_{\dagger}^{(mn)} &= Q_{\dagger 11}^{(uw)} \text{tr}(Q_{\dagger 11} \Pi_{+}^{uw}) + Q_{\dagger 22}^{(uw)} \text{tr}(Q_{\dagger 22} \Pi_{-}^{uw}) \\ \Upsilon^{klm} Q_{\dagger}^{(kn)} Q_{\dagger}^{(lo)} Q_{\dagger}^{(mv)} \Upsilon^{nov} &= Q_{\dagger 11}^{(uw)} \text{tr}(Q_{\dagger 11} \Pi_{+}^u Q_{\dagger 11} \Pi_{+}^w) + Q_{\dagger 22}^{(uw)} \text{tr}(Q_{\dagger 22} \Pi_{-}^u Q_{\dagger 22} \Pi_{-}^w) + 2Q_{\dagger 21}^{(uw)} \text{tr}(Q_{\dagger 12} \Pi_{-}^u Q_{\dagger 21} \Pi_{+}^w) \\ \Gamma_{\dagger}^{k;l;m,w} N_{\dagger}^{(kw)} Q_{\dagger}^{(lm)} &= \text{tr}(Q_{\dagger 11} \Delta_{W_{p;+} \mathcal{U}_{+} \mathcal{U}_{+}^{\top}} N_{\dagger 1}) + \text{tr}(Q_{\dagger 22} \Delta_{W_{p;-} \mathcal{U}_{-} \mathcal{U}_{-}^{\top}} N_{\dagger 2}) \\ \Gamma_{\dagger}^{k,u;l,w} Q_{\dagger}^{(kl)} O_{\dagger}^{(uw)} &= \text{tr}(\Delta_{W_{p;+}^2} Q_{\dagger 11} O_{\dagger}) + \text{tr}(\Delta_{W_{p;-}^2} Q_{\dagger 22} O_{\dagger}). \end{aligned}$$

By Lemma 1, $Q_{\dagger 11} = \Xi_1 / (\varphi \omega_{p;+}^{0,2}) + O(h)$, $Q_{\dagger 22} = \Xi_1 / (\varphi \omega_{p;+}^{0,2}) + O(h)$, $Q_{\dagger 21} = -\Xi_1 / (\varphi \omega_{p;+}^{0,2}) + O(h)$, $O_{\dagger} = (\omega_{p;+}^{0,2} / \varphi) \Xi_2 + O(h)$, $N_{\dagger 1} = \mu_{UU^{\top},+}^{-1} \Xi_2 / \varphi + O(h)$ and $N_{\dagger 2} = \mu_{UU^{\top},-}^{-1} \Xi_2 / \varphi + O(h)$. It follows that $\gamma_{p,1}^{\dagger} = \left((\omega_{p;+}^{0,4} / \omega_{p;+}^{0,2}) \Xi_1^{(uw)} \Psi_1^{uw} \right) / (2\varphi \omega_{p;+}^{0,2}) + O(h)$, $\gamma_{p,2}^{\dagger} = \left(-(\omega_{p;+}^{0,3} / \omega_{p;+}^{0,2})^2 \Xi_1^{(uw)} \Psi_2^{uw} \right) / (3\varphi \omega_{p;+}^{0,2}) + O(h)$, $\gamma_{p,3}^{\dagger} = (4\omega_{p;+}^{0,3} \text{tr}(\Xi_1 \Xi_2)) / (\varphi \omega_{p;+}^{0,2}) + O(h)$ and $\gamma_{p,4}^{\dagger} = (-2\omega_{p;+}^{0,2} \text{tr}(\Xi_1 \Xi_2)) / \varphi + O(h)$. Similar results hold for

$(\mathcal{V}_{p,1}^\dagger, \mathcal{V}_{p,2}^\dagger, \mathcal{V}_{p,3}^\dagger, \mathcal{V}_{p,4}^\dagger)$. By tedious algebra, it can be verified that $\mathcal{B}_p^\dagger = \mathbf{Q}^{(kl)} \Upsilon^k \Upsilon^l$. By (S7) in the supplement (ruc-econ.github.io/supplement_RD.pdf) and simple algebra,

$$\mathcal{B}_p^\dagger - \mathcal{B}_p^\dagger = \left\{ (\Delta_{\mathcal{M}_+}/\Delta_+ - \Delta_{\mathcal{M}_-}/\Delta_-) - (\Delta_{\mathcal{Z}_+}/\Delta_+ - \Delta_{\mathcal{Z}_-}/\Delta_-)^\top \gamma_\Delta \right\}^2 / \Sigma_\Delta = (\mathcal{B}_p^{\text{EL}} h^{p+1})^2 / \mathcal{V}_p^{\text{EL}} + O(h^{2p+3}). \quad (29)$$

It follows that $\mathcal{C}_p^{\text{pre}}(n, h) = \mathcal{C}_p(n, h) + O(\bar{n} \|\Delta_{\mathcal{U}}\|^2 h + n^{-1})$ and $\Pr[LR^* \leq x] = F_{\chi_1^2}(x) - \mathcal{C}_p(n, h) x f_{\chi_1^2}(x) + O(v_n^\# + \bar{n} \|\Delta_{\mathcal{U}}\|^2 h)$. By $LR_p(\vartheta_0 | h) = LR^* + O_p^*(\log(n)^{5/2} / \bar{n}^{3/2})$ and (S17) in the supplement with $(LR^*, \bar{n}(R_0 + R)^2)$ replaced by $(LR_p(\vartheta_0 | h), LR^*)$, we get the first conclusion. By using the same arguments, we can show that $\Pr[LR_{p+1}(\vartheta_0 | h) \leq x] = F_{\chi_1^2}(x) - \mathcal{C}_{p+1}^{\text{pre}}(n, h) x f_{\chi_1^2}(x) + O(v_n^\#)$. Then by Lemma 1 and similar calculations, $\Pr[LR_{p+1}(\vartheta_0 | h) \leq x] = F_{\chi_1^2}(x) - (nh)^{-1} \mathcal{V}_{p+1}^{\text{LR}} x f_{\chi_1^2}(x) + O(v_n^\diamond)$, where $v_n^\diamond := (\log(n) \|\Delta_{W_{p+1} \otimes U}\|) / \sqrt{\bar{n}} + \log(n)^{5/2} / \bar{n}^{3/2} + \bar{n} \|\Delta_{W_{p+1} \otimes U}\|^2 + h \|\Delta_{W_{p+1} \otimes U}\| + n^{-1}$. By Lemma 1, $\|\Delta_{W_{p+1} \otimes U}\| = O(h^{p+2})$. It is easy to check that the decay rate of $(\log(n) \|\Delta_{W_{p+1} \otimes U}\|) / \sqrt{\bar{n}} + \log(n)^{5/2} / \bar{n}^{3/2} + \bar{n} \|\Delta_{W_{p+1} \otimes U}\|^2 + h \|\Delta_{W_{p+1} \otimes U}\|$ is minimized if h is chosen to balance $\bar{n} \|\Delta_{W_{p+1} \otimes U}\|^2$ and $h \|\Delta_{W_{p+1} \otimes U}\|$. In this case, we have $h \asymp n^{-1/(p+2)}$ and $v_n^\diamond = O(n^{-1})$. The second conclusion follows from applying Taylor expansion to $\Pr[LR_p^c(\vartheta_0 | h) \leq x] = \Pr[LR_{p+1}(\vartheta_0 | h) \leq x (1 + (nh)^{-1} \mathcal{V}_{p+1}^{\text{LR}})]$. ■

Proof of Theorem 4. Let $\vartheta_3 := \mu_{Z,-}$. Now $\mathcal{T}_Z = \delta l_n$ is equivalent to $\vartheta_2 - \vartheta_3 = \delta_n$ where $\delta_n := (\mu_{D,+} - \mu_{D,-}) \delta l_n$ for simplicity. We redefine some notations for notational simplicity: $\theta := (\theta_0, \theta_1, \theta_2, \theta_3)$, $\vartheta := (\vartheta_0, \vartheta_1, \vartheta_2, \vartheta_3)$, $\mathcal{U}_i(\theta) := (W_{p+1,+;i} U_i(\theta_0, \theta_1, \theta_2)^\top, W_{p+1,-;i} U_i(\theta_0, \theta_1, \theta_3)^\top)^\top$ and $\mathcal{U}_i := \mathcal{U}_i(\vartheta)$ (\mathcal{U} defined similarly). Let Γ , $(\mathcal{V}_i(\theta), \mathcal{V}_i, \mathcal{V})$ and $(\mathcal{H}_i, \mathcal{H}, \mathcal{H}_{\dagger,i}, \mathcal{H}_{\dagger})$ be redefined accordingly using the redefined $(\mathcal{U}_i(\theta), \mathcal{U}_i, \mathcal{U})$. Also let $(A^k, A^{kl}, C^{k,n})$, $(\alpha^k, \alpha^{kl}, \alpha^{klm}, \gamma^{k,n}, \gamma^{k;l,n})$, $(M, J, P, \varpi, \Omega)$ and $(Q_{\dagger}, N_{\dagger}, O_{\dagger}, Q, N, O)$ be redefined accordingly. As in the proof of Theorem 3, we apply the rotation by Γ so that $\ell_{p+1}(\theta | h) = \sup_{\lambda} 2 \sum_i \log(1 + \lambda^\top \mathcal{V}_i(\theta))$. Let $\mathcal{V}_{\delta,i} := \mathcal{V}_i(\vartheta_0, \vartheta_1, \vartheta_2, \vartheta_3)$. It is easy to check that Lemma 6 still holds for $(\hat{\eta}_{p+1}, \hat{\lambda}_{p+1}, \tilde{\eta}_{p+1}, \tilde{\lambda}_{p+1})$ under $\mathcal{T}_Z = \delta l_n$. (27) and (28) hold for $(\hat{\eta}_{p+1}, \hat{\lambda}_{p+1})$ with \mathcal{V}_i replaced by $\mathcal{V}_{\delta,i}$ under $\mathcal{T}_Z = \delta l_n$. Similarly, (27) and (28) with $(\mathcal{V}_i, \mathcal{H}_i)$ replaced by $(\mathcal{V}_{\delta,i}, \mathcal{H}_{\dagger,i})$ hold for $(\tilde{\eta}_{p+1}, \tilde{\lambda}_{p+1})$ and $\bar{n}^{-1} \ell_{p+1}(\vartheta_0, \tilde{\vartheta}_{p+1} | h)$ under $\mathcal{T}_Z = \delta l_n$. Let $(\tilde{R}_1^\delta, \tilde{R}_2^\delta)$ be defined by the formulae of $(\tilde{R}_1, \tilde{R}_2)$ in the proof of Lemma 7 with \mathcal{V}_i replaced by $\mathcal{V}_{\delta,i}$. By arguments as in the proof of Lemma 7 and calculations in Ma (2017), $LR_{p+1}(\vartheta_0 | h) = \bar{n} (\tilde{R}_1^\delta + \tilde{R}_2^\delta)^2 + O_p^*(\log(n)^2 / \bar{n})$ under $\mathcal{T}_Z = \delta l_n$. Let $A_\delta^k := \bar{n}^{-1} \sum_i \mathcal{V}_{\delta,i}^{(k)}$ and $A_\delta^{kl} := \bar{n}^{-1} \sum_i \mathcal{V}_{\delta,i}^{(k)} \mathcal{V}_{\delta,i}^{(l)} - \alpha^{kl}$. Let $\tilde{\mathcal{H}}_i := \partial \mathcal{V}_i(\theta) / \partial \theta^\top$ and let $(\tilde{C}^{k,m}, \tilde{\gamma}^{k,m}, \tilde{C}^{k;l,m}, \tilde{\gamma}^{k;l,m})$ be defined by the formulae of $(C^{k,n}, \gamma^{k,n}, C^{k;l,n}, \gamma^{k;l,n})$ with $(\mathcal{H}_i, \mathcal{H})$ replaced by $(\tilde{\mathcal{H}}_i, \tilde{\mathcal{H}})$. It is easy to see that $A_\delta^k = A^k + \delta_n^{(a)} (\tilde{C}^{k,d_\theta+a} + \tilde{\gamma}^{k,d_\theta+a})$ and $A_\delta^{kl} = A^{kl} + \delta_n^{(a)} (\tilde{C}^{k;l,d_\theta+a} + \tilde{\gamma}^{k;l,d_\theta+a}) [k, l] + O_p^*(l_n^2)$. Note that by Lemma 1, $\alpha^k = e_{2d_u,k}^\top \Gamma \Delta_{\mathcal{U}} = O(n^{-1})$. By using these results and replacing A^k with $\hat{A}^k + \alpha^k$, we decompose $\tilde{R}_1^\delta = \tilde{R}_{11}^\delta + \tilde{R}_{10}^\delta + O(l_n^2 h)$, where $\tilde{R}_{11}^\delta := \varpi^{(k)} (A^k + \delta_n^{(a)} \tilde{C}^{k,d_\theta+a})$ and $\tilde{R}_{10}^\delta := \varpi^{(k)} \tilde{\gamma}^{k,d_\theta+a} \delta_n^{(a)}$.

Similarly, we write $\tilde{R}_2^\delta = \tilde{R}_{20}^\delta + \tilde{R}_{21}^\delta + R_2 + O_p^*(\log(n) l_n^3)$, where

$$\begin{aligned} \tilde{R}_{20}^\delta := & \left\{ \frac{1}{2} \varpi^{(m)} M^{(nl)} (\tilde{\gamma}^{m;n,d_\vartheta+a}[m,n]) \tilde{\gamma}^{l,d_\vartheta+b} - \varpi^{(n)} \tilde{\gamma}^{n;d_\vartheta+c,d_\vartheta+b} \tilde{\gamma}^{d_\vartheta+c,d_\vartheta+a} \right. \\ & + \left(\frac{1}{3} \varpi^{(k)} M^{(mv)} M^{(nl)} \alpha^{kmn} - \varpi^{(n)} M^{(mv)} P^{(ol)} \gamma^{m;n,k} \Omega^{(ko)} \right) \tilde{\gamma}^{v,d_\vartheta+a} \tilde{\gamma}^{l,d_\vartheta+b} \\ & \left. + \left((\gamma^{d_\vartheta+c;v,m}[d_\vartheta+c,v]) \Omega^{(mo)} P^{(ok)} \varpi^{(v)} - \alpha^{vm} d_\vartheta+c M^{(vk)} \varpi^{(m)} \right) \tilde{\gamma}^{k,d_\vartheta+b} \tilde{\gamma}^{d_\vartheta+c,d_\vartheta+a} \right\} \delta_n^{(a)} \delta_n^{(b)}, \quad (30) \end{aligned}$$

\tilde{R}_{21}^δ is defined by the sum of $\varpi^{(m)} M^{(nk)} (\tilde{\gamma}^{m;n,d_\vartheta+a}[m,n]) \tilde{A}^k \delta_n^{(a)} / 2 - \varpi^{(n)} (\tilde{\gamma}^{d_\vartheta+a;n,d_\vartheta+b}[d_\vartheta+a,n]) \tilde{A}^{d_\vartheta+a} \delta_n^{(b)}$ and the right hand side of (S11) in the supplement with α^k replaced by $\tilde{\gamma}^{k,d_\vartheta+a} \delta_n^{(a)}$ and R_2 is defined in the proof of Lemma 7. Let $R_0^\delta := \tilde{R}_{10}^\delta + \tilde{R}_{20}^\delta$, $R_1^\delta := \tilde{R}_{11}^\delta + \tilde{R}_{21}^\delta$ and $R^\delta := R_1^\delta + R_2$ so that we have $LR_{p+1}(\vartheta_0 | h) = \bar{n} (R_0^\delta + R^\delta)^2 + O_p^*(l_n h)$. Denote $\bar{\kappa}_0^{ab} := -M^{(kl)} \tilde{\gamma}^{l,d_\vartheta+a} \tilde{\gamma}^{k,d_\vartheta+b}$, $\beta_n^\delta := \bar{n} \bar{\kappa}_0^{ab} \delta_n^{(a)} \delta_n^{(b)}$ and $\bar{\sigma}_{p+1}^2 := \Sigma_\Delta / (\mu_{D,+} - \mu_{D,-})^2$, where Σ_Δ is defined by (11) with p changed to $p+1$. By (24) and (25),

$$\bar{\kappa}_0^{ab} \delta_n^{(a)} \delta_n^{(b)} = (\mu_{D,+} - \mu_{D,-})^2 (Q_\dagger - Q)^{(d_\vartheta+a, d_\vartheta+b)} \delta^{(d_\vartheta+a)} \delta^{(d_\vartheta+b)} \Delta_-^2 l_n^2 = \frac{(\gamma_\Delta^\top \delta)^2 l_n^2}{\bar{\sigma}_{p+1}^2}, \quad (31)$$

where γ_Δ is redefined by replacing p and \mathcal{Z}_- with $p+1$ and $W_{p+1;-}(Z - \vartheta_3)$, $\gamma_\Delta = \gamma_{\text{adj}} + O(h)$ and $\bar{\sigma}_{p+1}^2 = \gamma_{p+1}^{\text{EL}} + O(h)$. By (24), (25), the fact $\Omega P = \begin{bmatrix} 0_{d_\vartheta} & \mathbf{J}^\top \end{bmatrix}^\top$, tedious algebra and Lemma 1, $(R_0^\delta)^2 = \bar{\kappa}_0^{ab} \delta_n^{(a)} \delta_n^{(b)} + \bar{\kappa}_1^{abc} \delta_n^{(a)} \delta_n^{(b)} \delta_n^{(c)} + o(l_n^3)$ where

$$\begin{aligned} \bar{\kappa}_1^{abc} := & -\frac{2}{3} \alpha^{mnk} M^{(kl)} M^{(mv)} M^{(no)} \tilde{\gamma}^{v,d_\vartheta+a} \tilde{\gamma}^{o,d_\vartheta+b} \tilde{\gamma}^{l,d_\vartheta+c} + 2 \alpha^{mn} d_\vartheta+d M^{(mv)} M^{(no)} \tilde{\gamma}^{v,d_\vartheta+a} \tilde{\gamma}^{o,d_\vartheta+b} \tilde{\gamma}^{d_\vartheta+d,d_\vartheta+c} \\ & - 2 \alpha^{k d_\vartheta+e d_\vartheta+d} M^{kl} \tilde{\gamma}^{d_\vartheta+e,d_\vartheta+a} \tilde{\gamma}^{d_\vartheta+d,d_\vartheta+b} \tilde{\gamma}^{l,d_\vartheta+c}. \end{aligned}$$

Let $\bar{\kappa}_2^a := \alpha^{mnk} M^{(km)} M^{(nl)} \tilde{\gamma}^{l,d_\vartheta+a} / 3$, $\bar{\kappa}_3^a := -2 \gamma^{l;d_\vartheta+b,k} \Omega^{(km)} M^{(ml)} \tilde{\gamma}^{d_\vartheta+b,d_\vartheta+a}$, $\tilde{\kappa}_{2,n}^\delta := 1 + (\bar{\kappa}_2^a + \bar{\kappa}_3^a) \delta_n^{(a)}$ and $\bar{\kappa}_4^a := 2 \gamma^{o;n,l} \Omega^{(lv)} P^{(vo)} M^{(nk)} \tilde{\gamma}^{k,d_\vartheta+a}$. By calculation using arguments in the proof of Lemma 7, we have $\kappa_1(\sqrt{\bar{n}} R^\delta) = \tilde{\kappa}_{1,n} + o(l_n)$, $\kappa_2(\sqrt{\bar{n}} R^\delta) = \tilde{\kappa}_{2,n}^\delta + o(l_n)$ and $\kappa_3(\sqrt{\bar{n}} R^\delta) = o(l_n)$, where $\tilde{\kappa}_{1,n}$ is defined in the proof of Lemma 7. Then, $2\sqrt{\bar{n}} R_0^\delta \tilde{\kappa}_{1,n} = (-\bar{\kappa}_2^a + \bar{\kappa}_4^a) \delta_n^{(a)} + O(l_n^2)$. By arguments used to show (S13) and (S14) in the supplement (i.e., Skovgaard, 1981 with $s = p = q = 3$, $\beta_{s,n} = l_n$ and $\lambda_n = O(l_n)$),

$$\Pr \left[\bar{n} (R_0^\delta + R^\delta)^2 \leq x \right] = F \left(x / \bar{\kappa}_{2,n}^\delta \mid \left(\sqrt{\bar{n}} R_0^\delta + \tilde{\kappa}_{1,n} \right)^2 / \bar{\kappa}_{2,n}^\delta \right) + o(l_n). \quad (32)$$

Then by Taylor expansion,

$$F \left(x / \bar{\kappa}_{2,n}^\delta \mid \left(\sqrt{\bar{n}} R_0^\delta + \tilde{\kappa}_{1,n} \right)^2 / \bar{\kappa}_{2,n}^\delta \right) = F \left(x / \bar{\kappa}_{2,n}^\delta \mid \beta_n^\delta \right)$$

$$+ \left\{ \bar{n} \bar{\kappa}_1^{abc} \delta_n^{(a)} \delta_n^{(b)} \delta_n^{(c)} - \beta_n^\delta (\bar{\kappa}_2^a + \bar{\kappa}_3^a) \delta_n^{(a)} + (-\bar{\kappa}_2^a + \bar{\kappa}_4^a) \delta_n^{(a)} \right\} F^{(1)}(x | \beta_n^\delta) + o(l_n). \quad (33)$$

Let $f(\cdot | \eta)$ denote the $\chi_1^2(\eta)$ PDF. By using the recurrence properties of non-central χ^2 (Cohen, 1988), $-xf(x | \eta) = 2\eta F^{(2)}(x | \eta) + (\eta + 1) F^{(1)}(x | \eta)$. By these results and Taylor expansion,

$$\begin{aligned} F(x/\bar{\kappa}_{2,n}^\delta | \beta_n^\delta) &= F(x | \beta_n^\delta) + xf(x | \beta_n^\delta) (1/\bar{\kappa}_{2,n}^\delta - 1) + O(l_n^2) = F(x | \beta_n^\delta) \\ &+ \left(2\beta_n^\delta (\bar{\kappa}_2^a + \bar{\kappa}_3^a) \delta_n^{(a)} \right) F^{(2)}(x | \beta_n^\delta) + \left(\beta_n^\delta (\bar{\kappa}_2^a + \bar{\kappa}_3^a) \delta_n^{(a)} + (\bar{\kappa}_2^a + \bar{\kappa}_3^a) \delta_n^{(a)} \right) F^{(1)}(x | \beta_n^\delta) + o(l_n). \end{aligned} \quad (34)$$

By arguments as in the proof of Lemma 7 and Lemma 1,

$$(\bar{\kappa}_3^a + \bar{\kappa}_4^a) \delta_n^{(a)} = (\mu_{D,+} - \mu_{D,-}) \Delta_- \left(-\Gamma_{\dagger}^{k;l,u} N_{\dagger}^{(ku)} Q_{\dagger}^{(l d_\theta + a)} + \bar{\Gamma}^{a;b,s} \bar{N}^{(as)} \bar{Q}^{(b d_z + a)} \right) \delta^{(a)} l_n = O(l_n h),$$

and similarly, $\bar{\kappa}_3^a \delta_n^{(a)} = O(l_n h)$. It then follows from these results, (33) and (34) that

$$\Pr \left[\bar{n} (R_0^\delta + R^\delta)^2 \leq x \right] = F(x | \beta_n^\delta) + \left(\bar{n} \bar{\kappa}_1^{abc} \delta_n^{(a)} \delta_n^{(b)} \delta_n^{(c)} \right) F^{(1)}(x | \beta_n^\delta) + \left(2\beta_n^\delta \bar{\kappa}_2^a \delta_n^{(a)} \right) F^{(2)}(x | \beta_n^\delta) + o(l_n). \quad (35)$$

By (24), (25), simple algebra and Lemma 1, we can find constants \mathcal{K}_1^{abc} and \mathcal{K}_2^a such that $\bar{n} \bar{\kappa}_1^{abc} \delta_n^{(a)} \delta_n^{(b)} \delta_n^{(c)} = \mathcal{K}_1^{abc} \delta^{(a)} \delta^{(b)} \delta^{(c)} l_n + O(l_n h)$ and $\bar{\kappa}_2^a \delta_n^{(a)} = \mathcal{K}_2^a \delta^{(a)} l_n + O(l_n h)$. By (31) and Lemma 1, $\beta_n^\delta = H \left(\gamma_{\text{adj}}^\top \delta \right)^2 / \mathcal{V}_{p+1}^{\text{EL}} + O(h)$. Let $\mathcal{P}_1(\delta) := \mathcal{K}_1^{abc} \delta^{(a)} \delta^{(b)} \delta^{(c)}$ and $\mathcal{P}_2(\delta) := 2H \gamma_{\text{adj}}^{(a)} \gamma_{\text{adj}}^{(b)} \mathcal{K}_2^c \delta^{(a)} \delta^{(b)} \delta^{(c)} / \mathcal{V}_{p+1}^{\text{EL}}$. The conclusion follows from these results, (35), the fact that $LR_p^c(\vartheta_0 | h) = \bar{n} (R_0^\delta + R^\delta)^2 + O_p^*(l_n h)$ under $\mathcal{T}_Z = \delta l_n$ and (21). ■

Supplement for “Empirical Likelihood Covariate Adjustment for Regression Discontinuity Designs”

Jun Ma*

Zhengfei Yu†

Proof of Lemma 1. (a) follows from LIE and change of variables. (b) is a straightforward extension of [Bickel and Doksum \(2015, Proposition 11.3.1\)](#), which follows from LIE and $(p+1)$ -th order Taylor expansion. For (c), denote $\bar{q}(V, X | h) := h^{-1/2} W_{p;s}^k V$ and $\bar{\Omega} := \{\bar{q}(\cdot | h) : h \in \mathbb{H}\}$. Denote $\mathbb{P}_n^V f := n^{-1} \sum_i f(V_i, X_i)$, $\mathbb{P}^V f := \mathbb{E}[f(V, X)]$ and $\mathbb{G}_n^V := \sqrt{n}(\mathbb{P}_n^V - \mathbb{P}^V)$. Then we have

$$\|\mathbb{G}_n^V\|_{\bar{\Omega}} = \sup_{h \in \mathbb{H}} \left| \frac{1}{\sqrt{nh}} \sum_i (W_{p;s,i}^k V_i - \mathbb{E}[W_{p;s}^k V]) \right|.$$

Let $\sigma_{\bar{\Omega}}^2 := \sup_{f \in \bar{\Omega}} \mathbb{P}^V f^2$. It follows from LIE and change of variables that $\sigma_{\bar{\Omega}}^2 = \sup_{h \in \mathbb{H}} \mathbb{E}[h^{-1} W_{p;s}^{2k} g_{V^2}(X)] = O(1)$. Assume $s = +$ without loss of generality. By definition and the assumption that K is supported on $[-1, 1]$, $\bar{q}(v, x | h) = \mathcal{K}_{p;-}^k(x/h) h^{-1/2} 1(0 < x < h) v$. Since Assumption 3 also implies that $\mathcal{K}_{p;-}^k$ has bounded variation $\forall k \in \mathbb{N}$. By [Giné and Nickl \(2015, Proposition 3.6.12\)](#), $\{x \mapsto \mathcal{K}_{p;-}^k(x/h) : h \in \mathbb{H}\}$ is VC-type with respect to a constant envelope and its VC characteristics are independent of n . By [Kosorok \(2007, Lemma 9.6\)](#), $\{(x, v) \mapsto h^{-1/2} 1(0 < x < h) v : h \in \mathbb{H}\}$ is VC-subgraph with an envelope $(x, v) \mapsto \underline{h}^{-1/2} 1(0 < x < \bar{h}) |v|$ and VC index being at most 3. By [Kosorok \(2007, Theorem 9.3\)](#) and [Chernozhukov et al. \(2014, Corollary A.1\)](#), $\bar{\Omega}$ is VC-type with respect to an envelope $F_{\bar{\Omega}}(v, x) \propto \underline{h}^{-1/2} 1(0 < x < \bar{h}) |v|$. By [Chen and Kato \(2020, Corollary 5.5\)](#), $\mathbb{E}[\|\mathbb{G}_n^V\|_{\bar{\Omega}}] \lesssim \sigma_{\bar{\Omega}} \sqrt{\log(n)} + \log(n) (\mathbb{P}^V |F_{\bar{\Omega}}|^r)^{1/r} n^{1/r} / \sqrt{n}$, where $\mathbb{P}^V |F_{\bar{\Omega}}|^r = O(\bar{h}/\underline{h}^{r/2})$. (c) follows from Markov's inequality. ■

Proof of Lemma 2. Let $\mathcal{L}_{\#} := \{\lambda \in \mathbb{R}^{2d_u} : \|\lambda\| \leq \log(n) / \sqrt{nh}\}$. By

$$\max_i \|\mathcal{U}_i\| / \sqrt{nh} \lesssim \max_i 1(|X_i| \leq \bar{h}) \|U_i\| / \sqrt{n} \leq \left(\sum_i 1(|X_i| \leq \bar{h}) \|U_i\|^{12} \right)^{1/12} / \sqrt{n}$$

This version: May 7, 2022

*School of Economics, Renmin University of China

†Faculty of Humanities and Social Sciences, University of Tsukuba

and Markov's inequality, we have $\max_i \|\mathcal{U}_i\| / \sqrt{nh} = O_p(\bar{n}^{1/12} / \underline{n}^{1/2})$. It follows that $\max_i \sup_{\lambda \in \mathcal{L}_\#} |\lambda^\top \mathcal{U}_i| = O_p(\log(n) (\bar{n}^{1/12} / \underline{n}^{1/2}))$ and $\max_i \sup_{\lambda \in \mathcal{L}_\#} |\lambda^\top \mathcal{U}_i| < 1/2 \forall h \in \mathbb{H}$ wpa1. Therefore, $\mathcal{L}_\# \subseteq \mathcal{L}(\vartheta)$, $\forall h \in \mathbb{H}$ wpa1. Since $S(\cdot, \vartheta)$ is continuous and $\mathcal{L}_\#$ is compact, $\lambda_\# := \operatorname{argmax}_{\lambda \in \mathcal{L}_\#} S(\lambda, \vartheta)$ exists $\forall h \in \mathbb{H}$ wpa1. By the definition of $\lambda_\#$ and second-order Taylor expansion,

$$\begin{aligned} 0 = S(0_{2d_u}, \vartheta) &\leq S(\lambda_\#, \vartheta) = 2 \left(\sqrt{nh} \lambda_\# \right)^\top \bar{\mathcal{U}} - \left(\sqrt{nh} \lambda_\# \right)^\top \left(\frac{1}{nh} \sum_i \frac{\mathcal{U}_i \mathcal{U}_i^\top}{(1 + \dot{\lambda}_\#^\top \mathcal{U}_i)^2} \right) \left(\sqrt{nh} \lambda_\# \right) \\ &\leq 2 \left\| \sqrt{nh} \lambda_\# \right\| \left\| \bar{\mathcal{U}} \right\| - \left(\sqrt{nh} \lambda_\# \right)^\top \left(\frac{1}{nh} \sum_i \frac{\mathcal{U}_i \mathcal{U}_i^\top}{(1 + \max_i \sup_{\lambda \in \mathcal{L}_\#} |\lambda^\top \mathcal{U}_i|)^2} \right) \left(\sqrt{nh} \lambda_\# \right), \quad (\text{S1}) \end{aligned}$$

where $\dot{\lambda}_\#$ is the mean value that lies on the line joining 0_{2d_u} and $\lambda_\#$. Since $\max_i \sup_{\lambda \in \mathcal{L}_\#} |\lambda^\top \mathcal{U}_i| < 1/2 \forall h \in \mathbb{H}$ wpa1, by (S1),

$$0 \leq S(\lambda_\#, \vartheta) \leq 2 \left\| \sqrt{nh} \lambda_\# \right\| \left\| \bar{\mathcal{U}} \right\| - \frac{4}{9} \left(\sqrt{nh} \lambda_\# \right)^\top (\bar{\Delta}_{\mathcal{U}\mathcal{U}^\top} - \Delta_{\mathcal{U}\mathcal{U}^\top}) \left(\sqrt{nh} \lambda_\# \right) - \frac{4}{9} \left(\sqrt{nh} \lambda_\# \right)^\top \Delta_{\mathcal{U}\mathcal{U}^\top} \left(\sqrt{nh} \lambda_\# \right),$$

$\forall h \in \mathbb{H}$ wpa1 and therefore,

$$\varrho_{\min}(\Delta_{\mathcal{U}\mathcal{U}^\top}) \left\| \sqrt{nh} \lambda_\# \right\|^2 \leq \frac{9}{2} \left\| \sqrt{nh} \lambda_\# \right\| \left\| \bar{\mathcal{U}} \right\| + \left\| \bar{\Delta}_{\mathcal{U}\mathcal{U}^\top} - \Delta_{\mathcal{U}\mathcal{U}^\top} \right\| \left\| \sqrt{nh} \lambda_\# \right\|^2, \quad (\text{S2})$$

$\forall h \in \mathbb{H}$ wpa1. Since $\bar{\mathcal{U}} = (nh)^{-1/2} \sum_{i=1}^n (\mathcal{U}_i - \mathbb{E}[\mathcal{U}]) + \sqrt{nh} \Delta_{\mathcal{U}}$, it follows from Lemma 1 that $\left\| \bar{\mathcal{U}} \right\| = O_p(\sqrt{\log(n)})$. It also follows from Lemma 1 that $\bar{\Delta}_{\mathcal{U}\mathcal{U}^\top} - \Delta_{\mathcal{U}\mathcal{U}^\top} = O_p(\sqrt{\log(n)/\underline{n}} + \log(n) (\bar{n}^{1/6}/\underline{n}))$ and $\Delta_{\mathcal{U}\mathcal{U}^\top} = \operatorname{diag}(\psi_{UU^\top, +}, \psi_{UU^\top, -}) + O(\bar{h})$. Since $\operatorname{diag}(\psi_{UU^\top, +}, \psi_{UU^\top, -})$ is positive definite, $\varrho_{\min}(\Delta_{\mathcal{U}\mathcal{U}^\top})$ is bounded away from zero when n is sufficiently large. By assumption, $\left\| \sqrt{nh} \lambda_\# \right\| \leq \log(n)$. It follows from these results and (S2) that $\sqrt{nh} \lambda_\# = O_p(\sqrt{\log(n)})$. By this result, $\Pr[\sqrt{nh} \lambda_\# \leq \log(n)/2, \forall h \in \mathbb{H}] \rightarrow 1$ and therefore, wpa1, $\forall h \in \mathbb{H}$, $\lambda_\#$ is in the interior of $\mathcal{L}_\#$ and the first-order condition is satisfied: $\partial S(\lambda, \vartheta) / \partial \lambda|_{\lambda=\lambda_\#} = 0_{2d_u}$. Since $S(\cdot, \vartheta)$ is concave, $\lambda_\#$ attains $\sup_{\lambda \in \mathcal{L}(\vartheta)} S(\lambda, \vartheta) \forall h \in \mathbb{H}$ wpa1 and therefore, $\sup_{\lambda \in \mathcal{L}(\vartheta)} S(\lambda, \vartheta) = S(\lambda_\#, \vartheta) \leq 2 \left\| \sqrt{nh} \lambda_\# \right\| \left\| \bar{\mathcal{U}} \right\| = O_p(\log(n))$. Denote $\lambda_\natural := \sqrt{\log(n) / (nh)} \hat{\mathcal{U}} / \left\| \hat{\mathcal{U}} \right\|$. It can be shown by using similar arguments, boundedness of Θ and $\hat{\mathcal{U}}_i = \mathcal{U}_i - \mathcal{G}_i \hat{\eta}_p$ that $\max_i \left\| \hat{\mathcal{U}}_i \right\| / \sqrt{nh} = O_p(\bar{n}^{1/12} / \underline{n}^{1/2})$. By second-order Taylor expansion,

$$S(\lambda_\natural, \hat{\vartheta}_p) = 2 \left(\sqrt{nh} \lambda_\natural \right)^\top \hat{\mathcal{U}} - \left(\sqrt{nh} \lambda_\natural \right)^\top \left(\frac{1}{nh} \sum_i \frac{\hat{\mathcal{U}}_i \hat{\mathcal{U}}_i^\top}{(1 + \dot{\lambda}_\natural^\top \hat{\mathcal{U}}_i)^2} \right) \left(\sqrt{nh} \lambda_\natural \right)$$

$$\geq 2 \left(\sqrt{nh} \lambda_{\sharp} \right)^{\top} \hat{\mathcal{U}} - \left(\sqrt{nh} \lambda_{\sharp} \right)^{\top} \left(\frac{1}{nh} \sum_i \frac{\hat{\mathcal{U}}_i \hat{\mathcal{U}}_i^{\top}}{\left(1 - \sqrt{\log(n)/(nh)} \left(\max_i \|\hat{\mathcal{U}}_i\| \right) \right)^2} \right) \left(\sqrt{nh} \lambda_{\sharp} \right), \quad (\text{S3})$$

where $\hat{\lambda}_{\sharp}$ is the mean value that lies on the line joining 0_{2d_u} and λ_{\sharp} . Then, $\sqrt{\log(n)} \|\hat{\mathcal{U}}\| \leq S(\lambda_{\sharp}, \hat{\vartheta}_p) + 2 \left((nh)^{-1} \sum_i \|\hat{\mathcal{U}}_i\|^2 \right) \log(n)$, $\forall h \in \mathbb{H}$, wpa1. By $\hat{\mathcal{U}}_i = \mathcal{U}_i - \mathcal{G}_i \hat{\eta}_p$, Lemma 1 and boundedness of Θ , we have $(nh)^{-1} \sum_i \|\hat{\mathcal{U}}_i\|^2 = O_p(1)$. By the definition of $\hat{\vartheta}_p$, $S(\lambda_{\sharp}, \hat{\vartheta}_p) \leq \sup_{\lambda \in \mathcal{L}(\hat{\vartheta}_p)} S(\lambda, \hat{\vartheta}_p) \leq \sup_{\lambda \in \mathcal{L}(\vartheta)} S(\lambda, \vartheta)$. Since $\sup_{\lambda \in \mathcal{L}(\vartheta)} S(\lambda, \vartheta) = O_p(\log(n))$, it follows that $\hat{\mathcal{U}} = O_p(\sqrt{\log(n)})$. Since $\hat{\mathcal{U}} = \bar{\mathcal{U}} - \bar{\Delta}_{\mathcal{G}} \sqrt{nh} \hat{\eta}_p$, then,

$$\left\| \sqrt{nh} \hat{\eta}_p \right\| \sqrt{\varrho_{\min}(\bar{\Delta}_{\mathcal{G}}^{\top} \bar{\Delta}_{\mathcal{G}})} \leq \left\| \bar{\Delta}_{\mathcal{G}} \sqrt{nh} \hat{\eta}_p \right\| \leq \left\| \hat{\mathcal{U}} \right\| + \left\| \bar{\mathcal{U}} \right\|. \quad (\text{S4})$$

By Lemma 1, $\bar{\Delta}_{\mathcal{G}} = \begin{bmatrix} \mu_{G,+}^{\top} & \mu_{G,-}^{\top} \end{bmatrix}^{\top} + O_p(\sqrt{\log(n)/\underline{n}} + \bar{h})$. $\begin{bmatrix} \mu_{G,+}^{\top} & \mu_{G,-}^{\top} \end{bmatrix}^{\top}$ has full column rank, if $\mu_{D,+} \neq \mu_{D,-}$. By using the fact that $|\varrho_{\min}(A) - \varrho_{\min}(B)| \leq \|A - B\|$, $\varrho_{\min}(\bar{\Delta}_{\mathcal{G}}^{\top} \bar{\Delta}_{\mathcal{G}})$ is bounded away from zero $\forall h \in \mathbb{H}$, wpa1. (a) follows easily from this result, (S4) and the fact that $\|\hat{\mathcal{U}}\|$ and $\|\bar{\mathcal{U}}\|$ are both $O_p(\sqrt{\log(n)})$. By $\max_i \|\hat{\mathcal{U}}_i\|/\sqrt{nh} = O_p(\bar{n}^{1/12}/\underline{n}^{1/2})$ and the definition of \mathcal{L}_{\sharp} , $\max_i \sup_{\lambda \in \mathcal{L}_{\sharp}} |\lambda^{\top} \hat{\mathcal{U}}_i| = O_p(\log(n) (\bar{n}^{1/12}/\underline{n}^{1/2}))$ and therefore $\max_i \sup_{\lambda \in \mathcal{L}_{\sharp}} |\lambda^{\top} \hat{\mathcal{U}}_i| < 1/2 \forall h \in \mathbb{H}$ wpa1. Therefore, $\mathcal{L}_{\sharp} \subseteq \mathcal{L}(\hat{\vartheta}_p)$, $\forall h \in \mathbb{H}$ wpa1. Since $S(\cdot, \hat{\vartheta}_p)$ is continuous and \mathcal{L}_{\sharp} is compact, $\hat{\lambda}_{\sharp} := \operatorname{argmax}_{\lambda \in \mathcal{L}_{\sharp}} S(\lambda, \hat{\vartheta}_p)$ exists $\forall h \in \mathbb{H}$ wpa1. By the definition of $\hat{\lambda}_{\sharp}$ and similar arguments used to show (S2), we have $\varrho_{\min}(\Delta_{\mathcal{U}\mathcal{U}^{\top}}) \|\sqrt{nh} \hat{\lambda}_{\sharp}\| \leq \|\hat{\mathcal{U}}\| + \|\hat{\Delta}_{\mathcal{U}\mathcal{U}^{\top}} - \Delta_{\mathcal{U}\mathcal{U}^{\top}}\| \|\sqrt{nh} \hat{\lambda}_{\sharp}\|$. Since $\hat{\Delta}_{\mathcal{U}\mathcal{U}^{\top}} - \Delta_{\mathcal{U}\mathcal{U}^{\top}} = (nh)^{-1} \sum_i \{\mathcal{G}_i \hat{\eta}_p \mathcal{U}_i^{\top} + \mathcal{U}_i \hat{\eta}_p^{\top} \mathcal{G}_i^{\top} + \mathcal{G}_i \hat{\eta}_p \hat{\eta}_p^{\top} \mathcal{G}_i^{\top}\}$, it follows from Lemma 1 and (a) that $\hat{\Delta}_{\mathcal{U}\mathcal{U}^{\top}} - \Delta_{\mathcal{U}\mathcal{U}^{\top}} = O_p(\sqrt{\log(n)/\underline{n}})$ and therefore, $\hat{\Delta}_{\mathcal{U}\mathcal{U}^{\top}} - \Delta_{\mathcal{U}\mathcal{U}^{\top}} = O_p(\sqrt{\log(n)/\underline{n}} + \log(n) (\bar{n}^{1/6}/\underline{n}))$. Since $\|\sqrt{nh} \hat{\lambda}_{\sharp}\| \leq \log(n)$ by construction, it follows from these results that $\sqrt{nh} \hat{\lambda}_{\sharp} = O_p(\sqrt{\log(n)})$. Wpa1, $\forall h \in \mathbb{H}$, $\hat{\lambda}_{\sharp}$ is in the interior of \mathcal{L}_{\sharp} and the first-order condition is satisfied: $\partial S(\lambda, \hat{\vartheta}_p) / \partial \lambda|_{\lambda=\hat{\lambda}_{\sharp}} = 0_{2d_u}$. It follows from the concavity of $S(\cdot, \hat{\vartheta}_p)$ that $\hat{\lambda}_{\sharp}$ also attains $\sup_{\lambda \in \mathcal{L}(\hat{\vartheta}_p)} S(\lambda, \hat{\vartheta}_p) \forall h \in \mathbb{H}$ wpa1. Then (b) follows from setting $\hat{\lambda}_p = \hat{\lambda}_{\sharp}$. (c) and (d) follow from similar arguments. \blacksquare

Proof of Lemma 3. It is shown in the proof of Lemma 2 that $\hat{\lambda}_p$ satisfies the first-order condition which can be written as $\sum_i \hat{\mathcal{U}}_i / (1 + \hat{\lambda}_p^{\top} \hat{\mathcal{U}}_i) = 0_{2d_u} \forall h \in \mathbb{H}$ wpa1. We also showed that $\max_i \|\hat{\mathcal{U}}_i\|/\sqrt{nh} = O_p(\bar{n}^{1/12}/\underline{n}^{1/2})$ and $\sqrt{nh} \hat{\lambda}_p = O_p(\sqrt{\log(n)})$. Therefore, we have $\max_i |\hat{\lambda}_p^{\top} \hat{\mathcal{U}}_i| = o_p(1)$. By $\hat{\mathcal{U}}_i = \mathcal{U}_i - \mathcal{G}_i \hat{\eta}_p$, Lemma 1 and boundedness of Θ , we have $(nh)^{-1} \sum_i \|\hat{\mathcal{U}}_i\|^3 = O_p(1 + \log(n) (\bar{n}^{1/4}/\underline{n}))$ and $(nh)^{-1} \sum_i \|\hat{\mathcal{U}}_i\|^4 = O_p(1 + \log(n) (\bar{n}^{1/3}/\underline{n}))$. By these results, Lemma 2 and simple algebra, $(nh)^{-1} \sum_i \hat{\mathcal{U}}_i \hat{\mathcal{U}}_i^{\top} / (1 + \hat{\lambda}_p^{\top} \hat{\mathcal{U}}_i)^2 = \hat{\Delta}_{\mathcal{U}\mathcal{U}^{\top}} + o_p(1)$. It is shown in the proof of Lemma 2 that $\hat{\Delta}_{\mathcal{U}\mathcal{U}^{\top}} = \operatorname{diag}(\psi_{UU^{\top},+}, \psi_{UU^{\top},-}) + o_p(1)$. Therefore, $\varrho_{\min} \left((nh)^{-1} \sum_i \hat{\mathcal{U}}_i \hat{\mathcal{U}}_i^{\top} / (1 + \hat{\lambda}_p^{\top} \hat{\mathcal{U}}_i)^2 \right)$ is bounded away from zero $\forall h \in \mathbb{H}$, wpa1. By the implicit function

theorem, wpa1 $\forall h \in \mathbb{H}$, there exists a continuously differentiable function $\lambda(\cdot)$ defined on some open neighborhood $\mathbb{B}(\hat{\vartheta}_p)$ of $\hat{\vartheta}_p$ such that $\hat{\lambda}_p = \lambda(\hat{\vartheta}_p)$ and $(nh)^{-1} \sum_i \mathcal{U}_i(\theta) / (1 + \lambda(\theta)^\top \mathcal{U}_i(\theta)) = 0_{2d_u} \forall \theta \in \mathbb{B}(\hat{\vartheta}_p)$. Since $S(\cdot, \theta)$ is concave, $S(\lambda(\theta), \theta) = \sup_{\lambda \in \mathcal{L}(\theta)} S(\lambda, \theta)$ and $\hat{\vartheta}_p = \operatorname{argmin}_{\theta \in \mathbb{B}(\hat{\vartheta}_p)} S(\lambda(\theta), \theta)$. By the chain rule and $\sum_i \hat{\mathcal{U}}_i / (1 + \hat{\lambda}_p^\top \hat{\mathcal{U}}_i) = 0_{2d_u}$, the first-order condition for $\hat{\vartheta}_p$ can be written as $\sum_i \mathcal{G}_i^\top \hat{\lambda}_p / (1 + \hat{\lambda}_p^\top \hat{\mathcal{U}}_i) = 0_{2d_u}$, which holds $\forall h \in \mathbb{H}$ wpa1. By simple algebra we have

$$0_{2d_u} = \sum_i \left\{ \hat{\mathcal{U}}_i - \hat{\mathcal{U}}_i \hat{\mathcal{U}}_i^\top \hat{\lambda}_p + \frac{\hat{\mathcal{U}}_i (\hat{\mathcal{U}}_i^\top \hat{\lambda}_p)^2}{1 + \hat{\lambda}_p^\top \hat{\mathcal{U}}_i} \right\} \text{ and } 0_{d_\vartheta} = \sum_i \left\{ \mathcal{G}_i^\top \hat{\lambda}_p - \frac{\mathcal{G}_i^\top \hat{\lambda}_p (\hat{\lambda}_p^\top \hat{\mathcal{U}}_i)}{1 + \hat{\lambda}_p^\top \hat{\mathcal{U}}_i} \right\}. \quad (\text{S5})$$

By $\max_i |\hat{\lambda}_p^\top \hat{\mathcal{U}}_i| = o_p(1)$, $(nh)^{-1} \sum_i \|\hat{\mathcal{U}}_i\| = O_p(1)$, $(nh)^{-1} \sum_i \|\hat{\mathcal{U}}_i\|^3 = O_p(1 + \log(n) (\bar{n}^{1/4}/\underline{n}))$ and Lemma 2, $(nh)^{-1/2} \sum_i \left\{ \hat{\mathcal{U}}_i (\hat{\mathcal{U}}_i^\top \hat{\lambda}_p)^2 \right\} / (1 + \hat{\lambda}_p^\top \hat{\mathcal{U}}_i) = O_p(v_n^\dagger)$ and $(nh)^{-1/2} \sum_i \left\{ \mathcal{G}_i^\top \hat{\lambda}_p (\hat{\lambda}_p^\top \hat{\mathcal{U}}_i) \right\} / (1 + \hat{\lambda}_p^\top \hat{\mathcal{U}}_i) = O_p(v_n^\dagger)$, where $v_n^\dagger := \log(n) / \sqrt{\underline{n}} + \log(n)^2 (\bar{n}^{1/4}/\underline{n}^{3/2})$. By these results and $\hat{\mathcal{U}}_i = \mathcal{U}_i - \mathcal{G}_i \hat{\eta}_p$, (S5) can be written as $\hat{\Delta}_{\mathcal{U}\mathcal{U}^\top} \sqrt{nh} \hat{\lambda}_p + \bar{\Delta}_{\mathcal{G}} \sqrt{nh} \hat{\eta}_p = \bar{\mathcal{U}} + O_p(v_n^\dagger)$ and $\bar{\Delta}_{\mathcal{G}}^\top \sqrt{nh} \hat{\lambda}_p = O_p(\log(n) / \sqrt{\underline{n}})$. By $\hat{\Delta}_{\mathcal{U}\mathcal{U}^\top} - \Delta_{\mathcal{U}\mathcal{U}^\top} = O_p(\sqrt{\log(n)/\underline{n}} + \log(n) (\bar{n}^{1/6}/\underline{n}))$, $\bar{\Delta}_{\mathcal{G}} - \Delta_{\mathcal{G}} = O_p(\sqrt{\log(n)/\underline{n}})$ and Lemma 2, we have

$$\Delta_{\mathcal{U}\mathcal{U}^\top} \sqrt{nh} \hat{\lambda}_p + \Delta_{\mathcal{G}} \sqrt{nh} \hat{\eta}_p = \bar{\mathcal{U}} + O_p(v_n^\dagger) \text{ and } \Delta_{\mathcal{G}}^\top \sqrt{nh} \hat{\lambda}_p = O_p(v_n^\dagger). \quad (\text{S6})$$

Since it follows from Lemma 1 that $\Delta_{\mathcal{U}\mathcal{U}^\top} = \operatorname{diag}(\psi_{\mathcal{U}\mathcal{U}^\top, +}, \psi_{\mathcal{U}\mathcal{U}^\top, -}) + O(\bar{h})$ and $\Delta_{\mathcal{G}} = \begin{bmatrix} \mu_{G,+}^\top & \mu_{G,-}^\top \end{bmatrix}^\top + O(\bar{h}^{p+1})$, $\Delta_{\mathcal{U}\mathcal{U}^\top}$ and $\Delta_{\mathcal{G}}^\top \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1} \Delta_{\mathcal{G}}$ are invertible $\forall h \in \mathbb{H}$, when n is sufficiently large. (a) follows from

$$\begin{bmatrix} \Delta_{\mathcal{U}\mathcal{U}^\top} & \Delta_{\mathcal{G}} \\ \Delta_{\mathcal{G}}^\top & 0_{d_\vartheta \times d_\vartheta} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{Q} & \mathbf{N} \\ \mathbf{N}^\top & -\mathbf{O} \end{bmatrix}$$

and (S6). (b) follows from similar arguments. \blacksquare

Proof of Lemma 4. By Taylor expansion, $S(\hat{\lambda}_p, \hat{\vartheta}_p)$ is equal to the sum of $2\hat{\lambda}_p^\top (\sum_i \hat{\mathcal{U}}_i) - \sum_i (\hat{\lambda}_p^\top \hat{\mathcal{U}}_i)^2$ and a remainder term that is bounded up to a constant by $\sum_i |\hat{\lambda}_p^\top \hat{\mathcal{U}}_i|^3 / (1 - |\hat{\lambda}_p^\top \hat{\mathcal{U}}_i|)^3$. By using $(nh)^{-1} \sum_i \|\hat{\mathcal{U}}_i\|^3 = O_p(1 + \log(n) (\bar{n}^{1/4}/\underline{n}))$ and Lemma 2, $\sum_i |\hat{\lambda}_p^\top \hat{\mathcal{U}}_i|^3 = O_p(\sqrt{\log(n)} v_n^\dagger)$. By these results and $\max_i |\hat{\lambda}_p^\top \hat{\mathcal{U}}_i| = o_p(1)$, $S(\hat{\lambda}_p, \hat{\vartheta}_p) = 2\hat{\lambda}_p^\top (\sum_i \hat{\mathcal{U}}_i) - \sum_i (\hat{\lambda}_p^\top \hat{\mathcal{U}}_i)^2 + O_p(\sqrt{\log(n)} v_n^\dagger)$. It was shown in the proof of Lemma 3 that $\hat{\mathcal{U}} = \hat{\Delta}_{\mathcal{U}\mathcal{U}^\top} (\sqrt{nh} \hat{\lambda}_p) + O_p(v_n^\dagger)$. It follows from these results, Lemma 2 and $\hat{\Delta}_{\mathcal{U}\mathcal{U}^\top} - \Delta_{\mathcal{U}\mathcal{U}^\top} = O_p(\sqrt{\log(n)/\underline{n}} + \log(n) (\bar{n}^{1/6}/\underline{n}))$ that $S(\hat{\lambda}_p, \hat{\vartheta}_p) = (\sqrt{nh} \hat{\lambda}_p)^\top \Delta_{\mathcal{U}\mathcal{U}^\top} (\sqrt{nh} \hat{\lambda}_p) + O_p(\sqrt{\log(n)} v_n^\dagger)$. By Lemma 3 and $\bar{\mathcal{U}} = O_p(\sqrt{\log(n)})$, $S(\hat{\lambda}_p, \hat{\vartheta}_p) = \bar{\mathcal{U}}^\top \mathbf{Q} \bar{\mathcal{U}} + O_p(\sqrt{\log(n)} v_n^\dagger)$. Similarly, we have $S(\tilde{\lambda}_p, \vartheta_0, \tilde{\vartheta}_p) = \bar{\mathcal{U}}^\top \mathbf{Q}_\dagger \bar{\mathcal{U}} + O_p(\sqrt{\log(n)} v_n^\dagger)$. By definition, $LR_p(\vartheta_0 | h) = S(\tilde{\lambda}_p, \vartheta_0, \tilde{\vartheta}_p) - S(\hat{\lambda}_p, \hat{\vartheta}_p)$. Therefore, $LR_p(\vartheta_0 | h) =$

$\bar{U}^\top (Q_\dagger - Q) \bar{U} + O_p \left(\sqrt{\log(n)} v_n^\dagger \right)$. Then, by straightforward algebraic calculations,

$$\begin{aligned} Q_\dagger - Q &= \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1} \left\{ \Delta_{\mathcal{G}} \left(\Delta_{\mathcal{G}}^\top \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1} \Delta_{\mathcal{G}} \right)^{-1} \Delta_{\mathcal{G}}^\top - \Delta_{\mathcal{G}_\dagger} \left(\Delta_{\mathcal{G}_\dagger}^\top \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1} \Delta_{\mathcal{G}_\dagger} \right)^{-1} \Delta_{\mathcal{G}_\dagger}^\top \right\} \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1} \\ &= \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1} \left(\Delta_{\mathcal{G}_0} - \Delta_{\mathcal{G}_\dagger} \Phi_{\dagger\dagger}^{-1} \Phi_{\dagger 0} \right) \left(\Phi_{00} - \Phi_{0\dagger} \Phi_{\dagger\dagger}^{-1} \Phi_{\dagger 0} \right)^{-1} \left(\Delta_{\mathcal{G}_0}^\top - \Phi_{0\dagger} \Phi_{\dagger\dagger}^{-1} \Delta_{\mathcal{G}_\dagger}^\top \right) \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1}. \end{aligned} \quad (\text{S7})$$

Then by this result, (12), (14) and $\left(\Phi_{00} - \Phi_{0\dagger} \Phi_{\dagger\dagger}^{-1} \Phi_{\dagger 0} \right)^{-1} = \Sigma_\Delta / (\Delta_{\mathcal{D}_+} / \Delta_+ - \Delta_{\mathcal{D}_-} / \Delta_-)^2$

$$\begin{aligned} \bar{U}^\top (Q_\dagger - Q) \bar{U} &= \left\{ e_{d_u, 1}^\top \Phi_\pm^{-1} (\bar{U}_+ / \Delta_+ - \bar{U}_- / \Delta_-) \right\}^2 \Sigma_\Delta \\ &= \left\{ (\bar{\mathcal{M}}_+ / \Delta_+ - \bar{\mathcal{M}}_- / \Delta_-) - (\bar{\mathcal{Z}}_+ / \Delta_+ - \bar{\mathcal{Z}}_- / \Delta_-)^\top \gamma_\Delta \right\}^2 / \Sigma_\Delta. \end{aligned} \quad (\text{S8})$$

By using $\gamma_\Delta = \gamma_{\text{adj}} + O(\bar{h})$ and (11), $\Sigma_\Delta = \Delta_{\mathcal{E}^2} / \varphi^2 + O(\bar{h})$. By $\|\bar{U}\| = O_p \left(\sqrt{\log(n)} \right)$ and $\gamma_\Delta = \gamma_{\text{adj}} + O(\bar{h})$, the numerator on the right hand side of the second equality in (S8) is $\left\{ (nh)^{-1/2} \sum_i \mathcal{E}_i \right\}^2 + O_p \left(\log(n) \bar{h} \right)$. Let $\tilde{q}(T_i, X_i | h) := h^{-1/2} \mathcal{E}_i / \sqrt{\Delta_{\mathcal{E}^2}}$ and $\tilde{\mathfrak{Q}} := \{\tilde{q}(\cdot | h) : h \in \mathbb{H}\}$. Then it is clear that $\left\{ (nh)^{-1/2} \sum_i \mathcal{E}_i \right\}^2 / \Delta_{\mathcal{E}^2} = \left\{ \mathbb{G}_n^T \tilde{q}(\cdot | h) \right\}^2$ and therefore, $LR_p(\vartheta_0 | h) = \left\{ \mathbb{G}_n^T \tilde{q}(\cdot | h) \right\}^2 + O_p \left(\log(n) \bar{h} + \sqrt{\log(n)} v_n^\dagger \right)$. Also denote $\mathfrak{Q} := \{q(\cdot | h) : h \in \mathbb{H}\}$ and $\mathfrak{D} := \{q(\cdot | h) - \tilde{q}(\cdot | h) : h \in \mathbb{H}\}$. By similar arguments as in the proof of Lemma 1, $\tilde{\mathfrak{Q}}$ and \mathfrak{Q} are both VC-type with respect to the envelopes $(F_{\tilde{\mathfrak{Q}}}, F_{\mathfrak{Q}})$ satisfying $F_{\tilde{\mathfrak{Q}}}(T_i, X_i) \propto \underline{h}^{-1/2} 1(|X_i| \leq \bar{h}) |\epsilon_i - \mu_\epsilon| / \sqrt{\inf_{h \in \mathbb{H}} \Delta_{\mathcal{E}^2}}$ and $F_{\mathfrak{Q}}(T_i, X_i) \propto \underline{h}^{-1/2} 1(|X_i| \leq \bar{h}) |\epsilon_i - \mu_\epsilon| / \sqrt{\xi(|X_i|) f_{|X|}(|X_i|) \omega_{p;+}^{0,2}}$, respectively. By change of variables, $\mathbb{P}^T F_{\tilde{\mathfrak{Q}}}^{12} \asymp \mathbb{P}^T F_{\mathfrak{Q}}^{12} = O(\bar{h} / \underline{h}^6)$. By Chernozhukov et al. (2014, Lemma A.6), \mathfrak{D} is VC-type with respect to the envelope $F_{\mathfrak{D}} = F_{\tilde{\mathfrak{Q}}} + F_{\mathfrak{Q}}$. Let

$$\sigma_{\mathfrak{D}}^2 := \sup_{f \in \mathfrak{D}} \mathbb{P}^T f^2 = \sup_{h \in \mathbb{H}} \mathbb{E} \left[(q(T, X | h) - \tilde{q}(T, X | h))^2 \right].$$

By LIE and the fact that $(W_{p;+} + W_{p;-})^2 = \mathcal{K}_{p;+}(|X|/h)$,

$$\begin{aligned} \mathbb{E} \left[(q(T, X | h) - \tilde{q}(T, X | h))^2 \right] &= \mathbb{E} \left[\frac{1}{h} (W_{p;+} + W_{p;-})^2 (\epsilon - \mu_\epsilon)^2 \left(\frac{1}{\sqrt{\Delta_{\mathcal{E}^2}}} - \frac{1}{\sqrt{\xi(|X|) f_{|X|}(|X|) \omega_{p;+}^{0,2}}} \right)^2 \right] \\ &= \int_0^\infty \frac{1}{h} \mathcal{K}_{p;+} \left(\frac{z}{h} \right)^2 \left(\sqrt{\frac{\xi(z) f_{|X|}(z)}{\Delta_{\mathcal{E}^2}}} - \frac{1}{\sqrt{\omega_{p;+}^{0,2}}} \right)^2 dz. \end{aligned} \quad (\text{S9})$$

Note that $\Delta_{\mathcal{E}^2} = \int_0^\infty h^{-1} \mathcal{K}_{p;+}(z/h)^2 \xi(z) f_{|X|}(z) dz$ and therefore, it follows from mean value expansion and (S9) that $\sigma_{\mathfrak{D}}^2 = O(\bar{h}^2)$. By Chen and Kato (2020, Corollary 5.5), $\mathbb{E} [\|\mathbb{G}_n^T\|_{\mathfrak{D}}] \lesssim \sigma_{\mathfrak{D}} \sqrt{\log(n)} + \log(n) \|F_{\mathfrak{D}}\|_{\mathbb{P}^T, 12} n^{1/12} / \sqrt{n}$ and therefore, $\mathbb{E} [\|\mathbb{G}_n^T\|_{\mathfrak{D}}] = O \left(\sqrt{\log(n)} \cdot \bar{h} + \log(n) (\bar{n}^{1/12} / \underline{n}^{1/2}) \right)$. Let $\sigma_{\tilde{\mathfrak{Q}}}^2 := \sup_{f \in \tilde{\mathfrak{Q}}} \mathbb{P}^T f^2$ and $\sigma_{\mathfrak{Q}}^2 := \sup_{f \in \mathfrak{Q}} \mathbb{P}^T f^2$. It is easy to see that $\mathbb{P}^T f^2 = 1$, if $f \in \mathfrak{Q}$ or $f \in \tilde{\mathfrak{Q}}$ and therefore,

$\sigma_{\tilde{\Omega}}^2 = \sigma_{\Omega}^2 = 1$. Similarly, $\mathbb{E} \left[\|\mathbb{G}_n^T\|_{\tilde{\Omega}} \right] \lesssim \sigma_{\tilde{\Omega}} \sqrt{\log(n)} + \log(n) \|F_{\tilde{\Omega}}\|_{\mathbb{P}^T, 12} n^{1/12} / \sqrt{n}$ and a similar inequality with $\tilde{\Omega}$ replaced by Ω holds. Therefore, $\mathbb{E} \left[\|\mathbb{G}_n^T\|_{\tilde{\Omega}} \right] \asymp \mathbb{E} \left[\|\mathbb{G}_n^T\|_{\Omega} \right] = O \left(\sqrt{\log(n)} \right)$. Then it follows from Markov's inequality that $\left\{ \mathbb{G}_n^T \tilde{q}(\cdot | h) \right\}^2 - \left\{ \mathbb{G}_n^T q(\cdot | h) \right\}^2 = O_p \left(\log(n) \bar{h} + \log(n)^{3/2} (\bar{n}^{1/12} / \underline{n}^{1/2}) \right)$. The conclusion follows from this result and $LR_p(\vartheta_0 | h) = \left\{ \mathbb{G}_n^T \tilde{q}(\cdot | h) \right\}^2 + O_p \left(\log(n) \bar{h} + \sqrt{\log(n)} v_n^\dagger \right)$. ■

Proof of Lemma 5. Let $r_n := \sqrt{\bar{n}/\log(n)}$, $\bar{V}_i := V_i 1(V_i > r_n)$, $\underline{V}_i := V_i 1(V_i \leq r_n)$ and (\underline{V}, \bar{V}) be defined similarly. Then write $\bar{n}^{-1/2} \sum_i (W_{p;s,i}^k V_i - \mathbb{E}[W_{p;s,i}^k V_i]) = \bar{\mathcal{W}} + \underline{\mathcal{W}}$, where $\bar{\mathcal{W}} := \bar{n}^{-1/2} \sum_i (W_{p;s,i}^k \bar{V}_i - \mathbb{E}[W_{p;s,i}^k \bar{V}_i])$ and $\underline{\mathcal{W}} := \bar{n}^{-1/2} \sum_i (W_{p;s,i}^k \underline{V}_i - \mathbb{E}[W_{p;s,i}^k \underline{V}_i])$. Let $\sigma_{\underline{\mathcal{W}}}^2 := \text{Var}[h^{-1/2} W_{p;s,i}^k \underline{V}_i]$. By $\sigma_{\underline{\mathcal{W}}}^2 \leq \mathbb{E}[h^{-1} W_{p;s,i}^{2k} \underline{V}_i^2]$, LIE and change of variables, $\sigma_{\underline{\mathcal{W}}}^2 = O(1)$. $|W_{p;s,i}^k \underline{V}_i - \mathbb{E}[W_{p;s,i}^k \underline{V}_i]|$ is bounded by an upper bound that is proportional to r_n . Let $c > 0$ denote an arbitrary positive constant. By [Giné and Nickl \(2015, Theorem 3.1.7 and Equation 3.24\)](#) with $u = \log(n^c)$, $\Pr \left[|\underline{\mathcal{W}}| \geq \left(\sqrt{2c\sigma_{\underline{\mathcal{W}}}^2} + c/3 \right) \sqrt{\log(n)} \right] \leq 2n^{-c}$. By $\sigma_{\underline{\mathcal{W}}}^2 = O(1)$ and taking c to be sufficiently large, $\underline{\mathcal{W}} = O_p^* \left(\sqrt{\log(n)} \right)$. By Markov's inequality, the fact that $\bar{V}^2 \leq \bar{V}^2 |V/r_n|^3$ and change of variables, $\Pr \left[|\bar{\mathcal{W}}| \geq \sqrt{\log(n)} \right] \leq \mathbb{E} \left[h^{-1} W_{p;s,i}^{2k} \bar{V}^2 \right] / \log(n) \leq \mathbb{E} \left[h^{-1} W_{p;s,i}^{2k} |V|^5 \right] / (r_n^3 \cdot \log(n)) = O \left(\log(n) / \bar{n}^{3/2} \right)$ and therefore, $\bar{\mathcal{W}} = O_p^* \left(\sqrt{\log(n)} \right)$. ■

Proof of Lemma 6. By Markov's inequality, $\Pr \left[\bar{n}^{-1} \sum_i \|\mathcal{U}_i\|^5 > \Delta_{\|\mathcal{U}\|^5} + c \right]$ is bounded above by the fourth central moment of $\bar{n}^{-1} \sum_i \|\mathcal{U}_i\|^5$ divided by c^4 , where $c > 0$ is an arbitrary positive constant. By straightforward calculation and change of variables, its fourth central moment is bounded above by $3n^{-2} \left(\mathbb{E} \left[h^{-2} \|\mathcal{U}\|^{10} \right] \right)^2 + n^{-3} \mathbb{E} \left[h^{-4} \|\mathcal{U}\|^{20} \right] = O(\bar{n}^{-2})$. Therefore, $\bar{n}^{-1} \sum_i \|\mathcal{U}_i\|^5 = O_p^*(1)$ and by $\max_i \|\mathcal{U}_i\| \leq \left(\sum_i \|\mathcal{U}_i\|^5 \right)^{1/5}$, $\max_i \|\mathcal{U}_i\| = O_p^*(\bar{n}^{1/5})$. Then, by this result and the definition of $\mathcal{L}_\#$, $\Pr \left[\max_i \sup_{\lambda \in \mathcal{L}_\#} |\lambda^\top \mathcal{U}_i| \geq 1/2 \right]$ is bounded above by $\Pr \left[\max_i \|\mathcal{U}_i\| \geq (\sqrt{\bar{n}/\log(n)})/2 \right] = O(\bar{n}^{-2})$. Therefore, $\mathcal{L}_\# \subseteq \mathcal{L}(\vartheta)$ wp* and $\lambda_\# := \arg\max_{\lambda \in \mathcal{L}_\#} S(\lambda, \vartheta)$ exists wp*. By using $\bar{\mathcal{U}} = O_p^* \left(\sqrt{\log(n)} \right)$ and $\bar{\Delta}_{\mathcal{U}\mathcal{U}^\top} - \Delta_{\mathcal{U}\mathcal{U}^\top} = O_p^* \left(\sqrt{\log(n)/\bar{n}} \right)$, which follow from Lemma 5, and repeating the steps in the proof of Lemma 2, $\sqrt{\bar{n}} \lambda_\# = O_p^* \left(\sqrt{\log(n)} \right)$. Then, $\sqrt{\bar{n}} \lambda_\# \leq \log(n)/2$ wp* and $S(\lambda_\#, \vartheta) = \sup_{\lambda \in \mathcal{L}(\vartheta)} S(\lambda, \vartheta) = O_p^*(\log(n))$. By similar arguments, boundedness of Θ and $\hat{\mathcal{U}}_i = \mathcal{U}_i - \mathcal{G}_i \hat{\eta}_p$, $\max_i \|\hat{\mathcal{U}}_i\| = O_p^*(\bar{n}^{1/5})$ and $\bar{n}^{-1} \sum_i \|\hat{\mathcal{U}}_i\|^2 = O_p^*(1)$. By repeating the steps in the proof of Lemma 2, $\sqrt{\log(n)} \|\hat{\mathcal{U}}\| \leq \sup_{\lambda \in \mathcal{L}(\vartheta)} S(\lambda, \vartheta) + 2 \left(\bar{n}^{-1} \sum_i \|\hat{\mathcal{U}}_i\|^2 \right) \log(n) = O_p^*(\log(n))$. (a) follows from (S4), $\bar{\mathcal{U}} = O_p^* \left(\sqrt{\log(n)} \right)$, $\hat{\mathcal{U}} = O_p^* \left(\sqrt{\log(n)} \right)$ and the fact that $\varrho_{\min} \left(\bar{\Delta}_{\mathcal{G}}^\top \bar{\Delta}_{\mathcal{G}} \right)$ is bounded away from zero wp*, which follows from Lemmas 1 and 5. The proof of (b) parallels that of Lemma 2(b) and uses the fact $\hat{\Delta}_{\mathcal{U}\mathcal{U}^\top} - \Delta_{\mathcal{U}\mathcal{U}^\top} = O_p^* \left(\sqrt{\log(n)/\bar{n}} \right)$. (c) and (d) follow from similar arguments. ■

Proof of Lemma 7. A decomposition $LR^* = \bar{n} \left(\tilde{R}_1^2 + 2\tilde{R}_1 \tilde{R}_2 + 2\tilde{R}_1 \tilde{R}_3 + \tilde{R}_2^2 \right)$ can be derived. \tilde{R}_k is a homogeneous k -th order polynomial of $(A^k, A^{kl}, A^{klm}, C^{k,n}, C^{k;1,n})$ so that $\tilde{R}_1 = O_p^* \left(\sqrt{\log(n)/\bar{n}} \right)$, $\tilde{R}_2 = O_p^*(\log(n)/\bar{n})$ and $\tilde{R}_3 = O_p^* \left((\log(n)/\bar{n})^{3/2} \right)$. $-M$ is a projection matrix onto the orthogonal complement of the column space of Π_\dagger . Let ϖ_0 be a vector spanning the one-dimensional orthogonal complement of

the column space of Π_+ so that $-\mathbf{M} = \varpi_0 (\varpi_0^\top \varpi_0)^{-1} \varpi_0^\top$. Let $\varpi := \varpi_0 / \sqrt{\varpi_0^\top \varpi_0}$. Then, $\varpi^\top \varpi = 1$ and $-\mathbf{M} = \varpi \varpi^\top$. The expressions of $(\tilde{R}_1, \tilde{R}_2, \tilde{R}_3)$ can be readily obtained in a special case of [Ma \(2017\)](#). Algebraic calculations in [Ma \(2017\)](#) show that by setting $\tilde{R}_1 := \varpi^{(k)} A^k$,

$$\begin{aligned} \tilde{R}_2 := & \frac{1}{2} \mathbf{M}^{(mk)} \varpi^{(n)} A^{mn} A^k - \varpi^{(n)} A^{n d_\vartheta + a} A^{d_\vartheta + a} + \left\{ \frac{1}{3} \alpha^{vmn} \mathbf{M}^{(vl)} \mathbf{M}^{(mk)} \varpi^{(n)} - \gamma^{m;v,o} \Omega^{(on)} \mathbf{P}^{(nk)} \mathbf{M}^{(ml)} \varpi^{(v)} \right\} \\ & \times A^l A^k + \left\{ (\gamma^{d_\vartheta + a;v,m} [d_\vartheta + a, v]) \Omega^{(mo)} \mathbf{P}^{(ok)} \varpi^{(v)} - \alpha^{vm d_\vartheta + a} \mathbf{M}^{(vk)} \varpi^{(m)} \right\} A^k A^{d_\vartheta + a} - \Omega^{(ko)} \mathbf{P}^{(om)} \varpi^{(l)} \\ & \times C^{l,k} A^m + \left\{ \alpha^{v d_\vartheta + a d_\vartheta + b} \varpi^{(v)} - \gamma^{d_\vartheta + a; d_\vartheta + b, m} \Omega^{(mn)} \varpi^{(n)} \right\} A^{d_\vartheta + a} A^{d_\vartheta + b} + \Omega^{(km)} \varpi^{(m)} C^{d_\vartheta + a, k} A^{d_\vartheta + a}, \quad (\text{S10}) \end{aligned}$$

where $\gamma^{d_\vartheta + a;v,m} [d_\vartheta + a, v]$ denotes $\gamma^{d_\vartheta + a;v,m} + \gamma^{v;d_\vartheta + a,m}$ and \tilde{R}_3 to be given by the formula provided in [Ma \(2017, Appendix D.3\)](#), we have $LR^* = \bar{n} (\tilde{R}_1^2 + 2\tilde{R}_1 \tilde{R}_2 + 2\tilde{R}_1 \tilde{R}_3 + \tilde{R}_2^2)$. (S10) is formally the same as [Ma \(2017, \(D.2\)\)](#) with terms that depend on the second derivatives removed. The expression of \tilde{R}_3 is also essentially the same as that of R_3 in [Ma \(2017, Appendix D.3\)](#) with terms that depend on the higher-order derivatives removed and hence omitted for brevity.

Let $\alpha^k := \Delta_{\mathcal{V}^{(k)}}$ and $\mathring{A}^k := A^k - \alpha^k$. By replacing A^k with $\mathring{A}^k + \alpha^k$, we have $\tilde{R}_1 = \tilde{R}_{10} + \tilde{R}_{11}$, where $\tilde{R}_{10} := \varpi^{(k)} \alpha^k$ and $\tilde{R}_{11} := \varpi^{(k)} \mathring{A}^k$. Similarly, we replace A^k with $\mathring{A}^k + \alpha^k$ to decompose $\tilde{R}_2 = \tilde{R}_{22} + \tilde{R}_{21} + \tilde{R}_{20}$ so that \tilde{R}_{2k} is a homogeneous $(2-k)$ -th order polynomial of $\alpha^1, \dots, \alpha^{2d_u}$:

$$\begin{aligned} \tilde{R}_{21} := & \frac{1}{2} \mathbf{M}^{(mk)} \varpi^{(n)} A^{mn} \alpha^k - \varpi^{(n)} A^{n d_\vartheta + a} \alpha^{d_\vartheta + a} + \frac{2}{3} \alpha^{vmn} \mathbf{M}^{(vl)} \mathbf{M}^{(mk)} \varpi^{(n)} \mathring{A}^l \alpha^k - \gamma^{m;v,o} \Omega^{(on)} \mathbf{P}^{(nk)} \mathbf{M}^{(ml)} \varpi^{(v)} \\ & \times (\mathring{A}^l \alpha^k [l, k]) + \left\{ (\gamma^{d_\vartheta + a;v,m} [d_\vartheta + a, v]) \Omega^{(mo)} \mathbf{P}^{(ok)} \varpi^{(v)} - \alpha^{vm d_\vartheta + a} \mathbf{M}^{(vk)} \varpi^{(m)} \right\} (\alpha^k \mathring{A}^{d_\vartheta + a} [k, d_\vartheta + a]) \\ & - \Omega^{(ko)} \mathbf{P}^{(om)} \varpi^{(l)} C^{l,k} \alpha^m + \left\{ \alpha^{v d_\vartheta + a d_\vartheta + b} \varpi^{(v)} - \gamma^{d_\vartheta + a; d_\vartheta + b, m} \Omega^{(mn)} \varpi^{(n)} \right\} (\alpha^{d_\vartheta + a} \mathring{A}^{d_\vartheta + b} [d_\vartheta + a, d_\vartheta + b]) \\ & + \Omega^{(km)} \varpi^{(m)} C^{d_\vartheta + a, k} \alpha^{d_\vartheta + a}, \quad (\text{S11}) \end{aligned}$$

\tilde{R}_{22} is defined by the right hand side of (S10) with A^k replaced by \mathring{A}^k and $\tilde{R}_{20} := \tilde{R}_2 - \tilde{R}_{22} - \tilde{R}_{21} = O(\|\Delta_{\mathcal{U}}\|^2)$. Let $R_0 := \tilde{R}_{10} + \tilde{R}_{20}$, $R_1 := \tilde{R}_{11} + \tilde{R}_{21}$ and $R_2 := \tilde{R}_{22}$. We decompose $\tilde{R}_3 = \tilde{R}_{33} + \tilde{R}_{32} + \tilde{R}_{31} + \tilde{R}_{30}$ in a similar manner and let $R_3 := \tilde{R}_{33}$. R_3 is given by the formula of \tilde{R}_3 with A^k replaced by \mathring{A}^k . Then, let $R := R_1 + R_2 + R_3$. By Lemma 5, $\tilde{R}_1 + \tilde{R}_2 + \tilde{R}_3 = R_0 + R + O_p^*(\|\Delta_{\mathcal{U}}\| \log(n)/\bar{n})$ and therefore, $LR^* = \bar{n} (R_0 + R)^2 + O_p^*(v_n^\#)$.

Let $\mathcal{F} := (W_p \otimes U, W_p, (W_{p;+}^2, W_{p;-}^2)^\top \otimes (U, U^2), (W_{p;+}^3, W_{p;-}^3)^\top \otimes U^3)$. \mathcal{F}_i is defined analogously and let $d_{\mathcal{F}}$ denote the dimension of \mathcal{F} . It can be shown that $\sqrt{\bar{n}} R := h_n(\bar{\mathcal{F}})$, where $\bar{\mathcal{F}} := \bar{n}^{-1/2} \sum_i (\mathcal{F}_i - \mathbb{E}[\mathcal{F}])$ and h_n is a cubic polynomial. E.g., $\sqrt{\bar{n}} \varpi^{(k)} \mathring{A}^k = \tilde{\varpi}^\top \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1/2} (\bar{n}^{-1/2} \sum_i (\mathcal{U}_i - \mathbb{E}[\mathcal{U}]))$, where $\tilde{\varpi} := \mathbf{S} \begin{bmatrix} \varpi^\top & 0_{d_z}^\top \end{bmatrix}^\top$. It can be shown that other terms on the right hand side of (S11) can also be written as linear functions

of $\bar{\mathcal{F}}$. Similarly, it can be shown by tedious algebra that $\sqrt{\bar{n}}R_2$ and $\sqrt{\bar{n}}R_3$ are homogenous quadratic and cubic polynomials of $\bar{\mathcal{F}}$. A more lucid proof of this fact uses the observation that $\ell_p(\hat{\vartheta}_p | h) = \inf_{\theta_2} \sup_{\lambda_2} 2 \sum_i \log(1 + \lambda_2^\top \bar{\mathcal{U}}_i(\theta_2))$. Let $\mathcal{M}_i(\theta_0, \theta_1) := W_{p,i} \otimes (Y_i - \theta_0 D_i - \theta_1)$. By rearranging the moment conditions, $\ell_p(\hat{\vartheta}_p | h) = \inf_{\theta_0, \theta_1, \theta_2} \sup_{\lambda_1, \lambda_2} 2 \sum_i \log(1 + \lambda_1^\top \mathcal{M}_i(\theta_0, \theta_1) + \lambda_2^\top \bar{\mathcal{U}}_i(\theta_2))$. Let $\mathcal{W}_i(\theta) := (\mathcal{M}_i(\theta_0, \theta_1), \bar{\mathcal{U}}_i(\theta_2))$. $\hat{\vartheta}_p$ and $\hat{\lambda} := (\hat{\lambda}_1, \hat{\lambda}_2)$ satisfy the first-order conditions wp^* :

$$\sum_i \frac{\mathcal{M}_i(\hat{\vartheta}_{p,0}, \hat{\vartheta}_{p,1})}{1 + \hat{\lambda}^\top \mathcal{W}_i(\hat{\vartheta}_p)} = 0_2, \sum_i \frac{\bar{\mathcal{U}}_i(\hat{\vartheta}_{p,2})}{1 + \hat{\lambda}^\top \mathcal{W}_i(\hat{\vartheta}_p)} = 0_{2d_z}, \sum_i \frac{(W_{p,i} \otimes (D_i, 1))^\top \hat{\lambda}_1}{1 + \hat{\lambda}^\top \mathcal{W}_i(\hat{\vartheta}_p)} = 0_2, \sum_i \frac{\bar{\mathcal{G}}_i^\top \hat{\lambda}_2}{1 + \hat{\lambda}^\top \mathcal{W}_i(\hat{\vartheta}_p)} = 0_{d_z}.$$

The third condition implies that $\hat{\lambda}_1 = 0_2$ wp^* . Therefore, $\ell_p(\hat{\vartheta}_p | h) = 2 \sum_i \log(1 + \hat{\lambda}_2^\top \bar{\mathcal{U}}_i(\hat{\vartheta}_{p,2}))$ and the second and fourth conditions are $\sum_i \bar{\mathcal{U}}_i(\hat{\vartheta}_{p,2}) / (1 + \hat{\lambda}_2^\top \bar{\mathcal{U}}_i(\hat{\vartheta}_{p,2})) = 0_{2d_z}$ and $\sum_i \bar{\mathcal{G}}_i^\top \hat{\lambda}_2 / (1 + \hat{\lambda}_2^\top \bar{\mathcal{U}}_i(\hat{\vartheta}_{p,2})) = 0_{d_z}$, which coincide with the first-order conditions of $\inf_{\theta_2} \sup_{\lambda_2} 2 \sum_i \log(1 + \lambda_2^\top \bar{\mathcal{U}}_i(\theta_2))$. Therefore, we have $\ell_p(\hat{\vartheta}_p | h) = \inf_{\theta_2} \sup_{\lambda_2} 2 \sum_i \log(1 + \lambda_2^\top \bar{\mathcal{U}}_i(\theta_2))$. By expansion and Lemma 6, we get approximations for $\hat{\lambda}_2$, $\hat{\vartheta}_{p,2}$ and $\ell_p(\hat{\vartheta}_p | h)$ which are similar to (27) and (28). Then it is clear that by replacing sample averages with sums of their centered versions and population counterparts we can get further approximations which are polynomials in $\bar{n}^{-1/2} \sum_i (\bar{\mathcal{F}}_i - \mathbb{E}[\bar{\mathcal{F}}])$, where $(\bar{\mathcal{F}}_i, \bar{\mathcal{F}})$ are defined by the formulae of $(\mathcal{F}_i, \mathcal{F})$ with (U_i, U) replaced by (\bar{U}_i, \bar{U}) . Similarly, the stochastic expansion of $\ell_p(\vartheta_0, \tilde{\vartheta}_p | h)$ should involve only terms in $\bar{\mathcal{F}}$.

Let $\kappa_j(V)$ denote the j -th cumulant of a random variable V . We follow arguments in the proof of Calonico et al. (2022, Theorem S.1) and apply Skovgaard (1986, Theorem 3.4) with $s = 4$ to $S_n := B^{-1/2} \bar{\mathcal{F}}$ where $B := \text{Var}[\mathcal{F}] / h$. For any $t \in \mathbb{R}^{d_f}$ with $\|t\| = 1$, by change of variables and calculation of the moments (see, e.g., DiCiccio et al., 1988, Page 12), $\kappa_3(t^\top S_n) = \mathbb{E}[(t^\top S_n)^3] = O(\bar{n}^{-1/2})$, $\kappa_4(t^\top S_n) = \mathbb{E}[(t^\top S_n)^4] - 3(\mathbb{E}[(t^\top S_n)^2])^2 = O(\bar{n}^{-1})$ and $\rho_{s,n}(t) := \max\{|\kappa_s(t^\top S_n)| / s!, \sqrt{|\kappa_4(t^\top S_n)| / 4!}\} = O(\bar{n}^{-1/2})$, uniformly in t . Condition I and II of Skovgaard (1986, Theorem 3.4) are satisfied by taking $a_n(t) \propto \sqrt{\bar{n}}$ and $\epsilon_n = \bar{n}^{-3/2}$. Let $\hat{\psi}_V(t) := \mathbb{E}[\exp(it^\top V)]$ denote the characteristic function of a random vector V , where $i := \sqrt{-1}$. Let $\mathcal{F}_s := (W_{p;s}U, W_{p;s}, W_{p;s}^2(U, U^2), W_{p;s}^3(U^3))$, $s \in \{-, +\}$. Then, $\hat{\psi}_{\mathcal{F}}(t) = \mathbb{E}[\exp(it_+^\top \mathcal{F}_+) 1(X \geq 0)] + \mathbb{E}[\exp(it_-^\top \mathcal{F}_-) 1(X < 0)]$, where (t_-, t_+) denote corresponding coordinates of t . By change of variables, $\mathbb{E}[\exp(it_+^\top \mathcal{F}_+) 1(X \geq 0)] = h(f_X(0)E_+(t_+) + O(h)) + \Pr[X > h]$, where E_+ is the characteristic function of $\mathcal{K}_{p,+}(V)(U, 1)$, $\mathcal{K}_{p,+}(V)^2(U, U^2)$, $\mathcal{K}_{p,+}(V)^3(U^3)$, where (V, U) has the joint density given by $(v, u) \mapsto 1(0 \leq v \leq 1)f_{U|X}(u | 0)$. A similar result holds for $\mathbb{E}[\exp(it_-^\top \mathcal{F}_-) 1(X < 0)]$ with $E_-(t_-)$ defined similarly. Therefore, $\hat{\psi}_{\mathcal{F}}(t) = 1 - \Pr[-h < X \leq h] + hf_X(0)(E_+(t_+) + E_-(t_-)) + O(h^2)$. By Assumption 5, the vector-valued functions $(v, u) \mapsto (1, (\mathcal{K}_{p,+}(v), \mathcal{K}_{p,+}(v)^2, \mathcal{K}_{p,+}(v)^3) \otimes (1, u, u^2, u^3))$ are linearly independent. By invoking the same arguments as in the proof of Calonico et al. (2022, Lemma S.9), $\forall \varepsilon > 0$, $\exists c_\varepsilon > 0$ such that $\sup_{\|t\| > \varepsilon} |E_+(t_+)| < 1 - c_\varepsilon$. A similar result holds for E_- . Then by these results, $\forall \varepsilon > 0$, $\exists c_\varepsilon > 0$ such

that $\sup_{\|t\|>\varepsilon} \left| \hat{\Psi}_{\mathcal{F}}(t) \right| < 1 - c_\varepsilon h$, when n is sufficiently large. It follows from this result and arguments in the proof of [Calonico et al. \(2022, Theorem S.1\)](#) that $\forall \delta > 0, \exists c_\delta > 0$ such that $\sup_{\|t\|>\delta\sqrt{n}} \left| \hat{\Psi}_{S_n}(t) \right| \leq (1 - c_\delta h)^n$ when n is sufficiently large. It is also easy to see that $\forall \delta > 0, (1 - c_\delta h)^n \leq \epsilon_n^{d_f/2+2}$, when n is sufficiently large. Therefore, Condition III'' $_\alpha$ of [Skovgaard \(1986, Theorem 3.4 and Remark 3.5\)](#) is satisfied with $\alpha = 1$. Verification of Condition IV of [Skovgaard \(1986, Theorem 3.4\)](#) follows from essentially the same calculations and arguments in the proof of [Calonico et al. \(2022, Theorem S.1\)](#). Now all conditions for [Skovgaard \(1986, Theorem 3.4\)](#) are verified. It shows that S_n admits a valid Edgeworth expansion, i.e., conditions (3.1), (3.2) and (3.3) of [Skovgaard \(1981\)](#) are satisfied with $U_n = S_n$, $s = 4$, $\beta_{s,n} = \bar{n}^{-1}$ and the Edgeworth expansion holds uniformly over the class of all convex sets in \mathbb{R}^{d_f} . Note that we can write $\sqrt{\bar{n}}R = h_n(B^{1/2}S_n)$. Then we apply [Skovgaard \(1981\)](#) to show that the Edgeworth expansion is preserved by smooth transformations. Condition (3.4) of [Skovgaard \(1981\)](#) is satisfied with g_n taken to be $x \mapsto h_n(B^{1/2}x)$ whose the gradient at zero $\nabla g_n(0)$ is given by $\nabla g_n(0) = B^{1/2} \left(\tilde{\omega}^\top \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1/2}, 0_{d_f-2d_u}^\top \right)^\top + O(\|\Delta_{\mathcal{U}}\|)$ by the chain rule. Then we apply [Skovgaard \(1981, Theorem 3.2\)](#) to $f_n(S_n) := B_n^{-1}g_n(S_n)$, where $B_n^2 := \nabla g_n(0)^\top \nabla g_n(0)$. Then, $B_n^2 = \tilde{\omega}^\top \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1/2} (\text{Var}[\mathcal{U}]/h) \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1/2} \tilde{\omega} + O(\|\Delta_{\mathcal{U}}\|) = 1 + O(\|\Delta_{\mathcal{U}}\|)$. Condition I of [Skovgaard \(1981, Assumption 3.1\)](#) is satisfied with $p = 4$. Condition II of [Skovgaard \(1981, Assumption 3.1\)](#) is satisfied with $\lambda_n = O(\bar{n}^{-1/2})$ so that $\lambda_n^{p-1} = o(\bar{n}^{-1})$. Now all conditions for [Skovgaard \(1981, Theorem 3.2\)](#) are verified. It is left to compute the approximate cumulants.

Then we calculate the formal cumulants of $f_n(S_n) = B_n^{-1}\sqrt{\bar{n}}R$. In the calculations, we repeatedly use formulae for moments of products of sample averages (e.g., [DiCiccio et al., 1988, Page 12](#)) and Lemma 1. By definition, $E[R_1] = 0$. We calculate $E[R_2]$, let the remainder term absorb the terms that involve $\alpha^1, \dots, \alpha^{2d_u}$ and get $E[R_2] = \bar{n}^{-1}\bar{\kappa}_1 + O(\|\Delta_{\mathcal{U}}\|/n)$ where $\bar{\kappa}_1 := \alpha^{mnk}M^{(mk)}\varpi^{(n)}/6 - \Omega^{(ko)}P^{(om)}\varpi^{(l)}\gamma^{m;l,k}$. By formulae for third moments and Lemma 1, $E[R_3] = O(\bar{n}^{-2})$. Therefore, $\kappa_1(\sqrt{\bar{n}}R) = \tilde{\kappa}_{1,n} + O(\bar{n}^{-1/2}\|\Delta_{\mathcal{U}}\|h + \bar{n}^{-3/2})$ with $\tilde{\kappa}_{1,n} := \bar{n}^{-1/2}\bar{\kappa}_1$. For the second cumulant, by definition, $\kappa_2(R) = E[R^2] - (E[R])^2$ and by formulae for fifth and sixth moments and Lemma 1, $E[R^2] = E[R_1^2] + 2 \cdot E[R_1R_2] + 2 \cdot E[R_1R_3] + E[R_2^2] + O(\bar{n}^{-3})$. By $R_1 = \tilde{R}_{11} + \tilde{R}_{21}$ and calculation, $E[R_1^2] = E[\tilde{R}_{11}^2] + 2 \cdot E[\tilde{R}_{21}\tilde{R}_{11}] + O(\bar{n}^{-1}\|\Delta_{\mathcal{U}}\|^2)$, $E[R_1R_2] + E[R_1R_3] = E[\tilde{R}_{11}R_2] + E[\tilde{R}_{11}R_3] + O(\bar{n}^{-2}\|\Delta_{\mathcal{U}}\|)$. Then by calculation, $E[\tilde{R}_{11}^2] = \bar{n}^{-1} + O(\|\Delta_{\mathcal{U}}\|^2/n)$ and $2 \cdot E[\tilde{R}_{21}\tilde{R}_{11}] = \bar{n}^{-1}\tilde{\kappa}_{21,n} + O(\|\Delta_{\mathcal{U}}\|^2/n)$, where $\tilde{\kappa}_{21,n} := \alpha^{mno}M^{(no)}M^{(mk)}\alpha^k/3 - 2\gamma^{l;d_\vartheta+a,k}\Omega^{(km)}M^{(ml)}\alpha^{d_\vartheta+a}$. Then, $E[R_1^2] = \bar{n}^{-1}(1 + \tilde{\kappa}_{21,n}) + O(\bar{n}^{-1}\|\Delta_{\mathcal{U}}\|^2)$. Calculation of $2 \cdot E[\tilde{R}_{11}R_2] + 2 \cdot E[\tilde{R}_{11}R_3] + E[R_2^2]$ follows from replication of calculations in [Ma \(2017\)](#) and we can directly use the results therein. By calculations in [Ma \(2017\)](#), we have

$$2 \cdot E[\tilde{R}_{11}R_2] + 2 \cdot E[\tilde{R}_{11}R_3] + E[R_2^2] = \bar{n}^{-2} \sum_{j=1}^8 \bar{\kappa}_{2j} + O(\bar{n}^{-2}\|\Delta_{\mathcal{U}}\|h + \bar{n}^{-3}),$$

for some bounded constants $\bar{\kappa}_{21}, \dots, \bar{\kappa}_{28}$, e.g., $\bar{\kappa}_{21} := \alpha^{vmn} \mathbf{M}^{(vo)} \mathbf{M}^{(ml)} \mathbf{M}^{(nk)} \alpha^{klo} / 3 - \alpha^{vm} d_\theta + a \mathbf{M}^{(vo)} \mathbf{M}^{(mn)} \alpha^{on} d_\theta + a + \alpha^{n} d_\theta + a d_\theta + b \mathbf{M}^{(nm)} \alpha^{m} d_\theta + a d_\theta + b$ and the $O(\bar{n}^{-2} \|\Delta_{\mathcal{U}}\| h + \bar{n}^{-3})$ remainder collects terms that depend on $\alpha^1, \dots, \alpha^{2d_u}$ and higher-order terms from the fourth moment calculation. The expressions of $\bar{\kappa}_{22}, \dots, \bar{\kappa}_{28}$ are also easily obtained from [Ma \(2017\)](#) and hence omitted. Therefore, $\kappa_2(\sqrt{\bar{n}}R) = \tilde{\kappa}_{2,n} + O(\|\Delta_{\mathcal{U}}\|^2 + \bar{n}^{-1} \|\Delta_{\mathcal{U}}\| + \bar{n}^{-2})$, where $\tilde{\kappa}_{2,n} := 1 + \tilde{\kappa}_{21,n} + \tilde{\kappa}_{22,n}$ and $\tilde{\kappa}_{22,n} := \bar{n}^{-1} \left(\sum_{j=1}^8 \bar{\kappa}_{2j} - \bar{\kappa}_1^2 \right)$. By definition, $\kappa_3(R) = \mathbb{E}[R^3] - 3 \cdot \mathbb{E}[R] \mathbb{E}[R^2] + 2(\mathbb{E}[R])^3$ and by $\mathbb{E}[R] = \mathbb{E}[R_2] + O(\bar{n}^{-2})$, $\mathbb{E}[R_2] = O(\bar{n}^{-1})$, $\mathbb{E}[R^2] = \mathbb{E}[R_1^2] + O(\bar{n}^{-2})$ and $\mathbb{E}[R^3] = \mathbb{E}[R_1^3] + 3 \cdot \mathbb{E}[R_2 R_1^2] + O(\bar{n}^{-3})$, which follows from formulae for higher moments, we have $\kappa_3(R) = \mathbb{E}[R_1^3] - 3(\mathbb{E}[R_2 R_1^2] - \mathbb{E}[R_2] \mathbb{E}[R_1^2]) + O(\bar{n}^{-3})$. It is easy to check that $\mathbb{E}[R_1^3] = \mathbb{E}[\tilde{R}_{11}^3] + O(\bar{n}^{-2} \|\Delta_{\mathcal{U}}\|)$, $\mathbb{E}[R_2 R_1^2] = \mathbb{E}[R_2 \tilde{R}_{11}^2] + O(\bar{n}^{-2} \|\Delta_{\mathcal{U}}\|)$. By these results and $\mathbb{E}[R_1^2] = \mathbb{E}[\tilde{R}_{11}^2] + O(\bar{n}^{-1} \|\Delta_{\mathcal{U}}\|)$, $\kappa_3(R) = \mathbb{E}[\tilde{R}_{11}^3] - 3(\mathbb{E}[R_2 \tilde{R}_{11}^2] - \mathbb{E}[R_2] \mathbb{E}[\tilde{R}_{11}^2]) + O(\bar{n}^{-3} + \bar{n}^{-2} \|\Delta_{\mathcal{U}}\|)$. Calculation and expansion of $\mathbb{E}[\tilde{R}_{11}^3] - 3(\mathbb{E}[R_2 \tilde{R}_{11}^2] - \mathbb{E}[R_2] \mathbb{E}[\tilde{R}_{11}^2])$ follows from replication of calculations in [Ma \(2017\)](#). For example, by calculation using formulae for moments ([DiCiccio et al., 1988](#)),

$$\mathbb{E}[\tilde{R}_{11}^3] = n^{-2} \left(\mathbb{E} \left[\left(h^{-1} \varpi^{(k)} \mathcal{V}^{(k)} - \varpi^{(k)} \alpha^k \right)^3 \right] \right) = n^{-2} \mathbb{E} \left[\left(h^{-1} \varpi^{(k)} \mathcal{V}^{(k)} \right)^3 \right] + O(\bar{n}^{-2} \|\Delta_{\mathcal{U}}\| h),$$

and the $O(\bar{n}^{-2} \|\Delta_{\mathcal{U}}\| h)$ remainder collects all terms in the expansion of the third moment which depend on $\alpha^1, \dots, \alpha^{2d_u}$. Note that we can write $\mathbb{E} \left[h^{-1} (\varpi^{(k)} \mathcal{V}^{(k)})^3 \right] = \varpi^{(k)} \varpi^{(l)} \varpi^{(m)} \alpha^{klm}$ in coordinate notations. Similarly, we calculate $\mathbb{E}[R_2 \tilde{R}_{11}^2] - \mathbb{E}[R_2] \mathbb{E}[\tilde{R}_{11}^2]$. We note that coefficients of terms of order \bar{n}^{-2} in $\mathbb{E}[\tilde{R}_{11}^3] - 3(\mathbb{E}[R_2 \tilde{R}_{11}^2] - \mathbb{E}[R_2] \mathbb{E}[\tilde{R}_{11}^2])$ are formally the same as those of the leading terms in the calculation of the formal third cumulant in [Ma \(2017\)](#). Calculations in [Ma \(2017\)](#) show that the sum of these coefficients are exactly zero and therefore, the leading term vanishes so that $\kappa_3(\sqrt{\bar{n}}R) = O(\|\Delta_{\mathcal{U}}\|/\sqrt{\bar{n}} + \bar{n}^{-3/2})$. By this result, the fact that $\kappa_4(R) = \mathbb{E}[R^4] - 3(\mathbb{E}[R^2])^2 - 4 \cdot \mathbb{E}[R] \kappa_3(R) + 2(\mathbb{E}[R])^4$, $\mathbb{E}[R] = O(\bar{n}^{-1})$, $R = \tilde{R}_{11} + \tilde{R}_{21} + R_2 + R_3$ and standard calculations,

$$\begin{aligned} \kappa_4(R) &= \mathbb{E}[R^4] - 3(\mathbb{E}[R^2])^2 + O(\bar{n}^{-3} \|\Delta_{\mathcal{U}}\| + \bar{n}^{-4}) = \left\{ \mathbb{E}[\tilde{R}_{11}^4] - 3(\mathbb{E}[\tilde{R}_{11}^2])^2 \right\} \\ &\quad + 4 \left\{ \mathbb{E}[R_2 \tilde{R}_{11}^3] - 3 \cdot \mathbb{E}[R_2 \tilde{R}_{11}] \mathbb{E}[\tilde{R}_{11}^2] \right\} + 6 \left\{ \mathbb{E}[R_2^2 \tilde{R}_{11}^2] - \mathbb{E}[R_2^2] \mathbb{E}[\tilde{R}_{11}^2] \right\} \\ &\quad + 4 \left\{ \mathbb{E}[R_2 \tilde{R}_{11}^3] - 3 \cdot \mathbb{E}[R_2 \tilde{R}_{11}] \mathbb{E}[\tilde{R}_{11}^2] \right\} + O(\bar{n}^{-3} \|\Delta_{\mathcal{U}}\| + \bar{n}^{-4}). \quad (\text{S12}) \end{aligned}$$

And by standard calculations,

$$\begin{aligned} \mathbb{E}[\tilde{R}_{11}^4] - 3(\mathbb{E}[\tilde{R}_{11}^2])^2 &= n^{-3} \left(\mathbb{E} \left[\left(h^{-1} \varpi^{(k)} \mathcal{V}^{(k)} - \varpi^{(k)} \alpha^k \right)^4 \right] - 3 \left(\mathbb{E} \left[\left(h^{-1} \varpi^{(k)} \mathcal{V}^{(k)} - \varpi^{(k)} \alpha^k \right)^2 \right] \right)^2 \right) \\ &= n^{-3} \left(\mathbb{E} \left[\left(h^{-1} \varpi^{(k)} \mathcal{V}^{(k)} \right)^4 \right] - 3 \left(\mathbb{E} \left[\left(h^{-1} \varpi^{(k)} \mathcal{V}^{(k)} \right)^2 \right] \right)^2 \right) + O(\bar{n}^{-3} \|\Delta_{\mathcal{U}}\| h), \end{aligned}$$

and the $O(\bar{n}^{-3} \|\Delta_{\mathcal{U}}\| h)$ remainder collects all terms that depend on $\alpha^1, \dots, \alpha^{2d_u}$. Similarly, we also calculate $E[R_2 \tilde{R}_{11}^3] - 3 \cdot E[R_2 \tilde{R}_{11}] E[\tilde{R}_{11}^2]$, $E[R_2^2 \tilde{R}_{11}^2] - E[R_2^2] E[\tilde{R}_{11}^2]$ and $E[R_2 \tilde{R}_{11}^3] - 3 \cdot E[R_2 \tilde{R}_{11}] E[\tilde{R}_{11}^2]$ on the right hand side of the second equality in (S12), ignore small-order terms that depend on $\alpha^1, \dots, \alpha^{2d_u}$ and take the sum of the leading terms. We do not need to rework on the calculations since they are formally the same as those done in Ma (2017). Calculations in Ma (2017) show that the sum of the leading terms on the right hand side of (S12) is exactly zero so that it follows from this result and (S12) that $\kappa_4(\sqrt{\bar{n}}R) = O(\bar{n}^{-1} \|\Delta_{\mathcal{U}}\| + \bar{n}^{-2})$. By previous calculations and $B_n = 1 + O(\|\Delta_{\mathcal{U}}\|)$, we get the approximate cumulants for $f_n(S_n)$: $\kappa_1(f_n(S_n)) = B_n^{-1} \tilde{\kappa}_{1,n} + O(\bar{n}^{-1/2} \|\Delta_{\mathcal{U}}\| h + \bar{n}^{-3/2})$, $\kappa_2(f_n(S_n)) = B_n^{-2} \tilde{\kappa}_{2,n} + O(\|\Delta_{\mathcal{U}}\|^2 + \bar{n}^{-1} \|\Delta_{\mathcal{U}}\| + \bar{n}^{-2})$, $\kappa_3(f_n(S_n)) = O(\|\Delta_{\mathcal{U}}\|/\sqrt{\bar{n}} + \bar{n}^{-3/2})$ and $\kappa_4(f_n(S_n)) = O(\bar{n}^{-1} \|\Delta_{\mathcal{U}}\| + \bar{n}^{-2})$.

Let $\phi(\cdot | \mu, \sigma^2)$ denote the PDF of $N(\mu, \sigma^2)$. By applying Skovgaard (1981, Theorem 3.2) to $f_n(S_n) = B_n^{-1} \sqrt{\bar{n}} R$,

$$\Pr[\bar{n}(R_0 + R)^2 \leq x] = \int_{|t + (\sqrt{\bar{n}}R_0)/B_n| \leq \sqrt{x}/B_n} \phi(t | B_n^{-1} \tilde{\kappa}_{1,n}, B_n^{-2} \tilde{\kappa}_{2,n}) dt + O(\|\Delta_{\mathcal{U}}\|/\sqrt{\bar{n}} + \bar{n}^{-3/2}), \quad (\text{S13})$$

uniformly in $x > 0$. By using the recurrence properties of non-central χ^2 (Cohen, 1988) and mean value expansion, we have $\partial F(x | \lambda) / \partial \lambda|_{\lambda=\bar{\lambda}} = -x f_{\chi_1^2}(x) + O(\bar{\lambda})$. By this result, $B_n^2 = 1 + O(\|\Delta_{\mathcal{U}}\|)$, change of variables and mean value expansion,

$$\begin{aligned} & \int_{|t + (\sqrt{\bar{n}}R_0)/B_n| \leq \sqrt{x}/B_n} \phi(t | B_n^{-1} \tilde{\kappa}_{1,n}, B_n^{-2} \tilde{\kappa}_{2,n}) dt = \int_{|t| \leq \sqrt{x/\tilde{\kappa}_{2,n}}} \phi\left(t | \left(\sqrt{\bar{n}}R_0 + \tilde{\kappa}_{1,n}\right)/\sqrt{\tilde{\kappa}_{2,n}}, 1\right) dt \\ & = F\left(\frac{x}{\tilde{\kappa}_{2,n}} \mid \frac{(\sqrt{\bar{n}}R_0 + \tilde{\kappa}_{1,n})^2}{\tilde{\kappa}_{2,n}}\right) = F_{\chi_1^2}(x) - x f_{\chi_1^2}(x) \left(\left(\sqrt{\bar{n}}\tilde{R}_{10} + \tilde{\kappa}_{1,n}\right)^2 + \tilde{\kappa}_{21,n} + \tilde{\kappa}_{22,n}\right) + O(\nu_n^\sharp). \end{aligned} \quad (\text{S14})$$

By (S13) and (S14),

$$\Pr[\bar{n}(R_0 + R)^2 \leq x] = F_{\chi_1^2}(x) - \tilde{\mathcal{C}}_p^{\text{pre}}(n, h) x f_{\chi_1^2}(x) + O(\nu_n^\sharp), \quad (\text{S15})$$

where $\tilde{\mathcal{C}}_p^{\text{pre}}(n, h) := \bar{n}\tilde{R}_{10}^2 + 2\sqrt{\bar{n}}\tilde{R}_{10}\tilde{\kappa}_{1,n} + \tilde{\kappa}_{21,n} + \bar{n}^{-1} \sum_{j=1}^8 \bar{\kappa}_{2j}$. By tedious and lengthy algebra, we can directly show that $\tilde{R}_{10}^2 = \mathcal{B}_p^\dagger - \mathcal{B}_p^\ddagger$ and $\sum_{j=1}^8 \bar{\kappa}_{2j} = \sum_{j=1}^4 (\gamma_{p,j}^\dagger - \gamma_{p,j}^\ddagger) + O(h)$ and $2\sqrt{\bar{n}}\tilde{R}_{10}\tilde{\kappa}_{1,n} + \tilde{\kappa}_{21,n} = O(h\|\Delta_{\mathcal{U}}\|)$. By calculating $E[LR^*]$ with arguments used repeatedly in previous proofs, we find that $\tilde{\mathcal{C}}_p^{\text{pre}}(n, h)$ is just the leading term in the expansion $E[LR^*] - 1 = \tilde{\mathcal{C}}_p^{\text{pre}}(n, h) + o(\nu_n^\sharp)$, where $\nu_n^\sharp := \|\Delta_{\mathcal{U}}\| + \bar{n} \|\Delta_{\mathcal{U}}\|^2 + \bar{n}^{-1}$. We use the fact that $\ell_p(\hat{\vartheta}_p | h) = \inf_{\theta_2} \sup_{\lambda_2} 2 \sum_i \log(1 + \lambda_2^\top \bar{\mathcal{U}}_i(\theta_2))$ and an alternative expression for $LR^* = \bar{n}(\tilde{\ell}^* - \hat{\ell}^*)$ to get a more lucid proof.

We consider the singular value decomposition of $\Delta_{\bar{\mathcal{U}}\bar{\mathcal{U}}^\top}^{-1/2}(-\Delta_{\bar{\mathcal{G}}})$ such that $\bar{\mathbf{S}}^\top \Delta_{\bar{\mathcal{U}}\bar{\mathcal{U}}^\top}^{-1/2}(-\Delta_{\bar{\mathcal{G}}}) \bar{\mathbf{T}} = \begin{bmatrix} \bar{\Lambda} & 0_{d_z \times d_z} \end{bmatrix}^\top$

where $\bar{S}^\top \bar{S} = \mathbf{I}_{2d_z}$, $\bar{T}^\top \bar{T} = \mathbf{I}_{d_z}$ and $\bar{\Lambda}$ is a d_z -dimensional diagonal matrix. We apply the rotation by $\bar{\mathcal{V}}_i(\theta_2) := \bar{\Gamma} \bar{\mathcal{U}}_i(\theta_2)$ where $\bar{\Gamma} := \bar{S}^\top \Delta_{\bar{\mathcal{U}}\bar{\mathcal{U}}^\top}^{-1/2}$ so that $\ell_p(\hat{\vartheta}_p | h) = \inf_{\theta_2} \sup_{\lambda_2} 2 \sum_i \log(1 + \lambda_2^\top \bar{\mathcal{V}}_i(\theta_2))$ and calculations from Matsushita and Otsu (2013) can be applied. Also denote $\bar{\mathcal{V}}_i := \bar{\Gamma} \bar{\mathcal{U}}_i$, $\bar{\mathcal{H}}_i := \bar{\Gamma}(-\bar{\mathcal{G}}_i)$ ($\bar{\mathcal{V}}$ and $\bar{\mathcal{H}}$ defined similarly) and $\bar{\Omega} := (\bar{\Lambda} \bar{T}^\top)^{-1}$. Then it follows that $\Delta_{\bar{\mathcal{V}}\bar{\mathcal{V}}^\top} = \mathbf{I}_{2d_z \times 2d_z}$ and

$$\begin{bmatrix} -\Delta_{\bar{\mathcal{V}}\bar{\mathcal{V}}^\top} & \Delta_{\bar{\mathcal{H}}} \\ \Delta_{\bar{\mathcal{H}}}^\top & 0_{d_z \times d_z} \end{bmatrix}^{-1} = \begin{bmatrix} 0_{d_z \times d_z} & 0_{d_z \times d_z} & \bar{\Omega}^\top \\ 0_{d_z \times d_z} & -\mathbf{I}_{d_z} & 0_{d_z \times d_z} \\ \bar{\Omega} & 0_{d_z \times d_z} & \bar{\Omega} \bar{\Omega}^\top \end{bmatrix} = \begin{bmatrix} -(\bar{\Gamma}^\top)^{-1} \bar{\mathbf{Q}} \bar{\Gamma}^{-1} & -(\bar{\Gamma}^\top)^{-1} \bar{\mathbf{N}} \\ -\bar{\mathbf{N}}^\top \bar{\Gamma}^{-1} & \bar{\mathbf{O}} \end{bmatrix}. \quad (\text{S16})$$

Let $(A_\dagger^a, A_\dagger^{ab}, A_\dagger^{abc}, C_\dagger^{a,s}, C_\dagger^{a;b,s})$, $(\alpha_\dagger^a, \alpha_\dagger^{ab}, \alpha_\dagger^{abc}, \alpha_\dagger^{abcd})$ and $(\gamma_\dagger^{a,s}, \gamma_\dagger^{a;b,s}, \gamma_\dagger^{a;s;b,t}, \gamma_\dagger^{a;b;c,s})$ be defined by the same formulae as those of $(A^k, A^{kl}, A^{klm}, C^{k,n}, C^{k;l,n})$, $(\alpha^k, \alpha^{kl}, \alpha^{klm}, \alpha^{klmn})$ and $(\gamma^{k,n}, \gamma^{k;l,n}, \gamma^{k,n;l,o}, \gamma^{k;l;m,n})$, with $(\mathcal{V}, \mathcal{H}, \mathcal{V}_i, \mathcal{H}_i)$ replaced by $(\bar{\mathcal{V}}, \bar{\mathcal{H}}, \bar{\mathcal{V}}_i, \bar{\mathcal{H}}_i)$. The leading terms in the stochastic expansion of $\bar{n}^{-1} \ell_p(\hat{\vartheta}_p | h)$ is given by $\bar{n}^{-1} \hat{\ell}^\star = \tilde{R}_{\dagger 1}^{d_z+a} \tilde{R}_{\dagger 1}^{d_z+a} + 2\tilde{R}_{\dagger 1}^{d_z+a} \tilde{R}_{\dagger 2}^{d_z+a} + 2\tilde{R}_{\dagger 1}^{d_z+a} \tilde{R}_{\dagger 3}^{d_z+a} + \tilde{R}_{\dagger 2}^{d_z+a} \tilde{R}_{\dagger 2}^{d_z+a}$, where the expressions of $(\tilde{R}_{\dagger 1}^{d_z+a}, \tilde{R}_{\dagger 2}^{d_z+a}, \tilde{R}_{\dagger 3}^{d_z+a})$ are readily obtained in a special case of Matsushita and Otsu (2013) when the moment conditions are linear in parameters. E.g., $\tilde{R}_{\dagger 1}^{d_z+a} := A_\dagger^{d_z+a}$,

$$\tilde{R}_{\dagger 2}^{d_z+a} := -\frac{1}{2} A_\dagger^{d_z+b} A_\dagger^{d_z+a, d_z+b} + \frac{1}{3} \alpha_\dagger^{d_z+a, d_z+b, d_z+c} A_\dagger^{d_z+b} A_\dagger^{d_z+c} - \bar{\Omega}^{(st)} C_\dagger^{d_z+a, s} A_\dagger^t + \bar{\Omega}^{(st)} \gamma_\dagger^{d_z+a; d_z+b, s} A_\dagger^{d_z+b} A_\dagger^t$$

and the expression of $\tilde{R}_{\dagger 3}^{d_z+a}$ is omitted for brevity (see Matsushita and Otsu, 2013, A.1). Let $\hat{A}_\dagger^a := A_\dagger^a - \alpha_\dagger^a$. We again replace A_\dagger^a by $\hat{A}_\dagger^a + \alpha_\dagger^a$ to obtain $\tilde{R}_{\dagger 1}^{d_z+a} = \tilde{R}_{\dagger 11}^{d_z+a} + \tilde{R}_{\dagger 10}^{d_z+a}$, $\tilde{R}_{\dagger 2}^{d_z+a} = \tilde{R}_{\dagger 22}^{d_z+a} + \tilde{R}_{\dagger 21}^{d_z+a} + \tilde{R}_{\dagger 20}^{d_z+a}$ and $\tilde{R}_{\dagger 3}^{d_z+a} = \tilde{R}_{\dagger 33}^{d_z+a} + \tilde{R}_{\dagger 32}^{d_z+a} + \tilde{R}_{\dagger 31}^{d_z+a} + \tilde{R}_{\dagger 30}^{d_z+a}$. Then by standard calculations, $\mathbb{E}[\bar{n}^{-1} \hat{\ell}^\star]$ is equal to the sum of $\tilde{R}_{\dagger 10}^{d_z+a} \tilde{R}_{\dagger 10}^{d_z+a}$, $\tilde{R}_{\dagger 10}^{d_z+a} \mathbb{E}[\tilde{R}_{\dagger 22}^{d_z+a}]$, $\mathbb{E}[\tilde{R}_{\dagger 11}^{d_z+a} \tilde{R}_{\dagger 21}^{d_z+a}]$, $\mathbb{E}[\tilde{R}_{\dagger 11}^{d_z+a} \tilde{R}_{\dagger 11}^{d_z+a}]$ and $2 \cdot \mathbb{E}[\tilde{R}_{\dagger 11}^{d_z+a} \tilde{R}_{\dagger 22}^{d_z+a}] + 2 \cdot \mathbb{E}[\tilde{R}_{\dagger 11}^{d_z+a} \tilde{R}_{\dagger 33}^{d_z+a}] + \mathbb{E}[\tilde{R}_{\dagger 22}^{d_z+a} \tilde{R}_{\dagger 22}^{d_z+a}]$ with an $o(v_n^{\dagger})$ remainder term. By inverting using the second equality of (S16), $\tilde{R}_{\dagger 10}^{d_z+a} \tilde{R}_{\dagger 10}^{d_z+a} = \alpha_\dagger^{d_z+a} \alpha_\dagger^{d_z+a} = \bar{\mathbf{Q}}^{(ab)} \bar{\Upsilon}^a \bar{\Upsilon}^b$. By calculation and $\Delta_{\bar{\mathcal{V}}\bar{\mathcal{V}}^\top} = \mathbf{I}_{2d_z \times 2d_z}$, $\mathbb{E}[\tilde{R}_{\dagger 11}^{d_z+a} \tilde{R}_{\dagger 11}^{d_z+a}] = \bar{n}^{-1} d_z + O(\|\Delta_{\mathcal{U}}\|^2/n)$. It is easy to calculate that $\mathbb{E}[\tilde{R}_{\dagger 22}^{d_z+a}] = -\bar{n}^{-1} \alpha_\dagger^{d_z+a, d_z+b, d_z+b}/6 - \bar{\Omega}^{(st)} \gamma_\dagger^{t; d_z+a, s} + O(\|\Delta_{\mathcal{U}}\|/n)$. Then by (S16),

$$\tilde{R}_{\dagger 10}^{d_z+a} \mathbb{E}[\tilde{R}_{\dagger 22}^{d_z+a}] = -\bar{n}^{-1} \left(\frac{1}{6} \bar{\Upsilon}^{abc} \bar{\mathbf{Q}}^{(ab)} \bar{\mathbf{Q}}^{(cd)} \bar{\Upsilon}^d + \bar{\Gamma}^{a;b, s} \bar{\mathbf{N}}^{(as)} \bar{\mathbf{Q}}^{(bc)} \bar{\Upsilon}^c \right) + o(v_n^{\dagger}/\bar{n}).$$

By calculation and using (S16), $\mathbb{E}[\tilde{R}_{\dagger 11}^{d_z+a} \tilde{R}_{\dagger 21}^{d_z+a}] = \bar{n}^{-1} \bar{\Upsilon}^{abc} \bar{\mathbf{Q}}^{(ab)} \bar{\mathbf{Q}}^{(cd)} \bar{\Upsilon}^d/6 + o(v_n^{\dagger}/\bar{n})$. By calculation in Matsushita and Otsu (2013, A.4),

$$2 \cdot \mathbb{E}[\tilde{R}_{\dagger 11}^{d_z+a} \tilde{R}_{\dagger 22}^{d_z+a}] + 2 \cdot \mathbb{E}[\tilde{R}_{\dagger 11}^{d_z+a} \tilde{R}_{\dagger 33}^{d_z+a}] + \mathbb{E}[\tilde{R}_{\dagger 22}^{d_z+a} \tilde{R}_{\dagger 22}^{d_z+a}] = \bar{n}^{-2} \sum_{j=1}^8 \bar{\kappa}_{\dagger 2j} + o(v_n^{\dagger}/\bar{n}),$$

where the constants are defined by

$$\begin{aligned}
(\bar{\kappa}_{\dagger 21}, \bar{\kappa}_{\dagger 22}, \bar{\kappa}_{\dagger 23}, \bar{\kappa}_{\dagger 24}, \bar{\kappa}_{\dagger 25}, \bar{\kappa}_{\dagger 26}, \bar{\kappa}_{\dagger 27}, \bar{\kappa}_{\dagger 28}) &:= \left(\frac{1}{2} \bar{\Upsilon}^{abcd} \bar{Q}^{(ab)} \bar{Q}^{(cd)}, -\frac{1}{3} \bar{\Upsilon}^{abc} \bar{Q}^{(ad)} \bar{Q}^{(be)} \bar{Q}^{(cf)} \bar{\Upsilon}^{\text{def}}, \right. \\
&2\bar{\Gamma}^{a;b;c,s} \bar{N}^{(as)} \bar{Q}^{(bc)}, -\bar{\Gamma}^{a;b,s} \bar{Q}^{(ac)} \bar{Q}^{(bd)} \bar{N}^{(es)} \bar{\Upsilon}^{\text{cde}}, -\bar{\Gamma}^{a;s,b,t} \bar{Q}^{(ab)} \bar{O}^{(st)} \\
&\left. \bar{\Gamma}^{a;c,s} \bar{Q}^{(ab)} \bar{Q}^{(cd)} \bar{O}^{(st)} \bar{\Gamma}^{b;d,t}, -\bar{\Gamma}^{a;c,s} \bar{N}^{(at)} \bar{Q}^{(cd)} \bar{N}^{(bs)} \bar{\Gamma}^{b;d,t}, \bar{\Gamma}^{a;c,s} \bar{N}^{(as)} \bar{Q}^{(cd)} \bar{N}^{(bt)} \bar{\Gamma}^{b;d,t} \right).
\end{aligned}$$

Note that $(\bar{\kappa}_{\dagger 21}, \bar{\kappa}_{\dagger 22}, \bar{\kappa}_{\dagger 23}, \bar{\kappa}_{\dagger 25}) = (\mathcal{V}_{p,1}^\dagger, \mathcal{V}_{p,2}^\dagger, \mathcal{V}_{p,3}^\dagger, \mathcal{V}_{p,4}^\dagger)$. Therefore,

$$\mathbb{E} \left[\bar{n} \hat{\ell}^\star \right] = d_z + \bar{n} \mathcal{B}_p^\dagger - \bar{\Gamma}^{a;b,s} \bar{N}^{(as)} \bar{Q}^{(bc)} \bar{\Upsilon}^c + \bar{n}^{-1} \sum_{j=1}^8 \bar{\kappa}_{\dagger 2j} + o(v_n^\sharp).$$

Let $\bar{\kappa}_{\dagger 2j}$ be defined by the formula of $\bar{\kappa}_{\dagger 2j}$ with $(\bar{\Upsilon}, \bar{Q}, \bar{N}, \bar{O}, \bar{\Gamma})$ replaced by $(\Upsilon, Q_\dagger, N_\dagger, O_\dagger, \Gamma_\dagger)$ and also $(\bar{\kappa}_{\dagger 21}, \bar{\kappa}_{\dagger 22}, \bar{\kappa}_{\dagger 23}, \bar{\kappa}_{\dagger 25}) = (\mathcal{V}_{p,1}^\dagger, \mathcal{V}_{p,2}^\dagger, \mathcal{V}_{p,3}^\dagger, \mathcal{V}_{p,4}^\dagger)$. By following the same steps, we get a similar expansion for $\mathbb{E} \left[\bar{n} \tilde{\ell}^\star \right]$. And, then we have $\mathbb{E} \left[\bar{n} (\tilde{\ell}^\star - \hat{\ell}^\star) \right] - 1 = \tilde{\mathcal{C}}_p^{\text{pre}}(n, h) + o(v_n^\sharp)$

$$\tilde{\mathcal{C}}_p^{\text{pre}}(n, h) = \bar{n} (\mathcal{B}_p^\dagger - \mathcal{B}_p^\dagger) - \Gamma_\dagger^{k;l,u} N_\dagger^{(ku)} Q_\dagger^{(lm)} \Upsilon^m + \bar{\Gamma}^{a;b,s} \bar{N}^{(as)} \bar{Q}^{(bc)} \bar{\Upsilon}^c + \bar{n}^{-1} \sum_{j=1}^8 (\bar{\kappa}_{\dagger 2j} - \bar{\kappa}_{\dagger 2j}).$$

It is easy to see that by Lemma 1, $\Gamma_\dagger^{k;l,u} \asymp \bar{\Gamma}^{a;b,s} = O(h)$. Therefore, $\Gamma_\dagger^{k;l,u} N_\dagger^{(ku)} Q_\dagger^{(lm)} \Upsilon^m \asymp \bar{\Gamma}^{a;b,s} \bar{N}^{(as)} \bar{Q}^{(bc)} \bar{\Upsilon}^c = O(\|\Delta_{\mathcal{U}}\| h)$, $\bar{\kappa}_{\dagger 24} \asymp \bar{\kappa}_{\dagger 24} = O(h)$ and $\bar{\kappa}_{\dagger 26} \asymp \bar{\kappa}_{\dagger 27} \asymp \bar{\kappa}_{\dagger 28} \asymp \bar{\kappa}_{\dagger 26} \asymp \bar{\kappa}_{\dagger 27} \asymp \bar{\kappa}_{\dagger 28} = O(h^2)$. It follows from these results that $\tilde{\mathcal{C}}_p^{\text{pre}}(n, h) = \mathcal{C}_p^{\text{pre}}(n, h) + O(\|\Delta_{\mathcal{U}}\| h + n^{-1})$.

It is easily seen that the result (S13) with the weak inequality replaced by a strict inequality still holds (see Skovgaard, 1981, Theorem 3.2). By $LR^\star = \bar{n} (R_0 + R)^2 + O_p^\star(v_n^\sharp)$ and the fact (21),

$$\left| \Pr[LR^\star \leq x] - \Pr[\bar{n} (R_0 + R)^2 \leq x] \right| \leq \Pr \left[\left| \bar{n} (R_0 + R)^2 - x \right| \leq c_1 v_n^\sharp \right] + c_2 \left(\log(n) / \bar{n}^{3/2} \right) = O(v_n^\sharp), \quad (\text{S17})$$

where the equality follows from (S13) and boundedness of $\phi(\cdot \mid \tilde{\kappa}_{1,n}, \tilde{\kappa}_{2,n})$. The conclusion follows from (S15), (S17) and $\tilde{\mathcal{C}}_p^{\text{pre}}(n, h) = \mathcal{C}_p^{\text{pre}}(n, h) + O(\|\Delta_{\mathcal{U}}\| h + n^{-1})$. \blacksquare

References

- Bickel, P. J. and K. A. Doksum (2015). *Mathematical statistics: basic ideas and selected topics*, Volume 2. CRC Press.
- Calonico, S., M. D. Cattaneo, and M. H. Farrell (2022). Coverage error optimal confidence intervals for local

- polynomial regression. *Bernoulli*.
- Chen, X. and K. Kato (2020). Jackknife multiplier bootstrap: finite sample approximations to the u-process supremum with applications. *Probability Theory and Related Fields* 176(3-4), 1–67.
- Chernozhukov, V., D. Chetverikov, and K. Kato (2014). Gaussian approximation of suprema of empirical processes. *Annals of Statistics* 42(4), 1564–1597.
- Cohen, J. D. (1988). Noncentral chi-square: Some observations on recurrence. *The American Statistician* 42(2), 120–122.
- DiCiccio, T., P. Hall, and J. Romano (1988). Bartlett adjustments for empirical likelihood. Technical report No. 298, Department of Statistics, Stanford University.
- Giné, E. and R. Nickl (2015). *Mathematical foundations of infinite-dimensional statistical models*, Volume 40. Cambridge University Press.
- Kosorok, M. R. (2007). *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media.
- Ma, J. (2017). Second-order refinement of empirical likelihood ratio tests of nonlinear restrictions. *The Econometrics Journal* 20(1), 139–148.
- Matsushita, Y. and T. Otsu (2013). Second-order refinement of empirical likelihood for testing overidentifying restrictions. *Econometric Theory* 29(02), 324–353.
- Skovgaard, I. M. (1981). Transformation of an edgeworth expansion by a sequence of smooth functions. *Scandinavian Journal of Statistics*, 207–217.
- Skovgaard, I. M. (1986). On multivariate edgeworth expansions. *International Statistical Review/Revue Internationale de Statistique*, 169–186.