Empirical Likelihood Covariate Adjustment for Regression Discontinuity Designs

Jun Ma* Zhengfei Yu†

Abstract

This paper proposes a novel approach to incorporate information from covariates in regression discontinuity (RD) designs. We represent the covariate balance condition, which is assumed by the widely-used augmented local polynomial regression approach of Calonico, Cattaneo, Farrell, and Titiunik (2019), as over-identifying moment restrictions. The empirical likelihood (EL) estimator of the RD local average treatment effect (LATE) efficiently incorporates the information from covariate balance and achieves efficiency gain without additional assumptions. We resolve the indeterminacy raised by Calonico, Cattaneo, Farrell, and Titiunik (2019, Page 448) regarding the efficiency gain associated with covariate adjustment in RD. We propose a novel robust corrected EL confidence set which has several robustness properties. Our method complements the confidence interval of Calonico, Cattaneo, Farrell, and Titiunik (2019). It is particularly useful when the sample size is small but the researcher wishes to have the coverage error under control. We show that the EL confidence set with a simple data-driven correction 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 robust corrected EL confidence set is automatically robust against slight perturbation to the covariate balance condition, which may happen in cases such as data contamination and misspecified placebo outcomes used as covariates. We also show a novel uniform-in-bandwidth Wilks' theorem, which justifies correction for specification search, takes into account the effect of data-driven bandwidth choice and constructs a uniform confidence band used for sensitivity analysis in the sense of Armstrong and Kolesár (2018). We conduct Monte Carlo simulations to assess finite-sample performance of our method and also apply it to a real dataset to illustrate its usefulness.

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

JEL classification: C12, C14, C31, C36

This version: April 24, 2022

*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. In the RD context, local smoothing is preferred to global smoothing since the former has better boundary performances (Gelman and Imbens, 2019). The standard local smoothing method for RD is the local polynomial (LP) regression (Fan and Gijbels, 1996). 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 LP regression where the covariates enter linearly. The procedure does not require smoothing over covariates. See, e.g., Imbens and Lemieux (2008, Section 4.3) for earlier discussion on this procedure. Recently, 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 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) extends CCFT's approach to control for high-dimensional covariate vector by shrinkage. 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

[1] In 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).

weighting matrix.[2] We show in Theorem 1 and Remark 2 that the inclusion of those covariate-balancing-induced over-identifying moment conditions reduces the asymptotic variance of the RD estimator relative to the standard LP RD estimator without the 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 resolve the indeterminacy raised by Calonico, Cattaneo, Farrell, and Titiunik (2019, Page 448) regarding the efficiency gain associated with covariate adjustment in RD and show that regression adjustment leads to guaranteed efficiency gain. See Remarks 2 and 3. Theorem 2 shows a novel uniform-in-bandwidth extension of the standard Wilks' phenomenon (i.e., the EL ratio statistic is asymptotically χ^2). EL inference on the RD LATE is based on such a result and avoids calculation of standard error. Our uniform-in-bandwidth version adjusts for specification search over multiple bandwidths known as bandwidth snooping (Armstrong and Kolesár, 2018, AK, hereafter) and effects from data-dependent bandwidths in a robust manner (Remarks 4 and 7). It also provides a powerful tool for sensitivity analysis in the sense of AK (Remark 6).

We then investigate the second-order properties of the EL inference. Following Calonico et al. (2018b, 2020), we define the coverage error for inference as the discrepancy between the nominal and finite-sample coverage probabilities. Theorem 3 characterizes the leading coverage error 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 (Calonico et al., 2020) as the minimizer of this leading coverage error has a simple close form (Remark 10). Then by using the result of Theorem 3, we propose a simple yet novel robust corrected EL (RCEL) confidence set for the RD LATE with favorable robustness properties including "parametric" coverage accuracy and immunity to small perturbation to covariate balance, 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. The construction of the RCEL confidence set combines internalized bias removal (Calonico et al., 2014) and conventional Bartlett correction (i.e., a rescaling correction for improving the coverage accuracy) for EL (see, e.g., Chen and Cui, 2007). Compared with the expressions of correction factors for EL in other contexts (e.g., Chen and Cui, 2007; Matsushita and Otsu, 2013; Ma, 2017), the correction factor used by our method is very simple and thus can be estimated with good accuracy in finite samples, due to a special property of the moment conditions under consideration (i.e., asymptotic uncorrelatedness between the conditions and their derivatives).

[2]See, 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.

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. As an extension of Theorem 3, 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 often has access to observations on placebo outcomes, which are determined after treatment but unaffected by the treatment. Covariate balance should also hold for placebo outcomes. However, specification of the placebo outcomes is based on prior knowledge. Our method allows for flawed knowledge and thus placebo outcomes can be treated in the same way as pre-treatment covariates and used for enhancing efficiency. 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 actually drawn from a perturbed population (Kitamura et al., 2013) that slightly violates the balance condition. In such a situation, the coverage accuracy of our RCEL confidence set stays relatively unaffected even if covariate balance does not hold exactly, 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 snooping 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 of broader scope than the augmented regression approach of CCFT. In the literature, nonlinear estimators are proposed for RD with limited dependent (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 approach in these contexts is involved and the desired properties (consistency and efficiency gain) may no longer hold. Incorporating covariates by the EL probabilities (see, e.g., Brown and Newey, 2002) seems a simple solution and is able to deliver guaranteed efficiency gain under covariate balance. Such extensions are beyond the scope of this paper and left for future research.

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. Proofs are collected in the appendix. Section 8 concludes.

2 Regression discontinuity designs

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

$Y \in \mathbb{R}$ denotes the outcome variable, $D \in \{0,1\}$ denotes the binary treatment and $Z \in \mathbb{R}^{d_z}$ denotes pre-treatment covariates. 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 $D = I \coloneqq 1(X \geq 0)$ (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. Econometricians 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. For any individual, only one of the potential outcomes (treatments) is observed. 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., $(D_+, D_-) = (1, 0)$).[3] The RD model imposes only a few weak identifying assumptions. The model allows for direct causal effects of (X, Z) on (Y, D), arbitrary dimensionality of unobserved heterogeneity and self-selection into treatment based on the gain from treatment $Y(1) - Y(0)$ for given (X, Z).[4] Following CCFT, we let $(Z(1), Z(0))$ denote potential covariates and then $Z = DZ(1) + (1-D)Z(0)$. Let $\mathfrak{V}(k) \coloneqq (Y(k), Z(k))$, $\forall k \in \{0, 1\}$. Denote $g_{dd'}(x) \coloneqq \Pr[D_+ = d, D_- = d' \mid X = x]$ and $g_{\mathfrak{V}(k)|dd'}(x) \coloneqq \mathrm{E}[\mathfrak{V}(k) \mid D_+ = d, D_- = d', X = x]$. Similarly, let $\mathfrak{V} \coloneqq (Y, Z)$ and $g_{\mathfrak{V}|dd'}(x) \coloneqq \mathrm{E}[\mathfrak{V} \mid D_+ = d, D_- = d', X = x]$. By the law of iterated expectations (LIE), $g_{\mathfrak{V}} = \sum_{(d,d') \in \{0,1\}^2} g_{dd'} g_{\mathfrak{V}|dd'}$. Let $\mathscr{T}_Y \coloneqq \mathrm{E}[Y(1) - Y(0) \mid X = 0, D_+ > D_-]$ be the RD LATE (the average treatment effect for individuals with zero score in the complier group) and similarly, $\mathscr{T}_Z \coloneqq \mathrm{E}[Z(1) - Z(0) \mid X = 0, D_+ > D_-]$ 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)* $\Pr[D_- \leq D_+ \mid X = 0] = 1$*; (c)* $\Pr[D_+ > D_- \mid X = 0] \neq 0$*; (d)* $\mathscr{T}_Z = 0$.

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

[3] In the sharp RD case, the complier group is the same as the whole population. In the fuzzy RD case, some individuals with positive scores do not comply and $Y(0)$'s are observed and some with $X < 0$ receive the treatment and $Y(1)$'s are observed.

[4] RD 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 $D_- = h_-(X, Z, \eta)$.

(Arai et al., 2021). It can be shown that under these assumptions, $\mathscr{T}_Y$ is nonparametrically identified: $\mathscr{T}_Y = \vartheta_0 \coloneqq (\mu_{Y,+} - \mu_{Y,-}) / (\mu_{D,+} - \mu_{D,-})$ (see Hahn et al., 2001; Dong, 2018), where ϑ_0 is an observable population feature.[5] Similarly, under (a), (b) and (c), $\mathscr{T}_Z = (\mu_{Z,+} - \mu_{Z,-}) / (\mu_{D,+} - \mu_{D,-})$. 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 placebo outcomes. These variables are determined after the assignment but unaffected by the treatment. Placebo outcomes can be found in many applications. See, e.g., Cattaneo et al. (2019, Section 5.1) for discussion and real examples. Unlike pre-treatment variables, the assumption that the placebo outcomes satisfy (d) is based on our prior knowledge and understanding of the data generating mechanism. It is recommended in the literature (e.g., Cattaneo et al., 2019) that falsification analysis (i.e., testing $\mu_{Z,+} = \mu_{Z,-}$) should be carried out for placebo outcomes as well.

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 $K_h(t) \coloneqq h^{-1} K(t/h)$. Let $M \coloneqq Y - \vartheta_0 D$ and note that

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

Denote $\vartheta_1 \coloneqq \mu_{M,+} = \mu_{M,-}$. Let $p \geq 1$ be the integer-valued LP order. Denote $r_p(t) \coloneqq (1, t, \ldots, t^p)^\top$. According to Jiang and Doksum (2003), p-th order LP approximation of $(\psi_{M,-}, \psi_{M,+})$ can be derived from

[5] In 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: $\mathrm{E}[Y(1) - Y(0) \mid X = 0] = \vartheta_0$.

[6] While 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.

solving the following minimum contrast problem. Let $\mathrm{e}_{k,s}$ denote the s-th unit vector in $\mathbb{R}^k$. Let

$$\psi_- \coloneqq \mathrm{e}_{p+1,1}^\top \underset{z\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \int_{\underline{x}}^0 \left\{m_M(x) - z^\top r_p(x)\right\}^2 K_h(x)\,\mathrm{d}x \tag{1}$$

and ψ_+ be defined by the minimizer on the right hand side with the integral range $[\underline{x},0]$ replaced by $[0,\overline{x}]$. Denote $\mathrm{V}_{p;-} \coloneqq \int_{-1}^0 r_p(t)\,r_p(t)^\top K(t)\,\mathrm{d}t$ and $\mathcal{K}_{p;-}(t) \coloneqq \mathrm{e}_{p+1,1}^\top \mathrm{V}_{p;-}^{-1} r_p(t)\,K(t)$. Let $(\mathrm{V}_{p;+},\mathcal{K}_{p;+})$ 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 $W_{p;-,i} \coloneqq 1(X_i<0)\,\mathcal{K}_{p;-}(X_i/h)$, $W_{p;+,i} \coloneqq 1(X_i>0)\,\mathcal{K}_{p;+}(X_i/h)$ and $W_{p,i} \coloneqq (W_{p;+,i}, W_{p;-,i})^\top$. $(W_{p;-}, W_{p;+}, W_p)$ are defined by the same formulae with X_i replaced by X. Then, by solving the first-order conditions of (1), we have $\psi_{\mathrm{s}} = \mathrm{E}\left[h^{-1}W_{p;\mathrm{s}}g_M(X)\right] = \mathrm{E}\left[h^{-1}W_{p;\mathrm{s}}M\right]$. By Taylor expansion (see Jiang and Doksum, 2003), $\psi_{\mathrm{s}} = \psi_{M,\mathrm{s}} + O\left(h^{p+1}\right)$ under suitable smoothness assumptions imposed on g_M. From (1) with $m_M(x)$ replaced by $f_X(x)$, we have $\mathrm{E}\left[h^{-1}W_{p;\mathrm{s}}\right] = \varphi + O\left(h^{p+1}\right)$, $\forall \mathrm{s}\in\{-,+\}$. Therefore,

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

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 $\vartheta_2 \coloneqq \mu_{Z,+} = \mu_{Z,-}$ denote the common value. By solving (1) with m_M replaced by m_Z and f_X, we have

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

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:

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

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

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

and $\theta_\dagger \coloneqq (\theta_1, \theta_2)$.

We define the EL criterion function:

$$\begin{aligned} \ell_p(\theta \mid h) \quad &\coloneqq \quad \min_{w_1,\ldots,w_n} -2\sum_i \log(n \cdot w_i) \\ &\text{subject to } \sum_i w_i \ W_{p,i} \otimes \begin{pmatrix} Y_i - \theta_0 D_i - \theta_1 \\ Z_i - \theta_2 \end{pmatrix} = 0, \ \sum_i w_i = 1 \text{ and } w_i \geq 0, \forall i, \end{aligned} \tag{5}$$

where $\sum_i$ is understood as $\sum_{i=1}^n$. $-\sum_i \log(n \cdot w_i)/n$ is the Kullback-Leibler divergence from $(w_1, ..., w_n)$ to the uniform weights $1/n$. Denote $U_i(\theta) \coloneqq \left(Y_i - \theta_0 D_i - \theta_1, Z_i^\top - \theta_2^\top\right)^\top$, $U_i \coloneqq U_i(\vartheta)$ for notational simplicity and $d_u \coloneqq 1 + d_z$. $U(\theta)$ 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 $\widehat{\vartheta}_p \coloneqq \left(\widehat{\vartheta}_{p,0}, \widehat{\vartheta}_{p,1}, \widehat{\vartheta}_{p,2}\right) \coloneqq \operatorname{argmin}_{\theta \in \Theta} \ell_p(\theta \mid h)$, 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: $\widetilde{\vartheta}_p(\theta_0) \coloneqq \left(\widetilde{\vartheta}_{p,1}(\theta_0), \widetilde{\vartheta}_{p,2}(\theta_0)\right) \coloneqq \operatorname{argmin}_{\theta_\dagger \in \Theta_\dagger} \ell_p(\theta_0, \theta_\dagger \mid h)$, where $\Theta_\dagger \subseteq \mathbb{R}^{d_\dagger}$ is a compact constrained parameter space such that $\vartheta_\dagger$ is in the interior of $\Theta_\dagger$ and θ_0 is some hypothesized value. The EL ratio statistic is given by $LR_p(\theta_0 \mid h) \coloneqq \ell_p\left(\theta_0, \widetilde{\vartheta}_p(\theta_0) \mid h\right) - \ell_p\left(\widehat{\vartheta}_p \mid h\right)$, which is a function of θ_0. It is shown in the proof of Theorem 3 that $\ell_p\left(\widehat{\theta}_p \mid h\right) = \inf_{\theta_2} \sup_\lambda 2\sum_i \log\left(1 + \lambda^\top W_{p,i} \otimes (Z_i - \theta_2)\right)$ 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 $c_\tau \coloneqq F_{\chi_1^2}^{-1}(1-\tau)$ 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 $CS_\tau(h) \coloneqq \{\theta_0 : LR_p(\theta_0 \mid h) \leq c_\tau\}$. 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) $\widehat{w}_1, ..., \widehat{w}_n$ are those corresponding to the minimizer of the problem (5) with $\theta = \widehat{\vartheta}_p$ (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]

[8]It is shown in the proof of Theorem 3 that $n\widehat{w}_i = \left(1 + \widehat{\lambda}_p^\top W_{p,i} \otimes \left(Z_i - \widehat{\vartheta}_{p,2}\right)\right)^{-1}$ where $\widehat{\lambda}_p$ solves $\sum_i W_{p,i} \otimes \left(Z_i - \widehat{\vartheta}_{p,2}\right) / \left(1 + \widehat{\lambda}_p^\top W_{p,i} \otimes \left(Z_i - \widehat{\vartheta}_{p,2}\right)\right) = 0$. Indeed, it is clear that the EL estimator $\left(\widehat{\vartheta}_{p,0}, \widehat{\vartheta}_{p,1}\right)$ 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: $\sum_i \widehat{w}_i W_{p;\mathrm{s},i}\left(Y_i - \widehat{\vartheta}_{p,0} D_i - \widehat{\vartheta}_{p,1}\right) = 0$. We conjecture that the same reweighting adjustment can be extended to the nonlinear cases considered by Xu (2017, 2018).

4 Efficiency gain and uniform-in-bandwidth Wilks' phenomenon

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, $\mathrm{A}^{(jk)}$ denotes the jk-th element of a matrix A. By abuse of notation, for a d_v-dimensional random vector V, let V^2 denote $V \otimes V$ with duplicated coordinates removed, i.e., $V^2 \coloneqq \operatorname{vech}\left(VV^\top\right)$, where $\operatorname{vech}(\mathrm{A})$ denotes the half vectorization of a matrix A. Similarly, V^3 denotes the vector consisting of $V^{(1)}\operatorname{vech}\left(VV^\top\right), V^{(2)}\operatorname{vech}\left(\left(V^{(2)},...,V^{(d_v)}\right)^\top\left(V^{(2)},...,V^{(d_v)}\right)\right),...,\left(V^{(d_v)}\right)^3$. $\|x\|$ denotes the Euclidean norm of the vector x. We assume the following assumptions hold.

Assumption 2. *(a) On a neighborhood around* 0, $g_{\mathfrak{V}(k)|dd'}$ *and* $g_{dd'}$ *are* $(p+1)$*-times continuous differentiable with Lipschitz continuous* $(p+1)$*-th order derivatives,* $g_{\mathfrak{V}(k)^2|dd'}$ *is Lipschitz continuous and* $g_{\|\mathfrak{V}(k)\|^{12}|dd'}$ *is bounded and* f_X *is* $(p+1)$*-times continuous differentiable with Lipschitz continuous* $(p+1)$*-th order derivative; (b)* $\mathrm{E}\left[U(k)U(k)^\top \mid D_+ > D_-, X = 0\right]$ *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 $g_{\mathfrak{V}} = \sum_{(d,d')\in\{0,1\}^2} g_{dd'}g_{\mathfrak{V}|dd'}$, $\mathfrak{V} = D_+\mathfrak{V}(1) + (1-D_+)\mathfrak{V}(0)$ if $X > 0$ and $\mathfrak{V} = D_-\mathfrak{V}(1) + (1-D_-)\mathfrak{V}(0)$ if $X < 0$, Assumption 2(a) guarantees that $g_{\mathfrak{V}}$ and $m_{\mathfrak{V}}$ have continuous derivatives up to $(p+1)$-th order on the left and right neighborhoods of 0. Similarly, $g_{\mathfrak{V}^2}$ is continuous and $g_{\|\mathfrak{V}\|^{12}}$ is bounded on the left and right neighborhoods of 0, under Assumption 2(a). Denote $\gamma_{\mathsf{adj}} \coloneqq \Sigma^{-1}_{ZZ^\top,\pm}\Sigma_{ZM,\pm}$, $\epsilon \coloneqq M - Z^\top\gamma_{\mathsf{adj}}$ and $\sigma^2_{\mathsf{adj}} \coloneqq \Sigma_{M^2,\pm} - \Sigma_{MZ^\top,\pm}\gamma_{\mathsf{adj}}$. 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. Denote $\zeta_{p;\mathrm{s}} \coloneqq \omega^{p+1,1}_{p;\mathrm{s}}\left(\psi^{(p+1)}_{\epsilon,\mathrm{s}} - \mu_\epsilon\varphi^{(p+1)}\right)$, $\mathrm{s} \in \{-,+\}$, where $\omega^{j,k}_{p;+} \coloneqq \int_0^1 t^j\mathcal{K}_{p;+}(t)^k\,\mathrm{d}t$ and $\omega^{j,k}_{p;-} \coloneqq \int_{-1}^0 t^j\mathcal{K}_{p;-}(t)^k\,\mathrm{d}t$. 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 $\widehat{\vartheta}_{p,0}$, 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,* $\sqrt{nh}\left(\widehat{\vartheta}_{p,0} - \vartheta_0 - \mathscr{B}_p^{\mathsf{EL}} h^{p+1}\right) \to_d \mathrm{N}\left(0, \mathscr{V}_p^{\mathsf{EL}}\right)$, *where*

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

Remark 1. Terms in $\mathscr{B}_p^{\mathsf{EL}}$ that depend on the derivatives of f_X can be removed by using the bias-correcting moment conditions (Hall and Presnell, 1999). Denote $\bar{r}_p(u) := \left(u, u^2, \ldots, u^p\right)^\top$. For the EL estimator resulting from adding two more constraints $\sum_i w_i W_{p,i} \otimes \bar{r}_p(X_i/h) = 0$ to (5), the conclusion in Theorem 1 holds with $\mathscr{B}_p^{\mathsf{EL}} := \left(\omega_{p;+}^{p+1,1}\mu_{\epsilon,+}^{(p+1)} - \omega_{p;-}^{p+1,1}\mu_{\epsilon,-}^{(p+1)}\right) / \left\{\left(\mu_{D,+} - \mu_{D,-}\right)(p+1)!\right\}$.

Remark 2. We consider the special case of sharp RD and $p = 1$. The local linear estimator $\widehat{\vartheta}_0^{\mathsf{LL}}$ can be obtained from a single localized regression. CCFT's approach augments the regression to incorporate pre-treatment covariates. CCFT's covariate adjusted estimator $\widehat{\vartheta}_0^{\mathsf{CCFT}}$ is given by the regression coefficient of $I_i := 1(X_i \geq 0)$ in

$$\widehat{\vartheta}_0^{\mathsf{CCFT}} := \mathrm{e}_{4+d_z,3}^\top \underset{(a_0,b_0,a_1,b_1,d)\in\mathbb{R}^{4+d_z}}{\operatorname{argmin}} \sum_i K_h(X_i)\left\{Y_i - a_0 - b_0 X_i - a_1 I_i - b_1 X_i I_i - Z_i^\top d\right\}^2. \tag{6}$$

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

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

Next, we show that $\sigma_{\mathsf{adj}}^2 = \sigma_{\mathsf{CCFT}}^2$.[9] Observe that $\gamma_{\mathsf{adj}} = \gamma_{\mathsf{CCFT}}$ and also

$$\mathrm{Var}_{|0}\left[Z(0)^\top \gamma_{\mathsf{adj}}\right] + \mathrm{Var}_{|0}\left[Z(1)^\top \gamma_{\mathsf{adj}}\right] = \gamma_{\mathsf{adj}}^\top\left(\mathrm{Var}_{|0}[Z(0)] + \mathrm{Var}_{|0}[Z(1)]\right)\gamma_{\mathsf{adj}}$$

[9] Indeed, 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: $\widehat{\vartheta}_0^{\mathsf{CCFT}} - \widehat{\vartheta}_{1,0}^{\mathsf{EL}} = o_p\left((nh)^{-1/2}\right)$.

$$= \text{Cov}_{|0}\left[Y(1), Z(1)^\top \gamma_{\mathsf{adj}}\right] + \text{Cov}_{|0}\left[Y(0), Z(0)^\top \gamma_{\mathsf{adj}}\right] = \Sigma_{MZ^\top,\pm}\gamma_{\mathsf{adj}}. \quad (8)$$

Therefore,

$$\sigma^2_{\mathsf{adj}} = \text{Var}_{|0}[Y(0)] + \text{Var}_{|0}[Y(1)] - \text{Cov}_{|0}\left[Y(1), Z(1)^\top \gamma_{\mathsf{adj}}\right]$$
$$- \text{Cov}_{|0}\left[Y(0), Z(0)^\top \gamma_{\mathsf{adj}}\right] = \text{Var}_{|0}\left[Y(0) - Z(0)^\top \gamma_{\mathsf{adj}}\right] + \text{Var}_{|0}\left[Y(1) - Z(1)^\top \gamma_{\mathsf{adj}}\right] = \sigma^2_{\mathsf{CCFT}},$$

where the second equality follows from

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

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

Remark 3. We say that an estimator achieves efficiency gain if its asymptotic variance is smaller than that of the standard local linear estimator without covariates. $\sigma^2_{\mathsf{adj}} = \sigma^2_{\mathsf{CCFT}} \leq \sigma^2_{\mathsf{LL}}$ implies that both of the EL and CCFT estimators achieve efficiency gain without additional assumptions other than covariate balance. At the first glance, the guaranteed efficiency gain from the CCFT estimator may be surprising if one recalls that CCFT (on their Page 448) finds no definite ranking between σ^2_{CCFT} and σ^2_{LL} and interpret 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 running variable 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)) shows 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 brings efficiency gain. The assignment probability assumption is automatically fulfilled in RD designs.[10] Theorem 1 and the equivalence of the EL and CCFT estimators provide insight into this phenomenon of guaranteed efficiency gain from a different angle. CCFT's estimator can be interpreted as being efficiently incorporating the side

[10] Noack 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 covariate balance.

information from the covariate balance condition, which leads to efficiency gain.[11]

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 $f:\mathfrak{S}\to\mathbb{R}$ endowed with the sup-norm $\|f\|_{\mathfrak{S}}:=\sup_{s\in\mathfrak{S}}|f(s)|$. Let $\mathbb{H}:=\left[\underline{h},\overline{h}\right]$ be a compact bandwidth set where $\underline{h}=\underline{h}_n>0$ and $\overline{h}=\overline{h}_n>0$ $(\underline{h}<\overline{h})$ 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' phenomenon which states that $LR_p(\vartheta_0\mid h)\to_d\chi^2_1$. AK assumes that the outcome variable Y is bounded. Our result incorporates covariates and accommodates unbounded outcomes. The proof techniques we use also differ from those employed by AK.

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

$$\mathrm{E}\left[\Gamma_G(s)\,\Gamma_G(t)\right]=\sqrt{\frac{s}{t}}\frac{\int_0^{\infty}\mathcal{K}_{p;+}(z)\,\mathcal{K}_{p;+}((s/t)\,z)\,\mathrm{d}z}{\int_0^{\infty}\mathcal{K}_{p;+}(z)^2\,\mathrm{d}z}. \tag{9}$$

Then, $\Pr\left[LR_p(\vartheta_0\mid h)\le z_\tau\left(\overline{h}/\underline{h}\right)^2,\forall h\in\mathbb{H}\right]\to 1-\tau$, *as* $n\uparrow\infty$, *where* $z_\tau\left(\overline{h}/\underline{h}\right)$ *denotes the* $1-\tau$ *quantile of* $\|\Gamma_G\|_{\left[1,\overline{h}/\underline{h}\right]}$.

Remark 4. Theorem 2 generalizes the standard Wilks' theorem with a single bandwidth. It implies that when $h=\underline{h}=\overline{h}$, $\Pr\left[LR_p(\vartheta_0\mid h)\le c_\tau\right]=\Pr\left[\vartheta_0\in CS_\tau(h)\right]\to 1-\tau$. 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 $\hat{h}:=\mathrm{argmax}_{h\in\mathbb{H}}LR_p(0\mid h)$ is selected to maximize the p-value for the two-sided hypothesis test of $\vartheta_0=0$. AK shows that $z_\tau\left(\overline{h}/\underline{h}\right)^2>c_\tau$ when $\overline{h}/\underline{h}>1$ but $z_\tau\left(\overline{h}/\underline{h}\right)$ grows at a logarithmic speed as $\overline{h}/\underline{h}\uparrow\infty$. It is clear from Theorem 2 that $\Pr\left[\vartheta_0\in CS_\tau\left(\hat{h}\right)\right]\to 1-\tilde{\tau}$, where $\tilde{\tau}>\tau$ solves $z_{\tilde{\tau}}\left(\overline{h}/\underline{h}\right)^2=c_\tau$ 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 $z_\tau\left(\overline{h}/\underline{h}\right)^2$. Let $CS^{\mathsf{sc}}_\tau\left(h\mid\overline{h}/\underline{h}\right):=\left\{\theta_0:LR_p(\theta_0\mid h)\le z_\tau\left(\overline{h}/\underline{h}\right)^2\right\}$ be the snooping corrected confidence set. Then, $CS^{\mathsf{sc}}_\tau\left(h\mid\overline{h}/\underline{h}\right)$ has asymptotically correct coverage no matter how h is selected from $\mathbb{H}$, i.e., $\mathrm{liminf}_{n\uparrow\infty}\Pr\left[\vartheta_0\in CS^{\mathsf{sc}}_\tau\left(h\mid\overline{h}/\underline{h}\right)\right]\ge 1-\tau\ \forall h\in\mathbb{H}$.

Remark 5. The critical value $z_\tau\left(\overline{h}/\underline{h}\right)$ is based on direct Gaussian approximation and can be easily simu-

[11] Such 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.

lated, since the covariance structure (9) depends only on the kernel function.[12] If $\overline{h}/\underline{h} \uparrow \infty$ as $n \uparrow \infty$, then the critical value $z_\tau\left(\overline{h}/\underline{h}\right)$ can be replaced by its asymptotic counterpart $z_\tau^{\mathsf{asy}}\left(\overline{h}/\underline{h}\right)$, where $z_{1-\tau}^{\mathsf{asy}}\left(\overline{h}/\underline{h}\right) = -\log\left(-\log\left(1-\tau\right)\right)/a_n + b_n$ where (a_n, b_n) are constants that depend on $\overline{h}/\underline{h}$ and the kernel function. $z_\tau^{\mathsf{asy}}\left(\overline{h}/\underline{h}\right)$ is not recommended to be used in practice since its justification is based on the asymptotic theory of suprema of stationary Gaussian processes, which converges at a slow speed.

Remark 6. Theorem 4 shows that $\left\{CS_\tau^{\mathsf{sc}}\left(h \mid \overline{h}/\underline{h}\right) : h \in \mathbb{H}\right\}$ is an asymptotically valid confidence band for the constant function $\mathbb{H} \ni h \mapsto \vartheta_0$. By using the uniform confidence band 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 point-wise 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 $CS_\tau\left(h_{\mathsf{rf}}\right)$. In case of a statistically insignificant result (i.e., $0 \in CS_\tau\left(h_{\mathsf{rf}}\right)$), 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 ($CS_\tau\left(h\right) \subseteq (0,\infty)$ or $CS_\tau\left(h\right) \subseteq (-\infty, 0)$) corresponding to some $h \neq h_{\mathsf{rf}}$. In such a case, with suitable lower and upper bounds $\left(\underline{h}, \overline{h}\right)$ such that $\underline{h} < h_{\mathsf{rf}} < \overline{h}$, one may follow AK's approach and use the band $\left\{CS_\tau^{\mathsf{sc}}\left(h \mid \overline{h}/\underline{h}\right) : h \in \mathbb{H}\right\}$. If $\exists h \in \mathbb{H}$ such that $CS_\tau^{\mathsf{sc}}\left(h \mid \overline{h}/\underline{h}\right) \subseteq (0,\infty)$ or $CS_\tau^{\mathsf{sc}}\left(h \mid \overline{h}/\underline{h}\right) \subseteq (-\infty, 0)$, one may conclude that the RD LATE is different from zero and validity of such a result is guaranteed by Theorem 2. In case of $0 \notin CS_\tau\left(h_{\mathsf{rf}}\right)$, it is still necessary to examine the sensitivity of such a significant result to bandwidth choice (Imbens and Lemieux, 2008). As AK, with suitable $\left(\underline{h}, \overline{h}\right)$, one may conclude that $\vartheta_0 > 0$ in a robust sense if $\exists h \in \mathbb{H}$ such that $CS_\tau^{\mathsf{sc}}\left(h \mid \overline{h}/\underline{h}\right) \subseteq (0,\infty)$ and $\forall h \in \mathbb{H}$, $CS_\tau^{\mathsf{sc}}\left(h \mid \overline{h}/\underline{h}\right) \cap (0,\infty) \neq \emptyset$. Compared with AK, our confidence band incorporates information from covariates and the robust inference based on it is more powerful. The choice of $\left(\underline{h}, \overline{h}\right)$ follows AK (see Section 4.1.3 therein). We set $\underline{h}$ to be the value that gives approximately 50 effective observations (i.e., $n\underline{h} \approx 50$). On the other hand, $\overline{h}$ can be set proportionally to a commonly used reference bandwidth.

Remark 7. 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 $\widehat{h} \in \left[\underline{h}, \overline{h}\right]$, where $\left(\underline{h}, \overline{h}\right)$ are deterministic lower and upper bounds and $\widehat{h}$ is the minimizer of some data-dependent criterion function defined on $\left[\underline{h}, \overline{h}\right]$. Theorem 2 shows that snooping correction takes all noise in $\widehat{h}$ into account, by replacing the χ_1^2 quantile with $z_{1-\tau}\left(\overline{h}/\underline{h}\right)^2$. By Theorem 2, asymptotic validity of the robust confidence set $CS_\tau^{\mathsf{sc}}\left(\widehat{h} \mid \overline{h}/\underline{h}\right)$ is guaranteed without assuming that $\widehat{h}$ fulfills any property such as the stochastic order of

[12] See the `R` package `BWSnooping` from http://github.com/kolesarm/BWSnooping.

$\widehat{h}/h-1$ is sufficiently small so that the noise in $\widehat{h}$ is negligible, where h is some deterministic bandwidth that $\widehat{h}$ tries to capture.

Remark 8. As the main result of AK, Theorem 2 assumes deterministic upper and lower bounds $(\underline{h}, \overline{h})$. Let $\left(\underline{h}^*, \overline{h}^*\right)$ denote some deterministic bounds that some data-dependent bounds $(\underline{h}, \overline{h})$ capture. As argued by AK, the conclusion of Theorem 2 still holds under data-dependent bounds, if the orders of $\overline{h}/\overline{h}^* - 1$ and $\underline{h}/\underline{h}^* - 1$ are sufficiently small and $\left(\underline{h}^*, \overline{h}^*\right)$ satisfy the assumptions of Theorem 2.

5 Robust corrected empirical likelihood inference

In this section, we investigate the second-order properties of the EL ratio inference method. Theorem 3 provides the distributional expansion of $LR_p(\vartheta_0 \mid h)$ 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 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* $g_{\|\mathfrak{V}(k)\|^{2\mathfrak{o}}|dd'}$ *is bounded,* $\forall k \in \{0,1\}$.

Assumption 5. $\left(1, \mathcal{K}_{p;+}, \mathcal{K}_{p;+}^2, \mathcal{K}_{p;+}^3\right)$ *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 $\left(1, \mathcal{K}_{p;-}, \mathcal{K}_{p;-}^2, \mathcal{K}_{p;-}^3\right)$. Assumption 4 and 5 are used when establishing validity of the Edgeworth expansions in the proofs of Theorems 3 and 4. Let tr (A) denote the trace of a square matrix A. Denote $\Xi_1 \coloneqq \mu_{UU^\top,\pm}^{-1}$, $\Xi_2 \coloneqq \left(\mu_{UU^\top,+}^{-1} + \mu_{UU^\top,-}^{-1}\right)^{-1}$, $\Psi_1^{\mathsf{uw}} \coloneqq \operatorname{tr}\left(\Xi_1 \mu_{U^{(\mathsf{u})}U^{(\mathsf{w})}UU^\top,\pm}\right)$ and $\Psi_2^{\mathsf{uw}} \coloneqq \operatorname{tr}\left(\Xi_1 \mu_{U^{(\mathsf{u})}UU^\top,+} \Xi_1 \mu_{U^{(\mathsf{w})}UU^\top,+}\right) - 2 \cdot \operatorname{tr}\left(\Xi_1 \mu_{U^{(\mathsf{u})}UU^\top,-} \Xi_1 \mu_{U^{(\mathsf{w})}UU^\top,+}\right) + \operatorname{tr}\left(\Xi_1 \mu_{U^{(\mathsf{u})}UU^\top,-} \Xi_1 \mu_{U^{(\mathsf{w})}UU^\top,-}\right)$. Then, let

$$\mathscr{V}_p^\dagger \coloneqq \sum_{\mathsf{u},\mathsf{w}=1,\ldots,d_u} \left\{ \frac{1}{2} \frac{\omega_{p;+}^{0,4}}{\omega_{p;+}^{0,2}} \Xi_1^{(\mathsf{uw})} \Psi_1^{\mathsf{uw}} - \frac{1}{3} \frac{\left(\omega_{p;+}^{0,3}\right)^2}{\left(\omega_{p;+}^{0,2}\right)^2} \Xi_1^{(\mathsf{uw})} \Psi_2^{\mathsf{uw}} + \left(4\omega_{p;+}^{0,3} - 2\left(\omega_{p;+}^{0,2}\right)^2\right) \operatorname{tr}\left(\Xi_1 \Xi_2\right) \right\} / \left(\omega_{p;+}^{0,2} \varphi\right), \tag{10}$$

Let $\mathscr{V}_p^\ddagger$ be defined by the same formula with U replaced by $\bar{U} \coloneqq Z - \vartheta_2$ and the range changed to $\mathsf{u}, \mathsf{w} = 1, ..., d_z$ accordingly. Let $\mathscr{V}_p^{\mathsf{LR}} \coloneqq \mathscr{V}_p^\dagger - \mathscr{V}_p^\ddagger$. We write $a_n \asymp b_n$, if $a_n = O(b_n)$ and $b_n = O(a_n)$. The following theorem

is similar to Calonico et al. (2020, Theorem 3.1) and provides an asymptotic expansion of the coverage probability $\Pr\left[\vartheta_0 \in CS_\tau\left(h\right)\right] = \Pr\left[LR_p\left(\vartheta_0 \mid h\right) \leq c_\tau\right]$.

Theorem 3. *Suppose that Assumptions 1 - 5 hold. Suppose that h satisfies $nh^{2p+3} = o\left(1\right)$ and $\left(nh^3\right)^{-1} = O\left(1\right)$. Then, $\Pr\left[LR_p\left(\vartheta_0 \mid h\right) \leq x\right] = F_{\chi_1^2}\left(x\right) - \mathscr{C}_p\left(n,h\right) x f_{\chi_1^2}\left(x\right) + o\left(\upsilon_n^\star\right)$, where $\upsilon_n^\star \coloneqq nh^{2p+3} + h^{p+1} + \left(nh\right)^{-1}$, $\mathscr{C}_p\left(n,h\right) \coloneqq nh^{2p+3}\mathscr{B}_p^{\mathsf{LR}} + \left(nh\right)^{-1}\mathscr{V}_p^{\mathsf{LR}}$ and $\mathscr{B}_p^{\mathsf{LR}} \coloneqq \left(\mathscr{B}_p^{\mathsf{EL}}\right)^2 / \mathscr{V}_p^{\mathsf{EL}}$. Let the RCEL ratio be $LR_p^{\mathsf{rc}}\left(\theta_0 \mid h\right) \coloneqq LR_{p+1}\left(\theta_0 \mid h\right) / \left(1 + \left(nh\right)^{-1}\mathscr{V}_{p+1}^{\mathsf{LR}}\right)$. Then, $\Pr\left[LR_p^{\mathsf{rc}}\left(\vartheta_0 \mid h\right) \leq x\right] = F_{\chi_1^2}\left(x\right) + O\left(n^{-1}\right)$, if $h \asymp n^{-1/(p+2)}$.*

Remark 9. Theorem 3 shows that the coverage error $\Pr\left[\vartheta_0 \in CS_\tau\left(h\right)\right] - (1-\tau)$ is approximately equal to $-\mathscr{C}_p\left(n,h\right) c_\tau f_{\chi_1^2}\left(c_\tau\right)$. In the leading coverage error term, $nh^{2p+3}\mathscr{B}_p^{\mathsf{LR}}$ is the "bias" term that is contributed by the smoothing bias and $\left(nh\right)^{-1}\mathscr{V}_p^{\mathsf{LR}}$ is the "variance term" that stems from the stochastic variability. 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 $-n^{-(p+1)/(p+2)}C_1^{\mathsf{EL}}\left(H\right)$. where $C_1^{\mathsf{EL}}\left(H\right) \coloneqq \mathscr{B}_p^{\mathsf{LR}}H^{2p+3} + \mathscr{V}_p^{\mathsf{LR}}H^{-1}$. The coverage expansion for the EL confidence set takes a much simpler form than its Wald-type counterpart so that simple rescaling correction removes the leading term.[13] In addition, because the moment conditions $W_p \otimes \left(Y - \theta_0 D - \theta_1, Z^\top - \theta_2^\top\right)^\top$ (with $\theta = \vartheta$) is asymptotically uncorrelated with the derivatives with respect to (θ_1, θ_2) at $\theta = \vartheta$, the leading term in the distributional expansion further simplifies and as a result, $\mathscr{V}_p^{\mathsf{LR}}$ can be easily estimated by the plug-in estimator. We provide a matrix formula for $\mathscr{V}_p^{\mathsf{LR}}$ in the appendix. Parallel to Calonico et al., 2018b, we define the optimal constant H_{co} as the minimizer of the absolute value of the leading coverage error: $H_{\mathsf{co}} \coloneqq \operatorname{argmin}_{H>0}\left|C_1^{\mathsf{EL}}\left(H\right)\right|$, Hence the coverage optimal bandwidth is given by $h_{\mathsf{co}} = H_{\mathsf{co}}n^{-1/(p+2)}$.[14] where 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 of Theorem 3 that if $h \asymp n^{-1/(p+2)}$ the remainder term in the expansion of $\Pr\left[LR_p\left(\vartheta_0 \mid h\right) \leq x\right]$ is $O\left(n^{-1}\right)$ (up to a logarithmic term). Let $LR_p^{\mathsf{bc}}\left(\theta_0 \mid h\right) \coloneqq LR_p\left(\theta_0 \mid h\right) / \left(1 + \mathscr{C}_p\left(n,h\right)\right)$ be the standard Bartlett corrected EL ratio. One can show that Bartlett correction removes the leading coverage error term in the expansion in the first conclusion of Theorem 3: $\Pr\left[LR_p^{\mathsf{bc}}\left(\vartheta_0 \mid h\right) \leq x\right] = F_{\chi_1^2}\left(x\right) + o\left(\upsilon_n^\star\right)$. The infeasible Bartlett corrected EL confidence set $\left\{\theta_0 : LR_p^{\mathsf{bc}}\left(\theta_0 \mid h\right) \leq c_\tau\right\}$ has coverage accuracy with error rate $O\left(n^{-1}\right)$. However, $\left(\mathscr{B}_p^{\mathsf{LR}}, \mathscr{V}_p^{\mathsf{LR}}\right)$ depend on unknown parameters. One can replace these unknown

[13] Let $WS_p\left(\theta_0 \mid h\right)$ 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 $\bar{C}\left(H,x\right) n^{-(p+1)/(p+2)}$, where $\bar{C}\left(H,x\right) \coloneqq \left(C_1\left(H\right)x + C_3\left(H\right)x^3 + C_5\left(H\right)x^5\right) f_{\chi_1^2}\left(x\right)$, $C_k\left(H\right) \coloneqq c_{k,1}H^{2p+3} + c_{k,2}H^{p+1} + c_{k,3}H^{-1}$ and $\left(c_{k,1}, c_{k,2}, c_{k,3}\right)$ are constants. The distributional expansion corresponding to the EL ratio is similar but much simpler. Its leading error term satisfies $C_3\left(H\right) = C_5\left(H\right) = c_{1,2} = 0$.

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

quantities with their consistent nonparametric estimators to get the feasible Bartlett corrected EL confidence set. Note that $\mathscr{B}_p^{\mathsf{LR}}$ involves higher-order derivatives up to the order $p+1$ while $\mathscr{V}_p^{\mathsf{LR}}$ depends only on conditional expectations. Hence the latter can be estimated by a simple plug-in estimator $\widehat{\mathscr{V}}_p^{\mathsf{LR}}$ that is based on local linear regression with standard ROT bandwidths (Hansen, 2021, Chapter 21.6). By standard theory, $\widehat{\mathscr{V}}_p^{\mathsf{LR}} - \mathscr{V}_p^{\mathsf{LR}} = O_p\left(n^{-2/5}\right)$. On the other hand, a fully nonparametric estimator of $\mathscr{B}_p^{\mathsf{LR}}$ 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 $\mathscr{B}_p^{\mathsf{LR}}$ and for this reason, its coverage error decay rate can be much slower than $O\left(n^{-1}\right)$. This motivates us to propose a simple modification that avoids the estimation of $\mathscr{B}_p^{\mathsf{LR}}$, see the following Remark 11.

Remark 11. To avoid estimating $\mathscr{B}_p^{\mathsf{LR}}$ in the expansion for $\Pr\left[LR_p\left(\vartheta_0 \mid h\right) \leq x\right]$, we internalize bias removal by increase the order of LP by one, in the spirit of Calonico et al. (2014).[15] Under the smoothness assumptions, the smoothing bias is now of order h^{p+2}. EL automatically accounts for the change in variation by implicit studentization. In the proof of Theorem 3, we show that $\Pr\left[LR_{p+1}\left(\vartheta_0 \mid h\right) \leq x\right]$ is equal to the sum of $F_{\chi_1^2}\left(x\right) - \left(nh\right)^{-1} \mathscr{V}_{p+1}^{\mathsf{LR}} x f_{\chi_1^2}\left(x\right)$ and a remainder term that absorbs contribution from the smoothing bias. The leading "variance" term becomes $\left(nh\right)^{-1} \mathscr{V}_{p+1}^{\mathsf{LR}}$ and rescaling the EL ratio by $\left(1 + \left(nh\right)^{-1} \mathscr{V}_{p+1}^{\mathsf{LR}}\right)^{-1}$ eliminates it. Essentially this approach trades bias for variance, as the latter can be estimated with good accuracy. Removal of the term of order $\left(nh\right)^{-1}$ makes it feasible to choose a small bandwidth [Yu: what does the "small" mean? isn't the rate the same as before?]. The rate optimal bandwidth that minimizes the decay rate of the remainder term balances the terms of order nh^{2p+5} and h^{p+3}. The remainder term is $O\left(n^{-1}\right)$ if the rate of h is chosen optimally (i.e., $h \asymp n^{-1/(p+2)}$). Let the feasible RCEL ratio be $LR_p^{\mathsf{frc}}\left(\theta_0 \mid h\right) \coloneqq LR_{p+1}\left(\theta_0 \mid h\right) / \left(1 + \left(nh\right)^{-1} \widehat{\mathscr{V}}_{p+1}^{\mathsf{LR}}\right)$, where $\widehat{\mathscr{V}}_{p+1}^{\mathsf{LR}}$ is a plug-in estimator of $\mathscr{V}_{p+1}^{\mathsf{LR}}$. Since $\widehat{\mathscr{V}}_{p+1}^{\mathsf{LR}} - \mathscr{V}_{p+1}^{\mathsf{LR}} = O_p\left(n^{-2/5}\right)$, the distributions of $LR_p^{\mathsf{frc}}\left(\vartheta_0 \mid h\right)$ and $LR_p^{\mathsf{rc}}\left(\vartheta_0 \mid h\right)$ differ by an error of order $o\left(n^{-1}\right)$ and the second conclusion of Theorem 3 holds for $\Pr\left[LR_p^{\mathsf{frc}}\left(\vartheta_0 \mid h\right) \leq x\right]$. Then it follows that the feasible RCEL confidence set $CS_\tau^{\mathsf{frc}}\left(h\right) \coloneqq \left\{\theta_0 : LR_p^{\mathsf{frc}}\left(\theta_0 \mid h\right) \leq c_\tau\right\}$ has a coverage error of order $O\left(n^{-1}\right)$. Following Gelman and Imbens (2019), we recommend using lower order local polynomials and setting $p = 1$ or $p = 2$. In both cases, $CS_\tau^{\mathsf{frc}}\left(h\right)$ has coverage error decay rate $O\left(n^{-1}\right)$. 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 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 $CS_\tau^{\mathsf{frc}}\left(h\right)$ is of larger order of magnitude.

Remark 12. We now compare the feasible RCEL to CCFT's inference method. When $p = 1$, CCFT proposes

[15] Calonico 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.

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(b)) 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)). Similar to CCFT, we use local quadratic moment conditions ($p+1 = 2$) in (4) to reduce the smoothing bias. Meanwhule, our EL-based method analytically accounts for the effect of stochastic variability on the coverage error and then chooses the bandwidth rate optimally (i.e., setting $h \asymp n^{-1/3}$) to alleviate the effect of the smoothing bias. Our method achieves a faster $O\left(n^{-1}\right)$ coverage error decay rate. Note that the same smoothness assumption (i.e., twice differentiability of the conditional expectation functions) underlies such comparison. Our method has a coverage error of a smaller order but requires no more stringent assumptions. It is also computationally inexpensive since resampling for standard error estimation is not needed. If more smoothness is available (thrice differentiability, see Calonico et al., 2020, Theorem 3.1(b)), we can set $p = 2$. A cubic local regression (polynomial order $= p + 1$) increases variability but our method takes it into account through accurate estimation of the change in variability . The length of the resulting $CS_\tau^{\mathsf{frc}}(h)$ for $p = 2$ with the rate optimal bandwidth (i.e., $h \asymp n^{-1/4}$) has the same order as CCFT's confidence interval but $CS_\tau^{\mathsf{frc}}(h)$ enjoys a faster $O\left(n^{-1}\right)$ 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 $\widehat{H}$ and reports $CS_\tau^{\mathsf{frc}}\left(\widehat{h}\right)$ where $\widehat{h} \coloneqq \widehat{H} \cdot n^{-1/(p+2)}$. 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 rule-of-thumb (ROT) bandwidth proposed in Hansen (2021, Chapter 21.6). Our correction removes the leading coverage error term and makes coverage accuracy less sensitive to bandwidth choice. In simulations, we find that $CS_\tau^{\mathsf{frc}}\left(\widehat{h}\right)$ with $\widehat{h}$ taken to be any of the aforementioned data-driven bandwidth selectors has good coverage accuracy.

Remark 14. It can be easily verified that the conclusion of Theorem 2 with p changed to $p + 1$ still holds when $LR_{p+1}(\vartheta_0 \mid h)$ is replaced by $LR_p^{\mathsf{frc}}(\vartheta_0 \mid h)$ since they are first-order equivalent, uniformly in $h \in \mathbb{H}$. The "doubly corrected" confidence set $CS_\tau^{\mathsf{frc}}\left(h \mid \overline{h}/\underline{h}\right) \coloneqq \left\{\theta_0 : LR_p^{\mathsf{frc}}(\theta_0 \mid h) \leq z_\tau \left(\overline{h}/\underline{h}\right)^2\right\}$ takes into account

specification search over multiple bandwidths within $\mathbb{H}$. In the proofs, we show that the distribution of $\sup_{h\in\mathbb{H}} LR_p^{\mathsf{frc}}(\vartheta_0 \mid h))$ is approximated by the distribution of $\left\|\tilde{\Gamma}_G\right\|_{\mathbb{H}}^2 = \sup_{h\in\mathbb{H}} \tilde{\Gamma}_G(h)^2$ with a vanishing error, where $\left\{\tilde{\Gamma}_G(h) : h \in \mathbb{H}\right\}$ is a zero-mean Gaussian process whose definition is in the proof of Theorem 2 and $\tilde{\Gamma}_G(h)^2$ follows the χ_1^2 distribution $\forall h \in \mathbb{H}$.[16] Theorem 3 shows that for a single bandwidth h, the error of the distributional approximation $\Pr\left[LR_p^{\mathsf{frc}}(\vartheta_0 \mid h) \leq x\right] - \Pr\left[\tilde{\Gamma}_G(h)^2 \leq x\right]$ is $O\left(n^{-1}\right)$. We expect that the distributional approximation of $\sup_{h\in\mathbb{H}} \tilde{\Gamma}_G(h)^2 = \|\Gamma_G\|^2_{[1,\overline{h}/\underline{h}]}$ to $\sup_{h\in\mathbb{H}} LR_p^{\mathsf{frc}}(\vartheta_0 \mid h)$ inherits the favorable property of a fast decay rate of $\Pr\left[LR_p^{\mathsf{frc}}(\vartheta_0 \mid h) \leq x\right] - \Pr\left[\tilde{\Gamma}_G(h)^2 \leq x\right]$ with a single $h \in \mathbb{H}$. Therefore, we also expect a small coverage error for the confidence band $\left\{CS_\tau^{\mathsf{frc}}\left(h \mid \overline{h}/\underline{h}\right) : h \in \mathbb{H}\right\}$. Our method thus serves as a very effective tool for AK-type sensitivity analysis and robust inference.

We have shown that the RCEL confidence set has superb coverage accuracy, under the covariate balance assumption. We now consider the situation in which covariate balance fails to hold and analyze the sensitivity of the coverage accuracy of RCEL to the balance 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 include any Z with $\mathscr{T}_Z = 0$, regardless of pre-determined covariate or "unaffected" outcome. While expanding the set of covariates may improve the efficiency of estimation and inference, it bears the risk that the prior belief $\mathscr{T}_Z = 0$ is actually wrong for some "unaffected" outcome 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 of the RD model.. However, the usual falsification test sets $\mathscr{T}_Z = 0$ as the null hypothesis and may fail to reject if $\mathscr{T}_Z$ is close to the hypothesized value 0 under the null hypothesis. [17]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). $Z_1, ..., Z_n$ 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 for RD, the researcher may mistakenly include covariates that slightly violate the assumption $\mu_{Z,+} = \mu_{Z,-}$. Theorem 4 shows that our method is useful in the situations when our prior belief about the placebo outcomes is imperfect or the data on covariates are contaminated but the incurred imbalance is slight.

[16] $\left\{\tilde{\Gamma}_G(h) : h \in \mathbb{H}\right\}$ relates to $\left\{\Gamma_G(s) : s \in \left[1, \overline{h}/\underline{h}\right]\right\}$ by $\Gamma_G(s) = \tilde{\Gamma}_G(s \cdot \underline{h})\ \forall s \in \left[1, \overline{h}/\underline{h}\right]$ and clearly, $\left\|\tilde{\Gamma}_G\right\|_{\mathbb{H}} = \|\Gamma_G\|_{[1,\overline{h}/\underline{h}]}$.

[17] For example, 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 studying the effect on child mortality rate from causes affected by Head Start. However, the fact that the balance condition involving Z is not rejected empirically may be caused by a lack of sufficient power.

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 $\mathscr{T}_Z$ approaches the hypothesized value 0 under covariate balance at the rate of $(nh)^{-1/2}$ so that the coverage probability $\Pr\left[\vartheta_0 \in CS_\tau^{\mathsf{rc}}(h)\right] = \Pr\left[LR_p^{\mathsf{rc}}(\vartheta_0 \mid h) \leq c_\tau\right]$ 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 $l_n \coloneqq n^{-(p+1)/(2p+4)} \asymp (nh)^{-1/2}$. As the standard Pitman approach to analyzing power properties of tests of parametric hypotheses, we think of the local imbalance hypothesis $\mathscr{T}_Z = \delta l_n$ as reparametrization of values of $\mathscr{T}_Z$ that lie in a small neighborhood around 0, where $\delta \in \mathbb{R}^{d_z}$ denotes the localizing parameter.[18] $\mathscr{T}_Z = \delta l_n$ 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 $\mathrm{E}\left[h^{-1}W_{p;-}(Z - \vartheta_2)\right] = O(l_n)$. 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.[19] This is in accordance with fact that local imbalance with a large δ can be detected with a high probability in the first-stage falsification test.

We now consider $\Pr\left[\vartheta_0 \in CS_\tau^{\mathsf{rc}}(h)\right]$ as a function of δ under local imbalance. Theorem 3 shows that if $\delta = 0$, $\Pr\left[\vartheta_0 \in CS_\tau^{\mathsf{rc}}(h)\right] = 1 - \tau + O\left(n^{-1}\right)$. 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 $\Pr\left[\vartheta_0 \in CS_\tau^{\mathsf{rc}}(h)\right]$ as a function of δ at $\delta = 0$. We extend Theorem 3 and derive a two-term asymptotic expansion for $\Pr\left[\vartheta_0 \in CS_\tau^{\mathsf{rc}}(h)\right]$ and take the sum of the leading terms, denoted by $R(\delta)$, as approximation to $\Pr\left[\vartheta_0 \in CS_\tau^{\mathsf{rc}}(h)\right]$ in finite samples. We show that $R(0) = 1 - \tau$ and the gradient $\nabla R(\delta) \coloneqq \partial R(\delta)/\partial\delta$ 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 $Z_1, ..., Z_n$ are drawn from a locally perturbed population that satisfies $\mu_{Z,+} - \mu_{Z,-} = \delta l_n$. By using similar techniques as in Bravo (2003), we derive a second-order approximation to the distribution

[18] Our specification of $\mathscr{T}_Z = \delta l_n$ follows Gallant and White (1988, Chapter 7)'s "fixed data-generating process (DGP), drifting hypothesis" approach. $\mathscr{T}_Z = \delta l_n$ is understood as the assumption that our maintained hypothesized value 0 for the true RD LATE $\mathscr{T}_Z$ is chosen in such a way that $\mathscr{T}_Z - 0 = l_n\delta$ for some $\delta \neq 0$. For the alternative "fixed hypothesis, drifting DGP" approach, we let $G : \mathbb{R}^{5+2d_z} \to [0,1]$ denote the CDF of a joint distribution of the latent variables $(Y(0), Y(1), D_+, D_-, Z(0), Z(1))$ and the score X such that Assumptions 1, 2, 4 and $\mathscr{T}_Z = 0$ are satisfied and $\mathscr{T}_Z = \delta l_n$ is understood as the assumption that the latent variables and the score obey a joint distribution that is given by G with a location shift. By taking this alternative approach, we can show a result that is similar to Theorem 4 and leads to an identical conclusion. The proof requires more complicated arguments and suitable modification of the assumptions.

[19] The 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.

of $LR_p^{\mathsf{rc}}(\vartheta_0 \mid h)$ under $\mathscr{T}_Z = \delta l_n$. $F(\cdot \mid \eta)$ denotes the CDF of a $\chi_1^2(\eta)$ (non-central χ^2 with one degree of freedom and non-centrality parameter $\eta \geq 0$) random variable. Let $F^{(k)}(x \mid \eta) \coloneqq \partial^k F(x \mid \eta) / \partial \eta^k$ be the k-times partial derivative of $F(x \mid \eta)$ with respect to η.

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

Remark 15. In the approximation to $\Pr\left[\vartheta_0 \in CS_\tau^{\mathsf{rc}}(h)\right] = \Pr\left[LR_p^{\mathsf{rc}}(\vartheta_0 \mid h) \leq c_\tau\right]$, the first-order term $F\left(c_\tau \mid H\left(\gamma_\Delta^\top \delta / \bar{\sigma}_{p+1}\right)^2\right)$ is an even function of δ, $\partial F\left(c_\tau \mid H\left(\gamma_\Delta^\top \delta / \bar{\sigma}_{p+1}\right)^2\right) / \partial \delta\Big|_{\delta=0} = 0$ 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 $R(\delta) \coloneqq F\left(c_\tau \mid H\left(\gamma_\Delta^\top \delta / \bar{\sigma}_{p+1}\right)^2\right) + P(c_\tau, \delta) l_n$. 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 $\mathbb{S}_\nu$, where ν denotes a positive constant and $\mathbb{S}_\nu \coloneqq \left\{\delta \in \mathbb{R}^{d_z} : \|\delta\| = \nu\right\}$ represents perturbations with equal magnitude ν in all directions. $\delta_R^* \coloneqq \operatorname{argmin}_{\delta \in \mathbb{S}_\nu} R(\delta)$ 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.[20] 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 $\rho_\varsigma(x) \coloneqq \left(x^{-\varsigma} - 1\right) / \{\varsigma(1 + \varsigma)\}$ for $\varsigma \in \mathbb{R}$. We interpret $\rho_0(x) = -\log(x)$ as the limit of $\rho_\varsigma(x)$ as $\varsigma \to 0$. The nonparametric likelihood (NPL) criterion function $\ell_p^\varsigma(\theta \mid h)$ is defined by (5) with $\sum_i \rho_0(n \cdot w_i) = -\sum_i \log\left(w_i / n^{-1}\right)$ replaced by the more general Cressie-Read divergence $\sum_i \rho_\varsigma(n \cdot w_i)$. The NPL ratio $LR_{p+1}^\varsigma(\theta_0 \mid h)$ and confidence set are defined analogously. Under the same assumptions as in Theorem 4, we can show that $\Pr\left[LR_{p+1}^\varsigma(\vartheta_0 \mid h) \leq x\right]$ admits a similar two-term asymptotic expansion.[21] The first-order term in the expansion for $\Pr\left[LR_{p+1}^\varsigma(\vartheta_0 \mid h) \leq x\right]$ is still given by $F\left(x \mid H\left(\gamma_\Delta^\top \delta / \bar{\sigma}_{p+1}\right)^2\right)$. 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 the case of $\varsigma \neq 0$, since $\partial P_\varsigma(x, \delta) / \partial \delta|_{\delta=0} \neq 0$, there

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

[21] The proof of this result is omitted for brevity but available from the authors.

can be perturbation associated with a large drop in coverage probability of the NPL confidence set. Let $R_\varsigma(\delta) \coloneqq F\left(c_\tau \mid H\left(\gamma_\Delta^\top \delta/\bar{\sigma}_{p+1}\right)^2\right) + (P(x,\delta) + P_\varsigma(x,\delta))\, l_n$ and $\delta^*_{R_\varsigma} \coloneqq \operatorname{argmin}_{\delta\in\mathbb{S}_\nu} R_\varsigma(\delta)$. Then by similar arguments, $R_\varsigma\left(\delta^*_{R_\varsigma}\right) = (1-\tau) - \|\nabla R_\varsigma(0)\|\,\nu + o(\nu)$, 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 practical performance of the proposed corrected EL inference for sharp RD designs with covariates. The DGP of the outcome variable Y_i, the running variable X_i and the first covariate Z_{1i} is based on the simulation design of CCFT. Incorporation of additional covariates $Z_{2i}, ..., Z_{li}$ follows that of Arai et al. (2021). I.e., $Y_i = \mu_y(X_i, Z_{1i}) + \sum_{j=2}^{l} \pi^{j-1} Z_{ji} + \varepsilon_{y,i}$ and $Z_i = \mu_z(X_i) + \varepsilon_{z,i}$, where

$$\mu_y(x, z_1) \coloneqq \begin{cases} 0.36 + 0.96x + 5.47x^2 + 15.28x^3 + 15.87x^4 + 5.14x^5 + 0.22z_1 & \text{if } x < 0, \\ 0.38 + 0.62x - 2.84x^2 + 8.42x^3 - 10.24x^4 + 4.31x^5 + 0.28z_1 & \text{if } x \geq 0; \end{cases}$$

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

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 $(Z_{2i}, ..., Z_{li})$ have a multivariate normal distribution with mean zero and covariance matrix given by $\operatorname{Cov}[Z_{ji}, Z_{ki}] = 0.5^{|j-k|}$, $j, k \geq 2$. The coefficient $\pi = 0.2$. We consider three scenarios with $l = 0, 2$, and 4, which correspond to the total number of covariates $d_z = l + 1$ being $1, 3$, and 5. We take the LP order $p = 1$ and 2. The sample sizes are $n = 500, 1000$ and $2,000$. The number of Monte Carlo replications is $2,000$.

Table 1 presents the empirical coverage rates of the feasible RCEL confidence set $CS^{\mathsf{frc}}_\tau\left(\widehat{h}\right)$ proposed in Remark 11. The nominal coverage $1-\tau = 0.99$, 0.95 and 0.90. Following Remark 13, we consider bandwidth in the form of $\widehat{h} = \widehat{H} \cdot n^{-1/(p+2)}$ and two choices ofr the constant part $\widehat{H}$: ROT in the table corresponding to the rule-of-thumb recommended in Hansen (2021, Chapter 21.6) and CCFT corresponding to rescaled CCFT's bandwidth. Both ROT and CCFT bandwidths used for RCLE obey the optimal coverage rate discussed in Remark 10.[22] Table 1 also includes CCFT approach for comparison. For each (n, τ) combination,

[22] As Remark 12 notes, CCFT's coverage optimal bandwidth takes 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 8, our CCFT bandwidth used for "RCLE, $p = 1$" is $h_{CCFT} \cdot n^{-1/12}$ (rescaled to the coverage optimal rate $n^{-1/(p+2)}$ discussed in Remark 10). The CCFT bandwidth used for "RCLE, $p = 2$" is h_{CCFT} itself, as now the coverage

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 CCFT bandwidths perform reasonably well. When sample size is small ($n = 500$), RCLE exhibits a small advantage over CCFT, which is in line with the theoretical result that RCLE achieves a faster coverage error decay rate than CCFT, see Remark 12. We then examine the how the coverage rate of RCEL confidence set changes when the covariate balance condition is violated. We consider the case with one covariate ($d_z = 1$) and the modify the design of Z in the following way that the imbalance is characterized by a perturbation δ:

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

Obviously, $\mathscr{T}_Z = \delta$. Figure 8 plots the simulated coverage rates of RCEL and CCFT as a function of $\delta \in [-0.4, 0.4]$ for different 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 in Remark 15. Overall, our simulation results show that RCEL inference method can be a useful addition to practitioner's toolkit.

7 Empirical illustrations

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). The dataset has two appealing features: first, the sample size is large ($n = 154,543$), which provides a good example for illustrating non-parametric inference approaches. Second, it 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). The binary outcome variable Y indicates whether the candidate is elected in the next election, and the forcing variable X is the vote share margin in the previous election. Two covariates Z are included: candidates' age and gender. The main results are presented in Table 2. In the empirical section, the LP order $p = 1$. The ROT bandwidth is computed in the same way as the simulation exercise. The CCFT bandwidth and its rescaled version correspond to h_{CCFT} and $h_{CCFT} \cdot n^{-1/12}$ described in footnote 22. The columns of Table 2 present the estimates of RD treatment effect ϑ_0, p-values for testing

optimal rate for RCLE is $n^{-1/4}$.

$H_0 : \vartheta_0 = 0$, the 95% confidence intervals and the selected bandwidths. Both RCEL and CCFT deliver non-significant inference results comparable to the experiment benchmark. In contrast, as Hyytinen et al. (2018, columns (2) of Table 4) show, CCT using the MSE-optimal bandwidth rejects the null hypothesis of zero treatment effect, which is at odds with the experiment estimate. We also note that RCEL does not rejects the null hypothesis for all three bandwidth choices. This robustness is further confirmed by Figure 8, which conducts a sensitivity analysis of the RCEL inference method with respect to the bandwidth choice. 8 plots the confidence band (Remarks 6 and 14) as function of a range of bandwidth $h \in [\underline{h}, \overline{h}]$. Here we choose the lower bound of bandwidth range $\underline{h} = 0.12$, which would include about 3% of the sample. The upper bound $\overline{h} = 0.72$ is nearly two times the CCFT bandwidth in Table 2 and includes 17% of the total sample. The bandwidth snooping critical value $z_r = 2.413$ for the triangular kernel and bandwidth ratio $\overline{h}/\underline{h} = 6$ is calculated from `R` package `BWSnooping`. In Figure 8, the solid (or dotted) line corresponds to 95% uniform (or pointwise) confidence band. The vertical dashed lines indicate the three 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 for the entire bandwidth range we consider, which demonstrates the robustness of the finding of no incumbency advantage with respect to bandwidth choice. Overall, our example illustrates the practicality of the RCEL ratio inference approach.

8 Conclusion

This paper proposes a novel EL approach to covariate adjustment for regression discontinuity designs. Our approach incorporates covariates through novel over-identifying restrictions which represent the covariate balance condition. We show the first-order and second-order asymptotic properties of our method. We show that the widely-used regression estimator of CCFT achieves guaranteed efficiency gain. By establishing the first-order equivalence between our EL estimator and the regression estimator, we show that the efficiency gain can be attributed to incorporating the covariate balance condition as side information. We show a novel uniform-in-bandwidth Wilks' phenomenon, which can be used for sensitivity analysis and robust inference along the lines of AK. We derive the distributional expansion for the EL ratio statistic under the covariate balance condition and show that it 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 (2018a). Coverage error optimal confidence intervals for local polynomial regression. *arXiv preprint arXiv:1808.01398*.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2018b). 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, 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. 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.

DiCiccio, T., P. Hall, and J. Romano (1988). Bartlett adjustments for empirical likelihood. Technical report No. 298, Department of Statistics, Stanford University.

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.

Fan, J. and I. Gijbels (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, Volume 66. CRC 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.

Hall, P. and B. Presnell (1999). Intentionally biased bootstrap methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(1), 143–158.

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.

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.

Skovgaard, I. M. (1986). On multivariate edgeworth expansions. *International Statistical Review/Revue Internationale de Statistique*, 169–186.

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 coverage optimal bandwidth, d_Z = the number of covariates, p = the order of local polynomial moment conditions, $1-\tau$ = nominal coverage probability, n = sample size.

			$1-\tau=0.99$			$1-\tau=0.95$			$1-\tau=0.90$		
d_Z	Methods	h	$n=500$	$1,000$	$2,000$	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, $p=1$	CCFT	.9760	.9825	.9905	.9315	.9420	.9535	.8815	**.9005**	.9130
	CCFT, $p=2$	CCFT	.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, $p=1$	CCFT	.9670	.9767	.9860	.9130	**.9320**	**.9500**	.8580	.8695	**.9015**
	CCFT, $p=2$	CCFT	.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, $p=1$	CCFT	.9590	.9740	.9795	.8960	.9245	**.9345**	.8420	**.8820**	**.8805**
	CCFT, $p=2$	CCFT	.9630	.9690	.9760	.9060	.9185	.9295	.8475	.8745	.8745

Figure 1: Coverage rates when the balance condition violates by δ, n = sample size.

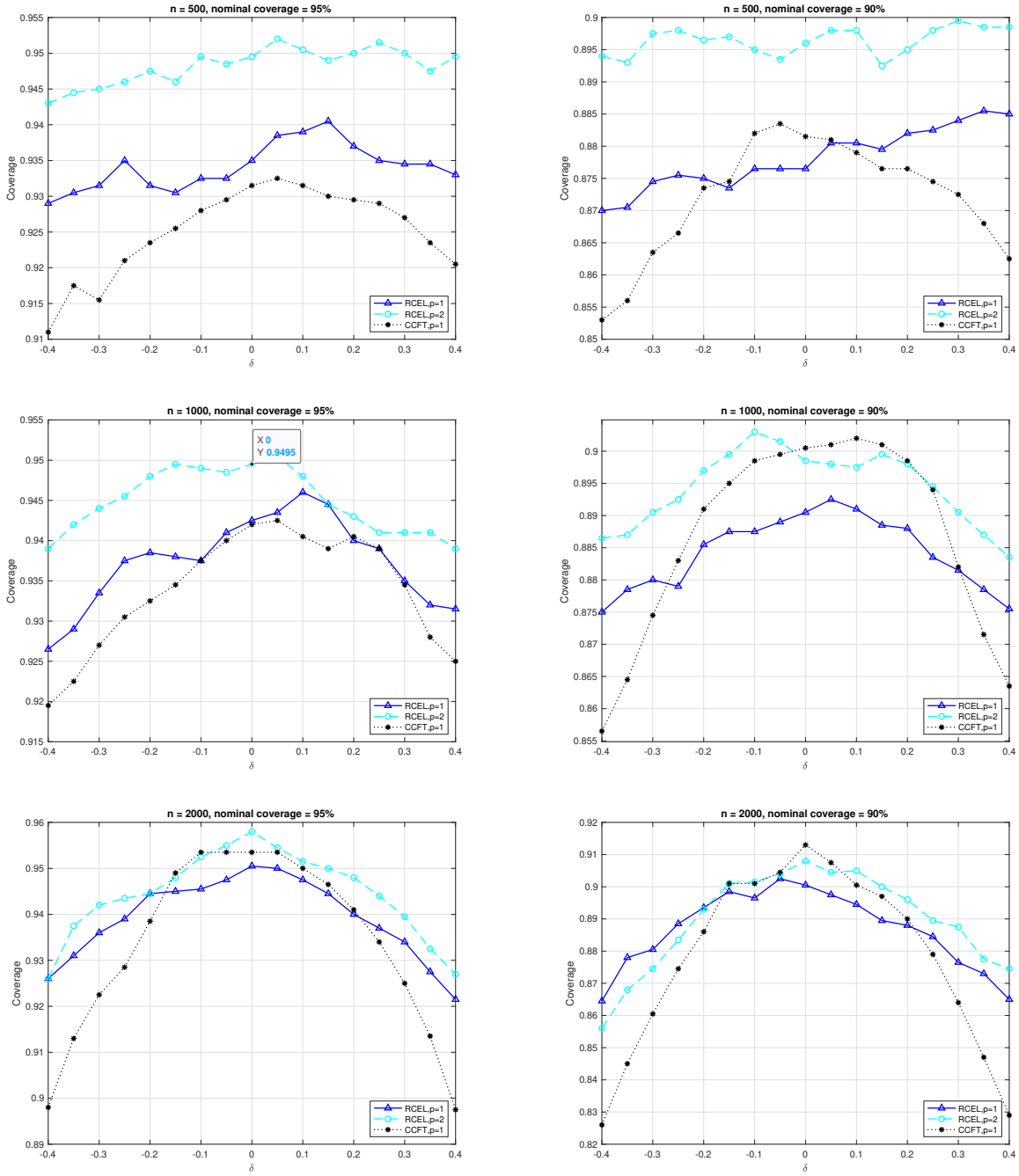
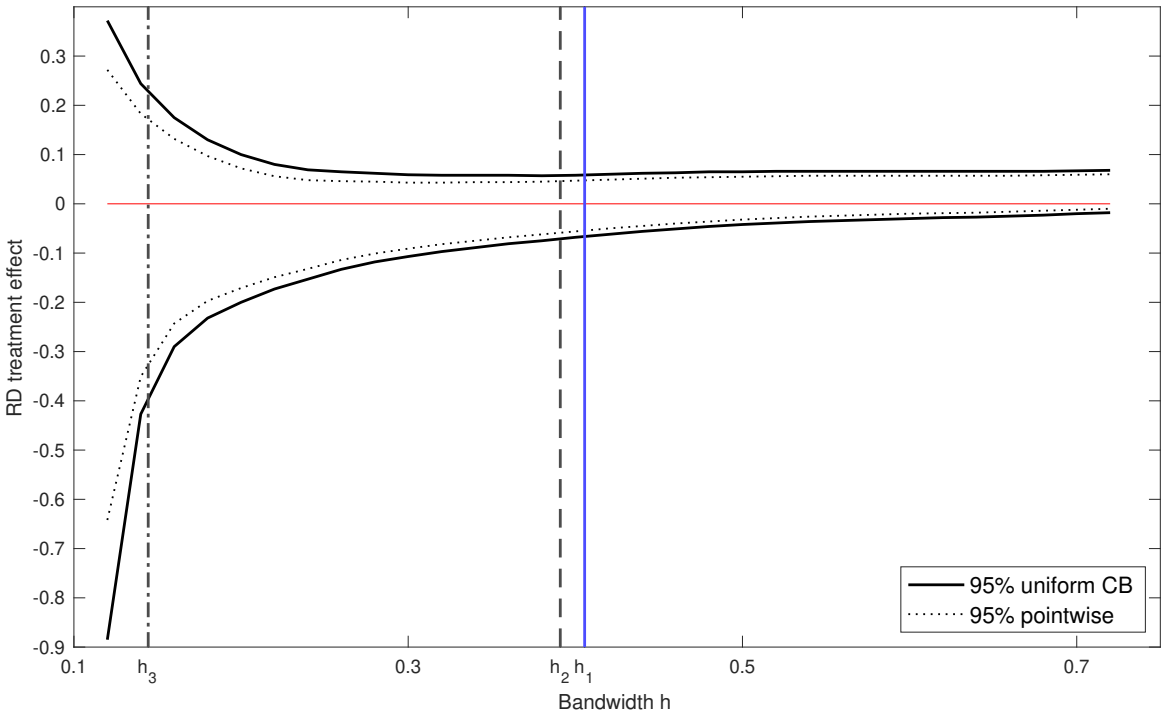


Table 2: Incumbency Advantage in Finnish Municipal Election: θ_0 = RD Treatment Effect, h = the bandwidth value.

	Methods	Bandwidth selector	$\hat{\theta}_0$	p-value	95% CI	CI length	h
RD with covariates							
$n = 154,543$	RCEL	ROT	-.003	.896	[-.054, .047]	.101	.406
	RCEL	CCFT	-.007	.807	[-.059 .046]	.105	.391
	RCEL	CCFT-rescaled	-.066	.587	[-.319, .171]	.490	.144
	CCFT	CCFT	-.012	.671	[-.068, .044]	.111	.391
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 as function of the bandwidth in the range $[0.12, 0.72]$. Bandwidth snooping critical value = 2.413^2. Vertical lines indicate bandwidth choices used in Table 2: ROT bandwidth $h_1 = 0.406$, CCFT bandwidth $h_2 = 0.391$, and rescaled CCFT bandwidth $h_3 = 0.144$.



Appendix A Proofs of Theorems 1 and 2

For a sequence of classes of $\mathbb{R}$-valued functions $\mathfrak{F}_n$ defined on $\mathscr{S}$ (a compact set in a finite-dimensional Euclidean space), let $\|f\|_{Q,2} := \left(\int f^2 \mathrm{d}Q\right)^{1/2}$ and $N\left(\varepsilon, \mathfrak{F}_n, \|\cdot\|_{Q,2}\right)$ 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 $F_{\mathfrak{F}_n}:\mathscr{S}\to\mathbb{R}_+$ 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 $\sup_{Q\in\mathcal{Q}_{\mathscr{S}}^{\mathtt{fd}}}N\left(\varepsilon\left\|F_{\mathfrak{F}_n}\right\|_{Q,2},\mathfrak{F}_n,\left\|\cdot\right\|_{Q,2}\right)\leq\left(A_{\mathfrak{F}_n}/\varepsilon\right)^{V_{\mathfrak{F}_n}}\ \forall\varepsilon\in(0,1]$ where $\mathcal{Q}_{\mathscr{S}}^{\mathtt{fd}}$ denotes the collection of all finitely discrete probability measures on $\mathscr{S}$. $\lesssim$ 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 $c>0$.

Lemma 1. *Let V denote a random variable and $\{V_1,...,V_n\}$ are i.i.d. copies of V. Let $\mathbb{B}(0)$ denote an open neighborhood of 0. Suppose that $(\underline{h},\overline{h})$ satisfy $\overline{h}=o(1)$. $\forall(\mathrm{s},k)\in\{-,+\}\times\mathbb{N}$, the following results hold uniformly in $h\in\mathbb{H}$: (a) if g_V is Lipschitz continuous on $\mathbb{B}(0)$, for $k\geq2$, $\mathrm{E}\left[h^{-1}W_{p;\mathrm{s}}^kV\right]=\psi_{V,\mathrm{s}}\omega_{p;\mathrm{s}}^{0,k}+O\left(\overline{h}\right)$; (b) if g_V is $(p+1)$-times continuously differentiable with Lipschitz continuous $g_V^{(p+1)}$ on $\mathbb{B}(0)$, $\mathrm{E}\left[h^{-1}W_{p;\mathrm{s}}V\right]=\psi_{V,\mathrm{s}}+\omega_{p;\mathrm{s}}^{p+1,1}\psi_{V,\mathrm{s}}^{(p+1)}h^{p+1}/(p+1)!+O\left(\overline{h}^{p+2}\right)$ and $\mathrm{E}\left[h^{-1}W_{p+1;\mathrm{s}}V\right]=\psi_{V,\mathrm{s}}+O\left(\overline{h}^{p+2}\right)$; (c) if $g_{|V|^r}$ is bounded on $\mathbb{B}(0)$ for some integer $r>2$, $(nh)^{-1/2}\sum_i\left(W_{p;\mathrm{s},i}^kV_i-\mathrm{E}\left[W_{p;\mathrm{s}}^kV\right]\right)=O_p\left(\sqrt{\log(n)}+\log(n)\left\{\left(n\overline{h}\right)^{1/r}/\left(n\underline{h}\right)^{1/2}\right\}\right)$.*

Proof. (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 $\bar{q}(V,X\mid h)\coloneqq h^{-1/2}W_{p;\mathrm{s}}^kV$ and $\bar{\mathfrak{Q}}\coloneqq\{\bar{q}(\cdot\mid h):h\in\mathbb{H}\}$. Denote $\mathbb{P}_n^Vf\coloneqq n^{-1}\sum_i f(V_i,X_i)$, $\mathbb{P}^Vf\coloneqq\mathrm{E}[f(V,X)]$ and $\mathbb{G}_n^V\coloneqq\sqrt{n}\left(\mathbb{P}_n^V-\mathbb{P}^V\right)$. Then we have $\left\|\mathbb{G}_n^V\right\|_{\bar{\mathfrak{Q}}}=\sup_{h\in\mathbb{H}}\left|(nh)^{-1/2}\sum_i\left(W_{p;\mathrm{s},i}^kV_i-\mathrm{E}\left[W_{p;\mathrm{s}}^kV\right]\right)\right|$. Let $\sigma_{\bar{\mathfrak{Q}}}^2\coloneqq\sup_{f\in\bar{\mathfrak{Q}}}\mathbb{P}^Vf^2$. It follows from LIE and change of variables that $\sigma_{\bar{\mathfrak{Q}}}^2=\sup_{h\in\mathbb{H}}\mathrm{E}\left[h^{-1}W_{p;\mathrm{s}}^{2k}g_{V^2}(X)\right]=O(1)$. Assume $\mathrm{s}=+$ without loss of generality. By definition and the assumption that K is supported on $[-1,1]$, $\bar{q}(v,x\mid h)=\mathcal{K}_{p;-}^k(x/h)h^{-1/2}1(0<x<h)v$. 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), $\left\{x\mapsto\mathcal{K}_{p;-}^k(x/h):h\in\mathbb{H}\right\}$ is VC-type with respect to a constant envelope and its VC characteristics are independent of n. By Kosorok (2007, Lemma 9.6), $\left\{(x,v)\mapsto h^{-1/2}1(0<x<h)v:h\in\mathbb{H}\right\}$ is VC-subgraph with an envelope $(x,v)\mapsto\underline{h}^{-1/2}1\left(0<x<\overline{h}\right)|v|$ and VC index being at most 3. By Kosorok (2007, Theorem 9.3) and Chernozhukov et al. (2014b, Corollary A.1), $\bar{\mathfrak{Q}}$ is VC-type with respect to an envelope $F_{\bar{\mathfrak{Q}}}(v,x)\propto\underline{h}^{-1/2}1\left(0<x<\overline{h}\right)|v|$. By Chen and Kato (2020, Corollary 5.5), $\mathrm{E}\left[\left\|\mathbb{G}_n^V\right\|_{\bar{\mathfrak{Q}}}\right]\lesssim\sigma_{\bar{\mathfrak{Q}}}\sqrt{\log(n)}+\log(n)\left(\mathbb{P}^V|F_{\bar{\mathfrak{Q}}}|^r\right)^{1/r}n^{1/r}/\sqrt{n}$, where $\mathbb{P}^V|F_{\bar{\mathfrak{Q}}}|^r=O\left(\overline{h}/\underline{h}^{r/2}\right)$. (c) follows from Markov's inequality. ■

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 $G_i\coloneqq\partial U_i(\theta)/\partial\theta^\top=\begin{bmatrix}G_{0,i} & G_{\dagger,i}\end{bmatrix}$, where $G_{0,i}\coloneqq\begin{bmatrix}D_i & 0_{d_z}^\top\end{bmatrix}^\top$ and $G_{\dagger,i}\coloneqq\mathrm{I}_{d_u}$. $(G,G_0,G_\dagger)$ are defined by the same formulae

with D_i replaced by D. Denote $\mathcal{U}_i(\theta) \coloneqq W_{p,i} \otimes U_i(\theta)$ and $\widehat{\mathcal{U}}_i \coloneqq W_{p,i} \otimes U_i\left(\widehat{\vartheta}_p\right)$. Let $\mathcal{U}_i \coloneqq W_{p,i} \otimes U_i$, $\mathcal{G}_i \coloneqq W_{p,i} \otimes G_i$, $\mathcal{G}_{0,i} \coloneqq W_{p,i} \otimes G_{0,i}$ and $\mathcal{G}_{\dagger,i} \coloneqq W_{p,i} \otimes G_{\dagger,i}$. $(\mathcal{U}, \mathcal{G}, \mathcal{G}_0, \mathcal{G}_\dagger)$ are defined similarly. Let $\mathcal{D}_\mathrm{s} \coloneqq W_{p;\mathrm{s}} D$. Denote $\Delta_\mathrm{s} \coloneqq \mathrm{E}\left[h^{-1} W_{p;\mathrm{s}}\right]$ and $\Delta_A \coloneqq \mathrm{E}\left[h^{-1} A\right]$ for a random variable/vector/matrix A. Let $\overline{\Delta}_{\mathcal{U}\mathcal{U}^\top} \coloneqq (nh)^{-1} \sum_i \mathcal{U}_i \mathcal{U}_i^\top$, $\widehat{\Delta}_{\mathcal{U}\mathcal{U}^\top} \coloneqq (nh)^{-1} \sum_i \widehat{\mathcal{U}}_i \widehat{\mathcal{U}}_i^\top$ and $\overline{\Delta}_{\mathcal{G}} \coloneqq (nh)^{-1} \sum_i \mathcal{G}_i$. Let $\overline{\mathcal{U}} \coloneqq (nh)^{-1/2} \sum_i \mathcal{U}_i$ and $\widehat{\mathcal{U}} \coloneqq (nh)^{-1/2} \sum_i \widehat{\mathcal{U}}_i$.

Denote $S(\lambda, \theta) \coloneqq 2 \sum_i \log\left(1 + \lambda^\top \mathcal{U}_i(\theta)\right)$ and $\widetilde{\vartheta}_p \coloneqq \widetilde{\vartheta}_p(\vartheta_0)$. Note that the dual form of the EL criterion function is $\ell_p(\theta \mid h) = \sup_{\lambda \in \mathcal{L}(\theta)} S(\lambda, \theta)$, where $\mathcal{L}(\theta) \coloneqq \left\{\lambda \in \mathbb{R}^{2d_u} : \lambda^\top \mathcal{U}_i(\theta) > -1,\ \forall i\right\}$. $\max_i$ is understood as $\max_{1 \le i \le n}$. For square matrices A and B, $\mathrm{diag}(\mathrm{A}, \mathrm{B})$ denotes the block diagonal matrix. $\|\mathrm{A}\|$ is understood as the spectral norm of A and $\varrho_{\min}(\mathrm{A})$ 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 $\overline{n} \coloneqq n\overline{h}$, $\underline{n} \coloneqq n\underline{h}$, $\widehat{\eta}_p \coloneqq \widehat{\vartheta}_p - \vartheta$ and $\widetilde{\eta}_p \coloneqq \widetilde{\vartheta}_p - \vartheta_\dagger$. "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 $\left(\underline{h}, \overline{h}\right)$ satisfy $n\overline{h}^{2p+3} = O(1)$ and $\log(n)\left(\overline{n}^{1/12}/\underline{n}^{1/2}\right) = o(1)$. Then, the following results hold uniformly in $h \in \mathbb{H}$: (a) $\sqrt{nh}\widehat{\eta}_p = O_p\left(\sqrt{\log(n)}\right)$; (b) $\widehat{\lambda}_p \coloneqq \mathrm{argmax}_{\lambda \in \mathcal{L}\left(\widehat{\vartheta}_p\right)} S\left(\lambda, \widehat{\vartheta}_p\right)$ exists wpa1 and $\sqrt{nh}\widehat{\lambda}_p = O_p\left(\sqrt{\log(n)}\right)$; (c) $\sqrt{nh}\widetilde{\eta}_p = O_p\left(\sqrt{\log(n)}\right)$; (d) $\widetilde{\lambda}_p \coloneqq \mathrm{argmax}_{\lambda \in \mathcal{L}\left(\vartheta_0, \widetilde{\vartheta}_p\right)} S\left(\lambda, \vartheta_0, \widetilde{\vartheta}_p\right)$ exists wpa1 and $\sqrt{nh}\widetilde{\lambda}_p = O_p\left(\sqrt{\log(n)}\right)$.*

Proof. Let $\mathcal{L}_\sharp \coloneqq \left\{\lambda \in \mathbb{R}^{2d_u} : \|\lambda\| \le \log(n)/\sqrt{nh}\right\}$. By $\max_i \|\mathcal{U}_i\|/\sqrt{nh} \lesssim \max_i 1\left(|X_i| \le \overline{h}\right)\|U_i\|/\sqrt{\underline{n}}$, $\max_i 1\left(|X_i| \le \overline{h}\right)\|U_i\| \le \left(\sum_i 1\left(|X_i| \le \overline{h}\right)\|U_i\|^{12}\right)^{1/12}$ and Markov's inequality, we have $\max_i \|\mathcal{U}_i\|/\sqrt{nh} = O_p\left(\overline{n}^{1/12}/\underline{n}^{1/2}\right)$. It follows that $\max_i \sup_{\lambda \in \mathcal{L}_\sharp} \left|\lambda^\top \mathcal{U}_i\right| = O_p\left(\log(n)\left(\overline{n}^{1/12}/\underline{n}^{1/2}\right)\right)$ and $\max_i \sup_{\lambda \in \mathcal{L}_\sharp} \left|\lambda^\top \mathcal{U}_i\right| < 1/2$ $\forall h \in \mathbb{H}$ wpa1. Therefore, $\mathcal{L}_\sharp \subseteq \mathcal{L}(\vartheta)$, $\forall h \in \mathbb{H}$ wpa1. Since $S(\cdot, \vartheta)$ is continuous and $\mathcal{L}_\sharp$ is compact, $\lambda_\sharp \coloneqq \mathrm{argmax}_{\lambda \in \mathcal{L}_\sharp} S(\lambda, \vartheta)$ exists $\forall h \in \mathbb{H}$ wpa1. By the definition of $\lambda_\sharp$ and second-order Taylor expansion,

$$
\begin{aligned}
0 = S(0_{2d_u}, \vartheta) \le S(\lambda_\sharp, \vartheta) &= 2\left(\sqrt{nh}\lambda_\sharp\right)^\top \overline{\mathcal{U}} - \left(\sqrt{nh}\lambda_\sharp\right)^\top \left(\frac{1}{nh}\sum_i \frac{\mathcal{U}_i \mathcal{U}_i^\top}{\left(1 + \dot{\lambda}_\sharp^\top \mathcal{U}_i\right)^2}\right)\left(\sqrt{nh}\lambda_\sharp\right) \\
&\le 2\left\|\sqrt{nh}\lambda_\sharp\right\|\|\overline{\mathcal{U}}\| - \left(\sqrt{nh}\lambda_\sharp\right)^\top \left(\frac{1}{nh}\sum_i \frac{\mathcal{U}_i \mathcal{U}_i^\top}{\left(1 + \max_i \sup_{\lambda \in \mathcal{L}_\sharp} \left|\lambda^\top \mathcal{U}_i\right|\right)^2}\right)\left(\sqrt{nh}\lambda_\sharp\right), \quad (11)
\end{aligned}
$$

where $\dot{\lambda}_\sharp$ is the mean value that lies on the line joining 0_{2d_u} and $\lambda_\sharp$. Since $\max_i \sup_{\lambda \in \mathcal{L}_\sharp} \left|\lambda^\top \mathcal{U}_i\right| < 1/2$ $\forall h \in \mathbb{H}$ wpa1, by (11),

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

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

$$\varrho_{\min}\left(\Delta_{\mathcal{U}\mathcal{U}^\top}\right)\left\|\sqrt{nh}\lambda_\sharp\right\|^2 \leq \frac{9}{2}\left\|\sqrt{nh}\lambda_\sharp\right\|\left\|\overline{\mathcal{U}}\right\| + \left\|\overline{\Delta}_{\mathcal{U}\mathcal{U}^\top} - \Delta_{\mathcal{U}\mathcal{U}^\top}\right\|\left\|\sqrt{nh}\lambda_\sharp\right\|^2, \tag{12}$$

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

$$\begin{aligned} S\left(\lambda_\natural, \widehat{\vartheta}_p\right) &= 2\left(\sqrt{nh}\lambda_\natural\right)^\top\widehat{\mathcal{U}} - \left(\sqrt{nh}\lambda_\natural\right)^\top\left(\frac{1}{nh}\sum_i\frac{\widehat{\mathcal{U}}_i\widehat{\mathcal{U}}_i^\top}{\left(1+\dot{\lambda}_\natural^\top\widehat{\mathcal{U}}_i\right)^2}\right)\left(\sqrt{nh}\lambda_\natural\right) \\ &\geq 2\left(\sqrt{nh}\lambda_\natural\right)^\top\widehat{\mathcal{U}} - \left(\sqrt{nh}\lambda_\natural\right)^\top\left(\frac{1}{nh}\sum_i\frac{\widehat{\mathcal{U}}_i\widehat{\mathcal{U}}_i^\top}{\left(1-\sqrt{\log(n)/(nh)}\left(\max_i\left\|\widehat{\mathcal{U}}_i\right\|\right)\right)^2}\right)\left(\sqrt{nh}\lambda_\natural\right), \end{aligned} \tag{13}$$

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

$$\left\|\sqrt{nh}\widehat{\eta}_p\right\|\sqrt{\varrho_{\min}\left(\overline{\Delta}_{\mathcal{G}}^\top\overline{\Delta}_{\mathcal{G}}\right)} \leq \left\|\overline{\Delta}_{\mathcal{G}}\sqrt{nh}\widehat{\eta}_p\right\| \leq \left\|\widehat{\mathcal{U}}\right\| + \left\|\overline{\mathcal{U}}\right\|. \tag{14}$$

By Lemma 1, $\overline{\Delta}_{\mathcal{G}} = \begin{bmatrix} \mu_{G,+}^\top & \mu_{G,-}^\top \end{bmatrix}^\top + O_p\left(\sqrt{\log(n)/\underline{n}} + \overline{h}\right)$. $\begin{bmatrix} \mu_{G,+}^\top & \mu_{G,-}^\top \end{bmatrix}^\top$ has full column rank, if $\mu_{D,+} \neq \mu_{D,-}$. By using the fact that $|\varrho_{\min}(\mathrm{A}) - \varrho_{\min}(\mathrm{B})| \leq \|\mathrm{A} - \mathrm{B}\|$, $\varrho_{\min}\left(\overline{\Delta}_{\mathcal{G}}^\top\overline{\Delta}_{\mathcal{G}}\right)$ is bounded away from zero $\forall h \in \mathbb{H}$, wpa1. (a) follows easily from this result, (14) and the fact that $\left\|\widehat{\mathcal{U}}\right\|$ and $\left\|\overline{\mathcal{U}}\right\|$ are both $O_p\left(\sqrt{\log(n)}\right)$. By $\max_i\left\|\widehat{\mathcal{U}}_i\right\|/\sqrt{nh} = O_p\left(\overline{n}^{1/12}/\underline{n}^{1/2}\right)$ and the definition of $\mathcal{L}_\sharp$, $\max_i\sup_{\lambda\in\mathcal{L}_\sharp}\left|\lambda^\top\widehat{\mathcal{U}}_i\right| = O_p\left(\log(n)\left(\overline{n}^{1/12}/\underline{n}^{1/2}\right)\right)$ and therefore $\max_i\sup_{\lambda\in\mathcal{L}_\sharp}\left|\lambda^\top\widehat{\mathcal{U}}_i\right| < 1/2$ $\forall h \in \mathbb{H}$ wpa1. Therefore, $\mathcal{L}_\sharp \subseteq \mathcal{L}\left(\widehat{\vartheta}_p\right)$,

$\forall h \in \mathbb{H}$ wpa1. Since $S\left(\cdot, \widehat{\vartheta}_p\right)$ is continuous and $\mathcal{L}_\sharp$ is compact, $\widehat{\lambda}_\sharp \coloneqq \operatorname{argmax}_{\lambda \in \mathcal{L}_\sharp} S\left(\lambda, \widehat{\vartheta}_p\right)$ exists $\forall h \in \mathbb{H}$ wpa1. By the definition of $\widehat{\lambda}_\sharp$ and similar arguments used to show (12), we have $\varrho_{\min}\left(\Delta_{\mathcal{U}\mathcal{U}^\top}\right)\left\|\sqrt{nh}\widehat{\lambda}_\sharp\right\| \lesssim \left\|\widehat{\mathcal{U}}\right\| + \left\|\widehat{\Delta}_{\mathcal{U}\mathcal{U}^\top} - \Delta_{\mathcal{U}\mathcal{U}^\top}\right\|\left\|\sqrt{nh}\widehat{\lambda}_\sharp\right\|$. Since $\widehat{\Delta}_{\mathcal{U}\mathcal{U}^\top} - \overline{\Delta}_{\mathcal{U}\mathcal{U}^\top} = (nh)^{-1}\sum_i \{\mathcal{G}_i\widehat{\eta}_p\mathcal{U}_i^\top + \mathcal{U}_i\widehat{\eta}_p^\top\mathcal{G}_i^\top + \mathcal{G}_i\widehat{\eta}_p\widehat{\eta}_p^\top\mathcal{G}_i^\top\}$, it follows from Lemma 1 and (a) that $\widehat{\Delta}_{\mathcal{U}\mathcal{U}^\top} - \overline{\Delta}_{\mathcal{U}\mathcal{U}^\top} = O_p\left(\sqrt{\log(n)/\underline{n}}\right)$ and therefore, $\widehat{\Delta}_{\mathcal{U}\mathcal{U}^\top} - \Delta_{\mathcal{U}\mathcal{U}^\top} = O_p\left(\sqrt{\log(n)/\underline{n}} + \log(n)\left(\overline{n}^{1/6}/\underline{n}\right)\right)$. Since $\left\|\sqrt{nh}\widehat{\lambda}_\sharp\right\| \leq \log(n)$ by construction, it follows from these results that $\sqrt{nh}\widehat{\lambda}_\sharp = O_p\left(\sqrt{\log(n)}\right)$. Wpa1, $\forall h \in \mathbb{H}$, $\widehat{\lambda}_\sharp$ is in the interior of $\mathcal{L}_\sharp$ and the first-order condition is satisfied: $\partial S\left(\lambda, \widehat{\vartheta}_p\right)/\partial\lambda\Big|_{\lambda=\widehat{\lambda}_\sharp} = 0_{2d_u}$. It follows from the concavity of $S\left(\cdot, \widehat{\vartheta}_p\right)$ that $\widehat{\lambda}_\sharp$ also attains $\sup_{\lambda \in \mathcal{L}\left(\widehat{\vartheta}_p\right)} S\left(\lambda, \widehat{\vartheta}_p\right)$ $\forall h \in \mathbb{H}$ wpa1. Then (b) follows from setting $\widehat{\lambda}_p = \widehat{\lambda}_\sharp$. (c) and (d) follow from similar arguments. ■

Denote $\mathrm{O} \coloneqq \left(\Delta_{\mathcal{G}}^\top\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1}\Delta_{\mathcal{G}}\right)^{-1}$, $\mathrm{N} \coloneqq \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1}\Delta_{\mathcal{G}}\mathrm{O}$ and $\mathrm{Q} \coloneqq \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1} - \mathrm{N}\Delta_{\mathcal{G}}^\top\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1}$. Let $(\mathrm{O}_\dagger, \mathrm{N}_\dagger, \mathrm{Q}_\dagger)$ 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) $\sqrt{nh}\left(\widehat{\lambda}_p^\top, \widehat{\eta}_p^\top\right) = \overline{\mathcal{U}}^\top\begin{bmatrix} \mathrm{Q} & \mathrm{N} \end{bmatrix} + O_p\left(\upsilon_n^\dagger\right)$; (b) $\sqrt{nh}\left(\widetilde{\lambda}_p^\top, \widetilde{\eta}_p^\top\right) = \overline{\mathcal{U}}^\top\begin{bmatrix} \mathrm{Q}_\dagger & \mathrm{N}_\dagger \end{bmatrix} + O_p\left(\upsilon_n^\dagger\right)$, where $\upsilon_n^\dagger \coloneqq \log(n)/\sqrt{\underline{n}} + \log(n)^{3/2}\left(\overline{n}^{1/6}/\underline{n}\right)$.*

Proof. It is shown in the proof of Lemma 2 that $\widehat{\lambda}_p$ satisfies the first-order condition which can be written as $\sum_i \widehat{\mathcal{U}}_i/\left(1 + \widehat{\lambda}_p^\top\widehat{\mathcal{U}}_i\right) = 0_{2d_u}$ $\forall h \in \mathbb{H}$ wpa1. We also showed that $\max_i \left\|\widehat{\mathcal{U}}_i\right\|/\sqrt{nh} = O_p\left(\overline{n}^{1/12}/\underline{n}^{1/2}\right)$ and $\sqrt{nh}\widehat{\lambda}_p = O_p\left(\sqrt{\log(n)}\right)$. Therefore, $\max_i \left|\widehat{\lambda}_p^\top\widehat{\mathcal{U}}_i\right| = o_p(1)$. By $\widehat{\mathcal{U}}_i = \mathcal{U}_i - \mathcal{G}_i\widehat{\eta}_p$, Lemma 1 and boundedness of Θ, $(nh)^{-1}\sum_i \left\|\widehat{\mathcal{U}}_i\right\|^3 = O_p\left(1 + \log(n)\left(\overline{n}^{1/4}/\underline{n}\right)\right)$ and $(nh)^{-1}\sum_i \left\|\widehat{\mathcal{U}}_i\right\|^4 = O_p\left(1 + \log(n)\left(\overline{n}^{1/3}/\underline{n}\right)\right)$. It follows from these result, Lemma 2 and simple algebra that $(nh)^{-1}\sum_i \widehat{\mathcal{U}}_i\widehat{\mathcal{U}}_i^\top/\left(1 + \widehat{\lambda}_p^\top\widehat{\mathcal{U}}_i\right)^2 = \widehat{\Delta}_{\mathcal{U}\mathcal{U}^\top} + o_p(1)$. It is shown in the proof of Lemma 2 that $\widehat{\Delta}_{\mathcal{U}\mathcal{U}^\top} = \operatorname{diag}\left(\psi_{UU^\top,+}, \psi_{UU^\top,-}\right) + o_p(1)$. Therefore, $\varrho_{\min}\left((nh)^{-1}\sum_i \widehat{\mathcal{U}}_i\widehat{\mathcal{U}}_i^\top/\left(1 + \widehat{\lambda}_p^\top\widehat{\mathcal{U}}_i\right)^2\right)$ 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 $\mathbb{B}\left(\widehat{\vartheta}_p\right)$ of $\widehat{\vartheta}_p$ such that $\widehat{\lambda}_p = \lambda\left(\widehat{\vartheta}_p\right)$ and $(nh)^{-1}\sum_i \mathcal{U}_i(\theta)/\left(1 + \lambda(\theta)^\top\mathcal{U}_i(\theta)\right) = 0_{2d_u}$ $\forall \theta \in \mathbb{B}\left(\widehat{\vartheta}_p\right)$. Since $S(\cdot, \theta)$ is concave, $S(\lambda(\theta), \theta) = \sup_{\lambda \in \mathcal{L}(\theta)} S(\lambda, \theta)$ and $\widehat{\vartheta}_p = \operatorname{argmin}_{\theta \in \mathbb{B}\left(\widehat{\vartheta}_p\right)} S(\lambda(\theta), \theta)$. By the chain rule and $\sum_i \widehat{\mathcal{U}}_i/\left(1 + \widehat{\lambda}_p^\top\widehat{\mathcal{U}}_i\right) = 0_{2d_u}$, the first-order condition for $\widehat{\vartheta}_p$ can be written as $\sum_i \mathcal{G}_i^\top\widehat{\lambda}_p/\left(1 + \widehat{\lambda}_p^\top\widehat{\mathcal{U}}_i\right) = 0_{2d_u}$, which holds $\forall h \in \mathbb{H}$ wpa1. By simple algebra we have

$$0_{2d_u} = \sum_i \left\{\widehat{\mathcal{U}}_i - \widehat{\mathcal{U}}_i\widehat{\mathcal{U}}_i^\top\widehat{\lambda}_p + \frac{\widehat{\mathcal{U}}_i\left(\widehat{\mathcal{U}}_i^\top\widehat{\lambda}_p\right)^2}{1 + \widehat{\lambda}_p^\top\widehat{\mathcal{U}}_i}\right\} \text{ and } 0_{d_\vartheta} = \sum_i \left\{\mathcal{G}_i^\top\widehat{\lambda}_p - \frac{\mathcal{G}_i^\top\widehat{\lambda}_p\left(\widehat{\lambda}_p^\top\widehat{\mathcal{U}}_i\right)}{1 + \widehat{\lambda}_p^\top\widehat{\mathcal{U}}_i}\right\}. \tag{15}$$

By $\max_i \left|\widehat{\lambda}_p^\top\widehat{\mathcal{U}}_i\right| = o_p(1)$, $(nh)^{-1}\sum_i \left\|\widehat{\mathcal{U}}_i\right\| = O_p(1)$, $(nh)^{-1}\sum_i \left\|\widehat{\mathcal{U}}_i\right\|^3 = O_p\left(1 + \log(n)\left(\overline{n}^{1/4}/\underline{n}\right)\right)$ and Lemma

2, $(nh)^{-1/2}\sum_i\left\{\widehat{\mathcal{U}}_i\left(\widehat{\mathcal{U}}_i^\top\widehat{\lambda}_p\right)^2\right\}/\left(1+\widehat{\lambda}_p^\top\widehat{U}_i\right) = O_p\left(\upsilon_n^\dagger\right)$ and $(nh)^{-1/2}\sum_i\left\{\mathcal{G}_i^\top\widehat{\lambda}_p\left(\widehat{\lambda}_p^\top\widehat{\mathcal{U}}_i\right)\right\}/\left(1+\widehat{\lambda}_p^\top\widehat{\mathcal{U}}_i\right) = O_p\left(\upsilon_n^\ddagger\right)$, where $\upsilon_n^\ddagger := \log(n)/\sqrt{\underline{n}} + \log(n)^2\left(\overline{n}^{1/4}/\underline{n}^{3/2}\right)$. By these results and $\widehat{\mathcal{U}}_i = \mathcal{U}_i - \mathcal{G}_i\widehat{\eta}_p$, (15) can be written as $\widehat{\Delta}_{\mathcal{U}\mathcal{U}^\top}\sqrt{nh}\widehat{\lambda}_p + \overline{\Delta}_{\mathcal{G}}\sqrt{nh}\widehat{\eta}_p = \overline{\mathcal{U}} + O_p\left(\upsilon_n^\ddagger\right)$ and $\overline{\Delta}_{\mathcal{G}}^\top\sqrt{nh}\widehat{\lambda}_p = O_p\left(\log(n)/\sqrt{\underline{n}}\right)$. By $\widehat{\Delta}_{\mathcal{U}\mathcal{U}^\top} - \Delta_{\mathcal{U}\mathcal{U}^\top} = O_p\left(\sqrt{\log(n)/\underline{n}} + \log(n)\left(\overline{n}^{1/6}/\underline{n}\right)\right)$, $\overline{\Delta}_{\mathcal{G}} - \Delta_{\mathcal{G}} = O_p\left(\sqrt{\log(n)/\underline{n}}\right)$ and Lemma 2, we have

$$\Delta_{\mathcal{U}\mathcal{U}^\top}\sqrt{nh}\widehat{\lambda}_p + \Delta_{\mathcal{G}}\sqrt{nh}\widehat{\eta}_p = \overline{\mathcal{U}} + O_p\left(\upsilon_n^\dagger\right) \text{ and } \Delta_{\mathcal{G}}^\top\sqrt{nh}\widehat{\lambda}_p = O_p\left(\upsilon_n^\dagger\right). \tag{16}$$

Since it follows from Lemma 1 that $\Delta_{\mathcal{U}\mathcal{U}^\top} = \operatorname{diag}\left(\psi_{UU^\top,+}, \psi_{UU^\top,-}\right) + O\left(\overline{h}\right)$ and $\Delta_{\mathcal{G}} = \begin{bmatrix} \mu_{G,+}^\top & \mu_{G,-}^\top \end{bmatrix}^\top + O\left(\overline{h}^{p+1}\right)$, $\Delta_{\mathcal{U}\mathcal{U}^\top}$ 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

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

and (16). (b) follows from similar arguments. ■

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

$$\begin{aligned}\Sigma_\Delta := \left(\mathrm{e}_{d_u,1}^\top\Phi_\pm^{-1}\mathrm{e}_{d_u,1}\right)^{-1} &= \left(\Delta_{\mathcal{M}_+^2}/\Delta_+^2 + \Delta_{\mathcal{M}_-^2}/\Delta_-^2\right) - \left(\Delta_{\mathcal{M}_+\mathcal{Z}_+^\top}/\Delta_+^2 + \Delta_{\mathcal{M}_-\mathcal{Z}_-^\top}/\Delta_-^2\right)\gamma_\Delta \\ &= \Delta_{\left(\mathcal{M}_+ - \mathcal{Z}_+^\top\gamma_\Delta\right)^2}/\Delta_+^2 + \Delta_{\left(\mathcal{M}_- - \mathcal{Z}_-^\top\gamma_\Delta\right)^2}/\Delta_-^2, \end{aligned} \tag{17}$$

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

$$\mathrm{e}_{d_u,1}^\top\Phi_\pm^{-1}\left(\overline{\mathcal{U}}_+/\Delta_+ - \overline{\mathcal{U}}_-/\Delta_-\right) = \left\{\left(\overline{\mathcal{M}}_+/\Delta_+ - \overline{\mathcal{M}}_-/\Delta_-\right) - \left(\overline{\mathcal{Z}}_+/\Delta_+ - \overline{\mathcal{Z}}_-/\Delta_-\right)^\top\gamma_\Delta\right\}/\Sigma_\Delta. \tag{18}$$

By simple algebra, $\left(\Phi_{00} - \Phi_{0\dagger}\Phi_{\dagger\dagger}^{-1}\Phi_{\dagger 0}\right)^{-1} = \Sigma_\Delta/\left(\Delta_{\mathcal{D}_+}/\Delta_+ - \Delta_{\mathcal{D}_-}/\Delta_-\right)^2$. 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,

$$\left(\Delta_{\mathcal{G}}^\top\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1}\Delta_{\mathcal{G}}\right)^{-1} = \left(\Delta_{\mathcal{D}_+}/\Delta_+ - \Delta_{\mathcal{D}_-}/\Delta_-\right)^{-2}\begin{bmatrix} \Sigma_\Delta & -\Sigma_\Delta\Phi_{0\dagger}\Phi_{\dagger\dagger}^{-1} \\ -\Phi_{\dagger\dagger}^{-1}\Phi_{\dagger 0}\Sigma_\Delta & \Phi_{\dagger\dagger}^{-1} + \Phi_{\dagger\dagger}^{-1}\Phi_{\dagger 0}\Sigma_\Delta\Phi_{0\dagger}\Phi_{\dagger\dagger}^{-1} \end{bmatrix}. \tag{19}$$

By straightforward algebraic calculation,

$$\left(\Delta_{\mathcal{G}_0}^\top - \Phi_{0\dagger}\Phi_{\dagger\dagger}^{-1}\Delta_{\mathcal{G}_\dagger}^\top\right)\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1}\overline{\mathcal{U}} = \left(\Delta_{\mathcal{D}_+}/\Delta_+ - \Delta_{\mathcal{D}_-}/\Delta_-\right)\mathrm{e}_{d_u,1}^\top\Phi_\pm^{-1}\left(\overline{\mathcal{U}}_+/\Delta_+ - \overline{\mathcal{U}}_-/\Delta_-\right). \tag{20}$$

Then, by this result, (18) and (19),

$$\begin{aligned}\mathrm{e}_{d_\vartheta,1}^\top\left(\Delta_{\mathcal{G}}^\top\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1}\Delta_{\mathcal{G}}\right)^{-1}\left(\Delta_{\mathcal{G}}^\top\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1}\overline{\mathcal{U}}\right) &= \left(\Delta_{\mathcal{D}_+}/\Delta_+ - \Delta_{\mathcal{D}_-}/\Delta_-\right)^{-2}\Sigma_\Delta\left(\Delta_{\mathcal{G}_0}^\top - \Phi_{0\dagger}\Phi_{\dagger\dagger}^{-1}\Delta_{\mathcal{G}_\dagger}^\top\right)\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1}\overline{\mathcal{U}} \\ &= \frac{\left(\overline{\mathcal{M}}_+/\Delta_+ - \overline{\mathcal{M}}_-/\Delta_-\right) - \left(\overline{\mathcal{Z}}_+/\Delta_+ - \overline{\mathcal{Z}}_-/\Delta_-\right)^\top\gamma_\Delta}{\Delta_{\mathcal{D}_+}/\Delta_+ - \Delta_{\mathcal{D}_-}/\Delta_-}.\end{aligned} \tag{21}$$

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

$$\begin{aligned}\sqrt{nh}\left(\widehat{\vartheta}_{p,0} - \vartheta_0\right) &= \mathrm{e}_{d_\vartheta,1}^\top\left(\Delta_{\mathcal{G}}^\top\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1}\Delta_{\mathcal{G}}\right)^{-1}\left(\Delta_{\mathcal{G}}^\top\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1}\overline{\mathcal{U}}\right) + o_p(1) \\ &= \left\{\left(\overline{\mathcal{M}}_+ - \overline{\mathcal{Z}}_+\gamma_{\mathsf{adj}}\right) - \left(\overline{\mathcal{M}}_- - \overline{\mathcal{Z}}_-\gamma_{\mathsf{adj}}\right)\right\}/\left(\psi_{D,+} - \psi_{D,-}\right) + o_p(1) \\ &= (nh)^{-1/2}\sum_i\left(W_{p;+,i} - W_{p;-,i}\right)\left(\epsilon_i - \mu_\epsilon\right)/\left(\psi_{D,+} - \psi_{D,-}\right) + o_p(1),\end{aligned} \tag{22}$$

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

$$\sqrt{nh}\left(\widehat{\vartheta}_{p,0} - \vartheta_0 - \frac{\Delta_\mathcal{E}}{\psi_{D,+} - \psi_{D,-}}\right) = \sum_i\left(\frac{\mathcal{E}_i}{\sqrt{nh}} - \mathrm{E}\left[\frac{\mathcal{E}}{\sqrt{nh}}\right]\right)/\left(\psi_{D,+} - \psi_{D,-}\right) + o_p(1) \tag{23}$$

follows from subtracting both sides of (22) by $\sqrt{nh}\Delta_\mathcal{E}/\left(\psi_{D,+} - \psi_{D,-}\right)$. By Lemma 1 with $\underline{h} = \overline{h} = h$, $\Delta_\mathcal{E} = \left(\zeta_{p;+} - \zeta_{p;-}\right)h^{p+1}/(p+1)! + O\left(h^{p+2}\right)$ and $\Delta_{\mathcal{E}^2} = \omega_{p;+}^{0,2}\sum_{\mathrm{s}\in\{-,+\}}\psi_{(\epsilon-\mu_\epsilon)^2,\mathrm{s}} + O(h)$. It follows from simple algebraic calculations that $\sum_{\mathrm{s}\in\{-,+\}}\psi_{(\epsilon-\mu_\epsilon)^2,\mathrm{s}} = \sigma_{\mathsf{adj}}^2\varphi$. Then, $\mathrm{Var}\left[\mathcal{E}/\sqrt{h}\right] = \Delta_{\mathcal{E}^2} - h\Delta_\mathcal{E}^2 = \omega_{p;+}^{0,2}\sigma_{\mathsf{adj}}^2\varphi + O(h)$. By LIE and change of variables, $\Delta_{\mathcal{E}^4} = O(1)$. Then,

$$\sum_i\mathrm{E}\left[\left(\frac{\mathcal{E}_i/\sqrt{nh} - \mathrm{E}\left[\mathcal{E}/\sqrt{nh}\right]}{\sqrt{\mathrm{Var}\left[\mathcal{E}/\sqrt{h}\right]}}\right)^4\right] \lesssim \frac{\Delta_{\mathcal{E}^4} + h^3\Delta_\mathcal{E}^4}{(nh)\left(\mathrm{Var}\left[\mathcal{E}/\sqrt{h}\right]\right)^2} = O\left((nh)^{-1}\right), \tag{24}$$

where the inequality follows from Loève's c_r inequality and the equality follows from $\mathrm{Var}\left[\mathcal{E}/\sqrt{h}\right] = \omega_{p;+}^{0,2}\sigma_{\mathsf{adj}}^2\varphi + O(h)$, $\Delta_{\mathcal{E}^4} = O(1)$ and $\Delta_\mathcal{E} = O\left(h^{p+1}\right)$. (24) verifies Lyapunov's condition. By Lyapunov's central limit theorem, $\sum_i\left(\mathcal{E}_i/\sqrt{nh} - \mathrm{E}\left[\mathcal{E}/\sqrt{nh}\right]\right)/\sqrt{\mathrm{Var}\left[\mathcal{E}/\sqrt{h}\right]} \to_d \mathrm{N}(0,1)$. The conclusion follows from this result, (23), $\mathrm{Var}\left[\mathcal{E}/\sqrt{h}\right] = \omega_{p;+}^{0,2}\sigma_{\mathsf{adj}}^2\varphi + O(h)$ and Slutsky's lemma. ■

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

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

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

$$\begin{aligned}
\mathrm{Q}_\dagger - \mathrm{Q} &= \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1}\left\{\Delta_{\mathcal{G}}\left(\Delta_{\mathcal{G}}^\top\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1}\Delta_{\mathcal{G}}\right)^{-1}\Delta_{\mathcal{G}}^\top - \Delta_{\mathcal{G}_\dagger}\left(\Delta_{\mathcal{G}_\dagger}^\top\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1}\Delta_{\mathcal{G}_\dagger}\right)^{-1}\Delta_{\mathcal{G}_\dagger}^\top\right\}\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1} \\
&= \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1}\left(\Delta_{\mathcal{G}_0} - \Delta_{\mathcal{G}_\dagger}\Phi_{\dagger\dagger}^{-1}\Phi_{\dagger 0}\right)\left(\Phi_{00} - \Phi_{0\dagger}\Phi_{\dagger\dagger}^{-1}\Phi_{\dagger 0}\right)^{-1}\left(\Delta_{\mathcal{G}_0}^\top - \Phi_{0\dagger}\Phi_{\dagger\dagger}^{-1}\Delta_{\mathcal{G}_\dagger}^\top\right)\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1}. \qquad (25)
\end{aligned}$$

Then by this result, (18), (20) and $\left(\Phi_{00} - \Phi_{0\dagger}\Phi_{\dagger\dagger}^{-1}\Phi_{\dagger 0}\right)^{-1} = \Sigma_\Delta / \left(\Delta_{\mathcal{D}_+}/\Delta_+ - \Delta_{\mathcal{D}_-}/\Delta_-\right)^2$

$$\begin{aligned}
\overline{\mathcal{U}}^\top(\mathrm{Q}_\dagger - \mathrm{Q})\overline{\mathcal{U}} &= \left\{\mathrm{e}_{d_u,1}^\top\Phi_\pm^{-1}\left(\overline{\mathcal{U}}_+/\Delta_+ - \overline{\mathcal{U}}_-/\Delta_-\right)\right\}^2\Sigma_\Delta \\
&= \left\{\left(\overline{\mathcal{M}}_+/\Delta_+ - \overline{\mathcal{M}}_-/\Delta_-\right) - \left(\overline{\mathcal{Z}}_+/\Delta_+ - \overline{\mathcal{Z}}_-/\Delta_-\right)^\top\gamma_\Delta\right\}^2/\Sigma_\Delta. \qquad (26)
\end{aligned}$$

By using $\gamma_\Delta = \gamma_{\mathsf{adj}} + O\left(\overline{h}\right)$ and (17), $\Sigma_\Delta = \Delta_{\mathcal{E}^2}/\varphi^2 + O\left(\overline{h}\right)$. By $\left\|\overline{\mathcal{U}}\right\| = O_p\left(\sqrt{\log(n)}\right)$ and $\gamma_\Delta = \gamma_{\mathsf{adj}} + O\left(\overline{h}\right)$, the numerator on the right hand side of the second equality in (26) is $\left\{(nh)^{-1/2}\sum_i\mathcal{E}_i\right\}^2 + O_p\left(\log(n)\overline{h}\right)$. Let $\tilde{q}(T_i, X_i \mid h) \coloneqq h^{-1/2}\mathcal{E}_i/\sqrt{\Delta_{\mathcal{E}^2}}$ and $\tilde{\mathfrak{Q}} \coloneqq \{\tilde{q}(\cdot \mid h) : h \in \mathbb{H}\}$. Then it is clear that $\left\{(nh)^{-1/2}\sum_i\mathcal{E}_i\right\}^2/\Delta_{\mathcal{E}^2} = \left\{\mathbb{G}_n^T\tilde{q}(\cdot \mid h)\right\}^2$ and therefore, $LR_p(\vartheta_0 \mid h) = \left\{\mathbb{G}_n^T\tilde{q}(\cdot \mid h)\right\}^2 + O_p\left(\log(n)\overline{h} + \sqrt{\log(n)}\upsilon_n^\dagger\right)$. Also denote $\mathfrak{Q} \coloneqq \{q(\cdot \mid h) : h \in \mathbb{H}\}$ and $\mathfrak{D} \coloneqq \{q(\cdot \mid h) - \tilde{q}(\cdot \mid h) : h \in \mathbb{H}\}$. 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 $\left(F_{\tilde{\mathfrak{Q}}}, F_{\mathfrak{Q}}\right)$ satisfying $F_{\tilde{\mathfrak{Q}}}(T_i, X_i) \propto$

$\underline{h}^{-1/2}1\left(|X_i|\leq\overline{h}\right)|\epsilon_i-\mu_\epsilon|/\sqrt{\inf_{h\in\mathbb{H}}\Delta_{\mathcal{E}^2}}$ and $F_{\mathfrak{Q}}\left(T_i,X_i\right)\propto\underline{h}^{-1/2}1\left(|X_i|\leq\overline{h}\right)|\epsilon_i-\mu_\epsilon|/\sqrt{\xi\left(|X_i|\right)f_{|X|}\left(|X_i|\right)}$, respectively. By change of variables, $\mathbb{P}^TF_{\tilde{\mathfrak{Q}}}^{12}\asymp\mathbb{P}^TF_{\mathfrak{Q}}^{12}=O\left(\overline{h}/\underline{h}^6\right)$. By Chernozhukov et al. (2014b, Lemma A.6), $\mathfrak{D}$ is VC-type with respect to the envelope $F_{\mathfrak{D}}=F_{\tilde{\mathfrak{Q}}}+F_{\mathfrak{Q}}$. Let $\sigma_{\mathfrak{D}}^2\coloneqq\sup_{f\in\mathfrak{D}}\mathbb{P}^Tf^2=\sup_{h\in\mathbb{H}}\mathrm{E}\left[\left(q\left(T,X\mid h\right)-\tilde{q}\left(T,X\mid h\right)\right)^2\right]$. By LIE and the fact that $\left(W_{p;+}+W_{p;-}\right)^2=\mathcal{K}_{p;+}\left(|X|/h\right)$,

$$\begin{aligned}\mathrm{E}\left[\left(q\left(T,X\mid h\right)-\tilde{q}\left(T,X\mid h\right)\right)^2\right]&=\mathrm{E}\left[\frac{1}{h}\left(W_{p;+}+W_{p;-}\right)^2\left(\epsilon-\mu_\epsilon\right)^2\left(\frac{1}{\sqrt{\Delta_{\mathcal{E}^2}}}-\frac{1}{\sqrt{\xi\left(|X|\right)f_{|X|}\left(|X|\right)\omega_{p;+}^{0,2}}}\right)^2\right]\\&=\int_0^\infty\frac{1}{h}\mathcal{K}_{p;+}\left(\frac{z}{h}\right)^2\left(\sqrt{\frac{\xi\left(z\right)f_{|X|}\left(z\right)}{\Delta_{\mathcal{E}^2}}}-\frac{1}{\sqrt{\omega_{p;+}^{0,2}}}\right)^2\mathrm{d}z.\end{aligned}\tag{27}$$

Note that $\Delta_{\mathcal{E}^2}=\int_0^\infty h^{-1}\mathcal{K}_{p;+}\left(z/h\right)^2\xi\left(z\right)f_{|X|}\left(z\right)\mathrm{d}z$ and therefore, it follows from mean value expansion and (27) that $\sigma_{\mathfrak{D}}^2=O\left(\overline{h}^2\right)$. By Chen and Kato (2020, Corollary 5.5), $\mathrm{E}\left[\left\|\mathbb{G}_n^T\right\|_{\mathfrak{D}}\right]\lesssim\sigma_{\mathfrak{D}}\sqrt{\log\left(n\right)}+\log\left(n\right)\|F_{\mathfrak{D}}\|_{\mathbb{P}^T,12}\,n^{1/12}/\sqrt{n}$ and therefore, $\mathrm{E}\left[\left\|\mathbb{G}_n^T\right\|_{\mathfrak{D}}\right]=O\left(\sqrt{\log\left(n\right)}\cdot\overline{h}+\log\left(n\right)\left(\overline{n}^{1/12}/\underline{n}^{1/2}\right)\right)$. Let $\sigma_{\tilde{\mathfrak{Q}}}^2\coloneqq\sup_{f\in\tilde{\mathfrak{Q}}}\mathbb{P}^Tf^2$ and $\sigma_{\mathfrak{Q}}^2\coloneqq\sup_{f\in\mathfrak{Q}}\mathbb{P}^Tf^2$. It is easy to see that $\mathbb{P}^Tf^2=1$, if $f\in\mathfrak{Q}$ or $f\in\tilde{\mathfrak{Q}}$ and therefore, $\sigma_{\tilde{\mathfrak{Q}}}^2=\sigma_{\mathfrak{Q}}^2=1$. Similarly, $\mathrm{E}\left[\left\|\mathbb{G}_n^T\right\|_{\tilde{\mathfrak{Q}}}\right]\lesssim\sigma_{\tilde{\mathfrak{Q}}}\sqrt{\log\left(n\right)}+\log\left(n\right)\left\|F_{\tilde{\mathfrak{Q}}}\right\|_{\mathbb{P}^T,12}n^{1/12}/\sqrt{n}$ and a similar inequality with $\tilde{\mathfrak{Q}}$ replaced by $\mathfrak{Q}$ holds. Therefore, $\mathrm{E}\left[\left\|\mathbb{G}_n^T\right\|_{\tilde{\mathfrak{Q}}}\right]\asymp\mathrm{E}\left[\left\|\mathbb{G}_n^T\right\|_{\mathfrak{Q}}\right]=O\left(\sqrt{\log\left(n\right)}\right)$. Then it follows from Markov's inequality that $\left\{\mathbb{G}_n^T\tilde{q}\left(\cdot\mid h\right)\right\}^2-\left\{\mathbb{G}_n^Tq\left(\cdot\mid h\right)\right\}^2=O_p\left(\log\left(n\right)\overline{h}+\log\left(n\right)^{3/2}\left(\overline{n}^{1/12}/\underline{n}^{1/2}\right)\right)$. The conclusion follows from this result and $LR_p\left(\vartheta_0\mid h\right)=\left\{\mathbb{G}_n^T\tilde{q}\left(\cdot\mid h\right)\right\}^2+O_p\left(\log\left(n\right)\overline{h}+\sqrt{\log\left(n\right)}\upsilon_n^\dagger\right)$. ■

Proof of Theorem 2. Denote $Z_{\mathfrak{Q}_\pm}\coloneqq\sup_{f\in\mathfrak{Q}_\pm}\mathbb{G}_n^Tf=\left\|\mathbb{G}_n^T\right\|_{\mathfrak{Q}}$. Since $F_{\mathfrak{Q}}$ is also an envelope of $\mathfrak{Q}_\pm\coloneqq\mathfrak{Q}\cup(-\mathfrak{Q})$ $(-\mathfrak{Q}\coloneqq\{-f:f\in\mathfrak{Q}\})$ and the covering number of $\mathfrak{Q}_\pm$ is at most twice that of $\mathfrak{Q}$, $\mathfrak{Q}_\pm$ is also VC-type with respect to $F_{\mathfrak{Q}}$. 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 $\left\{G^T\left(f\right):f\in\mathfrak{Q}_\pm\right\}$ that is a tight random element in $\ell^\infty\left(\mathfrak{Q}_\pm\right)$ and also satisfies $\mathrm{E}\left[G^T\left(f\right)G^T\left(g\right)\right]=\mathrm{Cov}\left[f\left(T,X\right),g\left(T,X\right)\right]$, $\forall f,g\in\mathfrak{Q}_\pm$.[23] By Giné and Nickl (2015, Theorem 3.7.28), almost surely the sample paths $\mathfrak{Q}_\pm\ni f\mapsto G^T\left(f\right)$ are prelinear and therefore, almost surely, $\forall f\in\mathfrak{Q}$, $G^T\left(f\right)+G^T\left(-f\right)=0$, and $\sup_{f\in\mathfrak{Q}_\pm}G^T\left(f\right)=\left\|G^T\right\|_{\mathfrak{Q}}$. Let $\bar{\Gamma}_G\left(h\right)\coloneqq G^T\left(q\left(\cdot\mid h\right)\right)$ and therefore, the zero-mean Gaussian process $\left\{\bar{\Gamma}_G\left(h\right):h\in\mathbb{H}\right\}$ is a tight random element in $\ell^\infty\left(\mathbb{H}\right)$ and has the covariance structure $\mathrm{E}\left[\bar{\Gamma}_G\left(h\right)\bar{\Gamma}_G\left(h'\right)\right]=\mathrm{Cov}\left[q\left(T,X\mid h\right),q\left(T,X\mid h'\right)\right]$, $\forall\left(h,h'\right)\in\mathbb{H}^2$. By definition, $\left\|\bar{\Gamma}_G\right\|_{\mathbb{H}}=\left\|G^T\right\|_{\mathfrak{Q}}$. By change of variables and LIE, $\sup_{f\in\mathfrak{Q}}\mathbb{P}^T\left|f\right|^3=\sup_{h\in\mathbb{H}}\mathrm{E}\left[\left|q\left(T,X\mid h\right)\right|^3\right]\lesssim\underline{h}^{-1/2}$ and similarly $\sup_{f\in\mathfrak{Q}}\mathbb{P}^T\left|f\right|^4\lesssim\underline{h}^{-1}$. Also, $\mathbb{P}^TF_{\mathfrak{Q}}^{12}\lesssim\overline{h}/\underline{h}^6$. By Chernozhukov et al. (2016, Theorem 2.1) with $B\left(f\right)=0$, $\mathcal{F}=\mathfrak{Q}_\pm$, $q=12$, $K_n=\log\left(n\right)$, $\sigma=1$,

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

$b \lesssim \underline{h}^{-1/2}$ and $\gamma = \log(n)^{-1}$, there exists $\widetilde{Z}_{\mathfrak{Q}_\pm} =_d \sup_{f\in\mathfrak{Q}_\pm} G^T(f) = \left\|G^T\right\|_{\mathfrak{Q}}$ which satisfies $Z_{\mathfrak{Q}_\pm} - \widetilde{Z}_{\mathfrak{Q}_\pm} = O_p(v_n^*)$, 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),

$$\begin{aligned}\mathrm{E}\left[\left\|G^T\right\|_{\mathfrak{Q}}\right] &\lesssim \int_0^{\sigma_{\mathfrak{Q}}\vee n^{-1/2}\|F_{\mathfrak{Q}}\|_{\mathbb{P}^T,2}} \sqrt{1+\log\left(N\left(\varepsilon,\mathfrak{Q},\|\cdot\|_{\mathbb{P}^T,2}\right)\right)}\mathrm{d}\varepsilon \\ &\lesssim \left(\sigma_{\mathfrak{Q}} \vee n^{-1/2}\|F_{\mathfrak{Q}}\|_{\mathbb{P}^T,2}\right)\sqrt{\log(n)} = O\left(\sqrt{\log(n)}\right). \quad (28)\end{aligned}$$

By Lemma 4, $\sup_{h\in\mathbb{H}} LR_p(\vartheta_0 \mid h) = \left\|\mathbb{G}_n^T\right\|_{\mathfrak{Q}}^2 + O_p\left(\log(n)\overline{h} + \log(n)^{3/2}\left(\overline{n}^{1/12}/\underline{n}^{1/2}\right)\right)$. By (28) and the fact that $\mathrm{E}\left[\left\|\mathbb{G}_n^T\right\|_{\mathfrak{Q}}\right] = O\left(\sqrt{\log(n)}\right)$, we have $Z_{\mathfrak{Q}_\pm}^2 - \widetilde{Z}_{\mathfrak{Q}_\pm}^2 = O_p\left(\sqrt{\log(n)}v_n^*\right)$. Therefore, $\sup_{h\in\mathbb{H}} LR_p(\vartheta_0 \mid h) = \widetilde{Z}_{\mathfrak{Q}_\pm}^2 + O_p\left(\sqrt{\log(n)}v_n^* + \log(n)\overline{h}\right)$. By Dudley (2002, Theorem 9.2.2) and $\sup_{h\in\mathbb{H}} LR_p(\vartheta_0 \mid h) - \widetilde{Z}_{\mathfrak{Q}_\pm}^2 = o_p\left(\log(n)^{-1}\right)$, there exists a null sequence $\varepsilon_n \downarrow 0$ such that $\Pr\left[\left|\sup_{h\in\mathbb{H}} LR_p(\vartheta_0 \mid h) - \widetilde{Z}_{\mathfrak{Q}_\pm}^2\right| > \varepsilon_n/\log(n)\right] \leq \varepsilon_n$ and by the fact that $(a-b)^2 \leq \left|a^2-b^2\right|\ \forall a,b\geq 0$,

$$\Pr\left[\left|\sqrt{\sup_{h\in\mathbb{H}} LR_p(\vartheta_0 \mid h)} - \widetilde{Z}_{\mathfrak{Q}_\pm}\right| > \sqrt{\varepsilon_n/\log(n)}\right] \leq \varepsilon_n. \quad (29)$$

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

$$|\Pr[V\leq t] - \Pr[W\leq t]| \leq \Pr[|W-t|\leq r_1] + r_2. \quad (30)$$

Then, by (29) and (30),

$$\begin{aligned}&\left|\Pr\left[\sup_{h\in\mathbb{H}} LR_p(\vartheta_0 \mid h) \leq z_{1-\tau}\left(\overline{h}/\underline{h}\right)^2\right] - \Pr\left[\widetilde{Z}_{\mathfrak{Q}_\pm}^2 \leq z_{1-\tau}\left(\overline{h}/\underline{h}\right)^2\right]\right| \\ &\qquad \leq \Pr\left[\left|\widetilde{Z}_{\mathfrak{Q}_\pm} - z_{1-\tau}\left(\overline{h}/\underline{h}\right)\right| \leq \sqrt{\varepsilon_n/\log(n)}\right] + \varepsilon_n. \quad (31)\end{aligned}$$

Since $\widetilde{Z}_{\mathfrak{Q}_\pm} =_d \left\|G^T\right\|_{\mathfrak{Q}}$ and $\left\{G^T(f) : f\in\mathfrak{Q}\right\}$ is a centered Gaussian process with $\mathrm{E}\left[G^T(f)^2\right] = 1$, $\forall f$, by using the Gaussian anti-concentration inequality (Chernozhukov et al., 2014a, Corollary 2.1) and (28),

$$\Pr\left[\left|\widetilde{Z}_{\mathfrak{Q}_\pm} - z_{1-\tau}\left(\overline{h}/\underline{h}\right)\right| \leq \sqrt{\varepsilon_n/\log(n)}\right] \lesssim \sqrt{\varepsilon_n/\log(n)}\left(\mathrm{E}\left[\left\|G^T\right\|_{\mathfrak{Q}}\right] + 1\right) = O\left(\sqrt{\varepsilon_n}\right). \quad (32)$$

It then follows from (31) and (32) that $\Pr\left[LR_p(\vartheta_0 \mid h) \leq z_{1-\tau}\left(\overline{h}/\underline{h}\right)^2, \forall h\in\mathbb{H}\right] = \Pr\left[\left\|\bar{\Gamma}_G\right\|_{\mathbb{H}} \leq z_{1-\tau}\left(\overline{h}/\underline{h}\right)\right] + o(1)$. Let N be an $\mathrm{N}(0,1)$ random variable that is independent of $\left\{\bar{\Gamma}_G(h) : h\in\mathbb{H}\right\}$. Let $\tilde{\Gamma}_G(h) :=$

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

Appendix B Proofs of Theorems 3 and 4

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

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

Proof. Let $r_n \coloneqq \sqrt{\bar{n}/\log(n)}$, $\overline{V}_i \coloneqq V_i 1(V_i > r_n)$, $\underline{V}_i \coloneqq V_i 1(V_i \leq r_n)$ and $\left(\underline{V},\overline{V}\right)$ be defined similarly. Then write $\bar{n}^{-1/2}\sum_i\left(W_{p;\mathrm{s},i}^k V_i - \mathrm{E}\left[W_{p;\mathrm{s}}^k V\right]\right) = \overline{\mathcal{W}} + \underline{\mathcal{W}}$, where $\overline{\mathcal{W}} \coloneqq \bar{n}^{-1/2}\sum_i\left(W_{p;\mathrm{s},i}^k \overline{V}_i - \mathrm{E}\left[W_{p;\mathrm{s}}^k \overline{V}\right]\right)$ and $\underline{\mathcal{W}} \coloneqq \bar{n}^{-1/2}\sum_i\left(W_{p;\mathrm{s},i}^k \underline{V}_i - \mathrm{E}\left[W_{p;\mathrm{s}}^k \underline{V}\right]\right)$. Let $\sigma_{\underline{\mathcal{W}}}^2 \coloneqq \mathrm{Var}\left[h^{-1/2}W_{p;\mathrm{s}}^k \underline{V}\right]$. By $\sigma_{\underline{\mathcal{W}}}^2 \leq \mathrm{E}\left[h^{-1}W_{p;\mathrm{s}}^{2k}\underline{V}^2\right]$, LIE and change of variables, $\sigma_{\underline{\mathcal{W}}}^2 = O(1)$. $\left|W_{p;\mathrm{s},i}^k \underline{V}_i - \mathrm{E}\left[W_{p;\mathrm{s}}^k \underline{V}\right]\right|$ 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, $\underline{\mathcal{W}} = O_p^\star\left(\sqrt{\log(n)}\right)$. By Markov's inequality, the fact that $\overline{V}^2 \leq \overline{V}^2|V/r_n|^3$ and change of variables, $\Pr\left[|\overline{\mathcal{W}}| \geq \sqrt{\log(n)}\right] \leq \mathrm{E}\left[h^{-1}W_{p;\mathrm{s}}^{2k}\overline{V}^2\right]/\log(n) \leq \mathrm{E}\left[h^{-1}W_{p;\mathrm{s}}^{2k}|V|^5\right]/\left(r_n^3\cdot\log(n)\right) = O\left(\log(n)/\bar{n}^{3/2}\right)$ and therefore, $\overline{\mathcal{W}} = O_p^\star\left(\sqrt{\log(n)}\right)$. ■

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)* $\sqrt{\bar{n}}\widehat{\eta}_p = O_p^\star\left(\sqrt{\log(n)}\right)$*; (b)* $\widehat{\lambda}_p := \operatorname{argmax}_{\lambda\in\mathcal{L}(\widehat{\vartheta}_p)} S\left(\lambda, \widehat{\vartheta}_p\right)$ *exists wp** *and* $\sqrt{\bar{n}}\widehat{\lambda}_p = O_p^\star\left(\sqrt{\log(n)}\right)$*; (c)* $\sqrt{\bar{n}}\widetilde{\eta}_p = O_p^\star\left(\sqrt{\log(n)}\right)$*; (d)* $\widetilde{\lambda}_p := \operatorname{argmax}_{\lambda\in\mathcal{L}(\vartheta_0,\widetilde{\vartheta}_p)} S\left(\lambda, \vartheta_0, \widetilde{\vartheta}_p\right)$ *exists wp** *and* $\sqrt{\bar{n}}\widetilde{\lambda}_p = O_p^\star\left(\sqrt{\log(n)}\right)$.

Proof. 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 $3n^{-2}\left(\mathrm{E}\left[h^{-2}\|\mathcal{U}\|^{10}\right]\right)^2 + n^{-3}\mathrm{E}\left[h^{-4}\|\mathcal{U}\|^{20}\right] = O\left(\bar{n}^{-2}\right)$. Therefore, $\bar{n}^{-1}\sum_i \|\mathcal{U}_i\|^5 = O_p^\star(1)$ and by $\max_i \|\mathcal{U}_i\| \leq \left(\sum_i \|\mathcal{U}_i\|^5\right)^{1/5}$, $\max_i \|\mathcal{U}_i\| = O_p^\star\left(\bar{n}^{1/5}\right)$. Then, by this result and the definition of $\mathcal{L}_\sharp$, $\Pr\left[\max_i \sup_{\lambda\in\mathcal{L}_\sharp} \left|\lambda^\top \mathcal{U}_i\right| \geq 1/2\right]$ is bounded above by $\Pr\left[\max_i \|\mathcal{U}_i\| \geq \left(\sqrt{\bar{n}}/\log(n)\right)/2\right] = O\left(\bar{n}^{-2}\right)$. Therefore, $\mathcal{L}_\sharp \subseteq \mathcal{L}(\vartheta)$ wp* and $\lambda_\sharp := \operatorname{argmax}_{\lambda\in\mathcal{L}_\sharp} S(\lambda,\vartheta)$ exists wp*. By using $\overline{\mathcal{U}} = O_p^\star\left(\sqrt{\log(n)}\right)$ and $\overline{\Delta}_{\mathcal{U}\mathcal{U}^\top} - \Delta_{\mathcal{U}\mathcal{U}^\top} = O_p^\star\left(\sqrt{\log(n)/\bar{n}}\right)$, which follow from Lemma 5, and repeating the steps in the proof of Lemma 2, $\sqrt{\bar{n}}\lambda_\sharp = O_p^\star\left(\sqrt{\log(n)}\right)$. Then, $\sqrt{\bar{n}}\lambda_\sharp \leq \log(n)/2$ wp* and $S(\lambda_\sharp,\vartheta) = \sup_{\lambda\in\mathcal{L}(\vartheta)} S(\lambda,\vartheta) = O_p^\star(\log(n))$. By similar arguments, boundedness of Θ and $\widehat{\mathcal{U}}_i = \mathcal{U}_i - \mathcal{G}_i\widehat{\eta}_p$, $\max_i \left\|\widehat{\mathcal{U}}_i\right\| = O_p^\star\left(\bar{n}^{1/5}\right)$ and $\bar{n}^{-1}\sum_i \left\|\widehat{\mathcal{U}}_i\right\|^2 = O_p^\star(1)$. By repeating the steps in the proof of Lemma 2, $\sqrt{\log(n)}\left\|\widehat{\mathcal{U}}\right\| \leq \sup_{\lambda\in\mathcal{L}(\vartheta)} S(\lambda,\vartheta) + 2\left(\bar{n}^{-1}\sum_i \left\|\widehat{\mathcal{U}}_i\right\|^2\right)\log(n) = O_p^\star(\log(n))$. (a) follows from (14), $\overline{\mathcal{U}} = O_p^\star\left(\sqrt{\log(n)}\right)$, $\widehat{\mathcal{U}} = O_p^\star\left(\sqrt{\log(n)}\right)$ and the fact that $\varrho_{\min}\left(\overline{\Delta}_{\mathcal{G}}^\top\overline{\Delta}_{\mathcal{G}}\right)$ 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 $\widehat{\Delta}_{\mathcal{U}\mathcal{U}^\top} - \Delta_{\mathcal{U}\mathcal{U}^\top} = O_p^\star\left(\sqrt{\log(n)/\bar{n}}\right)$. (c) and (d) follow from similar arguments. ■

Consider the singular value decomposition of $\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1/2}(-\Delta_{\mathcal{G}})$: $\mathrm{S}^\top\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1/2}(-\Delta_{\mathcal{G}})\,\mathrm{T} = \begin{bmatrix} \Lambda & 0_{d_\vartheta\times d_z}\end{bmatrix}^\top$, where $\mathrm{S}^\top\mathrm{S} = \mathrm{I}_{2d_u}$, $\mathrm{T}^\top\mathrm{T} = \mathrm{I}_{d_\vartheta}$ 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 $\Gamma := \mathrm{S}^\top\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1/2}$ 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 $\widehat{\mathcal{V}}_i := \Gamma\widehat{\mathcal{U}}_i$. Denote $\overline{\Delta}_{\mathcal{V}\mathcal{V}^\top} := \bar{n}^{-1}\sum_i \mathcal{V}_i\mathcal{V}_i^\top$ and $\overline{\Delta}_{\mathcal{H}} := \bar{n}^{-1}\sum_i \mathcal{H}_i$. Note that the EL criterion function is invariant to such a rotation, i.e., $\ell_p(\theta \mid h) = \sup_\lambda 2\sum_i \log\left(1+\lambda^\top\mathcal{V}_i(\theta)\right)$. For notational simplicity, we still use $\widehat{\lambda}_p$ and $\widetilde{\lambda}_p$ to denote the Lagrange multipliers. Clearly, (b) and (d) of Lemma 6 still hold. Let $\Pi := \Lambda\mathrm{T}^\top$ and $\Omega := \Pi^{-1}$. Then, $\Delta_{\mathcal{V}\mathcal{V}^\top} = \mathrm{I}_{2d_u}$, $\Delta_{\mathcal{H}} := \Gamma(-\Delta_{\mathcal{G}}) = \begin{bmatrix}\Pi^\top & 0_{d_\vartheta\times d_z}\end{bmatrix}^\top$ and $\Delta_{\mathcal{H}_\dagger} := \Gamma\left(-\Delta_{\mathcal{G}_\dagger}\right) = \begin{bmatrix}\Pi_\dagger^\top & 0_{d_\dagger\times d_z}\end{bmatrix}^\top$, where $\Pi_\dagger$ is a $d_\vartheta\times d_\dagger$ matrix collecting the last $d_\dagger$ columns of Π. Denote $\mathrm{J} := \left(\Pi_\dagger^\top\Pi_\dagger\right)^{-1}\Pi_\dagger^\top$, $\mathrm{P} := \Pi_\dagger\mathrm{J}$ and $\mathrm{M} := -\mathrm{I}_{d_\vartheta} + \mathrm{P}$. Then, by inverting the block matrices,

$$\begin{bmatrix} -\Delta_{\mathcal{V}\mathcal{V}^\top} & \Delta_{\mathcal{H}} \\ \Delta_{\mathcal{H}}^\top & 0_{d_\vartheta\times d_\vartheta}\end{bmatrix}^{-1} = \begin{bmatrix} -\Delta_{\mathcal{V}\mathcal{V}^\top}^{-1} + \Delta_{\mathcal{V}\mathcal{V}^\top}^{-1}\Delta_{\mathcal{H}}\left(\Delta_{\mathcal{H}}^\top\Delta_{\mathcal{V}\mathcal{V}^\top}^{-1}\Delta_{\mathcal{H}}\right)^{-1}\Delta_{\mathcal{H}}^\top\Delta_{\mathcal{V}\mathcal{V}^\top}^{-1} & \Delta_{\mathcal{V}\mathcal{V}^\top}^{-1}\Delta_{\mathcal{H}}\left(\Delta_{\mathcal{H}}^\top\Delta_{\mathcal{V}\mathcal{V}^\top}^{-1}\Delta_{\mathcal{H}}\right)^{-1} \\ \left(\Delta_{\mathcal{H}}^\top\Delta_{\mathcal{V}\mathcal{V}^\top}^{-1}\Delta_{\mathcal{H}}\right)^{-1}\Delta_{\mathcal{H}}^\top\Delta_{\mathcal{V}\mathcal{V}^\top}^{-1} & \left(\Delta_{\mathcal{H}}^\top\Delta_{\mathcal{V}\mathcal{V}^\top}^{-1}\Delta_{\mathcal{H}}\right)^{-1}\end{bmatrix}$$

$$= \begin{bmatrix} 0_{d_\vartheta \times d_\vartheta} & 0_{d_\vartheta \times d_z} & \Omega^\top \\ 0_{d_z \times d_\vartheta} & -\mathrm{I}_{d_z} & 0_{d_z \times d_\vartheta} \\ \Omega & 0_{d_\vartheta \times d_z} & \Omega\Omega^\top \end{bmatrix} = \begin{bmatrix} -\left(\Gamma^\top\right)^{-1} \mathrm{Q} \Gamma^{-1} & -\left(\Gamma^\top\right)^{-1} \mathrm{N} \\ -\mathrm{N}^\top \Gamma^{-1} & \mathrm{O} \end{bmatrix} \tag{33}$$

and

$$\begin{bmatrix} -\Delta_{\mathcal{V}\mathcal{V}^\top} & \Delta_{\mathcal{H}_\dagger} \\ \Delta_{\mathcal{H}_\dagger}^\top & 0_{d_\dagger \times d_\dagger} \end{bmatrix}^{-1} = \begin{bmatrix} \mathrm{M} & 0_{d_\vartheta \times d_z} & \mathrm{J}^\top \\ 0_{d_z \times d_\vartheta} & -\mathrm{I}_{d_z} & 0_{d_z \times d_\dagger} \\ \mathrm{J} & 0_{d_\dagger \times d_z} & \mathrm{J}\mathrm{J}^\top \end{bmatrix} = \begin{bmatrix} -\left(\Gamma^\top\right)^{-1} \mathrm{Q}_\dagger \Gamma^{-1} & -\left(\Gamma^\top\right)^{-1} \mathrm{N}_\dagger \\ -\mathrm{N}_\dagger^\top \Gamma^{-1} & \mathrm{O}_\dagger \end{bmatrix}. \tag{34}$$

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

$$\begin{aligned} 0 &= \sum_i \widehat{\mathcal{V}}_i \left\{ 1 - \widehat{\lambda}_p^\top \widehat{\mathcal{V}}_i + \left(\widehat{\lambda}_p^\top \widehat{\mathcal{V}}_i\right)^2 - \left(\widehat{\lambda}_p^\top \widehat{\mathcal{V}}_i\right)^3 + \frac{\left(\widehat{\lambda}_p^\top \widehat{\mathcal{V}}_i\right)^4}{1 + \widehat{\lambda}_p^\top \widehat{\mathcal{V}}_i} \right\} \\ 0 &= \sum_i \mathcal{H}_i^\top \widehat{\lambda}_p \left\{ 1 - \widehat{\lambda}_p^\top \widehat{\mathcal{V}}_i + \left(\widehat{\lambda}_p^\top \widehat{\mathcal{V}}_i\right)^2 - \frac{\left(\widehat{\lambda}_p^\top \widehat{\mathcal{V}}_i\right)^3}{1 + \widehat{\lambda}_p^\top \widehat{\mathcal{V}}_i} \right\}. \end{aligned} \tag{35}$$

By Lemma 6 and $\max_i \left\|\widehat{\mathcal{U}}_i\right\| = O_p^\star\left(\bar{n}^{1/5}\right)$, $\max_i \left|\widehat{\lambda}_p^\top \widehat{\mathcal{V}}_i\right| \lesssim \left\|\widehat{\lambda}_p\right\| \left(\max_i \left\|\widehat{\mathcal{V}}_i\right\|\right) = O_p^\star\left(\sqrt{\log(n)}/\bar{n}^{3/10}\right)$. Therefore, $\max_i \left|\widehat{\lambda}_p^\top \widehat{\mathcal{V}}_i\right| < 1/2$ wp*. By this result, Lemma 6 and $\bar{n}^{-1} \sum_i \left\|\widehat{\mathcal{U}}_i\right\|^4 = O_p^\star(1)$, which follows from boundedness of Θ, Markov's inequality and $\widehat{\mathcal{U}}_i = \mathcal{U}_i - \mathcal{G}_i \widehat{\eta}_p$, $\sum_i \left(\widehat{\lambda}_p^\top \widehat{\mathcal{V}}_i\right)^4 / \left(1 + \widehat{\lambda}_p^\top \widehat{\mathcal{V}}_i\right) = O_p^\star\left(\log(n)^2 / \bar{n}\right)$. Similarly, $\sum_i \mathcal{H}_i^\top \widehat{\lambda}_p \left(\widehat{\lambda}_p^\top \widehat{\mathcal{V}}_i\right)^3 / \left(1 + \widehat{\lambda}_p^\top \widehat{\mathcal{V}}_i\right) = O_p^\star\left(\log(n)^2 / \bar{n}\right)$. By $\widehat{\mathcal{V}}_i = \mathcal{V}_i + \mathcal{H}_i \widehat{\eta}_p$ and Lemma 6, $\sum_i \widehat{\lambda}_p^\top \widehat{\mathcal{V}}_i = \sum_i \widehat{\lambda}_p^\top \mathcal{V}_i + \sum_i \widehat{\lambda}_p^\top \mathcal{H}_i \widehat{\eta}_p$, $\sum_i \left(\widehat{\lambda}_p^\top \widehat{\mathcal{V}}_i\right)^2 = \sum_i \left(\widehat{\lambda}_p^\top \mathcal{V}_i\right)^2 + \sum_i \left(\widehat{\lambda}_p^\top \mathcal{H}_i \widehat{\eta}_p\right)^2 + 2 \sum_i \left(\widehat{\lambda}_p^\top \mathcal{V}_i\right) \left(\widehat{\lambda}_p^\top \mathcal{H}_i \widehat{\eta}_p\right)$ and $\sum_i \left(\widehat{\lambda}_p^\top \widehat{\mathcal{V}}_i\right)^3 = \sum_i \left(\widehat{\lambda}_p^\top \mathcal{V}_i\right)^3 + 3 \sum_i \left(\widehat{\lambda}_p^\top \mathcal{V}_i\right)^2 \left(\widehat{\lambda}_p^\top \mathcal{H}_i \widehat{\eta}_p\right) + O_p^\star\left(\log(n)^{5/2} / \bar{n}^{3/2}\right)$. By plugging these results into the right hand side of (35),

$$\begin{aligned} -\Delta_{\mathcal{V}\mathcal{V}^\top} \widehat{\lambda}_p + \Delta_{\mathcal{H}} \widehat{\eta}_p &= -\frac{1}{\bar{n}} \sum_i \mathcal{V}_i + \frac{1}{\bar{n}} \sum_i \mathcal{V}_i \left(\widehat{\lambda}_p^\top \mathcal{H}_i \widehat{\eta}_p\right) - \frac{1}{\bar{n}} \sum_i \mathcal{V}_i \left(\widehat{\lambda}_p^\top \mathcal{V}_i\right)^2 - \frac{2}{\bar{n}} \sum_i \mathcal{V}_i \left(\widehat{\lambda}_p^\top \mathcal{V}_i\right) \left(\widehat{\lambda}_p^\top \mathcal{H}_i \widehat{\eta}_p\right) \\ &\quad + \frac{1}{\bar{n}} \sum_i \mathcal{V}_i \left(\widehat{\lambda}_p^\top \mathcal{V}_i\right)^3 + \frac{1}{\bar{n}} \sum_i \mathcal{H}_i \widehat{\eta}_p \left(\widehat{\lambda}_p^\top \mathcal{V}_i\right) + \frac{1}{\bar{n}} \sum_i \mathcal{H}_i \widehat{\eta}_p \left(\widehat{\lambda}_p^\top \mathcal{H}_i \widehat{\eta}_p\right) \\ &\quad - \frac{1}{\bar{n}} \sum_i \mathcal{H}_i \widehat{\eta}_p \left(\widehat{\lambda}_p^\top \mathcal{V}_i\right)^2 + \left(\overline{\Delta}_{\mathcal{V}\mathcal{V}^\top} - \Delta_{\mathcal{V}\mathcal{V}^\top}\right) \widehat{\lambda}_p - \left(\overline{\Delta}_{\mathcal{H}} - \Delta_{\mathcal{H}}\right) \widehat{\eta}_p + O_p^\star\left(\left(\log(n)/\bar{n}\right)^2\right) \\ \Delta_{\mathcal{H}}^\top \widehat{\lambda}_p &= \frac{1}{\bar{n}} \sum_i \mathcal{H}_i^\top \widehat{\lambda}_p \left(\widehat{\lambda}_p^\top \mathcal{V}_i\right) + \frac{1}{\bar{n}} \sum_i \mathcal{H}_i^\top \widehat{\lambda}_p \left(\widehat{\lambda}_p^\top \mathcal{H}_i \widehat{\eta}_p\right) - \frac{1}{\bar{n}} \sum_i \mathcal{H}_i^\top \widehat{\lambda}_p \left(\widehat{\lambda}_p^\top \mathcal{V}_i\right)^2 \\ &\quad - \left(\overline{\Delta}_{\mathcal{H}}^\top - \Delta_{\mathcal{H}}^\top\right) \widehat{\lambda}_p + O_p^\star\left(\left(\log(n)/\bar{n}\right)^2\right). \end{aligned} \tag{36}$$

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

$$\begin{aligned}\bar{n}^{-1}\ell_p\left(\widehat{\vartheta}_p \mid h\right) = {} & \frac{2}{\bar{n}}\sum_i \widehat{\lambda}_p^\top \mathcal{V}_i + \frac{2}{\bar{n}}\sum_i \widehat{\lambda}_p^\top \mathcal{H}_i\widehat{\eta}_p - \frac{1}{\bar{n}}\sum_i \left(\widehat{\lambda}_p^\top \mathcal{V}_i\right)^2 - \frac{1}{\bar{n}}\sum_i \left(\widehat{\lambda}_p^\top \mathcal{H}_i\widehat{\eta}_p\right)^2 - \frac{2}{\bar{n}}\sum_i \left(\widehat{\lambda}_p^\top \mathcal{V}_i\right)\left(\widehat{\lambda}_p^\top \mathcal{H}_i\widehat{\eta}_p\right) \\ & + \frac{2}{3}\frac{1}{\bar{n}}\sum_i \left(\widehat{\lambda}_p^\top \mathcal{V}_i\right)^3 + \frac{2}{\bar{n}}\sum_i \left(\widehat{\lambda}_p^\top \mathcal{V}_i\right)^2\left(\widehat{\lambda}_p^\top \mathcal{H}_i\widehat{\eta}_p\right) - \frac{1}{2\bar{n}}\sum_i \left(\widehat{\lambda}_p^\top \mathcal{V}_i\right)^4 + O_p^\star\left(\log(n)^{5/2}/\bar{n}^{5/2}\right). \quad (37)\end{aligned}$$

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 (33) to invert (36) and get higher-order approximations for $\left(\widehat{\lambda}_p, \widehat{\eta}_p\right)$. We then replace all sample averages except $\bar{n}^{-1}\sum_i \mathcal{V}_i$ which is approximately centered ($\|\Delta_{\mathcal{V}}\| = O\left(h^{p+1}\right)$) with the sums of their population means and their centered versions. By iteratively replacing $\left(\widehat{\lambda}_p, \widehat{\eta}_p\right)$ on the right hand side of (36) with the approximations, using Lemmas 5 and 6 and dropping terms that are $O_p^\star\left((\log(n)/\bar{n})^2\right)$, we get cubic stochastic expansions of $\left(\widehat{\lambda}_p, \widehat{\eta}_p\right)$. By the same steps and plugging stochastic expansions of $\left(\widehat{\lambda}_p, \widehat{\eta}_p\right)$ into the right hand side of (37), we have a stochastic expansion of $\bar{n}^{-1}\ell_p\left(\widehat{\vartheta}_p \mid h\right)$ so that $\bar{n}^{-1}\ell_p\left(\widehat{\vartheta}_p \mid h\right) = \widehat{\ell}^\star + O_p^\star\left(\log(n)^{5/2}/\bar{n}^{5/2}\right)$, where the leading term $\widehat{\ell}^\star$ is a quartic polynomial of centered sample averages. Similarly, by using Lemmas 5 and 6, the first-order conditions and (34), we get cubic stochastic expansions of $\left(\widetilde{\lambda}_p, \widetilde{\eta}_p\right)$ and a quartic stochastic expansion of $\ell_p\left(\vartheta_0, \widetilde{\vartheta}_p \mid h\right)$ so that $\bar{n}^{-1}\ell_p\left(\vartheta_0, \widetilde{\vartheta}_p \mid h\right) = \widetilde{\ell}^\star + O_p^\star\left(\log(n)^{5/2}/\bar{n}^{5/2}\right)$. 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: $k, l, m, n, o, v, q = 1, ..., d_\vartheta$, $\mathsf{k}, \mathsf{l}, \mathsf{m}, \mathsf{n}, \mathsf{o}, \mathsf{v} = 1, ..., 2d_u$, $u, w = 1, ..., d_\dagger$, $\mathsf{a}, \mathsf{b}, \mathsf{c}, \mathsf{d}, \mathsf{e}, \mathsf{f} = 1, ..., 2d_z$, $s, t, a, b, c, d, e = 1, ..., d_z$ and $\mathsf{u}, \mathsf{w} = 1, ..., d_u$. Let $\alpha^{\mathsf{kl}} := \Delta_{\mathcal{V}^{(\mathsf{k})}\mathcal{V}^{(\mathsf{l})}}$, $\alpha^{\mathsf{klm}} := \Delta_{\mathcal{V}^{(\mathsf{k})}\mathcal{V}^{(\mathsf{l})}\mathcal{V}^{(\mathsf{m})}}$, $\alpha^{\mathsf{klmn}} := \Delta_{\mathcal{V}^{(\mathsf{k})}\mathcal{V}^{(\mathsf{l})}\mathcal{V}^{(\mathsf{m})}\mathcal{V}^{(\mathsf{n})}}$, $\gamma^{\mathsf{k},n} := \Delta_{\mathcal{H}^{(\mathsf{k}n)}}$, $\gamma^{\mathsf{k};\mathsf{l},n} := \Delta_{\mathcal{V}^{(\mathsf{k})}\mathcal{H}^{(\mathsf{l}n)}}$, $\gamma^{\mathsf{k};\mathsf{l};\mathsf{m},n} := \Delta_{\mathcal{V}^{(\mathsf{k})}\mathcal{V}^{(\mathsf{l})}\mathcal{H}^{(\mathsf{m}n)}}$, $\gamma^{\mathsf{k},n;\mathsf{l},o} := \Delta_{\mathcal{H}^{(\mathsf{k}n)}\mathcal{H}^{(\mathsf{l}o)}}$, $A^{\mathsf{k}} := \bar{n}^{-1}\sum_i \mathcal{V}_i^{(\mathsf{k})}$, $A^{\mathsf{kl}} := \bar{n}^{-1}\sum_i \mathcal{V}_i^{(\mathsf{k})}\mathcal{V}_i^{(\mathsf{l})} - \alpha^{\mathsf{kl}}$, $A^{\mathsf{klm}} := \bar{n}^{-1}\sum_i \mathcal{V}_i^{(\mathsf{k})}\mathcal{V}_i^{(\mathsf{l})}\mathcal{V}_i^{(\mathsf{m})} - \alpha^{\mathsf{klm}}$, $C^{\mathsf{k},n} := \bar{n}^{-1}\sum_i \mathcal{H}_i^{(\mathsf{k}n)} - \gamma^{\mathsf{k},n}$ and $C^{\mathsf{k};\mathsf{l},n} := \bar{n}^{-1}\sum_i \mathcal{V}_i^{(\mathsf{k})}\mathcal{H}_i^{(\mathsf{l}n)} - \gamma^{\mathsf{k};\mathsf{l},n}$. By Lemma 5, $\bar{n}^{-1}\sum_i \mathcal{V}_i^{(\mathsf{k})}\mathcal{V}_i^{(\mathsf{l})}\mathcal{V}_i^{(\mathsf{m})}\mathcal{V}_i^{(\mathsf{n})} - \alpha^{\mathsf{klmn}}$, $\bar{n}^{-1}\sum_i \mathcal{V}_i^{(\mathsf{k})}\mathcal{V}_i^{(\mathsf{l})}\mathcal{H}_i^{(\mathsf{m}n)} - \gamma^{\mathsf{k};\mathsf{l};\mathsf{m},n}$ and $\bar{n}^{-1}\sum_i \mathcal{H}_i^{(\mathsf{k}n)}\mathcal{H}_i^{(\mathsf{l}o)} - \gamma^{\mathsf{k},n;\mathsf{l},o}$ are all $O_p^\star\left(\sqrt{\log(n)/\bar{n}}\right)$. We can show that $\left(\widehat{\lambda}_p, \widehat{\eta}_p\right)$ and $\bar{n}^{-1}\ell_p\left(\widehat{\vartheta}_p \mid h\right)$ admit stochastic expansions with leading terms that are polynomials of $\left(A^{\mathsf{k}}, A^{\mathsf{kl}}, A^{\mathsf{klm}}, C^{\mathsf{k},n}, C^{\mathsf{k};\mathsf{l},n}\right)$ with coefficients given by $\left(\alpha^{\mathsf{kl}}, \alpha^{\mathsf{klm}}, \alpha^{\mathsf{klmn}}\right)$, $\left(\gamma^{\mathsf{k},n}, \gamma^{\mathsf{k};\mathsf{l},n}, \gamma^{\mathsf{k};\mathsf{l};\mathsf{m},n}, \gamma^{\mathsf{k},n;\mathsf{l},o}\right)$ 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 $\left(\widetilde{\lambda}_p, \widetilde{\eta}_p\right)$ and $\bar{n}^{-1}\ell_p\left(\vartheta_0, \widetilde{\vartheta}_p \mid h\right)$ that are polynomials of $\left(A^{\mathsf{k}}, A^{\mathsf{kl}}, A^{\mathsf{klm}}, C^{\mathsf{k},n}, C^{\mathsf{k};\mathsf{l},n}\right)$ 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 $LR^{\star} \coloneqq \bar{n}\left(\widetilde{\ell}^{\star} - \widehat{\ell}^{\star}\right)$ so that $LR_p\left(\vartheta_0 \mid h\right) = LR^{\star} + O_p^{\star}\left(\log\left(n\right)^{5/2}/\bar{n}^{3/2}\right)$.

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

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

Proof. A decomposition $LR^{\star} = \bar{n}\left(\tilde{R}_1^2 + 2\tilde{R}_1\tilde{R}_2 + 2\tilde{R}_1\tilde{R}_3 + \tilde{R}_2^2\right)$ can be derived. $\tilde{R}_k$ is a homogeneous k-th order polynomial of $\left(A^{\mathsf{k}}, A^{\mathsf{kl}}, A^{\mathsf{klm}}, C^{\mathsf{k},n}, C^{\mathsf{k};\mathsf{l},n}\right)$ so that $\tilde{R}_1 = O_p^{\star}\left(\sqrt{\log\left(n\right)/\bar{n}}\right)$, $\tilde{R}_2 = O_p^{\star}\left(\log\left(n\right)/\bar{n}\right)$ and $\tilde{R}_3 = O_p^{\star}\left(\left(\log\left(n\right)/\bar{n}\right)^{3/2}\right)$. $-\mathrm{M}$ is a projection matrix onto the orthogonal complement of the column space of $\Pi_{\dagger}$. Let ϖ_0 be a vector spanning the one-dimensional orthogonal complement of the column space of $\Pi_{\dagger}$ so that $-\mathrm{M} = \varpi_0\left(\varpi_0^{\top}\varpi_0\right)^{-1}\varpi_0^{\top}$. Let $\varpi \coloneqq \varpi_0/\sqrt{\varpi_0^{\top}\varpi_0}$. Then, $\varpi^{\top}\varpi = 1$ and $-\mathrm{M} = \varpi\varpi^{\top}$. The expressions of $\left(\tilde{R}_1, \tilde{R}_2, \tilde{R}_3\right)$ can be readily obtained in a special case of Ma (2017). Algebraic calculations in Ma (2017) show that by setting $\tilde{R}_1 \coloneqq \varpi^{(k)}A^k$,

$$\begin{aligned}\tilde{R}_2 \coloneqq{}& \frac{1}{2}\mathrm{M}^{(mk)}\varpi^{(n)}A^{mn}A^k - \varpi^{(n)}A^{n\,d_\vartheta+a}A^{d_\vartheta+a} + \left\{\frac{1}{3}\alpha^{vmn}\mathrm{M}^{(vl)}\mathrm{M}^{(mk)}\varpi^{(n)} - \gamma^{m;v,o}\Omega^{(on)}\mathrm{P}^{(nk)}\mathrm{M}^{(ml)}\varpi^{(v)}\right\} \\ &\times A^lA^k + \left\{\left(\gamma^{d_\vartheta+a;v,m}[d_\vartheta+a,v]\right)\Omega^{(mo)}\mathrm{P}^{(ok)}\varpi^{(v)} - \alpha^{vm\,d_\vartheta+a}\mathrm{M}^{(vk)}\varpi^{(m)}\right\}A^kA^{d_\vartheta+a} - \Omega^{(ko)}\mathrm{P}^{(om)}\varpi^{(l)} \\ &\times C^{l,k}A^m + \left\{\alpha^{v\,d_\vartheta+a\,d_\vartheta+b}\varpi^{(v)} - \gamma^{d_\vartheta+a;d_\vartheta+b,m}\Omega^{(mn)}\varpi^{(n)}\right\}A^{d_\vartheta+a}A^{d_\vartheta+b} + \Omega^{(km)}\varpi^{(m)}C^{d_\vartheta+a,k}A^{d_\vartheta+a}, \quad (38)\end{aligned}$$

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 $LR^\star = \bar{n}\left(\tilde{R}_1^2+2\tilde{R}_1\tilde{R}_2+2\tilde{R}_1\tilde{R}_3+\tilde{R}_2^2\right)$. (38) 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 $\alpha^{\mathsf{k}} \coloneqq \Delta_{\mathcal{V}^{(\mathsf{k})}}$ and $\mathring{A}^{\mathsf{k}} \coloneqq A^{\mathsf{k}} - \alpha^{\mathsf{k}}$. By replacing A^{k} with $\mathring{A}^{\mathsf{k}}+\alpha^{\mathsf{k}}$, we have $\tilde{R}_1 = \tilde{R}_{10}+\tilde{R}_{11}$, where $\tilde{R}_{10} \coloneqq \varpi^{(k)}\alpha^k$ and $\tilde{R}_{11} \coloneqq \varpi^{(k)}\mathring{A}^k$. Similarly, we replace A^{k} with $\mathring{A}^{\mathsf{k}}+\alpha^{\mathsf{k}}$ 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 $\alpha^1,...,\alpha^{2d_u}$:

$$\begin{aligned}\tilde{R}_{21} \coloneqq{}& \frac{1}{2}\mathrm{M}^{(mk)}\varpi^{(n)}A^{mn}\alpha^k - \varpi^{(n)}A^{n\,d_\vartheta+a}\alpha^{d_\vartheta+a} + \frac{2}{3}\alpha^{vmn}\mathrm{M}^{(vl)}\mathrm{M}^{(mk)}\varpi^{(n)}\mathring{A}^l\alpha^k - \gamma^{m;v,o}\Omega^{(on)}\mathrm{P}^{(nk)}\mathrm{M}^{(ml)}\varpi^{(v)} \\ &\times\left(\mathring{A}^l\alpha^k[l,k]\right) + \left\{\left(\gamma^{d_\vartheta+a;v,m}[d_\vartheta+a,v]\right)\Omega^{(mo)}\mathrm{P}^{(ok)}\varpi^{(v)} - \alpha^{vm\,d_\vartheta+a}\mathrm{M}^{(vk)}\varpi^{(m)}\right\}\left(\alpha^k\mathring{A}^{d_\vartheta+a}[k,d_\vartheta+a]\right) \\ &- \Omega^{(ko)}\mathrm{P}^{(om)}\varpi^{(l)}C^{l,k}\alpha^m + \left\{\alpha^{v\,d_\vartheta+a\,d_\vartheta+b}\varpi^{(v)} - \gamma^{d_\vartheta+a;d_\vartheta+b,m}\Omega^{(mn)}\varpi^{(n)}\right\}\left(\alpha^{d_\vartheta+a}\mathring{A}^{d_\vartheta+b}[d_\vartheta+a,d_\vartheta+b]\right) \\ &+ \Omega^{(km)}\varpi^{(m)}C^{d_\vartheta+a,k}\alpha^{d_\vartheta+a}, \quad (39)\end{aligned}$$

$\tilde{R}_{22}$ is defined by the right hand side of (38) with A^{k} replaced by $\mathring{A}^{\mathsf{k}}$ and $\tilde{R}_{20} \coloneqq \tilde{R}_2 - \tilde{R}_{22} - \tilde{R}_{21} = O\left(\|\Delta_{\mathcal{U}}\|^2\right)$. Let $R_0 \coloneqq \tilde{R}_{10}+\tilde{R}_{20}$, $R_1 \coloneqq \tilde{R}_{11}+\tilde{R}_{21}$ and $R_2 \coloneqq \tilde{R}_{22}$. We decompose $\tilde{R}_3 = \tilde{R}_{33}+\tilde{R}_{32}+\tilde{R}_{31}+\tilde{R}_{30}$ in a similar manner and let $R_3 \coloneqq \tilde{R}_{33}$. R_3 is given by the formula of $\tilde{R}_3$ with A^{k} replaced by $\mathring{A}^{\mathsf{k}}$. Then, let $R \coloneqq R_1+R_2+R_3$. By Lemma 5, $\tilde{R}_1+\tilde{R}_2+\tilde{R}_3 = R_0+R+O_p^\star\left(\|\Delta_{\mathcal{U}}\|\log(n)/\bar{n}\right)$ and therefore, $LR^\star = \bar{n}\left(R_0+R\right)^2 + O_p^\star\left(v_n^\sharp\right)$.

Let $\mathcal{F} \coloneqq \left(W_p\otimes U, W_p, \left(W_{p;+}^2, W_{p;-}^2\right)^\top \otimes \left(U, U^2\right), \left(W_{p;+}^3, W_{p;-}^3\right)^\top \otimes U^3\right)$. $\mathcal{F}_i$ is defined analogously and let d_f denote the dimension of $\mathcal{F}$. It can be shown that $\sqrt{\bar{n}}R \coloneqq h_n\left(\overline{\mathcal{F}}\right)$, where $\overline{\mathcal{F}} \coloneqq \bar{n}^{-1/2}\sum_i\left(\mathcal{F}_i - \mathrm{E}\left[\mathcal{F}\right]\right)$ and h_n is a cubic polynomial. E.g., $\sqrt{\bar{n}}\varpi^{(k)}\mathring{A}^k = \tilde{\varpi}^\top\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1/2}\left(\bar{n}^{-1/2}\sum_i\left(\mathcal{U}_i - \mathrm{E}\left[\mathcal{U}\right]\right)\right)$, where $\tilde{\varpi} \coloneqq \mathrm{S}\begin{bmatrix}\varpi^\top & 0_{d_z}^\top\end{bmatrix}^\top$. It can be shown that other terms on the right hand side of (39) can also be written as linear functions of $\overline{\mathcal{F}}$. 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 $\overline{\mathcal{F}}$. A more lucid proof of this fact uses the observation that $\ell_p\left(\widehat{\vartheta}_p \mid h\right) = \inf_{\theta_2}\sup_{\lambda_2} 2\sum_i \log\left(1+\lambda_2^\top\bar{\mathcal{U}}_i\left(\theta_2\right)\right)$. Let $\mathcal{M}_i\left(\theta_0,\theta_1\right) \coloneqq W_{p,i}\otimes\left(Y_i-\theta_0 D_i-\theta_1\right)$. By rearranging the moment conditions, $\ell_p\left(\widehat{\vartheta}_p \mid h\right) = \inf_{\theta_0,\theta_1,\theta_2}\sup_{\lambda_1,\lambda_2} 2\sum_i \log\left(1+\lambda_1^\top\mathcal{M}_i\left(\theta_0,\theta_1\right)+\lambda_2^\top\bar{\mathcal{U}}_i\left(\theta_2\right)\right)$. Let $\mathcal{W}_i\left(\theta\right) \coloneqq \left(\mathcal{M}_i\left(\theta_0,\theta_1\right), \bar{\mathcal{U}}_i\left(\theta_2\right)\right)$. $\widehat{\vartheta}_p$ and $\widehat{\lambda} \coloneqq \left(\widehat{\lambda}_1, \widehat{\lambda}_2\right)$ satisfy the first-order conditions wp*:

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

The third condition implies that $\widehat{\lambda}_1 = 0_2$ wp*. Therefore, $\ell_p\left(\widehat{\vartheta}_p \mid h\right) = 2\sum_i \log\left(1 + \widehat{\lambda}_2^\top \bar{\mathcal{U}}_i\left(\widehat{\vartheta}_{p,2}\right)\right)$ and the second and fourth conditions are $\sum_i \bar{\mathcal{U}}_i\left(\widehat{\vartheta}_{p,2}\right) / \left(1 + \widehat{\lambda}_2^\top \bar{\mathcal{U}}_i\left(\widehat{\vartheta}_{p,2}\right)\right) = 0_{2d_z}$ and $\sum_i \bar{\mathcal{G}}_i^\top \widehat{\lambda}_2 / \left(1 + \widehat{\lambda}_2^\top \bar{\mathcal{U}}_i\left(\widehat{\vartheta}_{p,2}\right)\right) = 0_{d_z}$, which coincide with the first-order conditions of $\inf_{\theta_2}\sup_{\lambda_2}\sum_i \log\left(1 + \lambda_2^\top \bar{\mathcal{U}}_i\left(\theta_2\right)\right)$. Therefore, we have $\ell_p\left(\widehat{\vartheta}_p \mid h\right) = \inf_{\theta_2}\sup_{\lambda_2} 2\sum_i \log\left(1 + \lambda_2 \bar{\mathcal{U}}_i\left(\theta_2\right)\right)$. By expansion and Lemma 6, we get approximations for $\widehat{\lambda}_2$, $\widehat{\vartheta}_{p,2}$ and $\ell_p\left(\widehat{\vartheta}_p \mid h\right)$ which are similar to (36) and (37). 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 $\bar{n}^{-1/2}\sum_i\left(\bar{\mathcal{F}}_i - \mathrm{E}\left[\bar{\mathcal{F}}\right]\right)$, where $\left(\bar{\mathcal{F}}_i, \bar{\mathcal{F}}\right)$ are defined by the formulae of $\left(\mathcal{F}_i, \mathcal{F}\right)$ with $\left(U_i, U\right)$ replaced by $\left(\bar{U}_i, \bar{U}\right)$. Similarly, the stochastic expansion of $\ell_p\left(\vartheta_0, \widetilde{\vartheta}_p \mid h\right)$ should involve only terms in $\overline{\mathcal{F}}$.

Let $\kappa_j\left(V\right)$ denote the j-th cumulant of a random variable V. We follow arguments in the proof of Calonico et al. (2018a, Theorem S.1) and apply Skovgaard (1986, Theorem 3.4) with $s = 4$ to $S_n \coloneqq \mathrm{B}^{-1/2}\overline{\mathcal{F}}$ where $\mathrm{B} \coloneqq \mathrm{Var}\left[\mathcal{F}\right]/h$. 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), $\kappa_3\left(t^\top S_n\right) = \mathrm{E}\left[\left(t^\top S_n\right)^3\right] = O\left(\bar{n}^{-1/2}\right)$, $\kappa_4\left(t^\top S_n\right) = \mathrm{E}\left[\left(t^\top S_n\right)^4\right] - 3\left(\mathrm{E}\left[\left(t^\top S_n\right)^2\right]\right)^2 = O\left(\bar{n}^{-1}\right)$ and $\rho_{s,n}\left(t\right) \coloneqq \max\left\{\left|\kappa_3\left(t^\top S_n\right)\right|/3!, \sqrt{\left|\kappa_4\left(t^\top S_n\right)\right|/4!}\right\} = O\left(\bar{n}^{-1/2}\right)$, uniformly in t. Condition I and II of Skovgaard (1986, Theorem 3.4) are satisfied by taking $a_n\left(t\right) \propto \sqrt{\bar{n}}$ and $\epsilon_n = \bar{n}^{-3/2}$. Let $\hat{\Psi}_V\left(t\right) \coloneqq \mathrm{E}\left[\exp\left(\mathrm{i}t^\top V\right)\right]$ denote the characteristic function of a random vector V, where $\mathrm{i} \coloneqq \sqrt{-1}$. Let $\mathcal{F}_{\mathrm{s}} \coloneqq \left(W_{p;\mathrm{s}}U, W_{p;\mathrm{s}}, W_{p;\mathrm{s}}^2\left(U, U^2\right), W_{p;\mathrm{s}}^3 U^3\right)$, $\mathrm{s} \in \{-,+\}$. Then, $\hat{\Psi}_{\mathcal{F}}\left(t\right) = \mathrm{E}\left[\exp\left(\mathrm{i}t_+^\top \mathcal{F}_+\right)1\left(X \geq 0\right)\right] + \mathrm{E}\left[\exp\left(\mathrm{i}t_-^\top \mathcal{F}_-\right)1\left(X < 0\right)\right]$, where $\left(t_-, t_+\right)$ denote corresponding coordinates of t. By change of variables, $\mathrm{E}\left[\exp\left(\mathrm{i}t_+^\top \mathcal{F}_+\right)1\left(X \geq 0\right)\right] = h\left(f_X\left(0\right)E_+\left(t_+\right) + O\left(h\right)\right) + \Pr\left[X > h\right]$, where E_+ is the characteristic function of $\mathcal{K}_{p;+}\left(V\right)\left(U, 1\right)$, $\mathcal{K}_{p;+}\left(V\right)^2\left(U, U^2\right)$, $\mathcal{K}_{p;+}\left(V\right)^3 U^3$, where $\left(V, U\right)$ has the joint density given by $\left(v, u\right) \mapsto 1\left(0 \leq v \leq 1\right) f_{U|X}\left(u \mid 0\right)$. A similar result holds for $\mathrm{E}\left[\exp\left(\mathrm{i}t_-^\top \mathcal{F}_-\right)1\left(X < 0\right)\right]$ with $E_-\left(t_-\right)$ defined similarly. Therefore, $\hat{\Psi}_{\mathcal{F}}\left(t\right) = 1 - \Pr\left[-h < X \leq h\right] + hf_X\left(0\right)\left(E_+\left(t_+\right) + E_-\left(t_-\right)\right) + O\left(h^2\right)$. By Assumption 5, the vector-valued functions $\left(v, u\right) \mapsto \left(1, \left(\mathcal{K}_{p;+}\left(v\right), \mathcal{K}_{p;+}\left(v\right)^2, \mathcal{K}_{p;+}\left(v\right)^3\right) \otimes \left(1, u, u^2, u^3\right)\right)$ are linearly independent. By invoking the same arguments as in the proof of Calonico et al. (2018a, Lemma S.9), $\forall \varepsilon > 0$, $\exists c_\varepsilon > 0$ such that $\sup_{\|t\| > \varepsilon}\left|E_+\left(t_+\right)\right| < 1 - c_\varepsilon$. A similar result holds for E_-. Then by these results, $\forall \varepsilon > 0$, $\exists c_\varepsilon > 0$ such that $\sup_{\|t\| > \varepsilon}\left|\hat{\Psi}_{\mathcal{F}}\left(t\right)\right| < 1 - c_\varepsilon h$, when n is sufficiently large. It follows from this result and arguments in the proof of Calonico et al. (2018a, Theorem S.1) that $\forall \delta > 0$, $\exists c_\delta > 0$ such that $\sup_{\|t\| > \delta\sqrt{\bar{n}}}\left|\hat{\Psi}_{S_n}\left(t\right)\right| \leq \left(1 - c_\delta h\right)^n$ when n is sufficiently large. It is also easy to see that $\forall \delta > 0$, $\left(1 - c_\delta h\right)^n \leq \epsilon_n^{d_f/2+2}$, 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. (2018a, 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 $\sqrt{\bar{n}}R = h_n\left(\mathrm{B}^{1/2}S_n\right)$. 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 $x \mapsto h_n\left(\mathrm{B}^{1/2}x\right)$ whose the gradient at zero $\nabla g_n(0)$ is given by $\nabla g_n(0) = \mathrm{B}^{1/2}\left(\tilde{\varpi}^\top \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1/2}, 0_{d_f-2d_u}^\top\right)^\top + O\left(\|\Delta_{\mathcal{U}}\|\right)$ by the chain rule. Then we apply Skovgaard (1981, Theorem 3.2) to $f_n(S_n) \coloneqq B_n^{-1}g_n(S_n)$, where $B_n^2 \coloneqq \nabla g_n(0)^\top \nabla g_n(0)$. Then, $B_n^2 = \tilde{\varpi}^\top \Delta_{\mathcal{U}\mathcal{U}^\top}^{-1/2}\left(\mathrm{Var}\left[\mathcal{U}\right]/h\right)\Delta_{\mathcal{U}\mathcal{U}^\top}^{-1/2}\tilde{\varpi} + O\left(\|\Delta_{\mathcal{U}}\|\right) = 1 + O\left(\|\Delta_{\mathcal{U}}\|\right)$. Condition I of Skovgaard (1981, Assumption 3.1) is satisfied with $p = 4$. Condition II of Skovgaard (1981, Assumption 3.1) is satisfied with $\lambda_n = O\left(\bar{n}^{-1/2}\right)$ so that $\lambda_n^{p-1} = o\left(\bar{n}^{-1}\right)$. 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, $\mathrm{E}\left[R_1\right] = 0$. We calculate $\mathrm{E}\left[R_2\right]$, let the remainder term absorb the terms that involve $\alpha^1, ..., \alpha^{2d_u}$ and get $\mathrm{E}\left[R_2\right] = \bar{n}^{-1}\bar{\kappa}_1 + O\left(\|\Delta_{\mathcal{U}}\|/n\right)$ where $\bar{\kappa}_1 \coloneqq \alpha^{mnk}\mathrm{M}^{(mk)}\varpi^{(n)}/6 - \Omega^{(ko)}\mathrm{P}^{(om)}\varpi^{(l)}\gamma^{m;l,k}$. By formulae for third moments and Lemma 1, $\mathrm{E}\left[R_3\right] = O\left(\bar{n}^{-2}\right)$. Therefore, $\kappa_1\left(\sqrt{\bar{n}}R\right) = \tilde{\kappa}_{1,n} + O\left(\bar{n}^{-1/2}\|\Delta_{\mathcal{U}}\|h + \bar{n}^{-3/2}\right)$ with $\tilde{\kappa}_{1,n} \coloneqq \bar{n}^{-1/2}\bar{\kappa}_1$. For the second cumulant, by definition, $\kappa_2(R) = \mathrm{E}\left[R^2\right] - \left(\mathrm{E}\left[R\right]\right)^2$ and by formulae for fifth and sixth moments and Lemma 1, $\mathrm{E}\left[R^2\right] = \mathrm{E}\left[R_1^2\right] + 2\cdot\mathrm{E}\left[R_1R_2\right] + 2\cdot\mathrm{E}\left[R_1R_3\right] + \mathrm{E}\left[R_2^2\right] + O\left(\bar{n}^{-3}\right)$. By $R_1 = \tilde{R}_{11} + \tilde{R}_{21}$ and calculation, $\mathrm{E}\left[R_1^2\right] = \mathrm{E}\left[\tilde{R}_{11}^2\right] + 2\cdot\mathrm{E}\left[\tilde{R}_{21}\tilde{R}_{11}\right] + O\left(\bar{n}^{-1}\|\Delta_{\mathcal{U}}\|^2\right)$, $\mathrm{E}\left[R_1R_2\right] + \mathrm{E}\left[R_1R_3\right] = \mathrm{E}\left[\tilde{R}_{11}R_2\right] + \mathrm{E}\left[\tilde{R}_{11}R_3\right] + O\left(\bar{n}^{-2}\|\Delta_{\mathcal{U}}\|\right)$. Then by calculation, $\mathrm{E}\left[\tilde{R}_{11}^2\right] = \bar{n}^{-1} + O\left(\|\Delta_{\mathcal{U}}\|^2/n\right)$ and $2\cdot\mathrm{E}\left[\tilde{R}_{21}\tilde{R}_{11}\right] = \bar{n}^{-1}\tilde{\kappa}_{21,n} + O\left(\|\Delta_{\mathcal{U}}\|^2/n\right)$, where $\tilde{\kappa}_{21,n} \coloneqq \alpha^{mno}\mathrm{M}^{(no)}\mathrm{M}^{(mk)}\alpha^k/3 - 2\gamma^{l;d_\vartheta+a,k}\Omega^{(km)}\mathrm{M}^{(ml)}\alpha^{d_\vartheta+a}$. Then, $\mathrm{E}\left[R_1^2\right] = \bar{n}^{-1}\left(1 + \tilde{\kappa}_{21,n}\right) + O\left(\bar{n}^{-1}\|\Delta_{\mathcal{U}}\|^2\right)$. Calculation of $2\cdot\mathrm{E}\left[\tilde{R}_{11}R_2\right] + 2\cdot\mathrm{E}\left[\tilde{R}_{11}R_3\right] + \mathrm{E}\left[R_2^2\right]$ follows from replication of calculations in Ma (2017) and we can directly use the results therein. By calculations in Ma (2017), we have

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

for some bounded constants $\bar{\kappa}_{21}, ..., \bar{\kappa}_{28}$, e.g., $\bar{\kappa}_{21} \coloneqq \alpha^{vmn}\mathrm{M}^{(vo)}\mathrm{M}^{(ml)}\mathrm{M}^{(nk)}\alpha^{klo}/3 - \alpha^{vm\,d_\vartheta+a}\mathrm{M}^{(vo)}\mathrm{M}^{(mn)}\alpha^{on\,d_\vartheta+a} + \alpha^{n\,d_\vartheta+a\,d_\vartheta+b}\mathrm{M}^{(nm)}\alpha^{m\,d_\vartheta+a\,d_\vartheta+b}$ and the $O\left(\bar{n}^{-2}\|\Delta_{\mathcal{U}}\|h + \bar{n}^{-3}\right)$ remainder collects terms that depend on $\alpha^1, ..., \alpha^{2d_u}$ and higher-order terms from the fourth moment calculation. The expressions of $\bar{\kappa}_{22}, ..., \bar{\kappa}_{28}$ are also easily obtained from Ma (2017) and hence omitted. Therefore, $\kappa_2\left(\sqrt{\bar{n}}R\right) = \tilde{\kappa}_{2,n} + O\left(\|\Delta_{\mathcal{U}}\|^2 + \bar{n}^{-1}\|\Delta_{\mathcal{U}}\| + \bar{n}^{-2}\right)$, where $\tilde{\kappa}_{2,n} \coloneqq 1 + \tilde{\kappa}_{21,n} + \tilde{\kappa}_{22,n}$ and $\tilde{\kappa}_{22,n} \coloneqq \bar{n}^{-1}\left(\sum_{j=1}^{8}\bar{\kappa}_{2j} - \bar{\kappa}_1^2\right)$. By definition, $\kappa_3(R) = \mathrm{E}\left[R^3\right] - 3\cdot\mathrm{E}\left[R\right]\mathrm{E}\left[R^2\right] + 2\left(\mathrm{E}\left[R\right]\right)^3$ and by $\mathrm{E}\left[R\right] = \mathrm{E}\left[R_2\right] + O\left(\bar{n}^{-2}\right)$, $\mathrm{E}\left[R_2\right] = O\left(\bar{n}^{-1}\right)$, $\mathrm{E}\left[R^2\right] = \mathrm{E}\left[R_1^2\right] + O\left(\bar{n}^{-2}\right)$ and $\mathrm{E}\left[R^3\right] = \mathrm{E}\left[R_1^3\right] + 3\cdot\mathrm{E}\left[R_2R_1^2\right] + O\left(\bar{n}^{-3}\right)$, which follows from formulae for higher moments, we have $\kappa_3(R) =$

$\mathrm{E}\left[R_1^3\right] - 3\left(\mathrm{E}\left[R_2 R_1^2\right] - \mathrm{E}\left[R_2\right]\mathrm{E}\left[R_1^2\right]\right) + O\left(\bar{n}^{-3}\right)$. It is easy to check that $\mathrm{E}\left[R_1^3\right] = \mathrm{E}\left[\tilde{R}_{11}^3\right] + O\left(\bar{n}^{-2}\left\|\Delta_{\mathcal{U}}\right\|\right)$, $\mathrm{E}\left[R_2 R_1^2\right] = \mathrm{E}\left[R_2 \tilde{R}_{11}^2\right] + O\left(\bar{n}^{-2}\left\|\Delta_{\mathcal{U}}\right\|\right)$. By these results and $\mathrm{E}\left[R_1^2\right] = \mathrm{E}\left[\tilde{R}_{11}^2\right] + O\left(\bar{n}^{-1}\left\|\Delta_{\mathcal{U}}\right\|\right)$, $\kappa_3\left(R\right) = \mathrm{E}\left[\tilde{R}_{11}^3\right] - 3\left(\mathrm{E}\left[R_2\tilde{R}_{11}^2\right] - \mathrm{E}\left[R_2\right]\mathrm{E}\left[\tilde{R}_{11}^2\right]\right) + O\left(\bar{n}^{-3} + \bar{n}^{-2}\left\|\Delta_{\mathcal{U}}\right\|\right)$. Calculation and expansion of $\mathrm{E}\left[\tilde{R}_{11}^3\right] - 3\left(\mathrm{E}\left[R_2\tilde{R}_{11}^2\right] - \mathrm{E}\left[R_2\right]\mathrm{E}\left[\tilde{R}_{11}^2\right]\right)$ follows from replication of calculations in Ma (2017). For example, by calculation using formulae for moments (DiCiccio et al., 1988),

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

and the $O\left(\bar{n}^{-2}\left\|\Delta_{\mathcal{U}}\right\| h\right)$ remainder collects all terms in the expansion of the third moment which depend on $\alpha^1, ..., \alpha^{2d_u}$. Note that we can write $\mathrm{E}\left[h^{-1}\left(\varpi^{(k)}\mathcal{V}^{(k)}\right)^3\right] = \varpi^{(k)}\varpi^{(l)}\varpi^{(m)}\alpha^{klm}$ in coordinate notations. Similarly, we calculate $\mathrm{E}\left[R_2\tilde{R}_{11}^2\right] - \mathrm{E}\left[R_2\right]\mathrm{E}\left[\tilde{R}_{11}^2\right]$. We note that coefficients of terms of order $\bar{n}^{-2}$ in $\mathrm{E}\left[\tilde{R}_{11}^3\right] - 3\left(\mathrm{E}\left[R_2\tilde{R}_{11}^2\right] - \mathrm{E}\left[R_2\right]\mathrm{E}\left[\tilde{R}_{11}^2\right]\right)$ 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 $\kappa_3\left(\sqrt{\bar{n}}R\right) = O\left(\left\|\Delta_{\mathcal{U}}\right\|/\sqrt{\bar{n}} + \bar{n}^{-3/2}\right)$. By this result, the fact that $\kappa_4\left(R\right) = \mathrm{E}\left[R^4\right] - 3\left(\mathrm{E}\left[R^2\right]\right)^2 - 4\cdot\mathrm{E}\left[R\right]\kappa_3\left(R\right) + 2\left(\mathrm{E}\left[R\right]\right)^4$, $\mathrm{E}\left[R\right] = O\left(\bar{n}^{-1}\right)$, $R = \tilde{R}_{11} + \tilde{R}_{21} + R_2 + R_3$ and standard calculations,

$$\begin{aligned}\kappa_4\left(R\right) = \mathrm{E}\left[R^4\right] - 3\left(\mathrm{E}\left[R^2\right]\right)^2 + O\left(\bar{n}^{-3}\left\|\Delta_{\mathcal{U}}\right\| + \bar{n}^{-4}\right) &= \left\{\mathrm{E}\left[\tilde{R}_{11}^4\right] - 3\left(\mathrm{E}\left[\tilde{R}_{11}^2\right]\right)^2\right\} \\ &+ 4\left\{\mathrm{E}\left[R_2\tilde{R}_{11}^3\right] - 3\cdot\mathrm{E}\left[R_2\tilde{R}_{11}\right]\mathrm{E}\left[\tilde{R}_{11}^2\right]\right\} + 6\left\{\mathrm{E}\left[R_2^2\tilde{R}_{11}^2\right] - \mathrm{E}\left[R_2^2\right]\mathrm{E}\left[\tilde{R}_{11}^2\right]\right\} \\ &+ 4\left\{\mathrm{E}\left[R_2\tilde{R}_{11}^3\right] - 3\cdot\mathrm{E}\left[R_2\tilde{R}_{11}\right]\mathrm{E}\left[\tilde{R}_{11}^2\right]\right\} + O\left(\bar{n}^{-3}\left\|\Delta_{\mathcal{U}}\right\| + \bar{n}^{-4}\right). \quad (40)\end{aligned}$$

And by standard calculations,

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

and the $O\left(\bar{n}^{-3}\left\|\Delta_{\mathcal{U}}\right\| h\right)$ remainder collects all terms that depend on $\alpha^1, ..., \alpha^{2d_u}$. Similarly, we also calculate $\mathrm{E}\left[R_2\tilde{R}_{11}^3\right] - 3\cdot\mathrm{E}\left[R_2\tilde{R}_{11}\right]\mathrm{E}\left[\tilde{R}_{11}^2\right]$, $\mathrm{E}\left[R_2^2\tilde{R}_{11}^2\right] - \mathrm{E}\left[R_2^2\right]\mathrm{E}\left[\tilde{R}_{11}^2\right]$ and $\mathrm{E}\left[R_2\tilde{R}_{11}^3\right] - 3\cdot\mathrm{E}\left[R_2\tilde{R}_{11}\right]\mathrm{E}\left[\tilde{R}_{11}^2\right]$ on the right hand side of the second equality in (40), ignore small-order terms that depend on $\alpha^1, ..., \alpha^{2d_u}$ 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 (40) is exactly zero so that it follows from this result and (40) that

$\kappa_4\left(\sqrt{\bar{n}}R\right) = O\left(\bar{n}^{-1}\left\|\Delta_{\mathcal{U}}\right\| + \bar{n}^{-2}\right)$. By previous calculations and $B_n = 1 + O\left(\left\|\Delta_{\mathcal{U}}\right\|\right)$, we get the approximate cumulants for $f_n\left(S_n\right)$: $\kappa_1\left(f_n\left(S_n\right)\right) = B_n^{-1}\tilde{\kappa}_{1,n} + O\left(\bar{n}^{-1/2}\left\|\Delta_{\mathcal{U}}\right\| h + \bar{n}^{-3/2}\right)$, $\kappa_2\left(f_n\left(S_n\right)\right) = B_n^{-2}\tilde{\kappa}_{2,n} + O\left(\left\|\Delta_{\mathcal{U}}\right\|^2 + \bar{n}^{-1}\left\|\Delta_{\mathcal{U}}\right\| + \bar{n}^{-2}\right)$, $\kappa_3\left(f_n\left(S_n\right)\right) = O\left(\left\|\Delta_{\mathcal{U}}\right\|/\sqrt{\bar{n}} + \bar{n}^{-3/2}\right)$ and $\kappa_4\left(f_n\left(S_n\right)\right) = O\left(\bar{n}^{-1}\left\|\Delta_{\mathcal{U}}\right\| + \bar{n}^{-2}\right)$.

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

$$\Pr\left[\bar{n}\left(R_0 + R\right)^2 \leq x\right] = \int_{\left|t + \left(\sqrt{\bar{n}}R_0\right)/B_n\right| \leq \sqrt{x}/B_n} \phi\left(t \mid B_n^{-1}\tilde{\kappa}_{1,n}, B_n^{-2}\tilde{\kappa}_{2,n}\right) \mathrm{d}t + O\left(\left\|\Delta_{\mathcal{U}}\right\|/\sqrt{\bar{n}} + \bar{n}^{-3/2}\right), \quad (41)$$

uniformly in $x > 0$. By using the recurrence properties of non-central χ^2 (Cohen, 1988) and mean value expansion, we have $\partial F\left(x \mid \lambda\right)/\partial\lambda\big|_{\lambda=\bar{\lambda}} = -x f_{\chi^2_1}\left(x\right) + O\left(\bar{\lambda}\right)$. By this result, $B_n^2 = 1 + O\left(\left\|\Delta_{\mathcal{U}}\right\|\right)$, change of variables and mean value expansion,

$$\begin{aligned}
&\int_{\left|t + \left(\sqrt{\bar{n}}R_0\right)/B_n\right| \leq \sqrt{x}/B_n} \phi\left(t \mid B_n^{-1}\tilde{\kappa}_{1,n}, B_n^{-2}\tilde{\kappa}_{2,n}\right) \mathrm{d}t = \int_{|t| \leq \sqrt{x/\tilde{\kappa}_{2,n}}} \phi\left(t \mid \left(\sqrt{\bar{n}}R_0 + \tilde{\kappa}_{1,n}\right)/\sqrt{\tilde{\kappa}_{2,n}}, 1\right) \mathrm{d}t \\
&= F\left(\frac{x}{\tilde{\kappa}_{2,n}} \mid \frac{\left(\sqrt{\bar{n}}R_0 + \tilde{\kappa}_{1,n}\right)^2}{\tilde{\kappa}_{2,n}}\right) = F_{\chi^2_1}\left(x\right) - x f_{\chi^2_1}\left(x\right)\left(\left(\sqrt{\bar{n}}\tilde{R}_{10} + \tilde{\kappa}_{1,n}\right)^2 + \tilde{\kappa}_{21,n} + \tilde{\kappa}_{22,n}\right) + O\left(\nu_n^\sharp\right). \quad (42)
\end{aligned}$$

By (41) and (42),

$$\Pr\left[\bar{n}\left(R_0 + R\right)^2 \leq x\right] = F_{\chi^2_1}\left(x\right) - \tilde{\mathscr{C}}_p^{\mathsf{pre}}\left(n, h\right) x f_{\chi^2_1}\left(x\right) + O\left(\nu_n^\sharp\right), \quad (43)$$

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

We consider the singular value decomposition of $\Delta_{\bar{\mathcal{U}}\bar{\mathcal{U}}^\top}^{-1/2}\left(-\Delta_{\bar{\mathcal{G}}}\right)$ such that $\bar{\mathrm{S}}^\top \Delta_{\bar{\mathcal{U}}\bar{\mathcal{U}}^\top}^{-1/2}\left(-\Delta_{\bar{\mathcal{G}}}\right)\bar{\mathrm{T}} = \begin{bmatrix} \bar{\Lambda} & 0_{d_z \times d_z} \end{bmatrix}^\top$ where $\bar{\mathrm{S}}^\top\bar{\mathrm{S}} = \mathrm{I}_{2d_z}$, $\bar{\mathrm{T}}^\top\bar{\mathrm{T}} = \mathrm{I}_{d_z}$ and $\bar{\Lambda}$ is a d_z-dimensional diagonal matrix. We apply the rotation by $\bar{\mathcal{V}}_i\left(\theta_2\right) \coloneqq \bar{\Gamma}\bar{\mathcal{U}}_i\left(\theta_2\right)$ where $\bar{\Gamma} \coloneqq \bar{\mathrm{S}}^\top \Delta_{\bar{\mathcal{U}}\bar{\mathcal{U}}^\top}^{-1/2}$ so that $\ell_p\left(\widehat{\vartheta}_p \mid h\right) = \inf_{\theta_2}\sup_{\lambda_2} 2\sum_i \log\left(1 + \lambda_2^\top \bar{\mathcal{V}}_i\left(\theta_2\right)\right)$ and calculations from Matsushita and Otsu (2013) can be applied. Also denote $\bar{\mathcal{V}}_i \coloneqq \bar{\Gamma}\bar{\mathcal{U}}_i$, $\bar{\mathcal{H}}_i \coloneqq \bar{\Gamma}\left(-\bar{\mathcal{G}}_i\right)$ ($\bar{\mathcal{V}}$ and $\bar{\mathcal{H}}$

defined similarly) and $\bar{\Omega} \coloneqq \left(\bar{\Lambda}\bar{\mathrm{T}}^{\top}\right)^{-1}$. Then it follows that $\Delta_{\bar{\mathcal{V}}\bar{\mathcal{V}}^{\top}} = \mathrm{I}_{2d_z \times 2d_z}$ and

$$\begin{bmatrix} -\Delta_{\bar{\mathcal{V}}\bar{\mathcal{V}}^{\top}} & \Delta_{\bar{\mathcal{H}}} \\ \Delta_{\bar{\mathcal{H}}}^{\top} & 0_{d_z \times d_z} \end{bmatrix}^{-1} = \begin{bmatrix} 0_{d_z \times d_z} & 0_{d_z \times d_z} & \bar{\Omega}^{\top} \\ 0_{d_z \times d_z} & -\mathrm{I}_{d_z} & 0_{d_z \times d_z} \\ \bar{\Omega} & 0_{d_z \times d_z} & \bar{\Omega}\bar{\Omega}^{\top} \end{bmatrix} = \begin{bmatrix} -\left(\bar{\Gamma}^{\top}\right)^{-1} \bar{\mathrm{Q}} \bar{\Gamma}^{-1} & -\left(\bar{\Gamma}^{\top}\right)^{-1} \bar{\mathrm{N}} \\ -\bar{\mathrm{N}}^{\top} \bar{\Gamma}^{-1} & \bar{\mathrm{O}} \end{bmatrix}. \tag{44}$$

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

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

and the expression of $\tilde{R}_{\ddagger 3}^{d_z+a}$ is omitted for brevity (see Matsushita and Otsu, 2013, A.1). Let $\mathring{A}_{\ddagger}^{\mathsf{a}} \coloneqq A_{\ddagger}^{\mathsf{a}} - \alpha_{\ddagger}^{\mathsf{a}}$. We again replace $A_{\ddagger}^{\mathsf{a}}$ by $\mathring{A}_{\ddagger}^{\mathsf{a}} + \alpha_{\ddagger}^{\mathsf{a}}$ to obtain $\tilde{R}_{\ddagger 1}^{d_z+a} = \tilde{R}_{\ddagger 11}^{d_z+a} + \tilde{R}_{\ddagger 10}^{d_z+a}$, $\tilde{R}_{\ddagger 2}^{d_z+a} = \tilde{R}_{\ddagger 22}^{d_z+a} + \tilde{R}_{\ddagger 21}^{d_z+a} + \tilde{R}_{\ddagger 20}^{d_z+a}$ and $\tilde{R}_{\ddagger 3}^{d_z+a} = \tilde{R}_{\ddagger 33}^{d_z+a} + \tilde{R}_{\ddagger 32}^{d_z+a} + \tilde{R}_{\ddagger 31}^{d_z+a} + \tilde{R}_{\ddagger 30}^{d_z+a}$. Then by standard calculations, $\mathrm{E}\left[\bar{n}^{-1}\widehat{\ell}^{\star}\right]$ is equal to the sum of $\tilde{R}_{\ddagger 10}^{d_z+a}\tilde{R}_{\ddagger 10}^{d_z+a}$, $\tilde{R}_{\ddagger 10}^{d_z+a}\mathrm{E}\left[\tilde{R}_{\ddagger 22}^{d_z+a}\right]$, $\mathrm{E}\left[\tilde{R}_{\ddagger 11}^{d_z+a}\tilde{R}_{\ddagger 21}^{d_z+a}\right]$, $\mathrm{E}\left[\tilde{R}_{\ddagger 11}^{d_z+a}\tilde{R}_{\ddagger 11}^{d_z+a}\right]$ and $2 \cdot \mathrm{E}\left[\tilde{R}_{\ddagger 11}^{d_z+a}\tilde{R}_{\ddagger 22}^{d_z+a}\right] + 2 \cdot \mathrm{E}\left[\tilde{R}_{\ddagger 11}^{d_z+a}\tilde{R}_{\ddagger 33}^{d_z+a}\right] + \mathrm{E}\left[\tilde{R}_{\ddagger 22}^{d_z+a}\tilde{R}_{\ddagger 22}^{d_z+a}\right]$ with an $o\left(\upsilon_n^{\natural}\right)$ remainder term. By inverting using the second equality of (44), $\tilde{R}_{\ddagger 10}^{d_z+a}\tilde{R}_{\ddagger 10}^{d_z+a} = \alpha_{\ddagger}^{d_z+a}\alpha_{\ddagger}^{d_z+a} = \bar{\mathrm{Q}}^{(\mathsf{ab})}\bar{\Upsilon}^{\mathsf{a}}\bar{\Upsilon}^{\mathsf{b}}$. By calculation and $\Delta_{\bar{\mathcal{V}}\bar{\mathcal{V}}^{\top}} = \mathrm{I}_{2d_z \times 2d_z}$, $\mathrm{E}\left[\tilde{R}_{\ddagger 11}^{d_z+a}\tilde{R}_{\ddagger 11}^{d_z+a}\right] = \bar{n}^{-1}d_z + O\left(\left\|\Delta_{\mathcal{U}}\right\|^2/n\right)$. It is easy to calculate that $\mathrm{E}\left[\tilde{R}_{\ddagger 22}^{d_z+a}\right] = -\bar{n}^{-1}\alpha_{\ddagger}^{d_z+a\,d_z+b\,d_z+b}/6 - \bar{\Omega}^{(st)}\gamma_{\ddagger}^{t;d_z+a,s} + O\left(\left\|\Delta_{\mathcal{U}}\right\|/n\right)$. Then by (44),

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

By calculation and using (44), $\mathrm{E}\left[\tilde{R}_{\ddagger 11}^{d_z+a}\tilde{R}_{\ddagger 21}^{d_z+a}\right] = \bar{n}^{-1}\bar{\Upsilon}^{\mathsf{abc}}\bar{\mathrm{Q}}^{(\mathsf{ab})}\bar{\mathrm{Q}}^{(\mathsf{cd})}\bar{\Upsilon}^{\mathsf{d}}/6 + o\left(\upsilon_n^{\natural}/\bar{n}\right)$. By calculation in Matsushita and Otsu (2013, A.4),

$$2 \cdot \mathrm{E}\left[\tilde{R}_{\ddagger 11}^{d_z+a}\tilde{R}_{\ddagger 22}^{d_z+a}\right] + 2 \cdot \mathrm{E}\left[\tilde{R}_{\ddagger 11}^{d_z+a}\tilde{R}_{\ddagger 33}^{d_z+a}\right] + \mathrm{E}\left[\tilde{R}_{\ddagger 22}^{d_z+a}\tilde{R}_{\ddagger 22}^{d_z+a}\right] = \bar{n}^{-2}\sum_{j=1}^{8}\bar{\kappa}_{\ddagger 2j} + o\left(\upsilon_n^{\natural}/\bar{n}\right),$$

where the constants are defined by

$$\left(\bar{\kappa}_{\dagger 21}, \bar{\kappa}_{\ddagger 22}, \bar{\kappa}_{\ddagger 23}, \bar{\kappa}_{\ddagger 24}, \bar{\kappa}_{\ddagger 25}, \bar{\kappa}_{\ddagger 26}, \bar{\kappa}_{\ddagger 27}, \bar{\kappa}_{\ddagger 28}\right) \coloneqq \left(\frac{1}{2}\bar{\Upsilon}^{\mathsf{abcd}}\bar{\mathrm{Q}}^{(\mathsf{ab})}\bar{\mathrm{Q}}^{(\mathsf{cd})}, -\frac{1}{3}\bar{\Upsilon}^{\mathsf{abc}}\bar{\mathrm{Q}}^{(\mathsf{ad})}\bar{\mathrm{Q}}^{(\mathsf{be})}\bar{\mathrm{Q}}^{(\mathsf{cf})}\bar{\Upsilon}^{\mathsf{def}},\right.$$

$$2\bar{\Gamma}^{\mathsf{a};\mathsf{b};\mathsf{c},s}\bar{\mathrm{N}}^{(\mathsf{a}s)}\bar{\mathrm{Q}}^{(\mathsf{bc})}, -\bar{\Gamma}^{\mathsf{a};\mathsf{b},s}\bar{\mathrm{Q}}^{(\mathsf{ac})}\bar{\mathrm{Q}}^{(\mathsf{bd})}\bar{\mathrm{N}}^{(\mathsf{e}s)}\bar{\Upsilon}^{\mathsf{cde}}, -\bar{\Gamma}^{\mathsf{a},s;\mathsf{b},t}\bar{\mathrm{Q}}^{(\mathsf{ab})}\bar{\mathrm{O}}^{(st)}$$
$$\bar{\Gamma}^{\mathsf{a};\mathsf{c},s}\bar{\mathrm{Q}}^{(\mathsf{ab})}\bar{\mathrm{Q}}^{(\mathsf{cd})}\bar{\mathrm{O}}^{(st)}\bar{\Gamma}^{\mathsf{b};\mathsf{d},t}, -\bar{\Gamma}^{\mathsf{a};\mathsf{c},s}\bar{\mathrm{N}}^{(\mathsf{a}t)}\bar{\mathrm{Q}}^{(\mathsf{cd})}\bar{\mathrm{N}}^{(\mathsf{b}s)}\bar{\Gamma}^{\mathsf{b};\mathsf{d},t}, \bar{\Gamma}^{\mathsf{a};\mathsf{c},s}\bar{\mathrm{N}}^{(\mathsf{a}s)}\bar{\mathrm{Q}}^{(\mathsf{cd})}\bar{\mathrm{N}}^{(\mathsf{b}t)}\bar{\Gamma}^{\mathsf{b};\mathsf{d},t}\Big).$$

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

$$\mathrm{E}\left[\bar{n}\widehat{\ell}^{\star}\right] = d_z + \bar{n}\mathscr{B}_p^{\ddagger} - \bar{\Gamma}^{\mathsf{a};\mathsf{b},s}\bar{\mathrm{N}}^{(\mathsf{a}s)}\bar{\mathrm{Q}}^{(\mathsf{bc})}\bar{\Upsilon}^{\mathsf{c}} + \bar{n}^{-1}\sum_{j=1}^{8}\bar{\kappa}_{\ddagger 2j} + o\left(\upsilon_n^{\natural}\right).$$

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

$$\tilde{\mathscr{C}}_p^{\mathsf{pre}}(n,h) = \bar{n}\left(\mathscr{B}_p^{\dagger} - \mathscr{B}_p^{\ddagger}\right) - \Gamma_{\dagger}^{\mathsf{k};\mathsf{l},u}\mathrm{N}_{\dagger}^{(\mathsf{k}u)}\mathrm{Q}_{\dagger}^{(\mathsf{lm})}\Upsilon^{\mathsf{m}} + \bar{\Gamma}^{\mathsf{a};\mathsf{b},s}\bar{\mathrm{N}}^{(\mathsf{a}s)}\bar{\mathrm{Q}}^{(\mathsf{bc})}\bar{\Upsilon}^{\mathsf{c}} + \bar{n}^{-1}\sum_{j=1}^{8}\left(\bar{\kappa}_{\dagger 2j} - \bar{\kappa}_{\ddagger 2j}\right).$$

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

It is easily seen that the result (41) with the weak inequality replaced by a strict inequality still holds (see Skovgaard, 1981, Theorem 3.2). By $LR^{\star} = \bar{n}\left(R_0 + R\right)^2 + O_p^{\star}\left(\upsilon_n^{\sharp}\right)$ and the fact (30),

$$\left|\Pr\left[LR^{\star} \leq x\right] - \Pr\left[\bar{n}\left(R_0 + R\right)^2 \leq x\right]\right| \leq \Pr\left[\left|\bar{n}\left(R_0 + R\right)^2 - x\right| \leq c_1\upsilon_n^{\sharp}\right] + c_2\left(\log(n)/\bar{n}^{3/2}\right) = O\left(\upsilon_n^{\sharp}\right), \tag{45}$$

where the equality follows from (41) and boundedness of $\phi(\cdot \mid \tilde{\kappa}_{1,n}, \tilde{\kappa}_{2,n})$. The conclusion follows from (43), (45) and $\tilde{\mathscr{C}}_p^{\mathsf{pre}}(n,h) = \mathscr{C}_p^{\mathsf{pre}}(n,h) + O\left(\|\Delta_{\mathcal{U}}\| h + n^{-1}\right)$. ■

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

$$\begin{aligned}
\Upsilon^{\mathsf{klmn}}\mathrm{Q}_{\dagger}^{(\mathsf{kl})}\mathrm{Q}_{\dagger}^{(\mathsf{mn})} &= \mathrm{Q}_{\dagger 11}^{(\mathsf{uw})}\mathrm{tr}\left(\mathrm{Q}_{\dagger 11}\Pi_{+}^{\mathsf{uw}}\right) + \mathrm{Q}_{\dagger 22}^{(\mathsf{uw})}\mathrm{tr}\left(\mathrm{Q}_{\dagger 22}\Pi_{-}^{\mathsf{uw}}\right) \\
\Upsilon^{\mathsf{klm}}\mathrm{Q}_{\dagger}^{(\mathsf{kn})}\mathrm{Q}_{\dagger}^{(\mathsf{lo})}\mathrm{Q}_{\dagger}^{(\mathsf{mv})}\Upsilon^{\mathsf{nov}} &= \mathrm{Q}_{\dagger 11}^{(\mathsf{uw})}\mathrm{tr}\left(\mathrm{Q}_{\dagger 11}\Pi_{+}^{\mathsf{u}}\mathrm{Q}_{\dagger 11}\Pi_{+}^{\mathsf{w}}\right) + \mathrm{Q}_{\dagger 22}^{(\mathsf{uw})}\mathrm{tr}\left(\mathrm{Q}_{\dagger 22}\Pi_{-}^{\mathsf{u}}\mathrm{Q}_{\dagger 22}\Pi_{-}^{\mathsf{w}}\right) + 2\mathrm{Q}_{\dagger 21}^{(\mathsf{uw})}\mathrm{tr}\left(\mathrm{Q}_{\dagger 12}\Pi_{-}^{\mathsf{u}}\mathrm{Q}_{\dagger 21}\Pi_{+}^{\mathsf{w}}\right)
\end{aligned}$$

$$\begin{aligned}\Gamma_{\dagger}^{\mathsf{k};\mathsf{l};\mathsf{m},w}\mathrm{N}_{\dagger}^{(\mathsf{k}w)}\mathrm{Q}_{\dagger}^{(\mathsf{l}\mathsf{m})} &= \operatorname{tr}\left(\mathrm{Q}_{\dagger 11}\Delta_{W_{p;+}\mathcal{U}_{+}\mathcal{U}_{+}^{\top}}\mathrm{N}_{\dagger 1}\right)+\operatorname{tr}\left(\mathrm{Q}_{\dagger 22}\Delta_{W_{p;-}\mathcal{U}_{-}\mathcal{U}_{-}^{\top}}\mathrm{N}_{\dagger 2}\right)\\ \Gamma_{\dagger}^{\mathsf{k},u;\mathsf{l},w}\mathrm{Q}_{\dagger}^{(\mathsf{k}\mathsf{l})}\mathrm{O}_{\dagger}^{(uw)} &= \operatorname{tr}\left(\Delta_{W_{p;+}^{2}}\mathrm{Q}_{\dagger 11}\mathrm{O}_{\dagger}\right)+\operatorname{tr}\left(\Delta_{W_{p;-}^{2}}\mathrm{Q}_{\dagger 22}\mathrm{O}_{\dagger}\right).\end{aligned}$$

By Lemma 1, $\mathrm{Q}_{\dagger 11}=\Xi_1/\left(\varphi\omega_{p;+}^{0,2}\right)+O\left(h\right)$, $\mathrm{Q}_{\dagger 22}=\Xi_1/\left(\varphi\omega_{p;+}^{0,2}\right)+O\left(h\right)$, $\mathrm{Q}_{\dagger 21}=-\Xi_1/\left(\varphi\omega_{p;+}^{0,2}\right)+O\left(h\right)$, $\mathrm{O}_{\dagger}=\left(\omega_{p;+}^{0,2}/\varphi\right)\Xi_2+O\left(h\right)$, $\mathrm{N}_{\dagger 1}=\mu_{UU^{\top},+}^{-1}\Xi_2/\varphi+O\left(h\right)$ and $\mathrm{N}_{\dagger 2}=\mu_{UU^{\top},-}^{-1}\Xi_2/\varphi+O\left(h\right)$. It follows that $\mathscr{V}_{p,1}^{\dagger}=\left(\left(\omega_{p;+}^{0,4}/\omega_{p;+}^{0,2}\right)\Xi_1^{(\mathsf{u}\mathsf{w})}\Psi_1^{\mathsf{u}\mathsf{w}}\right)/\left(2\varphi\omega_{p;+}^{0,2}\right)+O\left(h\right)$, $\mathscr{V}_{p,2}^{\dagger}=\left(-\left(\omega_{p;+}^{0,3}/\omega_{p;+}^{0,2}\right)^{2}\Xi_1^{(\mathsf{u}\mathsf{w})}\Psi_2^{\mathsf{u}\mathsf{w}}\right)/\left(3\varphi\omega_{p;+}^{0,2}\right)+O\left(h\right)$, $\mathscr{V}_{p,3}^{\dagger}=\left(4\omega_{p;+}^{0,3}\operatorname{tr}\left(\Xi_1\Xi_2\right)\right)/\left(\varphi\omega_{p;+}^{0,2}\right)+O\left(h\right)$ and $\mathscr{V}_{p,4}^{\dagger}=\left(-2\omega_{p;+}^{0,2}\operatorname{tr}\left(\Xi_1\Xi_2\right)\right)/\varphi+O\left(h\right)$. Similar results hold for $\left(\mathscr{V}_{p,1}^{\ddagger},\mathscr{V}_{p,2}^{\ddagger},\mathscr{V}_{p,3}^{\ddagger},\mathscr{V}_{p,4}^{\ddagger}\right)$. By tedious algebra, it can be verified that $\mathscr{B}_p^{\ddagger}=\mathrm{Q}^{(\mathsf{k}\mathsf{l})}\Upsilon^{\mathsf{k}}\Upsilon^{\mathsf{l}}$. By (25) and simple algebra,

$$\mathscr{B}_p^{\dagger}-\mathscr{B}_p^{\ddagger}=\left\{\left(\Delta_{\mathcal{M}_+}/\Delta_+-\Delta_{\mathcal{M}_-}/\Delta_-\right)-\left(\Delta_{\mathcal{Z}_+}/\Delta_+-\Delta_{\mathcal{Z}_-}/\Delta_-\right)^{\top}\gamma_{\Delta}\right\}^{2}/\Sigma_{\Delta}=\left(\mathscr{B}_p^{\mathsf{EL}}h^{p+1}\right)^{2}/\mathscr{V}_p^{\mathsf{EL}}+O\left(h^{2p+3}\right). \tag{46}$$

It follows that $\mathscr{C}_p^{\mathsf{pre}}\left(n,h\right)=\mathscr{C}_p\left(n,h\right)+O\left(\bar{n}\left\|\Delta_{\mathcal{U}}\right\|^{2}h+n^{-1}\right)$ and $\Pr\left[LR^{\star}\leq x\right]=F_{\chi_1^2}\left(x\right)-\mathscr{C}_p\left(n,h\right)xf_{\chi_1^2}\left(x\right)+O\left(v_n^{\sharp}+\bar{n}\left\|\Delta_{\mathcal{U}}\right\|^{2}h\right)$. By $LR_p\left(\vartheta_0\mid h\right)=LR^{\star}+O_p^{\star}\left(\log\left(n\right)^{5/2}/\bar{n}^{3/2}\right)$ and (45) with $\left(LR^{\star},\bar{n}\left(R_0+R\right)^{2}\right)$ replaced by $\left(LR_p\left(\vartheta_0\mid h\right),LR^{\star}\right)$, we get the first conclusion. The second conclusion follows from the first one, $\Pr\left[LR_p^{\mathsf{c}}\left(\vartheta_0\mid h\right)\leq x\right]=\Pr\left[LR_p\left(\vartheta_0\mid h\right)\leq x\left(1+\mathscr{C}_p\left(n,h\right)\right)\right]$ and Taylor expansion. ■

Proof of Theorem 4. We redefine some notations for notational simplicity: $\theta\coloneqq\left(\theta_0,\theta_1,\theta_2,\theta_3\right)$, $\vartheta\coloneqq\left(\vartheta_0,\vartheta_1,\vartheta_2,\vartheta_3\right)$, $\mathcal{U}_i\left(\theta\right)\coloneqq\left(W_{p;+,i}U_i\left(\theta_0,\theta_1,\theta_2\right)^{\top},W_{p;-,i}U_i\left(\theta_0,\theta_1,\theta_3\right)^{\top}\right)^{\top}$ and $\mathcal{U}_i\coloneqq\mathcal{U}_i\left(\vartheta\right)$ ($\mathcal{U}$ defined similarly). Terms that depend on $\left(\mathcal{V}_i\left(\theta\right),\mathcal{V}_i,\mathcal{V}\right)$ are redefined accordingly. As in the proof of Theorem 3, we apply the rotation by Γ so that $\ell_p\left(\theta\mid h\right)=\sup_{\lambda}2\sum_i\log\left(1+\lambda^{\top}\mathcal{V}_i\left(\theta\right)\right)$. Let $\mathcal{V}_{\delta,i}\coloneqq\mathcal{V}_i\left(\vartheta_0,\vartheta_1,\vartheta_2,\vartheta_2\right)$. It is easy to check that Lemma 6 still holds under $\mathscr{T}_Z=\delta l_n$. (36) and (37) with $\mathcal{V}_i$ replaced by $\mathcal{V}_{\delta,i}$ under $\mathscr{T}_Z=\delta l_n$. Similarly, (36) and (37) with $\left(\mathcal{V}_i,\mathcal{H}_i\right)$ replaced by $\left(\mathcal{V}_{\delta,i},\mathcal{H}_{\dagger,i}\right)$ hold for $\left(\widetilde{\lambda}_p,\widetilde{\eta}_p\right)$ and $\bar{n}^{-1}\ell_p\left(\vartheta_0,\widetilde{\vartheta}_p\mid h\right)$ under $\mathscr{T}_Z=\delta l_n$. Let $\left(\tilde{R}_1^{\delta},\tilde{R}_2^{\delta}\right)$ be defined by the formulae of $\left(\tilde{R}_1,\tilde{R}_2\right)$ 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), $LR_p\left(\vartheta_0\mid h\right)=\bar{n}\left(\tilde{R}_1^{\delta}+\tilde{R}_2^{\delta}\right)^{2}+O_p^{\star}\left(\log\left(n\right)^{2}/\bar{n}\right)$ under $\mathscr{T}_Z=\delta l_n$. Let $A_{\delta}^{\mathsf{k}}\coloneqq\bar{n}^{-1}\sum_i\mathcal{V}_{\delta,i}^{(\mathsf{k})}$ and $A_{\delta}^{\mathsf{k}\mathsf{l}}\coloneqq\bar{n}^{-1}\sum_i\mathcal{V}_{\delta,i}^{(\mathsf{k})}\mathcal{V}_{\delta,i}^{(\mathsf{l})}-\alpha^{\mathsf{k}\mathsf{l}}$. Let $\tilde{\mathcal{H}}_i\coloneqq\partial\mathcal{V}_i\left(\theta\right)/\partial\theta^{\top}$ and let $\left(\tilde{C}^{\mathsf{k},\mathsf{m}},\tilde{\gamma}^{\mathsf{k},\mathsf{m}},\tilde{C}^{\mathsf{k};\mathsf{l},\mathsf{m}},\tilde{\gamma}^{\mathsf{k};\mathsf{l},\mathsf{m}}\right)$ be defined by the formulae of $\left(C^{\mathsf{k},n}\gamma^{\mathsf{k},n},C^{\mathsf{k};\mathsf{l},n},\gamma^{\mathsf{k};\mathsf{l},n}\right)$ with $\left(\mathcal{H}_i,\mathcal{H}\right)$ replaced by $\left(\tilde{\mathcal{H}}_i,\tilde{\mathcal{H}}\right)$. Denote $\delta_n\coloneqq\left(\mu_{D,+}-\mu_{D,-}\right)\delta l_n$ for simplicity. It is easy to see that $A_{\delta}^{\mathsf{k}}=A^{\mathsf{k}}+\delta_n^{(a)}\left(\tilde{C}^{\mathsf{k},d_{\vartheta}+a}+\tilde{\gamma}^{\mathsf{k},d_{\vartheta}+a}\right)$ and $A_{\delta}^{\mathsf{k}\mathsf{l}}=A^{\mathsf{k}\mathsf{l}}+\delta_n^{(a)}\left(\tilde{C}^{\mathsf{k};\mathsf{l},d_{\vartheta}+a}+\tilde{\gamma}^{\mathsf{k};\mathsf{l},d_{\vartheta}+a}\right)[\mathsf{k},\mathsf{l}]+O_p^{\star}\left(l_n^{2}\right)$. By using these results and replacing A^{k} with $\mathring{A}^{\mathsf{k}}+\alpha^{\mathsf{k}}$, we decompose $\tilde{R}_1^{\delta}=\tilde{R}_{11}^{\delta}+\tilde{R}_{10}^{\delta}$, where $\tilde{R}_{11}^{\delta}\coloneqq\varpi^{(k)}\left(\mathring{A}^{k}+\delta_n^{(a)}\tilde{C}^{k,d_{\vartheta}+a}\right)$ and $\tilde{R}_{10}^{\delta}\coloneqq\varpi^{(k)}\tilde{\gamma}^{k,d_{\vartheta}+a}\delta_n^{(a)}$. Note that by Lemma 1, $\alpha^{\mathsf{k}}=\mathrm{e}_{2d_u,\mathsf{k}}^{\top}\Gamma\Delta_{\mathcal{U}}=O\left(n^{-1}\right)$. Similarly, we write $\tilde{R}_2^{\delta}=\tilde{R}_{20}^{\delta}+\tilde{R}_{21}^{\delta}+R_2+O_p^{\star}\left(\log\left(n\right)l_n^{3}\right)$, where

$$\tilde{R}_{20}^{\delta} \coloneqq \left\{ \frac{1}{2}\varpi^{(m)}\mathrm{M}^{(nl)} \left(\tilde{\gamma}^{m;n,d_\vartheta+a}[m,n]\right) \tilde{\gamma}^{l,d_\vartheta+b} - \varpi^{(n)}\tilde{\gamma}^{n;d_\vartheta+c,d_\vartheta+b}\tilde{\gamma}^{d_\vartheta+c,d_\vartheta+a} \right.$$
$$+ \left(\frac{1}{3}\varpi^{(k)}\mathrm{M}^{(mv)}\mathrm{M}^{(nl)}\alpha^{kmn} - \varpi^{(n)}\mathrm{M}^{(mv)}\mathrm{P}^{(ol)}\gamma^{m;n,k}\Omega^{(ko)} \right) \tilde{\gamma}^{v,d_\vartheta+a}\tilde{\gamma}^{l,d_\vartheta+b}$$
$$\left. \left(\left(\gamma^{d_\vartheta+c;v,m}[d_\vartheta+c,v]\right)\Omega^{(mo)}\mathrm{P}^{(ok)}\varpi^{(v)} - \alpha^{vm\,d_\vartheta+c}\mathrm{M}^{(vk)}\varpi^{(m)} \right) \tilde{\gamma}^{k,d_\vartheta+b}\tilde{\gamma}^{d_\vartheta+c,d_\vartheta+a} \right\} \delta_n^{(a)}\delta_n^{(b)}, \quad (47)$$

$\tilde{R}_{21}^{\delta}$ is defined by the sum of $\varpi^{(m)}\mathrm{M}^{(nk)} \left(\tilde{\gamma}^{m;n,d_\vartheta+a}[m,n]\right) \mathring{A}^k\delta_n^{(a)}/2 - \varpi^{(n)} \left(\tilde{\gamma}^{d_\vartheta+a;n,d_\vartheta+b}[d_\vartheta+a,n]\right) \mathring{A}^{d_\vartheta+a}\delta_n^{(b)}$ and the right hand side of (39) with α^{k} replaced by $\tilde{\gamma}^{\mathsf{k},d_\vartheta+a}\delta_n^{(a)}$ and R_2 is defined in the proof of Lemma 7. Let $R_0^{\delta} \coloneqq \tilde{R}_{10}^{\delta}+\tilde{R}_{20}^{\delta}$, $R_1^{\delta} \coloneqq \tilde{R}_{11}^{\delta}+\tilde{R}_{21}^{\delta}$ and $R^{\delta} \coloneqq R_1^{\delta}+R_2$ so that $LR_p\left(\vartheta_0 \mid h\right) = \bar{n}\left(R_0^{\delta}+R^{\delta}\right)^2 + O_p^{\star}\left(\log(n)^{3/2} l_n^2\right)$. Denote $\bar{\kappa}_0^{ab} \coloneqq -\mathrm{M}^{(kl)}\tilde{\gamma}^{l,d_\vartheta+a}\tilde{\gamma}^{k,d_\vartheta+b}$, $\beta_n^{\delta} \coloneqq \bar{n}\bar{\kappa}_0^{ab}\delta_n^{(a)}\delta_n^{(b)}$ and $\bar{\sigma}_p^2 \coloneqq \Sigma_{\Delta}/\left(\mu_{D,+}-\mu_{D,-}\right)^2$. By (33) and (34),

$$\bar{\kappa}_0^{ab}\delta_n^{(a)}\delta_n^{(b)} = \left(\mu_{D,+}-\mu_{D,-}\right)^2 \left(\mathrm{Q}_{\dagger}-\mathrm{Q}\right)^{(d_\vartheta+a\,d_\vartheta+b)} \delta^{(d_\vartheta+a)}\delta^{(d_\vartheta+b)}\Delta_{-}^2 l_n^2 = \frac{\left(\gamma_{\Delta}^{\top}\delta\right)^2 l_n^2}{\bar{\sigma}_p^2} \quad (48)$$

and $\bar{\sigma}_p^2 = \mathscr{V}_p^{\mathsf{EL}} + O(h)$. By (33), (34), the fact $\Omega\mathrm{P} = \left[\begin{array}{cc} 0_{d_\vartheta} & \mathrm{J} \end{array}\right]^{\top}$, tedious algebra and Lemma 1, $\left(R_0^{\delta}\right)^2 = \bar{\kappa}_0^{ab}\delta_n^{(a)}\delta_n^{(b)} + \bar{\kappa}_1^{abc}\delta_n^{(a)}\delta_n^{(b)}\delta_n^{(c)} + o\left(l_n^3\right)$ where

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

Let $\bar{\kappa}_2^a \coloneqq \alpha^{mnk}\mathrm{M}^{(km)}\mathrm{M}^{(nl)}\tilde{\gamma}^{l,d_\vartheta+a}/3$, $\bar{\kappa}_3^a \coloneqq -2\gamma^{l;d_\vartheta+b,k}\Omega^{(km)}\mathrm{M}^{(ml)}\tilde{\gamma}^{d_\vartheta+b,d_\vartheta+a}$, $\tilde{\kappa}_{2,n}^{\delta} \coloneqq 1+\left(\bar{\kappa}_2^a+\bar{\kappa}_3^a\right)\delta_n^{(a)}$ and $\bar{\kappa}_4^a \coloneqq 2\gamma^{o;n,l}\Omega^{(lv)}\mathrm{P}^{(vo)}\mathrm{M}^{(nk)}\tilde{\gamma}^{k,d_\vartheta+a}$. By calculation using arguments in the proof of Lemma 7, we have $\kappa_1\left(\sqrt{\bar{n}}R^{\delta}\right) = \tilde{\kappa}_{1,n} + o(l_n)$, $\kappa_2\left(\sqrt{\bar{n}}R^{\delta}\right) = \tilde{\kappa}_{2,n}^{\delta} + o(l_n)$ and $\kappa_3\left(\sqrt{\bar{n}}R^{\delta}\right) = o(l_n)$, where $\tilde{\kappa}_{1,n}$ is defined in the proof of Lemma 7. Then, $2\sqrt{\bar{n}}R_0^{\delta}\tilde{\kappa}_{1,n} = \left(-\bar{\kappa}_2^a+\bar{\kappa}_4^a\right)\delta_n^{(a)} + O\left(l_n^2\right)$. By arguments used to show (41) and (42) (i.e., Skovgaard, 1981 with $s=p=q=3$, $\beta_{s,n}=l_n$ and $\lambda_n = O(l_n)$),

$$\Pr\left[\bar{n}\left(R_0^{\delta}+R^{\delta}\right)^2 \leq x\right] = F\left(x/\tilde{\kappa}_{2,n}^{\delta} \mid \left(\sqrt{\bar{n}}R_0^{\delta}+\tilde{\kappa}_{1,n}\right)^2/\tilde{\kappa}_{2,n}^{\delta}\right) + o(l_n). \quad (49)$$

Then by Taylor expansion,

$$F\left(x/\tilde{\kappa}_{2,n}^{\delta} \mid \left(\sqrt{\bar{n}}R_0^{\delta}+\tilde{\kappa}_{1,n}\right)^2/\tilde{\kappa}_{2,n}^{\delta}\right) = F\left(x/\tilde{\kappa}_{2,n}^{\delta} \mid \beta_n^{\delta}\right)$$
$$+ \left\{ \bar{n}\bar{\kappa}_1^{abc}\delta_n^{(a)}\delta_n^{(b)}\delta_n^{(c)} - \beta_n^{\delta}\left(\bar{\kappa}_2^a+\bar{\kappa}_3^a\right)\delta_n^{(a)} + \left(-\bar{\kappa}_2^a+\bar{\kappa}_4^a\right)\delta_n^{(a)} \right\} F^{(1)}\left(x \mid \beta_n^{\delta}\right) + o(l_n). \quad (50)$$

Let $f\left(\cdot \mid \eta\right)$ denote the $\chi_1^2(\eta)$ PDF. By using the recurrence properties of non-central χ^2 (Cohen, 1988),

$-xf\left(x \mid \eta\right) = 2\eta F^{(2)}\left(x \mid \eta\right) + \left(\eta+1\right) F^{(1)}\left(x \mid \eta\right)$. By these results and Taylor expansion,

$$F\left(x/\tilde{\kappa}_{2,n}^{\delta} \mid \beta_n^{\delta}\right) = F\left(x \mid \beta_n^{\delta}\right) + xf\left(x \mid \beta_n^{\delta}\right)\left(1/\tilde{\kappa}_{2,n}^{\delta} - 1\right) + O\left(l_n^2\right) = F\left(x \mid \beta_n^{\delta}\right)$$
$$+ \left(2\beta_n^{\delta}\left(\bar{\kappa}_2^a + \bar{\kappa}_3^a\right)\delta_n^{(a)}\right) F^{(2)}\left(x \mid \beta_n^{\delta}\right) + \left(\beta_n^{\delta}\left(\bar{\kappa}_2^a + \bar{\kappa}_3^a\right)\delta_n^{(a)} + \left(\bar{\kappa}_2^a + \bar{\kappa}_3^a\right)\delta_n^{(a)}\right) F^{(1)}\left(x \mid \beta_n^{\delta}\right) + o\left(l_n\right). \quad (51)$$

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

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

and similarly, $\bar{\kappa}_3^a\delta_n^{(a)} = O\left(l_n h\right)$. It then follows from these results, (50) and (51) that

$$\Pr\left[\bar{n}\left(R_0^{\delta} + R^{\delta}\right)^2 \leq x\right] = F\left(x \mid \beta_n^{\delta}\right) + \left(\bar{n}\bar{\kappa}_1^{abc}\delta_n^{(a)}\delta_n^{(b)}\delta_n^{(c)}\right) F^{(1)}\left(x \mid \beta_n^{\delta}\right) + \left(2\beta_n^{\delta}\bar{\kappa}_2^a\delta_n^{(a)}\right) F^{(2)}\left(x \mid \beta_n^{\delta}\right) + o\left(l_n\right). \quad (52)$$

By (33), (34), algebra, we can find constants $\mathscr{K}_1^{abc}$ and $\mathscr{K}_2^a$ such that $\bar{n}\bar{\kappa}_1^{abc}\delta_n^{(a)}\delta_n^{(b)}\delta_n^{(c)} = \mathscr{K}_1^{abc}\delta^{(a)}\delta^{(b)}\delta^{(c)} l_n + O\left(l_n h\right)$ and $\bar{\kappa}_2^a\delta_n^{(a)} = \mathscr{K}_2^a\delta^{(a)} l_n + O\left(h\right)$. The conclusion with $\mathscr{P}_1\left(\delta\right) \coloneqq \mathscr{K}_1^{abc}\delta^{(a)}\delta^{(b)}\delta^{(c)}$ and $\mathscr{P}_2\left(\delta\right) \coloneqq 2H\gamma_{\mathsf{adj}}^{(a)}\gamma_{\mathsf{adj}}^{(b)}\mathscr{K}_2^c\delta^{(a)}\delta^{(b)}\delta^{(c)}/\mathscr{V}_p^{\mathsf{EL}}$ follows from these results, (48) and (52). ■