Journal of Econometrics 214 (2020) 379–412

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

On rank estimators in increasing dimensions

Yanqin Fan [a,*], Fang Han [b], Wei Li [c], Xiao-Hua Zhou [d,e]

[a] Department of Economics, University of Washington, Seattle, WA 98195, USA
[b] Department of Statistics, University of Washington, Seattle, WA 98195, USA
[c] School of Mathematical Sciences, Peking University, Beijing 100871, China
[d] Department of Biostatistics, University of Washington, Seattle, WA 98195, USA
[e] International Center for Mathematical Research, Peking University, Beijing, China

ARTICLE INFO

Article history:
Received 1 June 2018
Received in revised form 27 February 2019
Accepted 14 August 2019
Available online 30 August 2019

JEL classification:
C55
C14

Keywords:
Bahadur-type bounds
Degenerate U-processes
Maximal inequalities
Uniform bounds

ABSTRACT

The family of rank estimators, including Han's maximum rank correlation (Han, 1987) as a notable example, has been widely exploited in studying regression problems. For these estimators, although the linear index is introduced for alleviating the impact of dimensionality, the effect of large dimension on inference is rarely studied. This paper fills this gap via studying the statistical properties of a larger family of M-estimators, whose objective functions are formulated as U-processes and may be discontinuous in increasing dimension set-up where the number of parameters, p_n, in the model is allowed to increase with the sample size, n. First, we find that often in estimation, as $p_n/n \to 0$, $(p_n/n)^{1/2}$ rate of convergence is obtainable. Second, we establish Bahadur-type bounds and study the validity of normal approximation, which we find often requires a much stronger scaling requirement than $p_n^2/n \to 0$. Third, we state conditions under which the numerical derivative estimator of asymptotic covariance matrix is consistent, and show that the step size in implementing the covariance estimator has to be adjusted with respect to p_n. All theoretical results are further backed up by simulation studies.

1. Introduction

1.1. The general set-up, motivation, and main results

Let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n \in \mathbb{R}^{m_n}$ denote a random sample of size n from the probability measure $\mathbb{P}$. Let $\mathcal{F} := \{f(\cdot,\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^{p_n}\}$ be a class of real-valued, possibly asymmetric and *discontinuous*, functions on $\mathbb{R}^{m_n} \times \mathbb{R}^{m_n}$. This paper studies the following M-estimator with an objective function of a U-process structure,

$$\widehat{\boldsymbol{\theta}}_n := \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} \Gamma_n(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} \frac{1}{n(n-1)} \sum_{i\neq j=1}^{n} f(\boldsymbol{Z}_i, \boldsymbol{Z}_j; \boldsymbol{\theta}). \tag{1.1}$$

Let

$$\boldsymbol{\theta}_0 := \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} \Gamma(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} \mathbb{E}\Gamma_n(\boldsymbol{\theta}).$$

* Corresponding author.
E-mail addresses: fany88@uw.edu (Y. Fan), fanghan@uw.edu (F. Han), weylpeking@pku.edu.cn (W. Li), azhou@uw.edu (X.-H. Zhou).

https://doi.org/10.1016/j.jeconom.2019.08.003
0304-4076/

This paper aims to establish asymptotic properties of $\widehat{\boldsymbol{\theta}}_n$ as an estimator of $\boldsymbol{\theta}_0$ in situations with large or increasing dimensions $m_n \to \infty$ and $p_n \to \infty$ (with respect to the sample size n), to which existing results do not apply.

Members of (1.1) include the following notable examples proposed and studied in the current literature in fixed dimension, i.e., $m_n \equiv m$ and $p_n \equiv p$ for all n: (1) Han's maximum rank correlation (MRC) estimator for the generalized regression model (Han, 1987); (2) Cavanagh and Sherman's rank estimator for the same model as Han's (Cavanagh and Sherman, 1998); (3) Khan and Tamer's rank estimator for the semiparametric censored duration model (Khan and Tamer, 2007); and (4) Abrevaya and Shin's rank estimator for the generalized partially linear index model (Abrevaya and Shin, 2011). One common feature of these models is the presence of a linear index of the form $\boldsymbol{x}^\top\boldsymbol{\theta}$, where $\boldsymbol{x}$ represents covariates of dimension p which is typically large in many economic applications. The linear index structure is introduced to alleviate the "curse of dimensionality" associated with fully nonparametric models. Although motivated by possibly large dimension p, properties of $\widehat{\boldsymbol{\theta}}_n$ in these examples have only been established for fixed p when n approaches infinity (i.e., p does not change with n). Instead, this paper models the large p case by allowing p to go to infinity as $n \to \infty$, denoted as p_n, facilitating an explicit characterization of the effect of dimensionality on inference in these models.

More broadly, for the general set-up (1.1), we allow both m_n and p_n to go to infinity as $n \to \infty$ and establish the following properties of $\widehat{\boldsymbol{\theta}}_n$: (i) consistency; (ii) rate of convergence; (iii) normal approximation; and (iv) accuracy of normal approximation. The last property is also referred to as the "Bahadur–Kiefer representation" or simply the "Bahadur-type bound" (Bahadur, 1966; Kiefer, 1967; He and Shao, 1996), and is the major focus of this paper. Specifically, in Theorems 2.2, 2.3, and 2.4, under different scaling requirements for n, p_n, and ν_n, where ν_n characterizes the function complexity of $\mathcal{F}$, we prove consistency, efficient rate of convergence, and derive Bahadur-type bounds for the general M-estimator $\widehat{\boldsymbol{\theta}}_n$ of the form (1.1). To facilitate inference, we construct consistent estimators of the asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}_n$ similar to the numerical derivative estimators in Pakes and Pollard (1989), Sherman (1993), and Khan and Tamer (2007). The increasing dimension set-up in this paper reveals that for consistent variance–covariance matrix estimation, the step size in computing the numerical derivative should depend not only on the sample size n but also the dimensions m_n and p_n.

To provide further insight on the role of the dimension p_n, we apply our general results, Bahadur-type bounds especially, to the aforementioned rank estimators (1)-(4). Note that for these estimators $\nu_n = m_n = p_n$. Corollaries 3.1–3.4 provide sufficient conditions to guarantee consistency, efficient rate of convergence, and asymptotic normality (ASN) of the rank correlation estimators in increasing dimension. They demonstrate that, compared to competing alternatives such as simple linear regression, in terms of estimation, rank estimators are very appealing, maintaining the minimax optimal $(p_n/n)^{1/2}$ rates (Yu, 1997), while enjoying an additional robustness property to outliers and modeling assumptions. With regard to normal approximation, on the other hand, a much stronger scaling requirement might be needed, and a lower accuracy in normal approximation is anticipated. This observation also echoes a common belief in robust statistics that stronger scaling requirement than $p_n^2/n \to 0$ is needed for normal approximation validity (Jurečková et al., 2012).

All the theoretical results are further backed up by simulation studies. In particular, using Han's MRC estimator introduced below, we have demonstrated that for a given sample size, the accuracy of the normal approximation deteriorates quickly as the number of parameters p_n increases, indicating that our theoretical bound is difficult to improve further. Also, our simulation results suggest that for variance estimation, the step size needs to be adjusted with respect to p_n. Practically, our results indicate that although the linear index was introduced to alleviate the curse of dimensionality, one must be cautious in conducting inference using rank estimators when there are many covariates.

1.2. The generalized regression model and Han's MRC

Han's MRC in Example (1) is the first rank correlation estimator proposed to estimate the parameter $\boldsymbol{\beta}_0$ in the generalized regression model:

$$Y = D \circ F(\boldsymbol{X}^\top\boldsymbol{\beta}_0, \epsilon), \tag{1.2}$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^{p_n+1}$, $F(\cdot,\cdot)$ is a strictly increasing function of each of its arguments, and $D(\cdot)$ is a non-degenerate monotone increasing function of its argument. Important members of the generalized regression model in (1.2) include many widely known and extensively used econometrics models in diverse areas in empirical microeconomics such as the binary choice models, the ordered discrete response models, transformation models with unknown transformation functions, the censored regression models, and proportional and additive hazard models under the independence assumption and monotonicity constraints.

Han (1987) proposed estimating $\boldsymbol{\beta}_0$ in (1.2) with

$$\widehat{\boldsymbol{\beta}}_n^{\mathrm{H}} = \underset{\boldsymbol{\beta}:\beta_1=1}{\operatorname{argmax}} \Big\{ \frac{1}{n(n-1)} \sum_{i\neq j} \mathbb{1}(Y_i > Y_j)\mathbb{1}(\boldsymbol{X}_i^\top\boldsymbol{\beta} > \boldsymbol{X}_j^\top\boldsymbol{\beta}) \Big\}. \tag{1.3}$$

For model identification, following Sherman (1993), we assume the first component of $\boldsymbol{\beta}_0$ is equal to 1, and express $\boldsymbol{\beta}_0$ as $\boldsymbol{\beta}_0 = (1, \boldsymbol{\theta}_0^\top)^\top$. We consider estimating $\boldsymbol{\theta}_0$ by $\widehat{\boldsymbol{\theta}}_n^{\mathrm{H}} := \widehat{\boldsymbol{\beta}}_{n,-1}^{\mathrm{H}}$, the subvector of $\widehat{\boldsymbol{\beta}}_n^{\mathrm{H}}$ excluding its first component. We will use the generalized regression model (1.2) and Han's MRC $\widehat{\boldsymbol{\theta}}_n^{\mathrm{H}}$ to illustrate our notation, assumptions, and main results in Section 2. We defer a rigorous analysis of Han's estimator including verification of assumptions to Section 3 which also presents results for the other three rank correlation estimators.

Empirically, consider estimating the individual demand curve for a durable good such as a refrigerator. Let Y_i be whether the individual i buys a refrigerator and $\boldsymbol{X}_i$ be the vector of characteristics of the individual and the refrigerator included in the model. There are many potential candidates for the components of $\boldsymbol{X}_i$ such as personal income, marital status, the number of children, space of the kitchen, food habits; size of the refrigerator, temperature controls, lighting, shelves, dairy compartment, chiller, door styles. Assuming a single index form with $m_n = p_n + 1$, this binary choice model falls into our framework with (1.2). Our increasing dimension set-up allows more characteristics to be included in $\boldsymbol{X}_i$ as the sample size n increases and our results show that even with the single index form, estimation and inference are possible if p_n increases very mildly with n but otherwise are very challenging.

1.3. A brief review of related works and technical challenge

In contrast with the fixed dimension setting, where the model is assumed unchanged as n goes to infinity, the increasing dimension triangular array setting (Portnoy, 1984; Fan et al., 2015; Chernozhukov et al., 2015, 2017) makes our analysis different from and more challenging than most existing ones (cf. Theorem 3.2.16 and Example 3.2.22 in van der Vaart and Wellner (1996), or the main theorem in He and Shao (1996)). Technically, this paper builds on and contributes to two distinct literatures: the literature on estimation and inference in increasing dimension where existing works exclude *discontinuous* loss functions and the literature on rank estimation where existing works focus exclusively on finite dimensions. As a technical contribution, we establish a maximal inequality, yielding a uniform bound for degenerate U-processes in increasing dimensions which not only allows us to extend existing results on rank estimation in finite dimension to increasing dimensions but also establish Bahadur-type bounds. Besides the crucial role played by our new maximal inequality for degenerate U-processes in this paper, it should prove to be an indispensable tool in nonparametric and semiparametric econometrics in increasing dimensions where many estimators and test statistics are closely related to U-processes.

Since Huber's seminal paper (Huber, 1973), there has been a long history in statistics on evaluating the impact of parameter dimension on inference. Huber himself raised questions on the scaling limits of (n, p_n) for assuring M-estimation consistency and asymptotic normality in his 1973 paper (Huber, 1973). For addressing them, Portnoy (1984), Portnoy (1985), Mammen (1989, 1993) studied the linear regression model using smooth M-estimators such as the ordinary least squares. Their results revealed that, in response to Huber's question, for the simple linear regression model, asymptotic normality is usually attainable even when p_n^2/n is large. In contrast, Portnoy (1988) studied maximum likelihood estimators of generalized linear models, and proved that, for guaranteeing the validity of normal approximation, the requirement $p_n^2/n \to 0$ is in general unrelaxable. Different from the analysis in large p_n^2/n setting, the techniques in Portnoy (1988) are applicable to more general cases. For example, focusing on the general likelihood problem with a *differentiable* likelihood function, Spokoiny (2012a) has provided a finite-sample analysis of normal approximation accuracy. Related results have also been developed in He and Shao (2000). As a direct consequence, a set of regularity conditions could be derived for constructing Bahadur-type bounds, guaranteeing ASN provided some scaling requirements hold.

Extending existing works allowing for increasing parameter dimension, this paper studies asymptotic properties of $\widehat{\boldsymbol{\theta}}_n$ in (1.1), allowing both m_n and p_n to go to infinity as $n \to \infty$. The potential discontinuity and U-process structure of the objective function $\Gamma_n(\boldsymbol{\theta})$ prevent results or the proof strategy in the current literature on increasing parameter dimension from being directly applicable. On the other hand, for (1.1), the increasing dimension set-up in this paper poses technical challenges to the proof strategy adopted for fixed m_n and p_n exclusively studied in the current literature. To see this, recall that the main argument used in the current literature to establish asymptotic properties for estimators of the form (1.1) for fixed m_n and p_n follows Sherman (Sherman, 1993, 1994), which relies on the Hoeffding decomposition, a uniform bound for degenerate U-processes, and the classical M-estimation framework tracing back to Huber's seminal paper, Huber (1967). Specifically, for the statistic $\Gamma_n(\boldsymbol{\theta})$ in (1.1), Hoeffding (1948) derived the following well-known expansion now known as the Hoeffding decomposition:

$$\Gamma_n(\boldsymbol{\theta}) = \Gamma(\boldsymbol{\theta}) + \mathbb{P}_n g(\cdot; \boldsymbol{\theta}) + \mathbb{U}_n h(\cdot, \cdot; \boldsymbol{\theta}), \tag{1.4}$$

where

$$\begin{aligned} g(\boldsymbol{z}; \boldsymbol{\theta}) &:= \mathbb{E} f(\boldsymbol{z}, \cdot; \boldsymbol{\theta}) + \mathbb{E} f(\cdot, \boldsymbol{z}; \boldsymbol{\theta}) - 2\Gamma(\boldsymbol{\theta}), \\ h(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta}) &:= f(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta}) - \mathbb{E} f(\boldsymbol{z}_1, \cdot; \boldsymbol{\theta}) - \mathbb{E} f(\cdot, \boldsymbol{z}_2; \boldsymbol{\theta}) + \Gamma(\boldsymbol{\theta}), \\ \mathbb{P}_n g(\cdot; \boldsymbol{\theta}) &:= \sum_{i=1}^{n} g(\boldsymbol{Z}_i)/n, \text{ and} \\ \mathbb{U}_n h(\cdot, \cdot; \boldsymbol{\theta}) &:= \sum_{i \neq j = 1}^{n} h(\boldsymbol{Z}_i, \boldsymbol{Z}_j; \boldsymbol{\theta})/\{n(n-1)\}. \end{aligned} \tag{1.5}$$

Hoeffding (1948) further showed that for fixed m_n and p_n,

$$\Gamma_n(\boldsymbol{\theta}) \approx \underbrace{\Gamma(\boldsymbol{\theta}) + \mathbb{P}_n g(\cdot; \boldsymbol{\theta})}_{\widetilde{\Gamma}_n(\boldsymbol{\theta})}, \tag{1.6}$$

where the remainder term $\mathbb{U}_n h(\cdot,\cdot;\boldsymbol{\theta})$, formulated as a degenerate U-statistic, is asymptotically negligible in large samples. As a result, $\widehat{\boldsymbol{\theta}}_n$ is asymptotically equivalent to $\widetilde{\boldsymbol{\theta}}_n$ defined below:

$$\widetilde{\boldsymbol{\theta}}_n := \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} \widetilde{\Gamma}_n(\boldsymbol{\theta}). \tag{1.7}$$

Sherman (Sherman, 1993, 1994) was the first to notice that, by (1.4) and the negligibility of $\mathbb{U}_n h(\cdot,\cdot;\boldsymbol{\theta})$, the U-statistic formulation has intrinsically helped smooth the loss function in (1.1) from $\Gamma_n(\boldsymbol{\theta})$ to $\widetilde{\Gamma}_n(\boldsymbol{\theta})$, and hence renders an asymptotically normal estimator $\widehat{\boldsymbol{\theta}}_n$, even though the original loss function $\Gamma_n(\boldsymbol{\theta})$ may not be differentiable.

For increasing dimensions m_n and p_n, the Hoeffding decomposition of $\Gamma_n(\boldsymbol{\theta})$ takes the same form as in the case of fixed m_n and p_n. However existing maximal inequalities or uniform bounds for degenerate U-processes for finite dimensions crucial to Sherman (Sherman, 1993, 1994) and the classical M-estimation theory for finite dimensions are inapplicable. In response to the first challenge, this paper develops a maximal inequality, yielding a uniform bound for degenerate U-processes in increasing dimensions, which allows us to show that under regularity conditions, $\widehat{\boldsymbol{\theta}}_n$ is asymptotically equivalent to $\widetilde{\boldsymbol{\theta}}_n$. Due to the smoothness of $\widetilde{\Gamma}_n(\boldsymbol{\theta})$, we are able to build on and improve arguments used in the proofs of Spokoiny (2012a) on M-estimators with *differentiable* objective functions in increasing dimensions to establish asymptotic properties of $\widehat{\boldsymbol{\theta}}_n$.

1.4. Notation

For a set $\mathcal{S}$, denote its binary Cartesian product as $\mathcal{S}\otimes\mathcal{S}$. For a probability measure $\mathbb{P}$, denote its product measure as $\mathbb{P}\otimes\mathbb{P}$. For $q\in[1,\infty]$, the L_q-norm of a vector $\boldsymbol{\beta}$ is denoted by $\|\boldsymbol{\beta}\|_q$. The L_q -induced matrix operator norm of a matrix $\mathbf{A}$ is denoted by $\|\mathbf{A}\|_q$. One example is the spectral norm $\|\mathbf{A}\|_2$, which represents the maximal singular value of $\mathbf{A}$. In the sequel, when no confusion is possible, we will omit the subscript in the L_q-norm of $\boldsymbol{\beta}$ or $\mathbf{A}$ when $q=2$. The minimum and maximum eigenvalues of a real symmetric matrix are denoted by $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ respectively. Let $\mathbf{I}_p$ denote the $p\times p$ identity matrix. Let $\mathbb{S}^{p-1}$ denote the unit-sphere of $\mathbb{R}^p$ under $\|\cdot\|$. For a twice differentiable real-valued function $\tau(\boldsymbol{\theta})$, let $\nabla_1\tau(\boldsymbol{\theta})$ denote the vector of partial derivatives $(\partial\tau/\partial\theta_1,\ldots,\partial\tau/\partial\theta_p)^\top$ and $\nabla_2\tau(\boldsymbol{\theta})$ denote the Hessian matrix of $\tau(\boldsymbol{\theta})$. Let $\mathcal{B}(\boldsymbol{\theta}_0,r)=\{\boldsymbol{\theta}\in\Theta,\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|<r\}$ denote an open ball of radius $r>0$ centered at $\boldsymbol{\theta}_0\in\Theta$, and let $\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r)=\{\boldsymbol{\theta}\in\Theta,\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le r\}$ denote a closed ball of center $\boldsymbol{\theta}_0$ and radius r. For two real numbers a and b, we define $a\vee b=\max(a,b)$ and $a\wedge b=\min(a,b)$. We use $\xrightarrow{\mathbb{P}}$ to denote convergence in probability with respect to $\mathbb{P}$, and $\Rightarrow$ to denote convergence in distribution. For any two real sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n=O(b_n)$ if there exists an absolute positive constant C such that $|a_n|\le C|b_n|$ for any large enough n. We write $a_n\asymp b_n$ if both $a_n=O(b_n)$ and $b_n=O(a_n)$ hold. We write $a_n=o(b_n)$ if for any absolute positive constant C, we have $|a_n|\le C|b_n|$ for any large enough n. We write $a_n=O_{\mathbb{P}}(b_n)$ and $a_n=o_{\mathbb{P}}(b_n)$ if $a_n=O(b_n)$ and $a_n=o(b_n)$ hold stochastically. We let $C,C',C'',c,c',c'',\ldots$ be generic absolute positive constants, whose values will vary at different locations.

1.5. Paper organization

The rest of this paper is organized as follows. In Section 2, we introduce general methods for handling M-estimators of the particular format. In particular, Section 2.1 gives a new U-process bound in increasing dimensions, and Section 2.2 studies M-estimators of the form (1.1), whose loss functions are possibly discontinuous. Section 3 applies the results in Section 2 to the four motivating rank estimators. Section 4 offers detailed finite-sample studies, illustrating the impact of dimension on coverage probability and tuning parameter selection in the asymptotic covariance estimation. Concluding remarks and possible extensions are put in the end of the main text. All proofs are relegated to an Appendix.

2. Asymptotic theory for the M-estimator

Recall that $\boldsymbol{Z}_1,\boldsymbol{Z}_2,\ldots,\boldsymbol{Z}_n\in\mathbb{R}^{m_n}$ is a random sample from $\mathbb{P}$, rendering an empirical measure $\mathbb{P}_n$. Let $\mathcal{F}=\{f(\cdot,\cdot;\boldsymbol{\theta}):\boldsymbol{\theta}\in\Theta\subset\mathbb{R}^{p_n}\}$ be a VC-subgraph class of real-valued functions, with ν_n denoting the *VC*-dimension of $\mathcal{F}$ (see Section 2.6.2 in van der Vaart and Wellner (1996) for explicit definitions of VC-subgraph and VC-dimension of a VC-subgraph class). In addition, we assume the function class $\mathcal{F}$ to be uniformly bounded by an absolute constant. The family of bounded VC-subgraph classes includes, as subfamilies, those rank estimators proposed in Han (1987), Cavanagh and Sherman (1998), Khan and Tamer (2007), and Abrevaya and Shin (2011), and suffices for our purpose.

Without loss of generality, we assume that

$$f(\boldsymbol{z}_1,\boldsymbol{z}_2;\boldsymbol{\theta}_0)=0 \quad \text{for all } (\boldsymbol{z}_1,\boldsymbol{z}_2)\in\mathbb{R}^{m_n}\otimes\mathbb{R}^{m_n}, \tag{2.1}$$

which can always be arranged by working with $f(\boldsymbol{z}_1,\boldsymbol{z}_2;\boldsymbol{\theta})-f(\boldsymbol{z}_1,\boldsymbol{z}_2;\boldsymbol{\theta}_0)$ throughout.

The derivation of asymptotic properties of $\widehat{\boldsymbol{\theta}}_n$ can be understood in two steps. First we show the asymptotic equivalence of $\widehat{\boldsymbol{\theta}}_n$ and $\widetilde{\boldsymbol{\theta}}_n$ by proving negligibility of $\mathbb{U}_n h(\cdot,\cdot;\boldsymbol{\theta})$ and then establish asymptotic properties of $\widetilde{\boldsymbol{\theta}}_n$. Essential to the first step is an increasing dimension analogue of maximal inequalities for degenerate U-processes in finite dimensions. Because of increasing dimensions, we need to calculate an exact order of the decaying rate of $\sup_{\boldsymbol{\theta}}|\mathbb{U}_n h(\cdot,\cdot;\boldsymbol{\theta})|$ in a local neighborhood of $\boldsymbol{\theta}_0$, the proof of which requires a substantial amount of modifications to the decoupling arguments in Nolan and Pollard (1987). For the second step, we exploit Spokoiny's bracketing device technique (cf. Corollary 2.2 in Spokoiny (2012b)) on M-estimators with differentiable objective functions.

2.1. A maximal inequality for degenerate U-processes

For fixed dimensions, Sherman (Sherman, 1993, 1994) proved a maximal inequality for degenerate U-processes and used it to show that, when $\mathcal{F}$ is $\mathbb{P}$-Donsker (Dudley, 1999), uniformly over a small neighborhood Θ_0 surrounding $\boldsymbol{\theta}_0$,

$$\sup_{\boldsymbol{\theta}\in\Theta_0} |\Gamma_n(\boldsymbol{\theta}) - \widetilde{\Gamma}_n(\boldsymbol{\theta})| = \sup_{\boldsymbol{\theta}\in\Theta_0} |\mathbb{U}_n h(\cdot,\cdot;\boldsymbol{\theta})| = o_{\mathbb{P}}(1/n), \tag{2.2}$$

which, combined with the fact that $g(\cdot)$ is usually a smooth function by integration, is sufficient to guarantee that the stochastic differentiability condition (cf. Theorem 3.2.16 in van der Vaart and Wellner (1996)) holds. This suffices for establishing ASN in fixed dimension. However, when we allow the dimension to increase with the sample size, (2.2) is no longer correct.

To account for the effect of increasing dimension, we establish a new maximal inequality for degenerate U-processes in increasing dimensions. Theorem 2.1 works out an exact order of the rate of convergence of $\sup_{\boldsymbol{\theta}\in\Theta_0} |\mathbb{U}_n h(\cdot,\cdot;\boldsymbol{\theta})|$ as Θ_0 shrinks to the true point $\boldsymbol{\theta}_0$ at different rates $r_n \to 0$. It is formulated as two maximal inequalities, corresponding to the Glivenko–Cantelli and Donsker properties , for a degenerate U-process.

Theorem 2.1. *Suppose that $\mathcal{F}$ is uniformly bounded by an absolute constant, of VC-dimension ν_n, and $h(\cdot)$ is defined as in (1.5). Further recall that we have assumed $f(\cdot,\cdot;\boldsymbol{\theta}_0)$ satisfies (2.1). If $\nu_n/n \to 0$, then the following two claims hold.*

(i) Let r_n and ϵ_n be two sequences of nonnegative real numbers converging to zero. If

$$\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)} \mathbb{E}h^2(\cdot,\cdot;\boldsymbol{\theta}) \le \epsilon_n,$$

then there exists a sequence of nonnegative real numbers δ_n (only depending on ϵ_n, ν, n) converging to zero such that

$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)} |\mathbb{U}_n h(\cdot,\cdot;\boldsymbol{\theta})| \le \delta_n \nu_n/n\right\} = 1 - o(1).$$

(ii) Let $r_n := r(\nu_n, p_n, n)$ be a sequence of nonnegative real numbers converging to zero, and $\widetilde{\epsilon}_n = \epsilon(\nu_n, p_n, n, r_n)$ be a sequence of nonnegative real numbers (only depending on ν_n, p_n, n, r_n) converging to zero. Denote $\widetilde{\eta}_n = \eta(\nu_n, p_n, n, r_n) = \sqrt{\nu_n/n} \vee \widetilde{\epsilon}_n$. Suppose

$$\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)} \mathbb{E}h^2(\cdot,\cdot;\boldsymbol{\theta}) \le \widetilde{\epsilon}_n.$$

We then have

$$\mathbb{E}\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)} |\mathbb{U}_n h(\cdot,\cdot;\boldsymbol{\theta})| \le \frac{C\log(1/\widetilde{\eta}_n)\widetilde{\eta}_n^{1/2}\nu_n}{n} \tag{2.3}$$

holds for all sufficiently large n.

For deriving Theorem 2.1, one might consider employing the decoupling techniques as introduced in the proofs of the Main Corollary in Sherman (1994), or Theorem 5.3.7 in de la Pena and Giné (2012). However, since the considered U-process depends on an increasing number of covariates, the constants in the moment inequalities therein (e.g., $C(k,q)$ in Sherman (1994)) are no longer finite and are difficult to characterize in increasing dimensions. Instead, we resort to Nolan and Pollard's original treatment of degenerate U-processes.

Specifically, denoting

$$\mathbb{S}_n f(\cdot,\cdot;\boldsymbol{\theta}) = n(n-1)\mathbb{U}_n f(\cdot,\cdot;\boldsymbol{\theta}),$$

a modification to Theorem 6 in Nolan and Pollard (1987) will give us

$$\begin{aligned}\mathbb{E}\left\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)} |\mathbb{S}_n h(\cdot,\cdot;\boldsymbol{\theta})/(n\nu)|\right\} &\le CH\left(\left[\mathbb{E}\left\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)} \mathbb{U}_{2n}h^2(\cdot,\cdot;\boldsymbol{\theta})\right\}\right]^{1/2}\right)\\ &\le CH\left(\left[\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)} \mathbb{E}h^2(\cdot,\cdot;\boldsymbol{\theta}) + \mathbb{E}\left\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)} |\mathbb{P}_{2n}h_1(\cdot;\boldsymbol{\theta})|\right\}\right.\right.\\ &\quad \left.\left. + \mathbb{E}\left\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)} |\mathbb{P}_{2n}h_2(\cdot,\cdot;\boldsymbol{\theta})|\right\}\right]^{1/2}\right).\end{aligned} \tag{2.4}$$

Here $H(x) := x\{1+\log(1/x)\}$ for any $x \in (0,\infty)$, $\mathbb{U}_{2n}$ and $\mathbb{P}_{2n}$ have been introduced in (1.5), and $h_1(\boldsymbol{z},\boldsymbol{\theta}) := \mathbb{E}h^2(\boldsymbol{z},\cdot;\boldsymbol{\theta}) + \mathbb{E}h^2(\cdot,\boldsymbol{z};\boldsymbol{\theta}) - 2\mathbb{E}h^2(\cdot,\cdot;\boldsymbol{\theta})$ and $h_2(\boldsymbol{z}_1,\boldsymbol{z}_2;\boldsymbol{\theta}) := h^2(\boldsymbol{z}_1,\boldsymbol{z}_2;\boldsymbol{\theta}) - \mathbb{E}h^2(\boldsymbol{z}_1,\cdot;\boldsymbol{\theta}) - \mathbb{E}h^2(\cdot,\boldsymbol{z}_2;\boldsymbol{\theta}) + \mathbb{E}h^2(\cdot,\cdot;\boldsymbol{\theta})$ are two functions generated from $h(\cdot,\cdot;\boldsymbol{\theta})$. We have thus explicitly transformed the analysis of a degenerate U-process to that of a moment bound, and two empirical processes. Lastly, the bounds on the two empirical processes could be derived using, for example, Theorem 9.3 in Kosorok (2007).

2.2. Main results

We are now ready to state the main results in this section. For analyzing the statistical properties of the general M-estimator $\widehat{\boldsymbol{\theta}}_n$, three targets are in order: (i) consistency; (ii) rate of convergence; and (iii) Bahadur-type bounds. Of note, our analysis is under the increasing dimension triangular array setting where the true data generating process $\mathbb{P}$ is allowed to change with the sample size n.

We first establish consistency. This is via the following two assumptions.

Assumption 1. For each specified p_n, Θ is a compact subset of $\mathbb{R}^{p_n}$, and there exists an absolute constant $r_0 > 0$ such that $\mathcal{B}(\boldsymbol{\theta}_0, r_0) \subset \Theta$ and for any positive absolute constant $r < r_0$, there exists another absolute constant $\xi_0 > 0$ depending on r such that

$$\Gamma(\boldsymbol{\theta}_0) - \max_{\Theta \setminus \mathcal{B}(\boldsymbol{\theta}_0, r)} \Gamma(\boldsymbol{\theta}) \geq \xi_0. \tag{2.5}$$

Assumption 2. $\Gamma(\boldsymbol{\theta})$ is a continuous function at any $\boldsymbol{\theta} \in \Theta$, and $f(\cdot, \cdot; \boldsymbol{\theta})$ is almost everywhere continuous at $\boldsymbol{\theta}_0$.

Assumption 1 is the standard identifiability condition. Since $\Gamma(\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta} \in \mathbb{R}^{p_n}$ is also to change with n, it is regulated by a constant ξ_0 to eliminate the non-identifiable cases in large n. Assumption 2 enforces certain level of smoothness on Γ and f. Both are regular, and in particular, verifiable for all the considered examples of rank estimators using explicit expressions for Γ and f for these estimators. For example, for Han's MRC, Assumption 1 can be established using Taylor expansion applied to $\Gamma(\boldsymbol{\theta}) = \Gamma^{\mathsf{H}}(\boldsymbol{\theta}) = S^{\mathsf{H}}(\boldsymbol{\beta}) - S^{\mathsf{H}}(\boldsymbol{\beta}_0)$ with $S^{\mathsf{H}}(\boldsymbol{\beta}) := \mathbb{E}\{\mathbb{1}(Y_1 > Y_2)\mathbb{1}(\boldsymbol{X}_1^\top \boldsymbol{\beta} > \boldsymbol{X}_2^\top \boldsymbol{\beta})\}$.

With Assumptions 1 and 2, we immediately obtain the following theorem, establishing consistency for the studied M-estimator $\widehat{\boldsymbol{\theta}}_n$.

Theorem 2.2. *Suppose that Assumptions 1–2 hold. If $\nu_n/n \to 0$, then $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \xrightarrow{\mathbb{P}} 0$.*

It is of interest to point out that consistency is established solely based on an requirement of ν_n (which also intrinsically depends on m_n, p_n), since the uniform consistency of Γ_n to Γ can be determined solely by the relation between ν_n and n. For the four examples of rank correlation estimators (1)-(4), $\nu_n = p_n$ so consistency is ensured under Assumptions 1 and 2 as long as the number of parameters p_n increases at a slower rate than the sample size n.

For establishing rates of convergence and Bahadur-type bounds, on the other hand, more assumptions are needed. For each $\boldsymbol{z}$ in $\mathbb{R}^{m_n}$ and for each $\boldsymbol{\theta} \in \Theta$, define

$$\tau(\boldsymbol{z}; \boldsymbol{\theta}) = \mathbb{E}f(\boldsymbol{z}, \cdot; \boldsymbol{\theta}) + \mathbb{E}f(\cdot, \boldsymbol{z}; \boldsymbol{\theta}) \quad \text{and} \quad \zeta(\boldsymbol{z}; \boldsymbol{\theta}) = \tau(\boldsymbol{z}; \boldsymbol{\theta}) - \mathbb{E}\tau(\cdot; \boldsymbol{\theta}).$$

Here $\tau(\boldsymbol{z}; \boldsymbol{\theta})$ corresponds to $\widetilde{\Gamma}_n(\boldsymbol{\theta})$ in (1.6), and is the key for establishing ASN of $\widetilde{\boldsymbol{\theta}}_n$ in (1.7). The following assumption regulates $\tau(\cdot; \cdot)$.

Assumption 3. For each $r \leq r_0$, the following conditions hold.

(i) For each $\boldsymbol{z}$ in $\mathbb{R}^{m_n}$, all mixed second partial derivatives of $\tau(\boldsymbol{z}; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ exist on $\overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)$.

(ii) There exist two positive absolute constants $c_{\min}, c_{\max}$ such that $0 < c_{\min} \leq \lambda_{\min}(-\mathbf{V}) \leq \lambda_{\max}(-\mathbf{V}) \leq c_{\max}$, where $2\mathbf{V} := \mathbb{E}\nabla_2 \tau(\cdot; \boldsymbol{\theta}_0)$.

(iii) There exists a positive constant $\rho(r) < \frac{c_{\min}}{11 c_{\max}} \wedge cpr$ for some absolute constant $c > 0$, such that $\|\mathbf{I}_p - \mathbf{V}^{-1/2}\mathbf{V}(\boldsymbol{\theta})\mathbf{V}^{-1/2}\| \leq \rho(r)$ for any $\boldsymbol{\theta} \in \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)$, where $2\mathbf{V}(\boldsymbol{\theta}) := \mathbb{E}\nabla_2 \tau(\cdot; \boldsymbol{\theta})$.

(iv) Assume $0 < d_{\min} \leq \lambda_{\min}(\boldsymbol{\Delta}) \leq \lambda_{\max}(\boldsymbol{\Delta}) \leq d_{\max}$, where $\boldsymbol{\Delta} := \mathbb{E}\nabla_1 \tau(\cdot; \boldsymbol{\theta}_0)\{\nabla_1 \tau(\cdot; \boldsymbol{\theta}_0)\}^\top$ and $d_{\min}, d_{\max}$ are two positive absolute constants.

(v) There exist absolute constants $\nu_0 > 0$ and $\ell_0 > 0$ such that, for any $\boldsymbol{\theta} \in \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)$, the following holds:

$$\sup_{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in \mathbb{S}^{p_n - 1}} \log \mathbb{E} \exp\left\{\lambda \boldsymbol{\gamma}_1^\top \nabla_2 \zeta(\cdot; \boldsymbol{\theta}) \boldsymbol{\gamma}_2\right\} \leq \frac{\nu_0^2 \lambda^2}{2}, \quad \text{for all } |\lambda| \leq \ell_0.$$

Assumption 3 is the key assumption in order to establish Bahadur-type bounds for $\widehat{\boldsymbol{\theta}}_n$, and is posed for the M-estimation problem (1.6) of loss function $\widetilde{\Gamma}_n(\boldsymbol{\theta})$ corresponding to the function $\tau(\cdot)$. In the following we discuss more about this assumption. In detail, Assumption 3(i), (ii), and (iv) are regularity conditions to make sure that the studied problem is well posited, a condition corresponding to the local strong convexity condition in the high dimensional statistics literature (cf. Section 2.4 in Negahban et al. (2012)), and are verifiable for different methods. Consider, for example, Han's MRC estimator $\widehat{\boldsymbol{\theta}}_n^{\mathsf{H}}$ introduced in Section 1.2 for which $\tau = \tau^{\mathsf{H}}$:

$$\tau^{\mathsf{H}}(\boldsymbol{z}; \boldsymbol{\theta}) := \mathbb{E}f^{\mathsf{H}}(\boldsymbol{z}, \cdot; \boldsymbol{\theta}) + \mathbb{E}f^{\mathsf{H}}(\cdot, \boldsymbol{z}; \boldsymbol{\theta}),$$

where

$$f^{\mathsf{H}}(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta}) := \mathbb{1}(y_1 > y_2)\{\mathbb{1}(\boldsymbol{x}_1^\top \boldsymbol{\beta} > \boldsymbol{x}_2^\top \boldsymbol{\beta}) - \mathbb{1}(\boldsymbol{x}_1^\top \boldsymbol{\beta}_0 > \boldsymbol{x}_2^\top \boldsymbol{\beta}_0)\}.$$

Assumption 3(i), (ii), and (iv) then are immediately ensured by Theorem 4 and subsequent discussions in Sherman (1993). Assumption 3(iii) requires that $\mathbb{E}\tau(\cdot;\boldsymbol{\theta})$ is sufficiently smooth in $\boldsymbol{\theta}$, for example, $\mathbb{E}\tau(\cdot;\boldsymbol{\theta})$ has continuous and bounded mixed partial derivatives up to three. Assumption 3(v) requires the existence of exponential moments of the errors. They correspond to the "local identifiability condition": Assumption $(\mathcal{L}_0)$, and the "exponential moment condition", Assumption (ED_2), in Spokoiny (2012a, 2013) separately. These conditions are often implied by subgaussian designs. Particularly, in Theorem 3.1 in Section 3.1, we will verify Assumption 3(iii) and (v) for τ^{H}, i.e., Han's MRC under primitive conditions.

With the above assumptions, statistical properties of $\widehat{\boldsymbol{\theta}}_n$ could then be established as follows.

Theorem 2.3. *If $(\nu_n \vee p_n)/n \to 0$ and Assumptions 1–3 hold, we have*

$$\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|^2 = O_{\mathbb{P}}\Big(\frac{\nu_n \vee p_n}{n}\Big).$$

For the four examples of rank correlation estimators, $\nu_n = p_n$ so Theorem 2.3 leads to the minimax optimal rate $(p_n/n)^{1/2}$ under the condition: $p_n/n \to 0$. However, Theorem 2.4 implies that much stronger requirements on p_n are needed to establish Bahadur-type bounds, see Corollaries 3.1–3.4 for details.

Theorem 2.4. *Suppose Assumptions 1–3 hold, and there exists a constant $\epsilon_n = \epsilon(\nu_n, p_n, n)$ depending on ν_n, p_n, n such that, for any $c > 0$,*

$$\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}\{\boldsymbol{\theta}_0, c\sqrt{(\nu_n\vee p_n)/n}\}} \mathbb{E}h^2(\cdot,\cdot;\boldsymbol{\theta}) \le \widetilde{C}\epsilon_n,$$

where $\widetilde{C}$ only depends on c. Then, the following two statements hold.

(i) Denote $\eta_n = \eta(\nu_n, p_n, n) = \sqrt{\nu_n/n} \vee \epsilon_n$. If $\eta_n = o(1)$ and $\{(\nu_n \vee p_n)^{5/2}/n^{3/2}\} \vee \{\log(1/\eta_n)\eta_n^{1/2}\nu_n/n\} = o(1)$, we have

$$\big\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 + \mathbf{V}^{-1}\mathbb{P}_n\nabla_1\tau(\cdot;\boldsymbol{\theta}_0)\big\|^2 = O_{\mathbb{P}}\Big\{\frac{(\nu_n\vee p_n)^{5/2}}{n^{3/2}} + \frac{\log(1/\eta_n)\eta_n^{1/2}\nu_n}{n}\Big\}.$$

(ii) If we further have $\{(\nu_n \vee p_n)^{5/2}/n^{1/2}\} \vee \{\log(1/\eta_n)\eta_n^{1/2}\nu_n\} = o(1)$, then for any $\boldsymbol{\gamma} \in \mathbb{R}^{p_n}$,

$$\sqrt{n}\boldsymbol{\gamma}^\top(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)/(\boldsymbol{\gamma}^\top\mathbf{V}^{-1}\boldsymbol{\Delta}\mathbf{V}^{-1}\boldsymbol{\gamma})^{1/2} \Rightarrow N(0,1).$$

Remark 2.5. In the analysis, p_n and ν_n characterize the behavior of the smoothed estimator $\widetilde{\boldsymbol{\theta}}_n$ and the degenerate U-process $\{\mathbb{U}_n h(\cdot,\cdot;\boldsymbol{\theta}); \boldsymbol{\theta} \in \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r_n)\}$ separately. On the other hand, throughout the above three theorems, the dimension of data points, m_n, is not present. Instead, the impact of m_n on estimation and inference has been characterized by p_n and ν_n, both of which are usually of an order equal to or even greater than m_n. It is also noteworthy to point out that our analysis does allow an arbitrary subset of (m_n, p_n, ν_n) to be fixed, and the theory will directly proceed. In particular, when m_n, p_n, ν_n are all invariant with regard to n, we derived the conventional Bahadur representation for the studied class of M-estimators under the low-dimensional setting, which is a stronger result than asymptotic normality.

We conclude this section with a brief discussion on consistent estimation of the asymptotic covariance matrix in Theorem 2.4. For this, we are focused on the covariance estimator of a numerical derivative form, used in Pakes and Pollard (1989), Sherman (1993), and Khan and Tamer (2007).

First, for each $\boldsymbol{z}$ in $\mathbb{R}^{m_n}$ and for each $\boldsymbol{\theta}$ in Θ, define

$$\tau_n(\boldsymbol{z};\boldsymbol{\theta}) = \mathbb{P}_n f(\boldsymbol{z},\cdot;\boldsymbol{\theta}) + \mathbb{P}_n f(\cdot,\boldsymbol{z};\boldsymbol{\theta}).$$

Then, we define the numerical derivative of $\tau_n(\boldsymbol{z};\boldsymbol{\theta})$ as follows:

$$p_{ni}(\boldsymbol{z};\boldsymbol{\theta}) = \varepsilon_n^{-1}\{\tau_n(\boldsymbol{z};\boldsymbol{\theta} + \varepsilon_n\boldsymbol{u}_i) - \tau_n(\boldsymbol{z};\boldsymbol{\theta})\},$$

where ε_n denotes a sequence of real numbers converging to zero, and $\boldsymbol{u}_i$ denotes the unit vector in $\mathbb{R}^{p_n}$ with the ith component equal to one. Finally, we define the estimator of the matrix $\boldsymbol{\Delta}$ as $\widehat{\boldsymbol{\Delta}} = (\widehat{\delta}_{ij})$ with

$$\widehat{\delta}_{ij} := \mathbb{P}_n\{p_{ni}(\cdot;\widehat{\boldsymbol{\theta}}_n)p_{nj}(\cdot;\widehat{\boldsymbol{\theta}}_n)\}.$$

To estimate the matrix $\mathbf{V}$, we define the following function:

$$p_{nij}(\boldsymbol{z};\boldsymbol{\theta}) = \varepsilon_n^{-2}\{\tau_n(\boldsymbol{z};\boldsymbol{\theta} + \varepsilon_n(\boldsymbol{u}_i + \boldsymbol{u}_j)) - \tau_n(\boldsymbol{z};\boldsymbol{\theta} + \varepsilon_n\boldsymbol{u}_i) - \tau_n(\boldsymbol{z};\boldsymbol{\theta} + \varepsilon_n\boldsymbol{u}_j) + \tau_n(\boldsymbol{z};\boldsymbol{\theta})\}.$$

Then, we define the estimator of the matrix $\mathbf{V}$ as $\widehat{\mathbf{V}} = (\widehat{v}_{ij})$ with

$$\widehat{v}_{ij} := \frac{1}{2}\mathbb{P}_n p_{nij}(\cdot;\widehat{\boldsymbol{\theta}}_n).$$

Let $\widetilde{\mathcal{F}} = \{f(\boldsymbol{z},\cdot;\boldsymbol{\theta}) + f(\cdot,\boldsymbol{z};\boldsymbol{\theta}) : \boldsymbol{z} \in \mathbb{R}^m, \boldsymbol{\theta} \in \Theta\}$, and let $\widetilde{\nu}_n$ denote the VC-dimension of $\widetilde{\mathcal{F}}$. The following theorem establishes the consistency of the covariance estimator.

Theorem 2.6. *Suppose that Assumptions 1–3 hold and* $(\widetilde{\nu}_n \vee \nu_n \vee p_n)^{5/2}/n^{1/2} = o(1)$. *If the sequence* ε_n *satisfies:* $\varepsilon_n\sqrt{p_n} = o(1)$ *and* $\varepsilon_n^{-2}(\widetilde{\nu}_n \vee \nu_n \vee p_n)/\sqrt{n} = o(1)$, *then*

$$\|\widehat{\mathbf{V}}^{-1}\widehat{\mathbf{\Delta}}\widehat{\mathbf{V}}^{-1} - \mathbf{V}^{-1}\mathbf{\Delta}\mathbf{V}^{-1}\| \xrightarrow{\mathbb{P}} 0.$$

The increasing dimension set-up reveals that for consistent variance–covariance matrix estimation, the step size in computing the numerical derivative should depend not only on the sample size but also on the dimensions m_n and p_n.

3. Asymptotic properties of rank estimators

This section studies the four examples introduced in Introduction. In the sequel, the data points are understood to be independent and identically drawn from the considered model. Of note, throughout the following four examples, when the studied model is fixed, our result renders the conventional Bahadur representation for the corresponding estimator in fixed dimensions (see, for example, Subbotin (2008) for such a bound in fixed dimensions). Hence, we recover the asymptotic-normality-type theory in the corresponding paper, but under a stronger moment condition in order to take the impact of increasing dimension into consideration. In addition, it is worthwhile to point out that, for all studied methods, the dimension of the data points m_n and the VC dimensions ν_n and $\widetilde{\nu}_n$ of the studied function classes are all of the same order as p_n, the number of parameters to be estimated. Accordingly, in the following, we can use p_n to solely characterize the impact of dimension on inference.

3.1. Han's maximum rank correlation estimator

This section studies the generalized regression model (1.2) and Han's MRC estimator, as have been introduced in Section 1.2. Let $\mathcal{B}$ be a subset of $\{\boldsymbol{\beta} \in \mathbb{R}^{p_n+1} : \beta_1 = 1\}$. For any $\boldsymbol{\beta} \in \mathcal{B}$, let $\boldsymbol{\beta} = (1, \boldsymbol{\theta}^\top)^\top$, where $\boldsymbol{\theta} \in \Theta^{\mathsf{H}} \subset \mathbb{R}^{p_n}$. For any vector $\boldsymbol{z} = (y, \boldsymbol{x}^\top)^\top$, we define $\zeta^{\mathsf{H}}(\boldsymbol{z}; \boldsymbol{\theta}) = \tau^{\mathsf{H}}(\boldsymbol{z}; \boldsymbol{\theta}) - \mathbb{E}\tau^{\mathsf{H}}(\cdot; \boldsymbol{\theta})$,

$$\mathbf{\Delta}^{\mathsf{H}} = \mathbb{E}\nabla_1\tau^{\mathsf{H}}(\cdot; \boldsymbol{\theta}_0)\{\nabla_1\tau^{\mathsf{H}}(\cdot; \boldsymbol{\theta}_0)\}^\top, \quad \text{and} \quad 2\mathbf{V}^{\mathsf{H}} = \mathbb{E}\nabla_2\tau^{\mathsf{H}}(\cdot; \boldsymbol{\theta}_0).$$

Write $\mathit{\Gamma}_n^{\mathsf{H}}(\boldsymbol{\theta}) = S_n^{\mathsf{H}}(\boldsymbol{\beta}) - S_n^{\mathsf{H}}(\boldsymbol{\beta}_0)$ with

$$S_n^{\mathsf{H}}(\boldsymbol{\beta}) := \frac{1}{n(n-1)}\sum_{i\neq j}\mathbb{1}(Y_i > Y_j)\mathbb{1}(\boldsymbol{X}_i^\top\boldsymbol{\beta} > \boldsymbol{X}_j^\top\boldsymbol{\beta}).$$

Thus, Han's MRC estimator of $\boldsymbol{\theta}_0$, $\widehat{\boldsymbol{\theta}}_n^{\mathsf{H}}$, can be expressed as

$$\widehat{\boldsymbol{\theta}}_n^{\mathsf{H}} = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta^{\mathsf{H}}} \mathit{\Gamma}_n^{\mathsf{H}}(\boldsymbol{\theta}).$$

To conduct inference on $\boldsymbol{\theta}_0$ based on $\widehat{\boldsymbol{\theta}}_n^{\mathsf{H}}$, we further define

$$\tau_n^{\mathsf{H}}(\boldsymbol{z}; \boldsymbol{\theta}) = \mathbb{P}_n f^{\mathsf{H}}(\boldsymbol{z}, \cdot; \boldsymbol{\theta}) + \mathbb{P}_n f^{\mathsf{H}}(\cdot, \boldsymbol{z}; \boldsymbol{\theta}), \quad p_{ni}^{\mathsf{H}}(\boldsymbol{z}; \boldsymbol{\theta}) = \varepsilon_n^{-1}\{\tau_n^{\mathsf{H}}(\boldsymbol{z}; \boldsymbol{\theta} + \varepsilon_n\boldsymbol{u}_i) - \tau_n^{\mathsf{H}}(\boldsymbol{z}; \boldsymbol{\theta})\}, \quad \text{and}$$
$$p_{nij}^{\mathsf{H}}(\boldsymbol{z}; \boldsymbol{\theta}) = \varepsilon_n^{-2}\{\tau_n^{\mathsf{H}}(\boldsymbol{z}; \boldsymbol{\theta} + \varepsilon_n(\boldsymbol{u}_i + \boldsymbol{u}_j)) - \tau_n^{\mathsf{H}}(\boldsymbol{z}; \boldsymbol{\theta} + \varepsilon_n\boldsymbol{u}_i) - \tau_n^{\mathsf{H}}(\boldsymbol{z}; \boldsymbol{\theta} + \varepsilon_n\boldsymbol{u}_j) + \tau_n^{\mathsf{H}}(\boldsymbol{z}; \boldsymbol{\theta})\}.$$

Then, we define the estimator of the matrix $\mathbf{\Delta}^{\mathsf{H}}$ as $\widehat{\mathbf{\Delta}}^{\mathsf{H}} = (\widehat{\delta}_{ij}^{\mathsf{H}})$ and the estimator of the matrix $\mathbf{V}^{\mathsf{H}}$ as $\widehat{\mathbf{V}}^{\mathsf{H}} = (\widehat{v}_{ij}^{\mathsf{H}})$, where

$$\widehat{\delta}_{ij}^{\mathsf{H}} = \mathbb{P}_n\{p_{ni}^{\mathsf{H}}(\cdot; \widehat{\boldsymbol{\theta}}_n^{\mathsf{H}})p_{nj}^{\mathsf{H}}(\cdot; \widehat{\boldsymbol{\theta}}_n^{\mathsf{H}})\} \quad \text{and} \quad \widehat{v}_{ij}^{\mathsf{H}} = \frac{1}{2}\mathbb{P}_n p_{nij}^{\mathsf{H}}(\cdot; \widehat{\boldsymbol{\theta}}_n^{\mathsf{H}}).$$

Let $\boldsymbol{X} = (X_1, \widetilde{\boldsymbol{X}}^\top)^\top$, where $\widetilde{\boldsymbol{X}}$ denotes the last p components in $\boldsymbol{X}$. Assume the following assumption holds

Assumption 4. Assume

(i) Assumption 1 holds for Θ^{H} and $\mathit{\Gamma}^{\mathsf{H}}(\boldsymbol{\theta})$.
(ii) The random variables $\boldsymbol{X}$ and ϵ are independent.
(iii) Assume X_1 has an everywhere positive Lebesgue density, conditional on $\widetilde{\boldsymbol{X}}$.
(iv) Assumption 3 holds for $\tau^{\mathsf{H}}(\boldsymbol{z}; \boldsymbol{\theta})$ and $\zeta^{\mathsf{H}}(\boldsymbol{z}; \boldsymbol{\theta})$.

Assumption 5. For some absolute constant $C > 0$, $\sup_{i=2,\ldots,p+1}\mathbb{E}|X_i|^2 \leq C$.

Assumption 6. Let $f_0(\cdot \mid \widetilde{\boldsymbol{x}})$ denote the conditional density function of $\boldsymbol{X}^\top\boldsymbol{\beta}_0$ given $\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}$. Assume $f_0(\cdot \mid \widetilde{\boldsymbol{x}}) \leq C_0$ for any $\widetilde{\boldsymbol{x}}$ in the support of $\widetilde{\boldsymbol{X}}$, where $C_0 > 0$ is an absolute constant.

We then have the following corollary.

Corollary 3.1. *We have*

(i) Under Assumption 4(i)–(iii), if $p_n/n = o(1)$, *then* $\|\widehat{\theta}_n^{\mathsf{H}} - \theta_0\| \xrightarrow{\mathbb{P}} 0$.

(ii) Under Assumption 4, if $p_n/n = o(1)$, *then*

$$\|\widehat{\boldsymbol{\theta}}_n^{\mathsf{H}} - \boldsymbol{\theta}_0\|^2 = O_{\mathbb{P}}(p_n/n).$$

(iii) Under Assumptions 4–6, if $p_n^2/n = o(1)$ *and* $\log(n/p_n^2)p_n^{3/2}/n^{5/4} = o(1)$, *we have*

$$\|\widehat{\boldsymbol{\theta}}_n^{\mathsf{H}} - \boldsymbol{\theta}_0 + (\mathbf{V}^{\mathsf{H}})^{-1}\mathbb{P}_n\nabla_1\tau^{\mathsf{H}}(\cdot;\boldsymbol{\theta}_0)\|^2 = O_{\mathbb{P}}\big\{\log(n/p_n^2)p_n^{3/2}/n^{5/4}\big\}. \tag{3.1}$$

Furthermore, if $\log(n/p_n^2)p_n^{3/2}/n^{1/4} = o(1)$, *then for any* $\boldsymbol{\gamma} \in \mathbb{R}^{p_n}$,

$$\sqrt{n}\boldsymbol{\gamma}^\top(\widehat{\boldsymbol{\theta}}_n^{\mathsf{H}} - \boldsymbol{\theta}_0)/\{\boldsymbol{\gamma}^\top(\mathbf{V}^{\mathsf{H}})^{-1}\boldsymbol{\Delta}^{\mathsf{H}}(\mathbf{V}^{\mathsf{H}})^{-1}\boldsymbol{\gamma}\}^{1/2} \Rightarrow N(0,1).$$

(iv) Under conditions in (iii), if we further have $\varepsilon_n\sqrt{p_n} = o(1)$ *and* $\varepsilon_n^{-2}p_n/\sqrt{n} = o(1)$, *then*

$$\|(\widehat{\mathbf{V}}^{\mathsf{H}})^{-1}\widehat{\boldsymbol{\Delta}}^{\mathsf{H}}(\widehat{\mathbf{V}}^{\mathsf{H}})^{-1} - (\mathbf{V}^{\mathsf{H}})^{-1}\boldsymbol{\Delta}^{\mathsf{H}}(\mathbf{V}^{\mathsf{H}})^{-1}\| \xrightarrow{\mathbb{P}} 0.$$

In particular, we could choose $\epsilon_n \asymp (p_n/n)^{1/6}$, *which will render a consistent covariance estimator under the same scaling condition as (iii).*

In the following, we discuss more on the assumptions posed for Han's MRC estimator. Since the estimator takes pairwise differences as input, without loss of generality, the design is assumed to be zero-mean. First, Assumption 1 can be established using Assumption 3(ii), (iii), and Taylor expansion. Secondly, the conditions in Assumptions 2 and 3(i) are regular and can be satisfied. Then, Theorem 4 and subsequent discussions in Sherman (1993) ensure Assumption 3(ii) and (iv) hold. Lastly, we deal with Assumption 3(iii) and (v), which indeed deserve more discussion. In the following, we give sufficient conditions for guaranteeing Assumption 3(iii) and (v) hold.

More notation is needed. Let $f_0(\cdot \mid \widetilde{\boldsymbol{x}}, y)$ denote the conditional density function of X_1 given $\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}$ and $Y = y$. Let $f_0(\cdot)$ denote the marginal density function of $\boldsymbol{X}^\top\boldsymbol{\beta}_0$. Let

$$\kappa^{\mathsf{H}}(y,t) = \mathbb{E}\{\mathbb{1}(y > Y) - \mathbb{1}(y < Y) \mid \boldsymbol{X}^\top\boldsymbol{\beta}_0 = t\}, \quad \lambda^{\mathsf{H}}(y,t) = \kappa^{\mathsf{H}}(y,t)f_0(t),$$

$$\text{and} \quad \lambda_2^{\mathsf{H}}(y,t) = \frac{\partial}{\partial t}\lambda^{\mathsf{H}}(y,t).$$

We assume the following conditions on the design as well as the noisy hold.

Condition 1. *Suppose* $\boldsymbol{X}$ *is multivariate subgaussian, i.e., there exists an absolute constant* $c' > 0$ *such that* $\sup_{\boldsymbol{\gamma}\in\mathbb{S}^p}\|\boldsymbol{\gamma}^\top\boldsymbol{X}\|_{\psi_2} \le c'$, *where* $\|\boldsymbol{\gamma}^\top\boldsymbol{X}\|_{\psi_2} := \sup_{q\ge1} q^{-1/2}(\mathbb{E}|\boldsymbol{\gamma}^\top\boldsymbol{X}|^q)^{1/q}$.

Condition 2. *(i) Suppose that* $f_0(\cdot \mid \widetilde{\boldsymbol{x}}, y)$ *has uniformly bounded derivatives up to order three, i.e., there exists an absolute constant* $C'' > 0$ *such that* $|f_0^{(j)}(\cdot \mid \widetilde{\boldsymbol{x}}, y)| \le C''$ $(j = 1,2,3)$ *for any* $\widetilde{\boldsymbol{x}}$ *and* y *in the support of* $\widetilde{\boldsymbol{X}}$ *and* Y, *respectively; (ii)* $\lim_{|t|\to\infty} f_0^{(2)}(t \mid \widetilde{\boldsymbol{x}}, y) = 0$ *for any* $\widetilde{\boldsymbol{x}}$ *and* y; *(iii) Universally over the support of* Y *and any* $\boldsymbol{\theta} \in \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)$, $\int |f_0^{(3)}(t - \widetilde{\boldsymbol{x}}^\top\boldsymbol{\theta} \mid s, \widetilde{\boldsymbol{x}})|G_{\widetilde{\boldsymbol{X}}|Y=s}(\mathrm{d}\widetilde{\boldsymbol{x}}) \le c\{1 \wedge c'|t|^{-(1+c'')}\}$ *for some positive absolute constants* c, c', c'', *where* $G_{\widetilde{\boldsymbol{X}}|Y=s}(\cdot)$ *represents the probability measure of* $\widetilde{\boldsymbol{X}}$ *given* $Y = s$.

Condition 3. *Suppose that* $\lambda_2^{\mathsf{H}}(y,t)$ *is bounded, i.e., there exists an absolute constant* $c'' > 0$ *such that* $|\lambda_2^{\mathsf{H}}(y,t)| \le c''$ *for any* y *and* t *in the support of* Y *and* $\boldsymbol{X}^\top\boldsymbol{\beta}_0$, *respectively.*

We then have the following theorem, which states that the above conditions are sufficient ones to ensure Assumption 3(iii) and (v) hold.

Theorem 3.1. *Under Conditions 1–3, Assumption 3(iii) and (v) hold in this example.*

3.2. Cavanagh and Sherman's rank estimator

In contrast to Han's original proposal, Cavanagh and Sherman (1998) proposed estimating $\boldsymbol{\beta}_0$ in (1.2) using

$$\widehat{\boldsymbol{\beta}}_n^{\mathsf{C}} = \operatorname*{argmax}_{\boldsymbol{\beta}:\beta_1=1} S_n^{\mathsf{C}}(\boldsymbol{\beta}),$$

where

$$S_n^{\mathsf{C}}(\boldsymbol{\beta}) := \frac{1}{n(n-1)}\sum_{i\neq j} M(Y_i)\mathbb{1}(\boldsymbol{X}_i^\top\boldsymbol{\beta} > \boldsymbol{X}_j^\top\boldsymbol{\beta})$$

and one candidate function for $M(y)$ is

$$M(y) = a\mathbb{1}(y < a) + y\mathbb{1}(a \le y \le b) + b\mathbb{1}(y > b).$$

Here a and b are two absolute constants, and hence $M(y)$ is a trimming function for balancing the statistical efficiency and robustness to outliers. Let $\boldsymbol{\beta}_0 = (1, \boldsymbol{\theta}_0^\top)^\top$, and we aim to estimate $\boldsymbol{\theta}_0$.

We define the estimator $\widehat{\boldsymbol{\theta}}_n^{\mathsf{C}}$ and other parameters similarly as in Sections 1.2 and 3.1, with their explicit definitions relegated to the Appendix A.2.1. Then we have the following corollary.

Corollary 3.2. *We have*

(i) Under Assumption 7(i)–(iii) in the Appendix A.2.1, if $p_n/n = o(1)$, then $\|\widehat{\boldsymbol{\theta}}_n^{\mathsf{C}} - \boldsymbol{\theta}_0\| \xrightarrow{\mathbb{P}} 0$.

(ii) Suppose that Assumption 7 holds. If $p_n/n = o(1)$, then

$$\|\widehat{\boldsymbol{\theta}}_n^{\mathsf{C}} - \boldsymbol{\theta}_0\|^2 = O_{\mathbb{P}}(p_n/n).$$

(iii) Suppose that Assumptions 5–7 hold. If $p_n^2/n = o(1)$ and $\log(n/p_n^2)p_n^{3/2}/n^{5/4} = o(1)$, we have

$$\|\widehat{\boldsymbol{\theta}}_n^{\mathsf{C}} - \boldsymbol{\theta}_0 + (\mathbf{V}^{\mathsf{C}})^{-1}\mathbb{P}_n\nabla_1\tau^{\mathsf{C}}(\cdot;\boldsymbol{\theta}_0)\|^2 = O_{\mathbb{P}}\big\{\log(n/p_n^2)p_n^{3/2}/n^{5/4}\big\}.$$

If further $\log(n/p_n^2)p_n^{3/2}/n^{1/4} = o(1)$, then for any $\boldsymbol{\gamma} \in \mathbb{R}^{p_n}$,

$$\sqrt{n}\boldsymbol{\gamma}^\top(\widehat{\boldsymbol{\theta}}_n^{\mathsf{C}} - \boldsymbol{\theta}_0)/\{\boldsymbol{\gamma}^\top(\mathbf{V}^{\mathsf{C}})^{-1}\boldsymbol{\Delta}^{\mathsf{C}}(\mathbf{V}^{\mathsf{C}})^{-1}\boldsymbol{\gamma}\}^{1/2} \Rightarrow N(0,1).$$

(iv) Under conditions in (iii), if we further have $\varepsilon_n\sqrt{p_n} = o(1)$ and $\varepsilon_n^{-2}p_n/\sqrt{n} = o(1)$, then

$$\|(\widehat{\mathbf{V}}^{\mathsf{C}})^{-1}\widehat{\boldsymbol{\Delta}}^{\mathsf{C}}(\widehat{\mathbf{V}}^{\mathsf{C}})^{-1} - (\mathbf{V}^{\mathsf{C}})^{-1}\boldsymbol{\Delta}^{\mathsf{C}}(\mathbf{V}^{\mathsf{C}})^{-1}\| \xrightarrow{\mathbb{P}} 0.$$

In particular, we could choose $\epsilon_n \asymp (p_n/n)^{1/6}$, which will render a consistent covariance estimator under the same scaling condition as (iii).

3.3. Khan and Tamer's rank estimator for duration models

Consider Khan and Tamer's setting (Khan and Tamer, 2007), where the data are subject to censoring and the variable Y is no longer always observed. Use ξ to denote the random censoring variable, which can be arbitrarily correlated with $\boldsymbol{X}$. Let R be a binary variable indicating whether Y is uncensored or not. Let V denote a scalar random variable with $V = Y$ for uncensored observations, and $V = \xi$ otherwise. Consider the following right censored transformation model (Khan and Tamer, 2007):

$$T(V) = \min(\boldsymbol{X}^\top\boldsymbol{\beta}_0 + \epsilon, \xi),$$
$$R = \mathbb{1}(\boldsymbol{X}^\top\boldsymbol{\beta}_0 + \epsilon \le \xi),$$

where $T(\cdot)$ is assumed to be strictly monotonic. The (p_n+1)-dimensional vector $\boldsymbol{\beta}_0$ is unknown and is to be estimated.

Khan and Tamer (2007) proposed estimating $\boldsymbol{\beta}_0$ with $\widehat{\boldsymbol{\beta}}_n^{\mathsf{K}} = \operatorname{argmax}_{\boldsymbol{\beta}:\beta_1=1} S_n^{\mathsf{K}}(\boldsymbol{\beta})$, where

$$S_n^{\mathsf{K}}(\boldsymbol{\beta}) := \frac{1}{n(n-1)}\sum_{i\neq j} R_i\mathbb{1}(V_i < V_j)\mathbb{1}(\boldsymbol{X}_i^\top\boldsymbol{\beta} < \boldsymbol{X}_j^\top\boldsymbol{\beta}).$$

Let $\boldsymbol{\beta}_0 = (1, \boldsymbol{\theta}_0^\top)^\top$, and we consider estimation of $\boldsymbol{\theta}_0$.

We define the estimator $\widehat{\boldsymbol{\theta}}_n^{\mathsf{K}}$ and other parameters similarly as in Sections 1.2 and 3.1, with their explicit definitions relegated to Appendix A.2.2. Then we have the following corollary.

Corollary 3.3. *We have*

(i) Under Assumption 8(i)–(iii) in Appendix A.2.2, if $p_n/n = o(1)$, then $\|\widehat{\boldsymbol{\theta}}_n^{\mathsf{K}} - \boldsymbol{\theta}_0\| \xrightarrow{\mathbb{P}} 0$.

(ii) Under Assumption 8, if $p_n/n = o(1)$, then

$$\|\widehat{\boldsymbol{\theta}}_n^{\mathsf{K}} - \boldsymbol{\theta}_0\|^2 = O_{\mathbb{P}}(p_n/n).$$

(iii) Suppose that Assumptions 5–6 and 8 hold. If $p_n^2/n = o(1)$ and $\log(n/p_n^2)p_n^{3/2}/n^{5/4} = o(1)$, we have

$$\|\widehat{\boldsymbol{\theta}}_n^{\mathsf{K}} - \boldsymbol{\theta}_0 + (\mathbf{V}^{\mathsf{K}})^{-1}\mathbb{P}_n\nabla_1\tau^{\mathsf{K}}(\cdot;\boldsymbol{\theta}_0)\|^2 = O_{\mathbb{P}}\big\{\log(n/p_n^2)p_n^{3/2}/n^{5/4}\big\}.$$

If further $\log(n/p_n^2)p_n^{3/2}/n^{1/4} = o(1)$, then for any $\boldsymbol{\gamma} \in \mathbb{R}^{p_n}$,

$$\sqrt{n}\boldsymbol{\gamma}^\top(\widehat{\boldsymbol{\theta}}_n^{\mathsf{K}} - \boldsymbol{\theta}_0)/\{\boldsymbol{\gamma}^\top(\mathbf{V}^{\mathsf{K}})^{-1}\boldsymbol{\Delta}^{\mathsf{K}}(\mathbf{V}^{\mathsf{K}})^{-1}\boldsymbol{\gamma}\}^{1/2} \Rightarrow N(0,1).$$

(iv) Under conditions in (iii), if we further have $\varepsilon_n\sqrt{p_n} = o(1)$ *and* $\varepsilon_n^{-2}p_n/\sqrt{n} = o(1)$*, then*

$$\|(\widehat{\mathbf{V}}^{\mathsf{K}})^{-1}\widehat{\mathbf{\Delta}}^{\mathsf{K}}(\widehat{\mathbf{V}}^{\mathsf{K}})^{-1} - (\mathbf{V}^{\mathsf{K}})^{-1}\mathbf{\Delta}^{\mathsf{K}}(\mathbf{V}^{\mathsf{K}})^{-1}\| \xrightarrow{\mathbb{P}} 0.$$

In particular, we could choose $\epsilon_n \asymp (p_n/n)^{1/6}$*, which will render a consistent covariance estimator under the same scaling condition as (iii).*

3.4. Abrevaya and Shin's rank estimator for partially linear index models

Consider Abrevaya and Shin's partially linear index model (Abrevaya and Shin, 2011):

$$Y = T(\boldsymbol{X}^\top\boldsymbol{\beta}_0 + \eta(W) + \epsilon),$$

where $\boldsymbol{X} \in \mathbb{R}^{p_n+1}$, $W \in \mathbb{R}$, $T(\cdot)$ is a non-degenerate monotone function, $\eta(\cdot)$ is a smooth function, and ϵ is a random noisy independent of $(\boldsymbol{X}^\top, W)^\top$. Our primary interest is to estimate $\boldsymbol{\beta}_0 \in \mathbb{R}^{p_n+1}$. For this, Abrevaya and Shin (2011) proposed using $\widehat{\boldsymbol{\beta}}_n^{\mathsf{A}} = \mathrm{argmax}_{\boldsymbol{\beta}:\beta_1=1} S_n^{\mathsf{A}}(\boldsymbol{\beta})$, where

$$S_n^{\mathsf{A}}(\boldsymbol{\beta}) := \frac{1}{n(n-1)}\sum_{i\neq j}\mathbb{1}(Y_i > Y_j)\mathbb{1}(\boldsymbol{X}_i^\top\boldsymbol{\beta} > \boldsymbol{X}_j^\top\boldsymbol{\beta})K_b(W_i - W_j).$$

Here $K_b(u) := b^{-1}K(u/b)$ is a function facilitating pairwise comparison (Honoré and Powell, 2005). It involves a kernel function $K(\cdot)$ and a bandwidth parameter b. Let $\boldsymbol{\beta}_0 = (1, \boldsymbol{\theta}_0^\top)^\top$. Our aim is to estimate $\boldsymbol{\theta}_0$.

With the estimator $\widehat{\boldsymbol{\theta}}_n^{\mathsf{A}}$ and other parameters similarly defined as in Sections 1.2 and 3.1 and put in the Appendix A.2.3, we have the following corollary.

Corollary 3.4. *We have*

(i) Under Assumption 9(i)–(vii) in Appendix A.2.3, if $p_n/n^{1-2\delta} = o(1)$*, then* $\|\widehat{\boldsymbol{\theta}}_n^{\mathsf{A}} - \boldsymbol{\theta}_0\| \xrightarrow{\mathbb{P}} 0$.

(ii) Under Assumption 9, if $p_n/n^{1-\delta} \to 0$*, then*

$$\|\widehat{\boldsymbol{\theta}}_n^{\mathsf{A}} - \boldsymbol{\theta}_0\|^2 = O_{\mathbb{P}}\Big(\frac{p_n}{n^{1-\delta}} \wedge \frac{p_n^{3/2}}{n}\Big).$$

(iii) Under Assumptions 5 and 9–10, as $p_n^2/n^{1-\delta} = o(1)$ *and* $\log(n^{1-\delta}/p_n^2)p_n^{3/2}/n^{(5-5\delta)/4} = o(1)$*, we have*

$$\|\widehat{\boldsymbol{\theta}}_n^{\mathsf{A}} - \boldsymbol{\theta}_0 + (\mathbf{V}^{\mathsf{A}})^{-1}\mathbb{P}_n\nabla_1\tau^{\mathsf{A}}(\cdot;\boldsymbol{\theta}_0)\|^2 = O_{\mathbb{P}}\big\{n^{-\delta J} \vee \log(n^{1-\delta}/p_n^2)p_n^{3/2}/n^{(5-5\delta)/4}\big\}.$$

If further $\log(n^{1-\delta}/p_n^2)p_n^{3/2}/n^{(1-5\delta)/4} = o(1)$*, then for any* $\boldsymbol{\gamma} \in \mathbb{R}^{p_n}$*,*

$$\sqrt{n}\boldsymbol{\gamma}^\top(\widehat{\boldsymbol{\theta}}_n^{\mathsf{A}} - \boldsymbol{\theta}_0)/\{\boldsymbol{\gamma}^\top(\mathbf{V}^{\mathsf{A}})^{-1}\mathbf{\Delta}^{\mathsf{A}}(\mathbf{V}^{\mathsf{A}})^{-1}\boldsymbol{\gamma}\}^{1/2} \Rightarrow N(0,1).$$

(iv) Under conditions in (iii), if we further have $\varepsilon_n\sqrt{p_n} = o(1)$ *and* $\varepsilon_n^{-2}p_n/\sqrt{n^{1-2\delta}} = o(1)$*, then*

$$\|(\widehat{\mathbf{V}}^{\mathsf{A}})^{-1}\widehat{\mathbf{\Delta}}^{\mathsf{A}}(\widehat{\mathbf{V}}^{\mathsf{A}})^{-1} - (\mathbf{V}^{\mathsf{A}})^{-1}\mathbf{\Delta}^{\mathsf{A}}(\mathbf{V}^{\mathsf{A}})^{-1}\| \xrightarrow{\mathbb{P}} 0.$$

In particular, we could choose $\epsilon_n \asymp (p_n/n^{1-2\delta})^{1/6}$*. This will render a consistent covariance estimator under the scaling condition* $[p_n^4/n^{1-2\delta} \vee \{\log(n^{1-\delta}/p_n^2)\}^4 p_n^6/n^{1-5\delta}] = o(1)$*, which, at various cases, will be the same as the scaling condition in (iii).*

4. Simulation results

This section presents results from a small simulation study to illustrate two main implications of our theory. First for each fixed n, the normal approximation to the finite sample distribution of the studied rank correlation estimator will quickly become unreliable as p_n grows, suggesting that our theoretical bound is difficult to be improved in a significant way. Secondly, in estimating the asymptotic covariance based on the covariance estimator of the numerical derivative form, as n fixed, the tuning parameter that minimizes the Median Absolute Error (MAE) of the estimator will increase with the dimension p_n, echoing our theoretical observation.

In the simulation study, we focus on Han's MRC estimator of the form (1.3) and the following binary choice model:

$$Y_i = \mathbb{1}(\boldsymbol{X}_i^\top\boldsymbol{\beta}^* + \epsilon_i \geq 0),\ \ i = 1, \ldots, n,$$

where $\boldsymbol{X}_i \sim N(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma}_{jk} = 0.5^{|j-k|}$, $\epsilon_i \sim N(0,1)$, and $\boldsymbol{\beta}^* = (2, 4, 6, \ldots, 2(p+1))^\top$ representing the true regression coefficient. For each $n = 100, 200, 400$ and $p_n = 1, 2, 3, 4$, we simulate independent observations $\{Y_i, \boldsymbol{X}_i\}_{i=1}^n$ from the above model. Let $\boldsymbol{\beta}_0^* := \boldsymbol{\beta}^*/\boldsymbol{\beta}_1^*$ be the normalized regression coefficient. We aim to estimate $\boldsymbol{\beta}_0^*$ using Han's estimator $\widehat{\boldsymbol{\beta}}_n^{\mathsf{H}}$, which is implemented using the iterative marginal optimization algorithm proposed by Wang (2007), with the initial point chosen to be the truth.

Table 1
Coverage probability under the first projection direction.

n	p	Nominal coverage probability									
		0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
100	1	0.606	0.644	0.692	0.731	0.781	0.822	0.860	0.890	0.914	0.932
	2	0.806	0.829	0.844	0.862	0.881	0.895	0.907	0.920	0.930	0.948
	3	0.923	0.930	0.938	0.947	0.953	0.957	0.963	0.964	0.970	0.973
	4	0.877	0.892	0.905	0.920	0.926	0.939	0.945	0.948	0.956	0.964
200	1	0.518	0.561	0.619	0.672	0.719	0.763	0.809	0.861	0.903	0.939
	2	0.598	0.655	0.704	0.754	0.801	0.826	0.863	0.890	0.912	0.938
	3	0.702	0.746	0.788	0.820	0.846	0.874	0.893	0.911	0.930	0.953
	4	0.852	0.871	0.887	0.902	0.920	0.923	0.934	0.940	0.952	0.960
400	1	0.502	0.552	0.588	0.648	0.699	0.749	0.797	0.857	0.900	0.946
	2	0.500	0.555	0.604	0.663	0.724	0.766	0.819	0.858	0.905	0.945
	3	0.576	0.627	0.672	0.715	0.765	0.809	0.844	0.882	0.900	0.929
	4	0.613	0.672	0.711	0.737	0.782	0.833	0.870	0.890	0.920	0.944

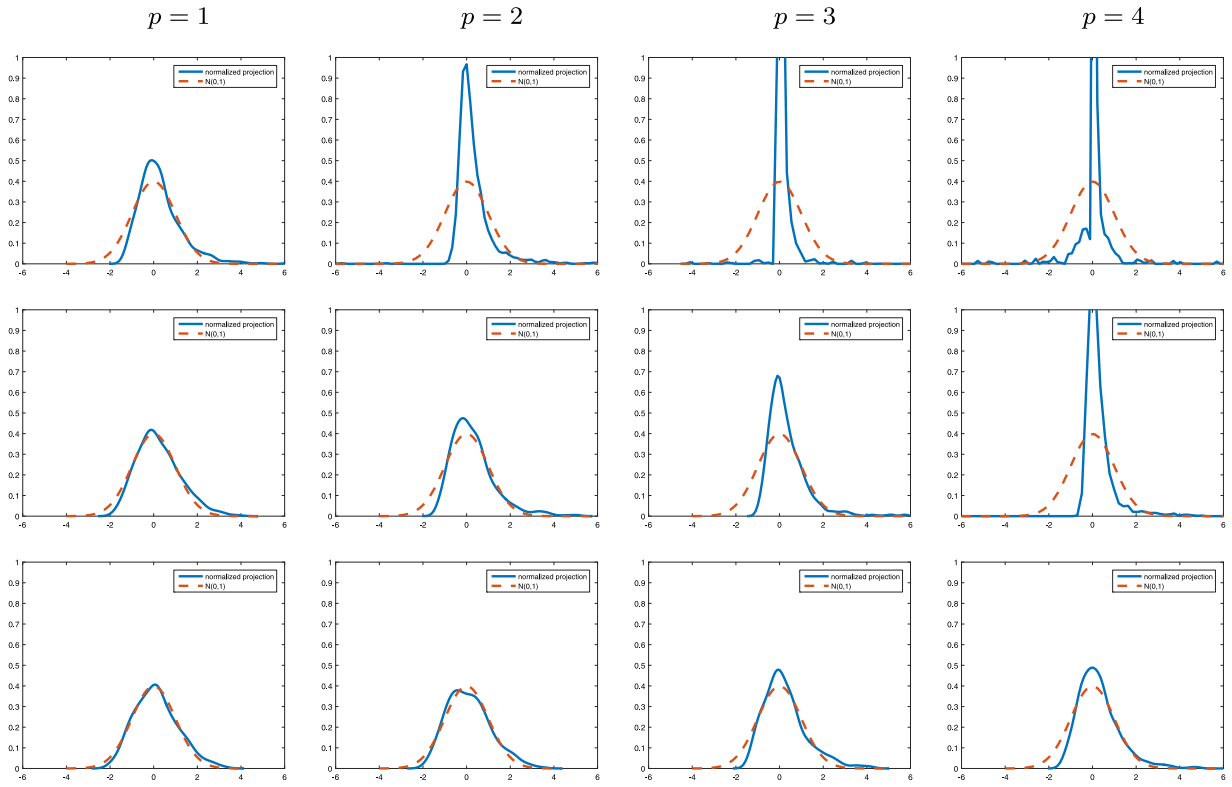

Fig. 1. Plots of the kernel density estimates of the normalized estimates (blue) v.s. $N(0, 1)$ (red) under the first projection direction ($n = 100, 200, 400$ from top to bottom).

Based on 1000 independent replications and using two-sided normal confidence interval, Tables 1–3 present the coverage probability as the nominal one varies from 0.5 to 0.95 for three projections of the same directions as $(1, 1, \ldots, 1)^\top$, $(1, 0, \ldots, 0)^\top$, and $(1, 2, \ldots, p_n)^\top$. For calculating the confidence intervals, we used the sample standard deviation of 1000 replications. We further plot the kernel estimates of the density functions of the normalized three projected estimates against the density function of $N(0, 1)$ in Figs. 1–3. The normalization is based on the true mean and the previous simulation-based standard deviation. In computing the kernel density estimates, we used normal kernel function and the bandwidth based on Silverman's rule-of-thumb.

Both the tables and figures reveal the same overall pattern that, for each fixed n, as p_n increases, the coverage probability will deviate more from the nominal, and the kernel estimates of the density function of the normalized estimator itself will deviate more from the standard normal. As observed, the deviation from normal has become very severe even for very small p_n. For example, for $p_n = 2$, we need n to be approximately 400 for achieving satisfactory coverage probability. This supports the theoretical observations in Theorem 2.4 and Corollary 3.1(iii). We further conduct

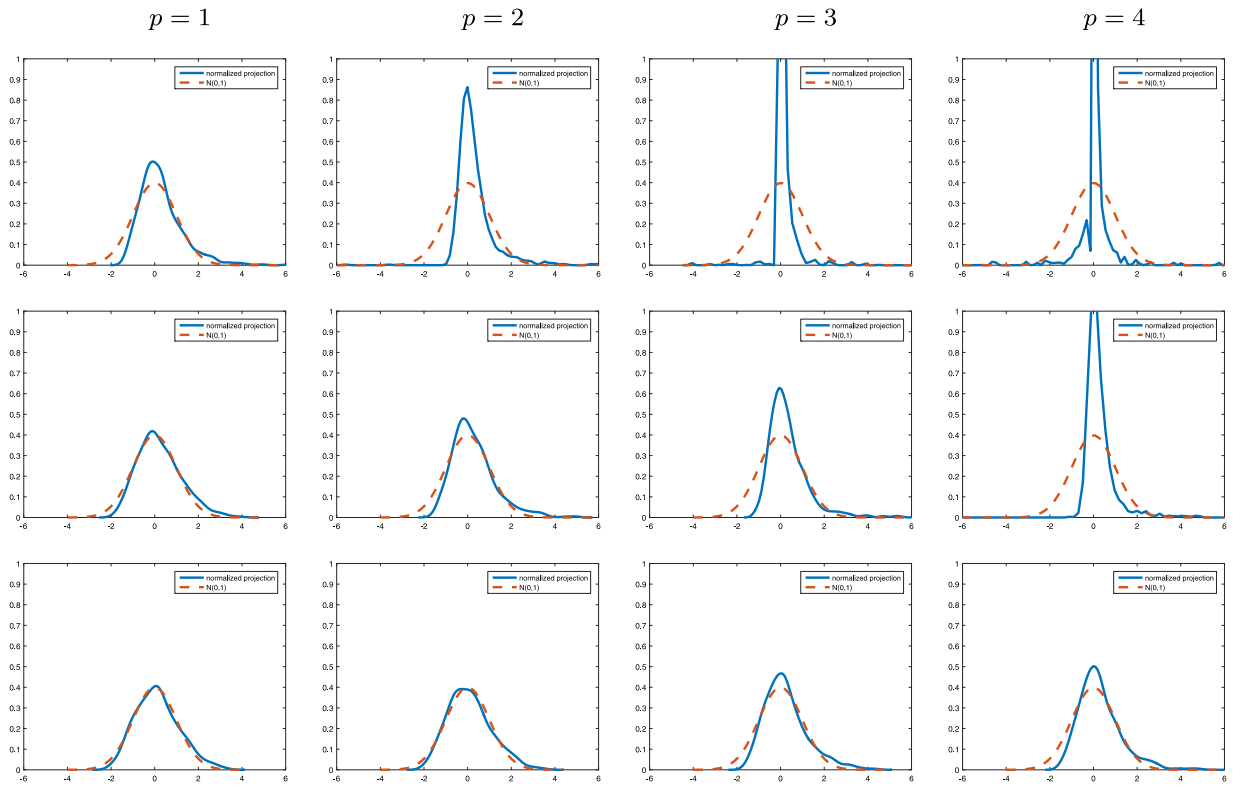

Fig. 2. Plots of the kernel density estimates of the normalized estimates (blue) v.s. $N(0, 1)$ (red) under the second projection direction ($n = 100, 200, 400$ from top to bottom).

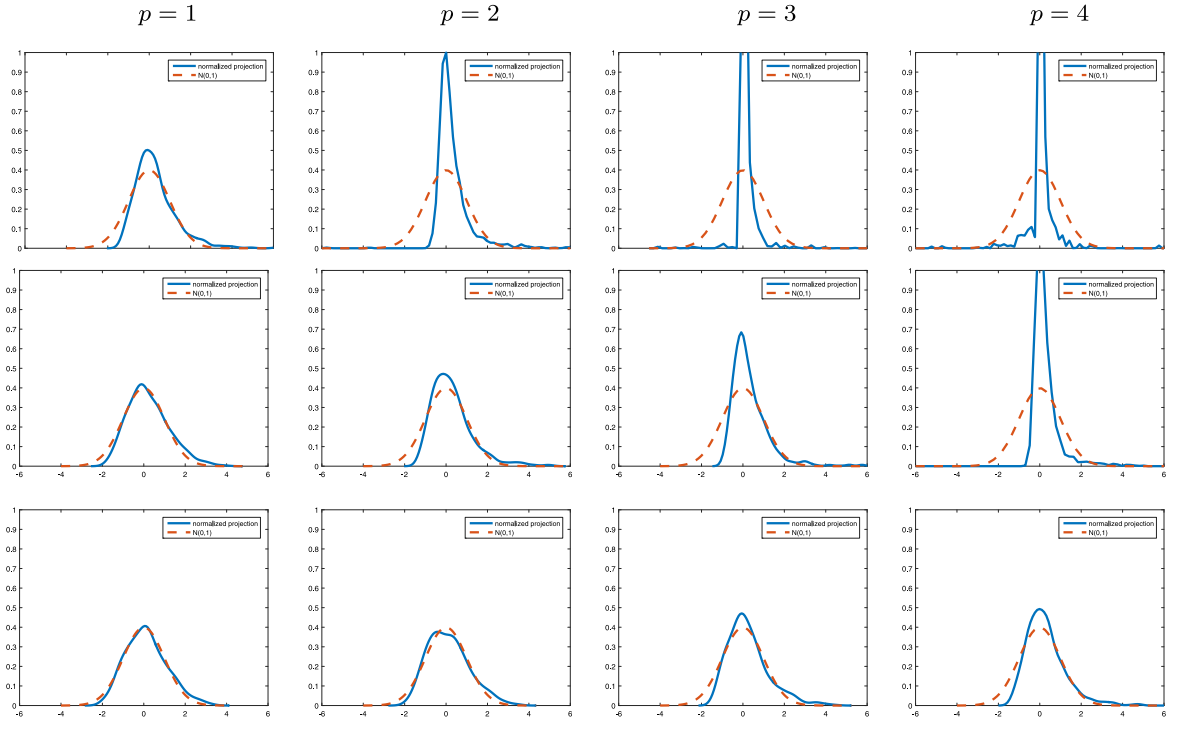

Fig. 3. Plots of the kernel density estimates of the normalized estimates (blue) v.s. $N(0, 1)$ (red) under the third projection direction ($n = 100, 200, 400$ from top to bottom).

Table 2
Coverage probability under the second projection direction.

n	p	Nominal coverage probability									
		0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
100	1	0.606	0.644	0.692	0.731	0.781	0.822	0.860	0.890	0.914	0.932
	2	0.790	0.820	0.839	0.858	0.875	0.890	0.905	0.914	0.929	0.945
	3	0.920	0.928	0.938	0.947	0.952	0.956	0.963	0.965	0.970	0.973
	4	0.876	0.890	0.903	0.918	0.926	0.939	0.944	0.949	0.956	0.965
200	1	0.518	0.561	0.619	0.672	0.719	0.763	0.809	0.861	0.903	0.939
	2	0.578	0.638	0.691	0.732	0.773	0.810	0.857	0.883	0.909	0.934
	3	0.699	0.735	0.770	0.801	0.831	0.869	0.889	0.912	0.929	0.947
	4	0.841	0.865	0.883	0.900	0.911	0.919	0.932	0.943	0.952	0.958
400	1	0.502	0.552	0.588	0.648	0.699	0.749	0.797	0.857	0.900	0.946
	2	0.519	0.573	0.623	0.661	0.701	0.754	0.810	0.861	0.901	0.947
	3	0.568	0.615	0.673	0.717	0.760	0.800	0.837	0.868	0.903	0.929
	4	0.592	0.637	0.675	0.732	0.774	0.817	0.856	0.881	0.911	0.937

Table 3
Coverage probability under the third projection direction.

n	p	Nominal coverage probability									
		0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
100	1	0.606	0.644	0.692	0.731	0.781	0.822	0.860	0.890	0.914	0.932
	2	0.804	0.828	0.846	0.861	0.880	0.897	0.907	0.921	0.931	0.948
	3	0.923	0.929	0.938	0.947	0.953	0.957	0.963	0.964	0.970	0.974
	4	0.877	0.892	0.904	0.920	0.926	0.939	0.945	0.948	0.956	0.964
200	1	0.518	0.561	0.619	0.672	0.719	0.763	0.809	0.861	0.903	0.939
	2	0.601	0.658	0.710	0.754	0.799	0.828	0.864	0.895	0.913	0.940
	3	0.712	0.749	0.787	0.820	0.843	0.874	0.893	0.913	0.930	0.954
	4	0.852	0.870	0.886	0.902	0.919	0.924	0.933	0.940	0.952	0.960
400	1	0.502	0.552	0.588	0.648	0.699	0.749	0.797	0.857	0.900	0.946
	2	0.502	0.547	0.602	0.661	0.720	0.771	0.813	0.861	0.908	0.944
	3	0.566	0.618	0.672	0.720	0.765	0.808	0.844	0.881	0.902	0.931
	4	0.617	0.663	0.708	0.738	0.789	0.835	0.871	0.892	0.920	0.946

Table 4
MAE of the covariance estimator. The results are obtained using 1000 replications.

ϵ_n	$n = 100$				$n = 200$				$n = 400$			
	$p = 1$	$p = 2$	$p = 3$	$p = 4$	$p = 1$	$p = 2$	$p = 3$	$p = 4$	$p = 1$	$p = 2$	$p = 3$	$p = 4$
$1.1n^{-1/6}$	0.476	1.509	2.198	**3.705**	0.170	**0.656**	**1.208**	**1.601**	0.081	**0.267**	**0.635**	**1.101**
$0.9n^{-1/6}$	**0.468**	1.555	**2.197**	3.786	**0.160**	0.663	1.269	1.695	0.073	0.275	0.671	1.144
$0.7n^{-1/6}$	0.494	1.433	2.247	3.870	**0.160**	0.690	1.257	1.755	**0.071**	0.303	0.722	1.214
$0.5n^{-1/6}$	0.521	1.402	2.473	3.802	0.175	0.755	1.334	1.774	0.081	0.339	0.764	1.261
$0.3n^{-1/6}$	0.503	**1.379**	2.665	3.867	0.235	0.814	1.445	1.916	0.121	0.408	0.843	1.343
$0.1n^{-1/6}$	0.657	1.464	2.962	4.762	0.329	0.874	1.452	2.161	0.201	0.475	0.869	1.379

different types of normality tests (Kolmogorov–Smirnov, Lilliefors, Jarque–Bera, Anderson–Darling, Henze–Zirkler) on the derived projected estimates as well as the original multi-dimensional estimates. They all reject the null hypothesis of normality except when $p_n = 1, n = 400$.

We then move on to study the estimation accuracy of the asymptotic covariance estimator discussed at the end of Section 2.2. For this, we focus on the same setup as previously conducted. Table 4 presents the MAE of the asymptotic covariance estimator for the projection direction $\{p_n^{-1/2}, \ldots, p_n^{-1/2}\}^\top$. There, it could be observed that, for each fixed n, the tuning parameter that attains the smallest MAE will in general become larger as p_n increases, supporting our observation in Theorem 2.6 and Corollary 3.1(iv).

Concluding remarks

This paper provided a first study of asymptotic properties of a general class of estimators defined as minimizers of possibly discontinuous objective functions of U-process structure allowing for the dimension of the parameter vector of interest to increase to infinity as the sample size n increases to infinity. Members of this class include important rank correlation estimators as detailed throughout this paper. Technically we have established a maximal inequality for degenerate U-processes in increasing dimensions which has played a critical role in deriving our theoretical results. We have also applied our general theory to the four motivating rank correlation estimators. Using Han's MRC estimator of

the form (1.3), we have provided numerical support to our theoretical findings that for a given sample size, the accuracy of the normal approximation deteriorates quickly as the number of parameters p_n increases and that for the variance estimation, the step size needs to be adjusted with respect to p_n.

This paper is focused on the setting that the parameter of interest itself is of an increasing dimension and inference has to be drawn on it. On the contrary, a growing literature studies the case that the parameter to be inferred is of a fixed dimension, but allows for a dimension-increasing (but still less than n) nuisance in the model. Substantial developments have been made along this line. For example, Cattaneo et al. (2018a,b) studied inferring the fixed-dimension linear component in a partially linear model, and Lei et al. (2018) established asymptotic normality of margins of linear and robust regression estimators in a simple linear model. Their set-up is fundamentally different from ours due to the difference of goals.[1]

We end this section with a brief discussion on further extensions. An immediate extension is on studying "penalized" rank estimators in ultra high dimensional settings where the dimension could be even larger than the sample size. For this much more challenging setting, to the authors' knowledge, most literature is still focused on simple structural statistical models (cf. Zhang and Zhang (2014), Van de Geer et al. (2014), Lee et al. (2016), and Javanmard and Montanari (2018) among many others). A notable exception is the post-selection inference framework proposed in Belloni et al. (2014, 2018), where a general set of regularization conditions has been posed for inference validity of Z-estimation. The authors believe that, combined with our local entropy analysis of the degenerate U-processes and the empirical process techniques developed by Talagrand and Spokoiny and specialized to rank estimators in this paper, the post-selection inference framework will prove useful in extending the current study to ultra high dimensional models. However, there are still many technical gaps, which we believe are fundamental and related to some key challenges in high dimensional probability in extending the scalar empirical processes to vector and matrix ones if no further smoothing (cf. Han et al. (2017)) is made. We will leave this for future research.

Acknowledgments

We thank Dr. Hansheng Wang for providing the code to implement the iterative marginal optimization algorithm, Mr. Shuo Jiang for helping conduct the simulations, and seminar/conference participants at Emory University, Peking University, and the 2019 Econometrics Workshop at Shanghai University of Finance and Economics for helpful comments. The research of Fang Han was supported in part by National Science Foundation, USA grant DMS-1712536. We are also grateful to the Associate Editor and two anonymous referees for instructive comments that have greatly improved the paper.

Appendix

A.1. Additional notation

For a vector $\boldsymbol{\alpha} \in \mathbb{R}^l$, we define $|\boldsymbol{\alpha}| = (|\alpha_1|, \ldots, |\alpha_l|)^\top$. For two sequences of real numbers a_n and b_n, $a_n \lesssim b_n$ means that $a_n \leq b_n$ up to a multiplicative constant. We use the symbol $a_n \sim b_n$ to denote that $a_n \lesssim b_n$ and $b_n \lesssim a_n$. In this appendix we drop the subscript n in m_n, ν_n, p_n.

A.2. Notation and assumptions in Section 3

Throughout this section, let $\boldsymbol{X} = (X_1, \widetilde{\boldsymbol{X}}^\top)^\top$, where $\widetilde{\boldsymbol{X}}$ denotes the last p components in $\boldsymbol{X}$.

A.2.1. Notation and assumptions in Section 3.2

The following definitions are similar to those in Section 1.2. We use $S^{\mathrm{C}}(\boldsymbol{\beta})$ to denote the expected value of $S_n^{\mathrm{C}}(\boldsymbol{\beta})$, and $S^{\mathrm{C}}(\boldsymbol{\beta}) = \mathbb{E}\{M(Y_1)\mathbb{1}(\boldsymbol{X}_1^\top\boldsymbol{\beta} > \boldsymbol{X}_2^\top\boldsymbol{\beta})\}$. Let $\boldsymbol{z} = (y, \boldsymbol{x}^\top)^\top$. We define

$$f^{\mathrm{C}}(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta}) = M(y_1)\{\mathbb{1}(\boldsymbol{x}_1^\top\boldsymbol{\beta} > \boldsymbol{x}_2^\top\boldsymbol{\beta}) - \mathbb{1}(\boldsymbol{x}_1^\top\boldsymbol{\beta}_0 > \boldsymbol{x}_2^\top\boldsymbol{\beta}_0)\},$$

$$\tau^{\mathrm{C}}(\boldsymbol{z}; \boldsymbol{\theta}) = \mathbb{E}f^{\mathrm{C}}(\boldsymbol{z}, \cdot; \boldsymbol{\theta}) + \mathbb{E}f^{\mathrm{C}}(\cdot, \boldsymbol{z}; \boldsymbol{\theta}), \quad \zeta^{\mathrm{C}}(\boldsymbol{z}; \boldsymbol{\theta}) = \tau^{\mathrm{C}}(\boldsymbol{z}; \boldsymbol{\theta}) - \mathbb{E}\tau^{\mathrm{C}}(\cdot; \boldsymbol{\theta}),$$

$$\boldsymbol{\Delta}^{\mathrm{C}} = \mathbb{E}\nabla_1\tau^{\mathrm{C}}(\cdot; \boldsymbol{\theta}_0)\{\nabla_1\tau^{\mathrm{C}}(\cdot; \boldsymbol{\theta}_0)\}^\top, \text{ and } 2\mathbf{V}^{\mathrm{C}} = \mathbb{E}\nabla_2\tau^{\mathrm{C}}(\cdot; \boldsymbol{\theta}_0).$$

Write $\Gamma^{\mathrm{C}}(\boldsymbol{\theta})$ for $S^{\mathrm{C}}(\boldsymbol{\beta}) - S^{\mathrm{C}}(\boldsymbol{\beta}_0)$ and $\Gamma_n^{\mathrm{C}}(\boldsymbol{\theta})$ for $S_n^{\mathrm{C}}(\boldsymbol{\beta}) - S_n^{\mathrm{C}}(\boldsymbol{\beta}_0)$. The estimator $\widehat{\boldsymbol{\theta}}_n^{\mathrm{C}}$ is defined as

$$\widehat{\boldsymbol{\theta}}_n^{\mathrm{C}} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta^{\mathrm{C}}} \Gamma_n^{\mathrm{C}}(\boldsymbol{\theta}).$$

[1] We note that our set-up is also fundamentally different from works on "many moment asymptotics" in GMM models such as Han and Phillips (2006), Newey and Windmeijer (2009), and Caner (2014), where the number of moment conditions increases but the number of parameters in such models is fixed as the sample size increases.

To conduct inference on $\boldsymbol{\theta}_0$ based on $\widehat{\boldsymbol{\theta}}_n^{\mathsf{C}}$, we further define

$$\tau_n^{\mathsf{C}}(\boldsymbol{z};\boldsymbol{\theta}) = \mathbb{P}_n f^{\mathsf{C}}(\boldsymbol{z},\cdot;\boldsymbol{\theta}) + \mathbb{P}_n f^{\mathsf{C}}(\cdot,\boldsymbol{z};\boldsymbol{\theta}), \quad p_{ni}^{\mathsf{C}}(\boldsymbol{z};\boldsymbol{\theta}) = \varepsilon_n^{-1}\{\tau_n^{\mathsf{C}}(\boldsymbol{z};\boldsymbol{\theta}+\varepsilon_n\boldsymbol{u}_i) - \tau_n^{\mathsf{C}}(\boldsymbol{z};\boldsymbol{\theta})\}, \quad \text{and}$$
$$p_{nij}^{\mathsf{C}}(\boldsymbol{z};\boldsymbol{\theta}) = \varepsilon_n^{-2}\{\tau_n^{\mathsf{C}}(\boldsymbol{z};\boldsymbol{\theta}+\varepsilon_n(\boldsymbol{u}_i+\boldsymbol{u}_j)) - \tau_n^{\mathsf{C}}(\boldsymbol{z};\boldsymbol{\theta}+\varepsilon_n\boldsymbol{u}_i) - \tau_n^{\mathsf{C}}(\boldsymbol{z};\boldsymbol{\theta}+\varepsilon_n\boldsymbol{u}_j) + \tau_n^{\mathsf{C}}(\boldsymbol{z};\boldsymbol{\theta})\}.$$

Then, we define the estimator of the matrix $\boldsymbol{\Delta}^{\mathsf{C}}$ as $\widehat{\boldsymbol{\Delta}}^{\mathsf{C}} = (\widehat{\delta}_{ij}^{\mathsf{C}})$ and the estimator of the matrix $\mathbf{V}^{\mathsf{C}}$ as $\widehat{\mathbf{V}}^{\mathsf{C}} = (\widehat{v}_{ij}^{\mathsf{C}})$, where

$$\widehat{\delta}_{ij}^{\mathsf{C}} = \mathbb{P}_n\{p_{ni}^{\mathsf{C}}(\cdot;\widehat{\boldsymbol{\theta}}_n^{\mathsf{C}})p_{nj}^{\mathsf{C}}(\cdot;\widehat{\boldsymbol{\theta}}_n^{\mathsf{C}})\}, \quad \text{and} \quad \widehat{v}_{ij}^{\mathsf{C}} = \frac{1}{2}\mathbb{P}_n p_{nij}^{\mathsf{C}}(\cdot;\widehat{\boldsymbol{\theta}}_n^{\mathsf{C}}).$$

We then make the following assumptions.

Assumption 7. Assume

(i) Assumption 1 holds for Θ^{C} and $\Gamma^{\mathsf{C}}(\boldsymbol{\theta})$.
(ii) The random variables $\boldsymbol{X}$ and ϵ are independent, and $\mathbb{E}\{M(Y) \mid \boldsymbol{X}\}$ depends on $\boldsymbol{X}$ only through $\boldsymbol{X}^\top\boldsymbol{\beta}_0$.
(iii) X_1 has an everywhere positive Lebesgue density, conditional on $\widetilde{\boldsymbol{X}}$.
(iv) Assumption 3 holds for $\tau^{\mathsf{C}}(\boldsymbol{z};\boldsymbol{\theta})$ and $\zeta^{\mathsf{C}}(\boldsymbol{z};\boldsymbol{\theta})$.

A.2.2. Notation and assumptions in Section 3.3

The following definitions are similar to those in Section 1.2. Let $S^{\mathsf{K}}(\boldsymbol{\beta})$ denote the expected value of $S_n^{\mathsf{K}}(\boldsymbol{\beta})$, and $S^{\mathsf{K}}(\boldsymbol{\beta}) = \mathbb{E}\{R_1\mathbb{1}(V_1 < V_2)\mathbb{1}(\boldsymbol{X}_1^\top\boldsymbol{\beta} < \boldsymbol{X}_2^\top\boldsymbol{\beta})\}$. Let $\boldsymbol{z} = (r, v, \boldsymbol{x}^\top)^\top$. We define

$$f^{\mathsf{K}}(\boldsymbol{z}_1,\boldsymbol{z}_2;\boldsymbol{\theta}) = r_1\mathbb{1}(v_1 < v_2)\{\mathbb{1}(\boldsymbol{x}_1^\top\boldsymbol{\beta} < \boldsymbol{x}_2^\top\boldsymbol{\beta}) - \mathbb{1}(\boldsymbol{x}_1^\top\boldsymbol{\beta}_0 < \boldsymbol{x}_2^\top\boldsymbol{\beta}_0)\},$$
$$\tau^{\mathsf{K}}(\boldsymbol{z};\boldsymbol{\theta}) = \mathbb{E}f^{\mathsf{K}}(\boldsymbol{z},\cdot;\boldsymbol{\theta}) + \mathbb{E}f^{\mathsf{K}}(\cdot,\boldsymbol{z};\boldsymbol{\theta}), \quad \zeta^{\mathsf{K}}(\boldsymbol{z};\boldsymbol{\theta}) = \tau^{\mathsf{K}}(\boldsymbol{z};\boldsymbol{\theta}) - \mathbb{E}\tau^{\mathsf{K}}(\cdot;\boldsymbol{\theta}),$$
$$\boldsymbol{\Delta}^{\mathsf{K}} = \mathbb{E}\nabla_1\tau^{\mathsf{K}}(\cdot;\boldsymbol{\theta}_0)\{\nabla_1\tau^{\mathsf{K}}(\cdot;\boldsymbol{\theta}_0)\}^\top, \quad \text{and} \quad 2\mathbf{V}^{\mathsf{K}} = \mathbb{E}\nabla_2\tau^{\mathsf{K}}(\cdot;\boldsymbol{\theta}_0).$$

Write $\Gamma^{\mathsf{K}}(\boldsymbol{\theta})$ for $S^{\mathsf{K}}(\boldsymbol{\beta}) - S^{\mathsf{K}}(\boldsymbol{\beta}_0)$ and $\Gamma_n^{\mathsf{K}}(\boldsymbol{\theta})$ for $S_n^{\mathsf{K}}(\boldsymbol{\beta}) - S_n^{\mathsf{K}}(\boldsymbol{\beta}_0)$. The estimator $\widehat{\boldsymbol{\theta}}_n^{\mathsf{K}}$ is defined as

$$\widehat{\boldsymbol{\theta}}_n^{\mathsf{K}} = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta^{\mathsf{K}}} \Gamma_n^{\mathsf{K}}(\boldsymbol{\theta}).$$

To conduct inference on $\boldsymbol{\theta}_0$ based on $\widehat{\boldsymbol{\theta}}_n^{\mathsf{K}}$, we further define

$$\tau_n^{\mathsf{K}}(\boldsymbol{z};\boldsymbol{\theta}) = \mathbb{P}_n f^{\mathsf{K}}(\boldsymbol{z},\cdot;\boldsymbol{\theta}) + \mathbb{P}_n f^{\mathsf{K}}(\cdot,\boldsymbol{z};\boldsymbol{\theta}), \quad p_{ni}^{\mathsf{K}}(\boldsymbol{z};\boldsymbol{\theta}) = \varepsilon_n^{-1}\{\tau_n^{\mathsf{K}}(\boldsymbol{z};\boldsymbol{\theta}+\varepsilon_n\boldsymbol{u}_i) - \tau_n^{\mathsf{K}}(\boldsymbol{z};\boldsymbol{\theta})\}, \quad \text{and}$$
$$p_{nij}^{\mathsf{K}}(\boldsymbol{z};\boldsymbol{\theta}) = \varepsilon_n^{-2}\{\tau_n^{\mathsf{K}}(\boldsymbol{z};\boldsymbol{\theta}+\varepsilon_n(\boldsymbol{u}_i+\boldsymbol{u}_j)) - \tau_n^{\mathsf{K}}(\boldsymbol{z};\boldsymbol{\theta}+\varepsilon_n\boldsymbol{u}_i) - \tau_n^{\mathsf{K}}(\boldsymbol{z};\boldsymbol{\theta}+\varepsilon_n\boldsymbol{u}_j) + \tau_n^{\mathsf{K}}(\boldsymbol{z};\boldsymbol{\theta})\}.$$

Then, we define the estimator of the matrix $\boldsymbol{\Delta}^{\mathsf{K}}$ as $\widehat{\boldsymbol{\Delta}}^{\mathsf{K}} = (\widehat{\delta}_{ij}^{\mathsf{K}})$ and the estimator of the matrix $\mathbf{V}^{\mathsf{K}}$ as $\widehat{\mathbf{V}}^{\mathsf{K}} = (\widehat{v}_{ij}^{\mathsf{K}})$, where

$$\widehat{\delta}_{ij}^{\mathsf{K}} = \mathbb{P}_n\{p_{ni}^{\mathsf{K}}(\cdot;\widehat{\boldsymbol{\theta}}_n^{\mathsf{K}})p_{nj}^{\mathsf{K}}(\cdot;\widehat{\boldsymbol{\theta}}_n^{\mathsf{K}})\} \quad \text{and} \quad \widehat{v}_{ij}^{\mathsf{K}} = \frac{1}{2}\mathbb{P}_n p_{nij}^{\mathsf{K}}(\cdot;\widehat{\boldsymbol{\theta}}_n^{\mathsf{K}}).$$

We then make the following assumptions.

Assumption 8. Assume

(i) Assumption 1 holds for Θ^{K} and $\Gamma^{\mathsf{K}}(\boldsymbol{\theta})$.
(ii) The random variables $(\xi, \boldsymbol{X})$ and ϵ are independent, and $\mathbb{E}(\xi \mid \boldsymbol{X})$ depends on $\boldsymbol{X}$ only through $\boldsymbol{X}^\top\boldsymbol{\beta}_0$.
(iii) X_1 has an everywhere positive Lebesgue density, conditional on $\widetilde{\boldsymbol{X}}$.
(iv) Assumption 3 holds for $\tau^{\mathsf{K}}(\boldsymbol{z};\boldsymbol{\theta})$ and $\zeta^{\mathsf{K}}(\boldsymbol{z};\boldsymbol{\theta})$.

A.2.3. Notation and assumptions in Section 3.4

Let $\phi(\cdot)$ denote the density of W. Let $\boldsymbol{z} = (y, \boldsymbol{x}^\top, w)^\top$. We define

$$f^{\mathsf{A}}(\boldsymbol{z}_1,\boldsymbol{z}_2;\boldsymbol{\theta}) = \mathbb{1}(y_1 > y_2)\{\mathbb{1}(\boldsymbol{x}_1^\top\boldsymbol{\beta} > \boldsymbol{x}_2^\top\boldsymbol{\beta}) - \mathbb{1}(\boldsymbol{x}_1^\top\boldsymbol{\beta}_0 > \boldsymbol{x}_2^\top\boldsymbol{\beta}_0)\}K\{(w_1-w_2)/b\},$$
$$m(\boldsymbol{z}_1,\boldsymbol{z}_2;\boldsymbol{\theta}) = \mathbb{1}(y_1 > y_2)\mathbb{1}(\boldsymbol{x}_1^\top\boldsymbol{\beta} > \boldsymbol{x}_2^\top\boldsymbol{\beta}),$$
$$\psi(w_1,w_2;\boldsymbol{\theta}) = \mathbb{E}\{m(\boldsymbol{Z}_1,\boldsymbol{Z}_2;\boldsymbol{\theta}) - m(\boldsymbol{Z}_1,\boldsymbol{Z}_2;\boldsymbol{\theta}_0) \mid W_1 = w_1, W_2 = w_2\},$$
$$\Gamma^{\mathsf{A}}(\boldsymbol{\theta}) = \mathbb{E}_W\{\psi(W,W;\boldsymbol{\theta})\phi(W)\},$$
$$\tau^{\mathsf{A}}(\boldsymbol{z};\theta) = \mathbb{E}\{m(\boldsymbol{z},\boldsymbol{Z}_2;\boldsymbol{\theta}) \mid W_2 = w\}\phi(w) + \mathbb{E}\{m(\boldsymbol{Z}_1,\boldsymbol{z};\boldsymbol{\theta}) \mid W_1 = w\}\phi(w),$$
$$\zeta^{\mathsf{A}}(\boldsymbol{z};\boldsymbol{\theta}) = \tau^{\mathsf{A}}(\boldsymbol{z};\boldsymbol{\theta}) - \mathbb{E}\tau^{\mathsf{A}}(\cdot;\boldsymbol{\theta}), \quad \boldsymbol{\Delta}^{\mathsf{A}} = \mathbb{E}\nabla_1\tau^{\mathsf{A}}(\cdot;\boldsymbol{\theta}_0)\{\nabla_1\tau^{\mathsf{A}}(\cdot;\boldsymbol{\theta}_0)\}^\top, \quad \text{and} \quad 2\mathbf{V}^{\mathsf{A}} = \mathbb{E}\nabla_2\tau^{\mathsf{A}}(\cdot;\boldsymbol{\theta}_0).$$

Write $\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta})$ for $S_n^{\mathsf{A}}(\boldsymbol{\beta}) - S_n^{\mathsf{A}}(\boldsymbol{\beta}_0)$. The estimator $\widehat{\boldsymbol{\theta}}_n^{\mathsf{A}}$ is defined as

$$\widehat{\boldsymbol{\theta}}_n^{\mathsf{A}} = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta^{\mathsf{A}}} \Gamma_n^{\mathsf{A}}(\boldsymbol{\theta}).$$

Note that $\mathbb{E}\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta}) \neq \Gamma^{\mathsf{A}}(\boldsymbol{\theta})$. This is different from the general set-up in Section 2.2. However, by Taylor expansion, we show that $\sup_{\boldsymbol{\theta}\in\Theta^{\mathsf{A}}} \left|\mathbb{E}\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta}) - \Gamma^{\mathsf{A}}(\boldsymbol{\theta})\right|$ is negligible under the assumptions adopted in this section. Then, following the proof of the general method, we can similarly establish the consistency and asymptotic normality of $\widehat{\boldsymbol{\theta}}_n^{\mathsf{A}}$.

To conduct inference on $\boldsymbol{\theta}_0$ based on $\widehat{\boldsymbol{\theta}}_n^{\mathsf{A}}$, we further define

$$\tau_n^{\mathsf{A}}(\boldsymbol{z};\boldsymbol{\theta}) = \mathbb{P}_n f^{\mathsf{A}}(\boldsymbol{z},\cdot;\boldsymbol{\theta}) + \mathbb{P}_n f^{\mathsf{A}}(\cdot,\boldsymbol{z};\boldsymbol{\theta}), \quad p_{ni}^{\mathsf{A}}(\boldsymbol{z};\boldsymbol{\theta}) = \varepsilon_n^{-1}\{\tau_n^{\mathsf{A}}(\boldsymbol{z};\boldsymbol{\theta}+\varepsilon_n\boldsymbol{u}_i) - \tau_n^{\mathsf{A}}(\boldsymbol{z};\boldsymbol{\theta})\}, \quad \text{and}$$
$$p_{nij}^{\mathsf{A}}(\boldsymbol{z};\boldsymbol{\theta}) = \varepsilon_n^{-2}\{\tau_n^{\mathsf{A}}(\boldsymbol{z};\boldsymbol{\theta}+\varepsilon_n(\boldsymbol{u}_i+\boldsymbol{u}_j)) - \tau_n^{\mathsf{A}}(\boldsymbol{z};\boldsymbol{\theta}+\varepsilon_n\boldsymbol{u}_i) - \tau_n^{\mathsf{A}}(\boldsymbol{z};\boldsymbol{\theta}+\varepsilon_n\boldsymbol{u}_j) + \tau_n^{\mathsf{A}}(\boldsymbol{z};\boldsymbol{\theta})\}.$$

Then, we define the estimator of the matrix $\boldsymbol{\Delta}^{\mathsf{A}}$ as $\widehat{\boldsymbol{\Delta}}^{\mathsf{A}} = (\widehat{\delta}_{ij}^{\mathsf{A}})$ and the estimator of the matrix $\mathbf{V}^{\mathsf{A}}$ as $\widehat{\mathbf{V}}^{\mathsf{A}} = (\widehat{v}_{ij}^{\mathsf{A}})$, where

$$\widehat{\delta}_{ij}^{\mathsf{A}} = \mathbb{P}_n\{p_{ni}^{\mathsf{A}}(\cdot;\widehat{\boldsymbol{\theta}}_n^{\mathsf{A}})p_{nj}^{\mathsf{A}}(\cdot;\widehat{\boldsymbol{\theta}}_n^{\mathsf{A}})\}, \quad \text{and} \quad \widehat{v}_{ij}^{\mathsf{A}} = \frac{1}{2}\mathbb{P}_n p_{nij}^{\mathsf{A}}(\cdot;\widehat{\boldsymbol{\theta}}_n^{\mathsf{A}}).$$

We make the following assumptions.

Assumption 9. Assume

(i) Assumption 1 holds for Θ^{A} and $\Gamma^{\mathsf{A}}(\boldsymbol{\theta})$;
(ii) The random variables $(\boldsymbol{X}, W)$ and ϵ are independent.
(iii) X_1 has an everywhere positive Lebesgue density, conditional on $\widetilde{\boldsymbol{X}}$ and W.
(iv) W is continuously distributed on a compact subset $\mathcal{W}$ of $\mathbb{R}$.
(v) The kernel function $K(\cdot)$ satisfies: (1) $K(\cdot)$ is twice continuously differential with compact interval $[-C, C] \supseteq \mathcal{W}$; (2) $K(\cdot)$ is symmetric about 0 and integrates to 1; (3) for some integer $J \geq 6$, $\int u^j K(u)\,\mathrm{d}u = 0$ with $j = 1, \ldots, J-1$ and $\int u^J K(u)\,\mathrm{d}u$ is bounded.
(vi) The bandwidth b is defined as $b = cn^{-\delta}$ for constants $c > 0$ and $\frac{1}{J} < \delta < \frac{1}{5}$.
(vii) For any w_2, the Jth derivative of $\psi(w_1, w_2;\boldsymbol{\theta})\cdot\phi(w_1)$ with respect to w_1 is continuous and bounded for all $\boldsymbol{\theta} \in \Theta^{\mathsf{A}}$.
(viii) Assumption 3 holds for $\tau^{\mathsf{A}}(\boldsymbol{z};\boldsymbol{\theta})$ and $\zeta^{\mathsf{A}}(\boldsymbol{z};\boldsymbol{\theta})$.

Assumption 10. Let $f_0(\cdot \mid \widetilde{\boldsymbol{x}}, w)$ denote the conditional density function of $\boldsymbol{X}^\top\boldsymbol{\beta}_0$ given $(\widetilde{\boldsymbol{X}}, W) = (\widetilde{\boldsymbol{x}}, w)$. Assume $f_0(\cdot \mid \widetilde{\boldsymbol{x}}, w) \leq C_1$ for any $\widetilde{\boldsymbol{x}}$ and w in the support of $\widetilde{\boldsymbol{X}}$ and W, respectively, where C_1 is an absolute positive constant.

A.3. Proofs in Section 2

For each $\boldsymbol{\theta} \in \Theta$, define measures

$$\mathbb{S}_n f(\cdot,\cdot;\boldsymbol{\theta}) = n(n-1)\mathbb{U}_n f(\cdot,\cdot;\boldsymbol{\theta})$$

and

$$\mathbb{T}_n f(\cdot,\cdot;\boldsymbol{\theta}) = \sum_{i\neq j}\{f(\boldsymbol{Z}_{2i},\boldsymbol{Z}_{2j};\boldsymbol{\theta}) + f(\boldsymbol{Z}_{2i},\boldsymbol{Z}_{2j-1};\boldsymbol{\theta}) + f(\boldsymbol{Z}_{2i-1},\boldsymbol{Z}_{2j};\boldsymbol{\theta}) + f(\boldsymbol{Z}_{2i-1},\boldsymbol{Z}_{2j-1};\boldsymbol{\theta})\}.$$

To prove Theorems 2.1–2.4 in Section 2, we need several lemmas. For simplicity, we omit the parameter $\boldsymbol{\theta}$ in each function $f(\cdot,\cdot;\boldsymbol{\theta}) \in \mathcal{F}$ in the lemmas. Let F denote the envelope function of $\mathcal{F}$ for which $0 < \mathbb{E}F^r < \infty$, for any $r \geq 1$. The covering number $N_r(\varepsilon, \mathbb{P}\otimes\mathbb{P}, \mathcal{F}, F)$ is defined as the smallest cardinality for a subclass $\mathcal{F}^*$ of $\mathcal{F}$ such that $\min_{f^*\in\mathcal{F}^*}\mathbb{E}|f - f^*|^r \leq \varepsilon^r\mathbb{E}F^r$, for each $f \in \mathcal{F}$.

A.3.1. Some auxiliary lemmas

Lemma A.1. *Suppose that $\mathcal{F}$ is b-uniformly bounded, then the class $\mathcal{F}^2 = \{f^2 : f \in \mathcal{F}\}$ with envelope b^2 satisfies $N_r(2\varepsilon, \mathbb{P}\otimes\mathbb{P}, \mathcal{F}^2, b^2) \leq N_r(\varepsilon, \mathbb{P}\otimes\mathbb{P}, \mathcal{F}, b)$.*

Proof. Find functions $f_1, \ldots, f_m$ such that

$$\min_i \mathbb{E}|f - f_i|^r \leq \varepsilon^r b^r, \quad \text{for each } f \in \mathcal{F}.$$

Then, with the appropriate i,

$$\mathbb{E}|f^2 - f_i^2|^r \leq (2b)^r\mathbb{E}|f - f_i|^r \leq (2b)^r\varepsilon^r b^r = (2\epsilon)^r(b^2)^r.$$

This implies that $N_r(2\varepsilon, \mathbb{P}\otimes\mathbb{P}, \mathcal{F}^2, b^2) \leq N_r(\varepsilon, \mathbb{P}\otimes\mathbb{P}, \mathcal{F}, b)$. □

Lemma A.2. *Suppose that $\mathcal{F}$ is b-uniformly bounded. Then $\mathbb{E}\sup_{g\in\mathbb{P}\mathcal{F}}|\mathbb{P}_n g - \mathbb{E}g| \lesssim \sqrt{v/n}$, where $\mathbb{P}\mathcal{F} := \{\mathbb{E}_{\mathbb{P}} f(\boldsymbol{z},\cdot) : f \in \mathcal{F}\}$.*

Proof. With a little abuse of notation, let $\epsilon_1, \epsilon_2, \ldots$ be the Rademacher sequence, where $\epsilon_i \in \{-1, 1\}$ is symmetric around 0. By the classic symmetrization theorem (cf. Theorem 8.8 in Kosorok, 2007), we have

$$\mathbb{E} \sup_{g \in \mathbb{P}\mathcal{F}} |\mathbb{P}_n g - \mathbb{E} g| \leq \mathbb{E}_{\boldsymbol{Z}} \mathbb{E}_{\epsilon} \sup_{g \in \mathbb{P}\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(\boldsymbol{Z}_i) \right|. \tag{A.1}$$

Next, we try to bound $\mathbb{E}_{\epsilon} \sup_{g \in \mathbb{P}\mathcal{F}} |\sum_{i=1}^{n} \epsilon_i g(\boldsymbol{z}_i)/n|$ for fixed $\boldsymbol{z}_i$. To that end, consider the stochastic process $\{\sum_{i=1}^{n} \epsilon_i g(\boldsymbol{z}_i)/\sqrt{n} : g \in \mathbb{P}\mathcal{F}\}$. It is easy to verify that $\sum_{i=1}^{n} \epsilon_i \{g_1(\boldsymbol{z}_i) - g_2(\boldsymbol{z}_i)\}/\sqrt{n}$ is sub-gaussian with parameter $\|g_1 - g_2\|_{L_2(\mathbb{P}_n)}^2 := \sum_{i=1}^{n} \{g_1(\boldsymbol{z}_i) - g_2(\boldsymbol{z}_i)\}^2/n$, where $g_1, g_2 \in \mathbb{P}\mathcal{F}$. Consequently, Dudley's entropy integral, combined with the fact that $\sup_{g_1, g_2 \in \mathbb{P}\mathcal{F}} \|g_1 - g_2\|_{L_2(\mathbb{P}_n)} \leq 2b$, implies that

$$\mathbb{E}_{\epsilon} \sup_{g \in \mathbb{P}\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(\boldsymbol{z}_i) \right| \leq \frac{24}{\sqrt{n}} \int_0^{2b} \sqrt{\log N_2(t/b, \mathbb{P}_n, \mathbb{P}\mathcal{F}, b)}\, \mathrm{d}t. \tag{A.2}$$

By Theorem 9.3 in Kosorok (2007) and Lemma 20 in Nolan and Pollard (1987), there exists a universal constant K such that $N_2(t/b, \mathbb{P}_n, \mathbb{P}\mathcal{F}, b) \leq K\nu(16e)^{\nu}(b/t)^{2(\nu-1)}$. Substituting this bound into (A.2), we find that there exist constants c_0, c_1, only depending on K, b but not on (ν, n), such that

$$\mathbb{E}_{\epsilon} \sup_{g \in \mathbb{P}\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(\boldsymbol{z}_i) \right| \leq c_0 \sqrt{\frac{\nu}{n}} \left\{ 1 + \int_0^{2b} \sqrt{\log(b/t)}\, \mathrm{d}t \right\} \leq c_1 \sqrt{\frac{\nu}{n}}.$$

Combining this with (A.1) implies that $\mathbb{E} \sup_{g \in \mathbb{P}\mathcal{F}} |\mathbb{P}_n g - \mathbb{E} g| \leq c_1 \sqrt{\nu/n}$. This completes the proof. □

Lemma A.3. *Suppose that $\mathcal{F}$ is $\mathbb{P}$-degenerate and b-uniformly bounded. Then $\mathbb{E} \sup_{f \in \mathcal{F}} |\mathbb{U}_n f| \lesssim \nu/n$.*

Proof. First, by the relationship between $\mathbb{S}_n$ and $\mathbb{U}_n$: $\mathbb{S}_n = n(n-1)\mathbb{U}_n$, we just need to show that $\mathbb{E} \sup_{f \in \mathcal{F}} |\mathbb{S}_n f/(n\nu)|$ is bounded. Apply Theorem 6 in Nolan and Pollard (1987) to get

$$\mathbb{E} \sup_{f \in \mathcal{F}} |\mathbb{S}_n f| \leq C\mathbb{E} \left\{ \sigma_n + \tau_n J_n\left(\frac{\sigma_n}{\tau_n}\right) \right\}, \tag{A.3}$$

where C is a universal constant, $\sigma_n = \sup_{f \in \mathcal{F}} (\mathbb{T}_n f^2)^{1/2}/4$, $\tau_n = (\mathbb{T}_n b^2)^{1/2}$, and $J_n(x) = \int_0^x \log N_2(t, \mathbb{T}_n, \mathcal{F}, b)\, \mathrm{d}t$. By Theorem 9.3 in Kosorok (2007), we have $N_2(t, \mathbb{T}_n, \mathcal{F}, 1) \leq K\nu(4e)^{\nu}(2/t)^{2(\nu-1)}$, and thus $J_n(x) \leq cH(x)\nu$ for some constant c depending on K, where $H(x) = x\{1 + \log(1/x)\}$.

Since $\mathcal{F}$ is b-uniformly bounded, it holds that $\sigma_n/\tau_n \in [0, 1/4]$. Note also that $H(x)$ is bounded when $x \in [0, 1]$. We immediately have $H(\sigma_n/\tau_n)$ is bounded. Additionally, by the definition of $\mathbb{T}_n$, we see that $\tau_n = \{4n(n-1)\}^{1/2} \lesssim n$. Combining all these points with (A.3) implies that there exists some constant c' depending on C, c such that

$$\frac{\mathbb{E} \sup_{f \in \mathcal{F}} |\mathbb{S}_n f|}{n\nu} \leq c' \mathbb{E} H\left(\frac{\sigma_n}{\tau_n}\right) \frac{\tau_n}{n} < C'$$

for some large enough absolute constant C'. This completes the proof. □

Lemma A.4. *If for each $\varepsilon > 0$, (i) $\log N_1(\varepsilon, \mathbb{T}_n, \mathcal{F}, F) = O_{\mathbb{P}}(n)$, (ii) $\log N_1(\varepsilon, \mathbb{P}_n \otimes \mathbb{P}, \mathcal{F}, F) = o_{\mathbb{P}}(n)$, (iii) $\log N_1(\varepsilon, \mathbb{P} \otimes \mathbb{P}, \mathcal{F}, F) = o(n)$, then $\sup_{f \in \mathcal{F}} |\mathbb{U}_n f - \mathbb{E} f| \to 0$ almost surely.*

The proof of this lemma follows along the same lines as the proof of Theorem 7 in Nolan and Pollard (1987), though the condition (iii) in this lemma is different from there.

A.3.2. Proof of Theorem 2.1

Proof. (i) It is equivalent to showing that there exists a sequence of nonnegative real numbers δ_n converging to zero such that

$$\mathbb{P}\left\{ \sup_{\boldsymbol{\theta} \in \overline{B}(\boldsymbol{\theta}_0, r_n)} |\mathbb{U}_n h(\cdot, \cdot; \boldsymbol{\theta})| \geq \delta_n \nu/n \right\} = o(1),$$

or

$$\mathbb{P}\left\{ \sup_{\boldsymbol{\theta} \in \overline{B}(\boldsymbol{\theta}_0, r_n)} |\mathbb{S}_n h(\cdot, \cdot; \boldsymbol{\theta})/(n\nu)| \geq \delta_n \right\} = o(1).$$

By Chebyshev's inequality, it suffices to show that

$$\mathbb{E}\left\{ \sup_{\boldsymbol{\theta} \in \overline{B}(\boldsymbol{\theta}_0, r_n)} |\mathbb{S}_n h(\cdot, \cdot; \boldsymbol{\theta})/(n\nu)| \right\} / \delta_n = o(1).$$

We try to bound $\mathbb{E}\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}|\mathbb{S}_n h(\cdot,\cdot;\boldsymbol{\theta})/(n\nu)|\}$. Without loss of generality, assume $\mathcal{F}$ is uniformly bounded by $b = 1/4$. Thus, for any $\boldsymbol{\theta}\in\Theta$, $h(\cdot,\cdot;\boldsymbol{\theta})\le 1$, i.e., the class of functions $\mathscr{H} := \{h^2(\cdot,\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)\}$ is 1-uniformly bounded. Similar to the proof of Lemma A.3, we apply Theorem 6 in Nolan and Pollard (1987) here to get

$$\begin{aligned}\mathbb{E}\Big\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}|\mathbb{S}_n h(\cdot,\cdot;\boldsymbol{\theta})/(n\nu)|\Big\} &\le C_1\mathbb{E}H\Big(\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}\{\mathbb{T}_n h^2(\cdot,\cdot;\boldsymbol{\theta})\}^{1/2}/(2n)\Big)\\ &\le C_1 H\Big(\mathbb{E}\Big[\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}\{\mathbb{T}_n h^2(\cdot,\cdot;\boldsymbol{\theta})\}^{1/2}/(2n)\Big]\Big)\\ &= C_1 H\Big(\mathbb{E}\Big\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}\mathbb{T}_n h^2(\cdot,\cdot;\boldsymbol{\theta})/(2n)^2\Big\}^{1/2}\Big),\end{aligned} \tag{A.4}$$

where C_1 is some constant. The second inequality holds because $H(x)$ is concave in x.

Note that $\mathbb{T}_n h^2(\cdot,\cdot;\boldsymbol{\theta})/(2n)^2 = \mathbb{T}_n h^2(\cdot,\cdot;\boldsymbol{\theta})/\{2n(2n-1)\}\cdot\{2n(2n-1)\}/(2n)^2 \le \mathbb{U}_{2n}h^2(\cdot,\cdot;\boldsymbol{\theta})\le 1$ and that $H(x)$ is increasing in $(0,1]$. Thus, from (A.4), we additionally have

$$\begin{aligned}\mathbb{E}\Big\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}|\mathbb{S}_n h(\cdot,\cdot;\boldsymbol{\theta})/(n\nu)|\Big\} &\le C_1 H\Big(\mathbb{E}\Big\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}\mathbb{U}_{2n}h^2(\cdot,\cdot;\boldsymbol{\theta})\Big\}^{1/2}\Big)\\ &\le C_1 H\Big(\Big[\mathbb{E}\Big\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}\mathbb{U}_{2n}h^2(\cdot,\cdot;\boldsymbol{\theta})\Big\}\Big]^{1/2}\Big),\end{aligned} \tag{A.5}$$

where the last inequality holds because $x^{1/2}$ is concave in x. Now, we need only to consider $\mathbb{E}\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}\mathbb{U}_{2n}h^2(\cdot,\cdot;\boldsymbol{\theta})\}$.

By a decomposition of $\mathbb{U}_{2n}h^2(\cdot,\cdot,\boldsymbol{\theta})$ into a sum of its expected value, plus a smoothly parameterized, zero-mean empirical process, plus a degenerate U-process of order two, we have

$$\begin{aligned}\mathbb{E}\Big\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}\mathbb{U}_{2n}h^2(\cdot,\cdot;\boldsymbol{\theta})\Big\} \le{}& \sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}\mathbb{E}h^2(\cdot,\cdot;\boldsymbol{\theta}) + \mathbb{E}\Big\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}|\mathbb{P}_{2n}h_1(\cdot;\boldsymbol{\theta})|\Big\}\\ &+\mathbb{E}\Big\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}|\mathbb{P}_{2n}h_2(\cdot,\cdot;\boldsymbol{\theta})|\Big\},\end{aligned} \tag{A.6}$$

where $h_1(\boldsymbol{z},\boldsymbol{\theta}) = \mathbb{E}h^2(\boldsymbol{z},\cdot;\boldsymbol{\theta}) + \mathbb{E}h^2(\cdot,\boldsymbol{z};\boldsymbol{\theta}) - 2\mathbb{E}h^2(\cdot,\cdot;\boldsymbol{\theta})$ and $h_2(\boldsymbol{z}_1,\boldsymbol{z}_2;\boldsymbol{\theta}) = h^2(\boldsymbol{z}_1,\boldsymbol{z}_2;\boldsymbol{\theta}) - \mathbb{E}h^2(\boldsymbol{z}_1,\cdot;\boldsymbol{\theta}) - \mathbb{E}h^2(\cdot,\boldsymbol{z}_2;\boldsymbol{\theta}) + \mathbb{E}h^2(\cdot,\cdot;\boldsymbol{\theta})$.

By the condition in (i), it holds that $\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}\mathbb{E}h^2(\cdot,\cdot;\boldsymbol{\theta})\le\epsilon_n$. By Lemmas 16 and 20 in Nolan and Pollard (1987), and Lemma A.1, we have $N_r(\varepsilon,\mathbb{Q},\mathscr{H},1)\le N_r(\varepsilon/16,\mathbb{Q},\mathcal{F},1/4)^4$. Then, following the proof of Lemma A.2, we have $\mathbb{E}\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}|\mathbb{P}_{2n}h_1(\cdot;\boldsymbol{\theta})|\}\le C_2\sqrt{\nu/n}$ for some constant C_2. Additionally, following the proof of Lemma A.3, we have $\mathbb{E}\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}|\mathbb{P}_{2n}h_2(\cdot,\cdot;\boldsymbol{\theta})|\}\le C_3\nu/n$ for some constant C_3.

Take $\delta_n = H^{1/2}((\epsilon_n + C_2\sqrt{\nu/n} + C_3\nu/n)^{1/2})$. If $\epsilon_n\to 0$ and $\nu/n\to 0$, then

$$\mathbb{E}\Big\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}|\mathbb{S}_n h(\cdot,\cdot;\boldsymbol{\theta})/(n\nu)|\Big\}/\delta_n \le C_1 H^{1/2}((\epsilon_n + C_2\sqrt{\nu/n} + C_3\nu/n)^{1/2}) = o(1),$$

because $H(x)\to 0$ as $x\to 0$. This completes proof of (i).

(ii) The proof is based on (A.4)–(A.6) in the proof of (i). First, by the condition in (ii), it holds that $\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}\mathbb{E}h^2(\cdot,\cdot;\boldsymbol{\theta})\le\widetilde{\epsilon}_n$. Then, similar to the proof of (i), $\mathbb{E}\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}|\mathbb{P}_{2n}h_1(\cdot;\boldsymbol{\theta})|\}\le c'\sqrt{\nu/n}$ for some constant c', and $\mathbb{E}\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}|\mathbb{P}_{2n}h_2(\cdot,\cdot;\boldsymbol{\theta})|\}\le C'\nu/n$ for some constant C'. Since $\widetilde{\eta}_n = \sqrt{\nu/n}\vee\widetilde{\epsilon}_n$ and $\nu/n\to 0$, there exists a constant c'' depending on c', C' such that

$$\mathbb{E}\Big\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}\mathbb{U}_{2n}h^2(\cdot,\cdot;\boldsymbol{\theta})\Big\}\le c''\widetilde{\eta}_n$$

holds for sufficiently large n. Combining this with (A.5) implies that

$$\mathbb{E}\Big\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}|\mathbb{S}_n h(\cdot,\cdot;\boldsymbol{\theta})/(n\nu)|\Big\}\le C''\log(1/\widetilde{\eta}_n)\widetilde{\eta}_n^{1/2}$$

for some constant C''. Finally, by the relationship between $\mathbb{U}_n$ and $\mathbb{S}_n$, we conclude that

$$\mathbb{E}\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}|\mathbb{U}_n h(\cdot,\cdot;\boldsymbol{\theta})|\le C''\log(1/\widetilde{\eta}_n)\widetilde{\eta}_n^{1/2}\nu/n$$

holds for sufficiently large n. □

A.3.3. Proof of Theorem 2.2

Proof. The proof is twofold. We first show the uniform convergence of $\Gamma_n(\boldsymbol{\theta})$, and then establish the consistency of $\widehat{\boldsymbol{\theta}}_n$.

Step 1. By Theorem 9.3 in Kosorok (2007), we have $\log N_1(\varepsilon, \mu, \mathcal{F}, F) \lesssim v$ for any $\varepsilon > 0$ and any finite measure μ. If $v/n \to 0$, then all the three conditions in Lemma A.4 hold. Apply this lemma here to get that $\Gamma_n(\boldsymbol{\theta})$ converges almost surely to $\Gamma(\boldsymbol{\theta})$ uniformly in $\boldsymbol{\theta} \in \Theta$.
Step 2. Let $\Theta_0(r) = \mathcal{B}(\boldsymbol{\theta}_0, r)$, where $r \le r_0$. By Assumption 1, we see that $\Theta_1 := \Theta - \Theta_0(r)$ is compact. By Assumption 2, $\Gamma(\boldsymbol{\theta})$ is continuous. Combining these two pieces yields that $\max_{\boldsymbol{\theta}\in\Theta_1} \Gamma(\boldsymbol{\theta})$ exists. Again, by Assumption 1, we know that $\Gamma(\boldsymbol{\theta}_0) - \max_{\boldsymbol{\theta}\in\Theta_1} \Gamma(\boldsymbol{\theta}) \ge \xi_0$.

By *Step 1*, we can find a sufficiently large N such that for all $n > N$,

$$\sup_{\boldsymbol{\theta}\in\Theta} |\Gamma_n(\boldsymbol{\theta}) - \Gamma(\boldsymbol{\theta})| < \xi_0/2$$

holds almost surely. Combining this with the definition of $\widehat{\boldsymbol{\theta}}_n$ yields that

$$\Gamma(\boldsymbol{\theta}_0) < \Gamma_n(\boldsymbol{\theta}_0) + \xi_0/2 \le \Gamma_n(\widehat{\boldsymbol{\theta}}_n) + \xi_0/2 < \Gamma(\widehat{\boldsymbol{\theta}}_n) + \xi_0.$$

This implies that $\widehat{\boldsymbol{\theta}}_n \notin \Theta_1$, i.e., $\widehat{\boldsymbol{\theta}}_n \in \Theta_0(r)$ for all $n > N$. Since this is true for any $r < r_0$, we have

$$\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \to 0 \quad \text{almost surely,}$$

and hence also in probability. This completes the proof. □

A.3.4. Proof of Theorem 2.3

Proof. The proof is conducted in four steps. Based on the Hoeffding decomposition of $\Gamma_n(\boldsymbol{\theta})$, we consider $\Gamma(\boldsymbol{\theta})$, $\mathbb{P}_n g(\cdot; \boldsymbol{\theta})$ and $\mathbb{U}_n h(\cdot, \cdot; \boldsymbol{\theta})$ separately in the first three steps. We finally obtain the convergence rate of $\widehat{\boldsymbol{\theta}}_n$ in the last step.

Step 1. Fixing $\boldsymbol{\theta} \in \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)$, define

$$\omega(\boldsymbol{\theta}) = \mathbb{E}\tau(\cdot; \boldsymbol{\theta}) - \mathbb{E}\tau(\cdot, \boldsymbol{\theta}_0) - (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbf{V}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) = 2\Gamma(\boldsymbol{\theta}) - (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbf{V}(\boldsymbol{\theta} - \boldsymbol{\theta}_0). \tag{A.7}$$

Additionally, expand $\omega(\boldsymbol{\theta})$ about $\boldsymbol{\theta}_0$ to get

$$\omega(\boldsymbol{\theta}) = (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \nabla_1 \omega(\boldsymbol{\theta}'), \tag{A.8}$$

where $\boldsymbol{\theta}'$ is a point on the line connecting $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}$, and $\nabla_1\omega(\boldsymbol{\theta}') = \nabla_1 \mathbb{E}\tau(\cdot; \boldsymbol{\theta}') - 2\mathbf{V}(\boldsymbol{\theta}' - \boldsymbol{\theta}_0)$. Expand $\nabla_1 \mathbb{E}\tau(\cdot; \boldsymbol{\theta}')$ in $\nabla_1\omega(\boldsymbol{\theta}')$ about $\boldsymbol{\theta}_0$ to get

$$\nabla_1\omega(\boldsymbol{\theta}') = 2\mathbf{V}(\boldsymbol{\theta}'')(\boldsymbol{\theta}' - \boldsymbol{\theta}_0) - 2\mathbf{V}(\boldsymbol{\theta}' - \boldsymbol{\theta}_0) = 2\{\mathbf{V}(\boldsymbol{\theta}'') - \mathbf{V}\}(\boldsymbol{\theta}' - \boldsymbol{\theta}_0)$$

for $\boldsymbol{\theta}''$ between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}'$. By Assumption 3(ii) and (iii), we have

$$\begin{aligned} \sup_{\boldsymbol{\theta}'\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)} \|\nabla_1\omega(\boldsymbol{\theta}')\| &\le 2 \sup_{\boldsymbol{\theta}'\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)} \|\mathbf{V}^{1/2}\{\mathbf{I}_p - \mathbf{V}^{-1/2}\mathbf{V}(\boldsymbol{\theta}'')\mathbf{V}^{-1/2}\}\mathbf{V}^{1/2}\| \|\boldsymbol{\theta}' - \boldsymbol{\theta}_0\| \\ &\le 2c_{\max}\rho(r)\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|. \end{aligned} \tag{A.9}$$

Combining this with (A.7) and (A.8) yields

$$\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)} \left|\Gamma(\boldsymbol{\theta}) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbf{V}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right| \le c_{\max}\rho(r)\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2. \tag{A.10}$$

Step 2. Fixing $\boldsymbol{z}$ in $\mathbb{R}^m$ and $\boldsymbol{\theta}$ in $\overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)$, define

$$\psi(\boldsymbol{z}; \boldsymbol{\theta}) = \tau(\boldsymbol{z}; \boldsymbol{\theta}) - \tau(\boldsymbol{z}; \theta_0) - (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \nabla_1\tau(\boldsymbol{z}; \boldsymbol{\theta}_0) - (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbf{V}(\boldsymbol{\theta} - \boldsymbol{\theta}_0).$$

With a little abuse of notation, we still use $\boldsymbol{\theta}'$ to denote some point between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}$ below. Expand $\psi(\boldsymbol{z}; \boldsymbol{\theta})$ about $\boldsymbol{\theta}_0$ to get

$$\psi(\boldsymbol{z}; \boldsymbol{\theta}) = (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \nabla_1\psi(\boldsymbol{z}; \boldsymbol{\theta}') = (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \{\nabla_1\tau(\boldsymbol{z}; \boldsymbol{\theta}') - \nabla_1\tau(\boldsymbol{z}; \boldsymbol{\theta}_0) - 2\mathbf{V}(\boldsymbol{\theta}' - \boldsymbol{\theta}_0)\}.$$

Note that $\tau(\boldsymbol{z}; \boldsymbol{\theta}') = \zeta(\boldsymbol{z}; \boldsymbol{\theta}') + \mathbb{E}\tau(\cdot; \boldsymbol{\theta}')$. It then follows from the above equation and $\nabla_1 \mathbb{E}\tau(\cdot; \boldsymbol{\theta}_0) = 0$ that

$$\begin{aligned} \mathbb{P}_n\psi(\cdot; \boldsymbol{\theta}) &= (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \{\mathbb{P}_n\nabla_1\zeta(\cdot; \boldsymbol{\theta}') - \mathbb{P}_n\nabla_1\zeta(\cdot; \boldsymbol{\theta}_0) + \nabla_1\mathbb{E}\tau(\cdot; \boldsymbol{\theta}') - 2\mathbf{V}(\boldsymbol{\theta}' - \boldsymbol{\theta}_0)\} \\ &= (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbb{P}_n\{\nabla_1\zeta(\cdot; \boldsymbol{\theta}') - \nabla_1\zeta(\cdot; \boldsymbol{\theta}_0)\} + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \{\nabla_1\mathbb{E}\tau(\cdot; \boldsymbol{\theta}') - 2\mathbf{V}(\boldsymbol{\theta}' - \boldsymbol{\theta}_0)\}. \end{aligned}$$

By *Step 1*, we have that

$$\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)} \|(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \{\nabla_1\mathbb{E}\tau(\cdot; \boldsymbol{\theta}') - 2\mathbf{V}(\boldsymbol{\theta}' - \boldsymbol{\theta}_0)\}\| \le 2c_{\max}\rho(r)\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2. \tag{A.11}$$

Next, we try to bound $\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)} \|(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbb{P}_n\{\nabla_1\zeta(\cdot; \boldsymbol{\theta}') - \nabla_1\zeta(\cdot; \boldsymbol{\theta}_0)\}\|$. Consider the vector process

$$\Lambda(\boldsymbol{\theta}) = \sqrt{n}\mathbb{P}_n\{\nabla_1\zeta(\cdot; \boldsymbol{\theta}) - \nabla_1\zeta(\cdot; \boldsymbol{\theta}_0)\}.$$

According to Assumption 3(v), it holds, for any $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in \mathbb{S}^{p-1}$, that

$$\log \mathbb{E} \exp\left\{\lambda \boldsymbol{\gamma}_1^\top \nabla_1 \Lambda(\boldsymbol{\theta}) \boldsymbol{\gamma}_2\right\} = n \log \mathbb{E} \exp\left\{\frac{\lambda}{\sqrt{n}} \boldsymbol{\gamma}_1^\top \nabla_2 \zeta(\cdot; \boldsymbol{\theta}) \boldsymbol{\gamma}_2\right\} \le \frac{\nu_0^2 \lambda^2}{2}$$

for any $|\lambda| \le g_n$ with $g_n = \sqrt{n}\ell_0$. It then follows from Theorem A.3 in Spokoiny (2013) that for any $0 < \varepsilon < 1$,

$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta} \in \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)} \|\Lambda(\boldsymbol{\theta})\| > 6\nu_0 r d_p(\varepsilon)\right\} \le \varepsilon,$$

where

$$d_p(\varepsilon) = \begin{cases} \sqrt{4p - 2\log\varepsilon} & \text{if } 4p - 2\log\varepsilon \le g_n^2, \\ g_n^{-1} \log\varepsilon + \frac{1}{2}(4p g_n^{-1} + g_n) & \text{if } 4p - 2\log\varepsilon > g_n^2. \end{cases}$$

Thus,

$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta} \in \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)} \left|(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbb{P}_n\{\nabla_1 \zeta(\cdot; \boldsymbol{\theta}') - \nabla_1 \zeta(\cdot; \boldsymbol{\theta}_0)\}\right| > \frac{6\nu_0 r}{\sqrt{n}} d_p(\varepsilon) \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|\right\} \le \varepsilon. \tag{A.12}$$

This, combined with (A.11), implies that

$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta} \in \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)} \|\mathbb{P}_n \psi(\cdot; \boldsymbol{\theta})\| > 2c_{\max}\rho(r)\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 + \frac{6\nu_0 r}{\sqrt{n}} d_p(\varepsilon)\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|\right\} \le \varepsilon. \tag{A.13}$$

Note that $\tau(\boldsymbol{z}; \boldsymbol{\theta}_0) = 0$ and

$$\begin{aligned} g(\boldsymbol{z}; \boldsymbol{\theta}) &= \tau(\boldsymbol{z}; \theta) - \tau(\boldsymbol{z}; \boldsymbol{\theta}_0) - 2\Gamma(\boldsymbol{\theta}) \\ &= (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \nabla_1 \tau(\boldsymbol{z}; \boldsymbol{\theta}_0) + \psi(\boldsymbol{z}; \boldsymbol{\theta}) - \{2\Gamma(\boldsymbol{\theta}) - (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbf{V}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\}. \end{aligned}$$

Apply (A.10) and (A.13) to see that

$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta} \in \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)} \left|\mathbb{P}_n g(\cdot; \boldsymbol{\theta}) - \frac{1}{\sqrt{n}}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \boldsymbol{W}_n\right| > 4c_{\max}\rho(r)\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 + \frac{6\nu_0 r}{\sqrt{n}} d_p(\varepsilon)\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|\right\} \le \varepsilon, \tag{A.14}$$

where $\boldsymbol{W}_n = \sqrt{n}\mathbb{P}_n \nabla_1 \tau(\cdot; \boldsymbol{\theta}_0)$.

Step 3. By Assumption 2, $f(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta})$ is continuous at $\boldsymbol{\theta}_0$ almost surely. Since $\mathcal{F}$ is uniformly bounded, a dominated convergence argument implies that the same holds true for $h(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta})$. In view of $f(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{\theta}_0) = 0$ for all $\boldsymbol{z}_1, \boldsymbol{z}_2$, it holds that $h(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta}_0) = 0$. Thus, the boundedness of h and the dominated convergence theorem establish

$$\mathbb{E}h^2(\cdot, \cdot; \boldsymbol{\theta}) \to 0 \quad \text{as} \quad \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \to 0. \tag{A.15}$$

Equivalently, there exists a constant $\alpha(r) > 0$ such that $\sup_{\boldsymbol{\theta} \in \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)} \mathbb{E}h^2(\cdot, \cdot; \boldsymbol{\theta}) \le \alpha(r)$ and $\alpha(r) \to 0$ as $r \to 0$. By Theorem 2.1, there exists a sequence of nonnegative real numbers δ_n (depending on $\alpha(r), \nu, n$) converging to zero as $r \to 0$ and $n \to \infty$, such that

$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta} \in \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)} |\mathbb{U}_n h(\cdot, \cdot; \boldsymbol{\theta})| > \delta_n \nu / n\right\} \le \epsilon \tag{A.16}$$

holds for sufficiently large n.

Step 4. The Hoeffding decomposition, combined with (A.10), (A.14), and (A.16) in the above three steps, implies that

$$\begin{aligned} \mathbb{P}\Bigg\{\sup_{\boldsymbol{\theta} \in \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)} &\left|\Gamma_n(\boldsymbol{\theta}) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbf{V}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \frac{1}{\sqrt{n}}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \boldsymbol{W}_n\right| > \\ &5c_{\max}\rho(r)\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 + \frac{6\nu_0 r}{\sqrt{n}} d_p(\varepsilon)\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + \delta_n \frac{\nu}{n}\Bigg\} \le 2\varepsilon. \end{aligned} \tag{A.17}$$

In view of $\boldsymbol{W}_n = \sqrt{n}\mathbb{P}_n \nabla_1 \tau(\cdot; \boldsymbol{\theta}_0)$, it holds that $(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \boldsymbol{W}_n / \{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \boldsymbol{\Delta}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\}^{1/2} \Rightarrow N(0, 1)$. This, combined with Assumption 3(iv), implies that there exists a constant b_ε depending on $d_{\max}$ such that

$$\mathbb{P}\left\{|(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top W_n| > b_\varepsilon \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|\right\} \le \varepsilon$$

holds for sufficiently large n. Define the set

$$\mathcal{A}_{n,\varepsilon} = \Big\{ \boldsymbol{Z} : \sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r)} \big|\Gamma_n(\boldsymbol{\theta}) - \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top \mathbf{V}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\big| \le \frac{b_\varepsilon}{\sqrt{n}}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| + 5c_{\max}\rho(r)\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 + \frac{6\nu_0 r}{\sqrt{n}} d_p(\varepsilon)\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| + \delta_n\frac{\nu}{n} \Big\},$$

then $\mathbb{P}(\mathcal{A}_{n,\varepsilon}) \ge 1-3\varepsilon$ holds for sufficiently large n. The following analysis is on the set $\mathcal{A}_{n,\varepsilon}$.

By Theorem 2.2, $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \to 0$ almost surely. Thus, for sufficiently large n, $\widehat{\boldsymbol{\theta}}_n \in \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)$. This implies that

$$\Gamma_n(\widehat{\boldsymbol{\theta}}_n) \le \frac{1}{2}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)^\top \mathbf{V}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0) + \frac{b_\varepsilon}{\sqrt{n}}\|\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\| + 5c_{\max}\rho(r)\|\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|^2 + \frac{6\nu_0 r}{\sqrt{n}} d_p(\varepsilon)\|\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\| + \delta_n\frac{\nu}{n}.$$

In view of $\Gamma_n(\widehat{\boldsymbol{\theta}}_n) \ge \Gamma_n(\boldsymbol{\theta}_0) = 0$, it holds that

$$0 \le \frac{1}{2}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)^\top \mathbf{V}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0) + \frac{b_\varepsilon}{\sqrt{n}}\|\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\| + 5c_{\max}\rho(r)\|\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|^2 + \frac{6\nu_0 r}{\sqrt{n}} d_p(\varepsilon)\|\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\| + \delta_n\frac{\nu}{n}.$$

This, combined with Assumption 3(ii) and $\rho(r) < \frac{c_{\min}}{11c_{\max}}$, implies that

$$\frac{1}{2}\kappa\|\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|^2 \le \frac{b_\varepsilon}{\sqrt{n}}\|\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\| + \frac{6\nu_0 r}{\sqrt{n}} d_p(\varepsilon)\|\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\| + \delta_n\frac{\nu}{n}, \tag{A.18}$$

where $\kappa = c_{\min} - 10c_{\max}\rho(r) > 0$. By the definition of $d_p(\varepsilon)$ and $p/n \to 0$, there exists a constant c_ε such that $d_p(\varepsilon) \le c_\varepsilon\sqrt{p}$ for sufficiently large n. Combining this with (A.18) yields that

$$\frac{1}{2}\kappa\Big(\|\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\| - \frac{b_\varepsilon + 6\nu_0 r c_\varepsilon\sqrt{p}}{\kappa\sqrt{n}}\Big)^2 \le \frac{(b_\varepsilon + 6\nu_0 r c_\varepsilon\sqrt{p})^2}{2\kappa n} + \delta_n\frac{\nu}{n}.$$

Solving the above equation establishes that

$$\|\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\| \le C_\varepsilon\sqrt{\frac{\nu\vee p}{n}}$$

holds for sufficiently large n, where C_ε is some constant depending only on $c_{\min}, c_{\max}, \rho(r), d_{\max}, \varepsilon$, but not depending on ν, p, n. Thus,

$$\mathbb{P}\Big\{\|\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\| \le C_\varepsilon\sqrt{\frac{\nu\vee p}{n}}\Big\} \ge 1-3\varepsilon$$

holds for sufficiently large n. This completes the proof. □

A.3.5. Proof of Theorem 2.4

Proof. The proof is based on the proof of Theorem 2.3. We first define $\widehat{\boldsymbol{t}}_n = \sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ and $\boldsymbol{t}_n^* = -\mathbf{V}^{-1}\boldsymbol{W}_n$. By Theorem 2.3, for any $\varepsilon > 0$, there exists a constant $C'_\varepsilon > 0$ such that

$$\mathbb{P}\Big\{\|\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\| > C'_\varepsilon\sqrt{\frac{\nu\vee p}{n}}\Big\} \le \varepsilon \tag{A.19}$$

holds for sufficiently large n. By the definition of $\boldsymbol{W}_n$, $\boldsymbol{W}_n = \sqrt{n}\mathbb{P}_n\nabla_1\tau(\cdot;\boldsymbol{\theta}_0)$, it holds that for any $\boldsymbol{\gamma} \in \mathbb{R}^p$, $\boldsymbol{\gamma}^\top\boldsymbol{t}_n^*/(\boldsymbol{\gamma}^\top\mathbf{V}^{-1}\boldsymbol{\Delta}\mathbf{V}^{-1}\boldsymbol{\gamma})^{1/2} \Rightarrow N(0,1)$. This, combined with Assumption 3(ii) and (iv), implies that there exists a constant C''_ε such that

$$\mathbb{P}\big\{\|\boldsymbol{t}_n^*\| \ge C''_\varepsilon\big\} \le \varepsilon \tag{A.20}$$

holds for sufficiently large n. Thus, by (A.19) and (A.20), there exists a constant c'_ε depending on $C'_\varepsilon, C''_\varepsilon$ such that

$$\mathbb{P}\big(\mathcal{A}'_{n,\varepsilon}\big) \ge 1-2\varepsilon \tag{A.21}$$

holds for sufficiently large n, where $\mathcal{A}'_{n,\varepsilon} := \{\boldsymbol{Z} : \widehat{\boldsymbol{\theta}}_n \in \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r_n), \boldsymbol{t}_n^*/\sqrt{n} + \boldsymbol{\theta}_0 \in \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r_n)\}$ and $r_n := c'_\varepsilon\sqrt{(\nu\vee p)/n}$.

Fix $\boldsymbol{\theta} \in \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r_n)$. Then following the proofs of *Steps 1–2* in Theorem 2.3, we have

$$\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)} \big|\Gamma(\boldsymbol{\theta}) - \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top\mathbf{V}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\big| \le c_{\max}\rho(r_n)\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2, \tag{A.22}$$

and

$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}\left|\mathbb{P}_n g(\cdot;\boldsymbol{\theta})-\frac{1}{\sqrt{n}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top \boldsymbol{W}_n\right|>4c_{\max}\rho(r_n)\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2+\frac{6\nu_0 r_n}{\sqrt{n}}d_p(\varepsilon)\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\right\}\le\varepsilon. \tag{A.23}$$

Since $\mathcal{F}$ is uniformly bounded and $\sup_{\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}\mathbb{E}h^2(\cdot,\cdot;\boldsymbol{\theta})\le\widetilde{C}\epsilon_n$, Theorem 2.1(ii) implies that there exists a constant C_ε such that

$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}\left|\mathbb{U}_n h(\cdot,\cdot;\boldsymbol{\theta})\right|>C_\varepsilon\log(1/\eta_n)\eta_n^{1/2}\frac{\nu}{n}\right\}\le\varepsilon \tag{A.24}$$

holds for sufficiently large n. This, together with (A.22) and (A.23), implies that

$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}\left|\Gamma_n(\boldsymbol{\theta})-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top\mathbf{V}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)-\frac{1}{\sqrt{n}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top\boldsymbol{W}_n\right|>5c_{\max}\rho(r_n)\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2+\frac{6\nu_0 r_n}{\sqrt{n}}d_p(\varepsilon)\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|+C_\varepsilon\log(1/\eta_n)\eta_n^{1/2}\frac{\nu}{n}\right\}\le 2\varepsilon. \tag{A.25}$$

In view of $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le r_n$, $\rho(r_n)\le pr_n$ and $d_p(\varepsilon)\le c_\varepsilon\sqrt{p}$, it holds that

$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}\left|\Gamma_n(\boldsymbol{\theta})-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top\mathbf{V}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)-\frac{1}{\sqrt{n}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top\boldsymbol{W}_n\right|>5c_{\max}pr_n^3+\frac{6\nu_0 c_\varepsilon r_n^2\sqrt{p}}{\sqrt{n}}+C_\varepsilon\log(1/\eta_n)\eta_n^{1/2}\frac{\nu}{n}\right\}\le 2\varepsilon \tag{A.26}$$

for sufficiently large n. Define the set

$$\mathcal{A}''_{n,\varepsilon}=\left\{\boldsymbol{Z}:\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)}\left|\Gamma_n(\boldsymbol{\theta})-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top\mathbf{V}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)-\frac{1}{\sqrt{n}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top\boldsymbol{W}_n\right|\le\phi_\varepsilon(\nu,p,n)\right\}, \tag{A.27}$$

where

$$\phi_\varepsilon(\nu,p,n):=5c_{\max}pr_n^3+\frac{6\nu_0 c_\varepsilon r_n^2\sqrt{p}}{\sqrt{n}}+C_\varepsilon\log(1/\eta_n)\eta_n^{1/2}\frac{\nu}{n}.$$

Then, $\mathbb{P}(\mathcal{A}''_{n,\varepsilon})\ge 1-2\varepsilon$. Additionally, $\mathbb{P}(\mathcal{A}'_{n,\varepsilon}\cap\mathcal{A}''_{n,\varepsilon})\ge 1-4\varepsilon$. The following analysis is on the set $\mathcal{A}'_{n,\varepsilon}\cap\mathcal{A}''_{n,\varepsilon}$.

By definition, $\Gamma_n(\widehat{\boldsymbol{\theta}}_n)=\Gamma_n(\widehat{\boldsymbol{t}}_n/\sqrt{n}+\boldsymbol{\theta}_0)\ge\Gamma_n(\boldsymbol{t}_n^*/\sqrt{n}+\boldsymbol{\theta}_0)$. Apply the inequality in (A.27) twice, then multiply through by n, consolidate terms, and use the fact that $\mathbf{V}$ is negative definite to get that

$$0\le-\frac{1}{2}(\widehat{\boldsymbol{t}}_n-\boldsymbol{t}_n^*)^\top\mathbf{V}(\widehat{\boldsymbol{t}}_n-\boldsymbol{t}_n^*)\le 2n\phi_\varepsilon(\nu,p,n). \tag{A.28}$$

Note that $\phi_\varepsilon(\nu,p,n)\lesssim(\nu\vee p)^{5/2}/n^{3/2}+\log(1/\eta_n)\eta_n^{1/2}\nu/n$. This, combined with (A.28) and Assumption 3(ii), implies that

$$\|\widehat{\boldsymbol{t}}_n-\boldsymbol{t}_n^*\|\lesssim\left\{\frac{(\nu\vee p)^{5/2}}{n^{1/2}}\vee\log(1/\eta_n)\eta_n^{1/2}\nu\right\}^{1/2}.$$

Recall the definition of $\widehat{\boldsymbol{t}}_n$ and $\boldsymbol{t}_n^*$, we immediately have

$$\left\|\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0+\mathbf{V}^{-1}\mathbb{P}_n\nabla_1\tau(\cdot;\boldsymbol{\theta}_0)\right\|^2=O_{\mathbb{P}}\left\{\frac{(\nu\vee p)^{5/2}}{n^{3/2}}+\frac{\log(1/\eta_n)\eta_n^{1/2}\nu}{n}\right\}.$$

Furthermore, if $\{(\nu\vee p)^{5/2}/n^{1/2}\}\vee\{\log(1/\eta_n)\eta_n^{1/2}\nu\}\to 0$, then by Assumption 3(iv) and Slutsky's Theorem, it holds that for any $\boldsymbol{\gamma}\in\mathbb{R}^p$, $\boldsymbol{\gamma}^\top\widehat{\boldsymbol{t}}_n/\{\boldsymbol{\gamma}^\top\mathbf{V}^{-1}\boldsymbol{\Delta}\mathbf{V}^{-1}\boldsymbol{\gamma}\}^{1/2}\Rightarrow N(0,1)$. This completes the proof. □

A.3.6. Proof of Theorem 2.6

Proof. Note that the function class $\mathcal{F}$ is uniformly bounded by an absolute constant. We immediately have that $\widetilde{\mathcal{F}}$ is also uniformly bounded by an absolute constant. In addition, $\mathbb{E}\tau_n(\boldsymbol{z};\boldsymbol{\theta})=\tau(\boldsymbol{z};\boldsymbol{\theta})$. It then follows from Lemma A.2 that

$$\sup_{\mathbb{R}^m\otimes\Theta}|\tau_n(\boldsymbol{z};\boldsymbol{\theta})-\tau(\boldsymbol{z};\boldsymbol{\theta})|=O_{\mathbb{P}}\left(\sqrt{\widetilde{\nu}/n}\right). \tag{A.29}$$

Since $\varepsilon_n^{-1}\sqrt{\widetilde{\nu}/n}\to 0$, we just need to consider

$$\widetilde{\delta}_{ij}:=\mathbb{P}_n\{\widetilde{p}_{ni}(\cdot;\widehat{\boldsymbol{\theta}}_n)\widetilde{p}_{nj}(\cdot;\widehat{\boldsymbol{\theta}}_n)\},$$

where

$$\widetilde{p}_{ni}(\boldsymbol{z};\boldsymbol{\theta}) := \varepsilon_n^{-1}\{\tau(\boldsymbol{z};\boldsymbol{\theta}+\varepsilon_n\boldsymbol{u}_i)-\tau(\boldsymbol{z};\boldsymbol{\theta})\}.$$

Expand $\widetilde{p}_{ni}(\boldsymbol{z};\widehat{\boldsymbol{\theta}}_n)$ about $\boldsymbol{\theta}_0$ to get

$$\widetilde{p}_{ni}(\boldsymbol{z};\widehat{\boldsymbol{\theta}}_n) = \widetilde{p}_{ni}(\boldsymbol{z};\boldsymbol{\theta}_0) + \varepsilon_n^{-1}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)^\top \underbrace{\{\nabla_1\tau(\boldsymbol{z};\boldsymbol{\theta}^*+\varepsilon_n\boldsymbol{u}_i)-\nabla_1\tau(\boldsymbol{z};\boldsymbol{\theta}^*)\}}_{R_n}, \tag{A.30}$$

where $\boldsymbol{\theta}^*$ denotes some point between $\widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_0$. Note that $\tau(\boldsymbol{z};\boldsymbol{\theta}) = \zeta(\boldsymbol{z};\boldsymbol{\theta}) + \mathbb{E}\tau(\cdot;\boldsymbol{\theta})$. We can rewrite R_n in the above equation as follows:

$$R_n = \underbrace{\{\nabla_1\zeta(\boldsymbol{z};\boldsymbol{\theta}^*+\varepsilon_n\boldsymbol{u}_i)-\nabla_1\zeta(\boldsymbol{z};\boldsymbol{\theta}^*)\}}_{R_{n1}} + \underbrace{\{\nabla_1\mathbb{E}\tau(\cdot;\boldsymbol{\theta}^*+\varepsilon_n\boldsymbol{u}_i)-\nabla_1\mathbb{E}\tau(\cdot;\boldsymbol{\theta}^*)\}}_{R_{n2}}.$$

We discuss R_{n1} and R_{n2} separately. First, following the calculations in *Step 2* of the proof of Theorem 2.3, we have

$$\sup_{\boldsymbol{\theta}^*\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)} \|\mathbb{P}_n\{\nabla_1\zeta(\cdot;\boldsymbol{\theta}^*+\varepsilon_n\boldsymbol{u}_i)-\nabla_1\zeta(\cdot;\boldsymbol{\theta}^*)\}\| = O_{\mathbb{P}}\big(r_n\sqrt{p/n}\big),$$

where $r_n := \sqrt{(\nu\vee p)/n}$. In view of $R_{n1} = \mathbb{P}_n\{\nabla_1\zeta(\cdot;\boldsymbol{\theta}^*+\varepsilon_n\boldsymbol{u}_i)-\nabla_1\zeta(\cdot;\boldsymbol{\theta}^*)\}$, it then holds that

$$\sup_{\boldsymbol{\theta}^*\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)} \|R_{n1}\| = O_{\mathbb{P}}\big(r_n\sqrt{p/n}\big). \tag{A.31}$$

We now turn to consider R_{n2}. Following similar arguments as in *Step 1* of the proof of Theorem 2.3, we have

$$\sup_{\boldsymbol{\theta}^*\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)} \|R_{n2}\| = O\big(r_n\rho(r_n)+\varepsilon_n\big) = O\big(pr_n^3+\varepsilon_n\big).$$

Combining this with (A.31) and (A.30) implies that

$$\widetilde{p}_{ni}(\boldsymbol{z};\widehat{\boldsymbol{\theta}}_n) = \widetilde{p}_{ni}(\boldsymbol{z};\boldsymbol{\theta}_0) + O_{\mathbb{P}}\big(\varepsilon_n^{-1}r_n\big)O_{\mathbb{P}}\big(r_n\sqrt{p/n}+pr_n^3+\varepsilon_n\big). \tag{A.32}$$

Next, we consider $\widetilde{p}_{ni}(\cdot;\boldsymbol{\theta}_0) = \varepsilon_n^{-1}\{\tau(\boldsymbol{z};\boldsymbol{\theta}_0+\varepsilon_n\boldsymbol{u}_i)-\tau(\boldsymbol{z};\boldsymbol{\theta}_0)\}$. Expand $\tau(\boldsymbol{z};\boldsymbol{\theta}_0+\varepsilon_n\boldsymbol{u}_i)-\tau(\boldsymbol{z};\boldsymbol{\theta}_0)$ about $\varepsilon_n = 0$ to get

$$\tau(\boldsymbol{z};\boldsymbol{\theta}_0+\varepsilon_n\boldsymbol{u}_i)-\tau(\boldsymbol{z};\boldsymbol{\theta}_0) = \varepsilon_n\boldsymbol{u}_i^\top\nabla_1\tau(\boldsymbol{z};\boldsymbol{\theta}_0) + \varepsilon_n^2\boldsymbol{u}_i^\top\nabla_2\tau(\boldsymbol{z};\boldsymbol{\theta}_0+\alpha\varepsilon_n\boldsymbol{u}_i)\boldsymbol{u}_i, \tag{A.33}$$

where $\alpha\in(0,1)$. Again using the equality $\tau(\boldsymbol{z};\boldsymbol{\theta}) = \zeta(\boldsymbol{z};\boldsymbol{\theta}) + \mathbb{E}\tau(\cdot;\boldsymbol{\theta})$, we have

$$\boldsymbol{u}_i^\top\nabla_2\tau(\boldsymbol{z};\boldsymbol{\theta}_0+\alpha\varepsilon_n\boldsymbol{u}_i)\boldsymbol{u}_i = \underbrace{\boldsymbol{u}_i^\top\mathbf{V}(\boldsymbol{\theta}_0+\alpha\varepsilon_n\boldsymbol{u}_i)\boldsymbol{u}_i}_{T_{n1}} + \underbrace{\boldsymbol{u}_i^\top\nabla_2\zeta(\boldsymbol{z};\boldsymbol{\theta}_0+\alpha\varepsilon_n\boldsymbol{u}_i)\boldsymbol{u}_i}_{T_{n2}}. \tag{A.34}$$

By Assumption 3(ii) and (iii), we have

$$\sup_{\alpha\in(0,1)} |T_{n1}| = \boldsymbol{u}_i^\top\{\mathbf{V}(\boldsymbol{\theta}_0+\alpha\varepsilon_n\boldsymbol{u}_i)-\mathbf{V}\}\boldsymbol{u}_i + \boldsymbol{u}_i^\top\mathbf{V}\boldsymbol{u}_i = O(1). \tag{A.35}$$

By Assumption 3(v), we know that T_{n2} is zero-mean subexponential. Thus, by the equivalent definitions of zero-mean subexponential variables, it holds that

$$\sup_{\alpha\in(0,1)} \mathbb{E}|T_{n2}| \le \sup_{\alpha\in(0,1)} (\mathbb{E}T_{n2}^2)^{1/2} \tag{A.36}$$

is bounded. That is, $\sup_{\alpha\in(0,1)}|T_{n2}| = O_{\mathbb{P}}(1)$. Put (A.33)–(A.36) together. We then have

$$\widetilde{p}_{ni}(\boldsymbol{z};\boldsymbol{\theta}_0) = \boldsymbol{u}_i^\top\nabla_1\tau(\boldsymbol{z};\boldsymbol{\theta}_0) + O_{\mathbb{P}}(\varepsilon_n).$$

This, combined with (A.32), implies that

$$\widetilde{p}_{ni}(\boldsymbol{z};\widehat{\boldsymbol{\theta}}_n) = \boldsymbol{u}_i^\top\nabla_1\tau(\boldsymbol{z};\boldsymbol{\theta}_0) + O_{\mathbb{P}}\big\{\varepsilon_n^{-1}\sqrt{\widetilde{v}/n}+\varepsilon_n^{-1}r_n(r_n\sqrt{p/n}+pr_n^3+\varepsilon_n)+\varepsilon_n\big\}.$$

Additionally, combining this with (A.29) implies that

$$\begin{aligned}\widehat{\delta}_{ij} &= \mathbb{P}_n\{\boldsymbol{u}_i^\top\nabla_1\tau(\cdot;\boldsymbol{\theta}_0)\boldsymbol{u}_j^\top\nabla_1\tau(\cdot;\boldsymbol{\theta}_0)\} + O_{\mathbb{P}}\big[\{\varepsilon_n^{-1}\sqrt{\widetilde{v}/n}+\varepsilon_n^{-1}r_n(r_n\sqrt{p/n}+pr_n^3+\varepsilon_n)+\varepsilon_n\}^2\big]\\ &= \delta_{ij} + O_{\mathbb{P}}\big(1/\sqrt{n}\big) + O_{\mathbb{P}}\big[\{\varepsilon_n+\varepsilon_n^{-1}(r_n^2\sqrt{p/n}+pr_n^4+\sqrt{\widetilde{v}/n})+r_n\}^2\big].\end{aligned}$$

Thus,

$$\|\widehat{\boldsymbol{\Delta}}-\boldsymbol{\Delta}\| = O_{\mathbb{P}}\big(p/\sqrt{n}\big) + O_{\mathbb{P}}\big[p\{\varepsilon_n+\varepsilon_n^{-1}(r_n^2\sqrt{p/n}+pr_n^4+\sqrt{\widetilde{v}/n})+r_n\}^2\big].$$

Similarly,

$$\|\widehat{\mathbf{V}}-\mathbf{V}\| = O_{\mathbb{P}}\big(p/\sqrt{n}\big) + O_{\mathbb{P}}\big[p\{\varepsilon_n+\varepsilon_n^{-2}(r_n^2\sqrt{p/n}+pr_n^4+\sqrt{\widetilde{v}/n})+\varepsilon_n^{-1}r_n\}^2\big].$$

By assumption, $(\widetilde{\nu} \vee \nu \vee p)^{5/2}/n^{1/2} = o(1)$, $\varepsilon_n\sqrt{p} = o(1)$, and $\varepsilon_n^{-2}(\widetilde{\nu} \vee \nu \vee p)/\sqrt{n} = o(1)$. It can then be easy to verify that

$$\|\widehat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\| = o_{\mathbb{P}}(1) \quad \text{and} \quad \|\widehat{\mathbf{V}} - \mathbf{V}\| = o_{\mathbb{P}}(1).$$

This, combined with Assumption 3(ii) and (iv), implies that $\|\widehat{\boldsymbol{\Delta}}\| = O_{\mathbb{P}}(1)$, $\|\widehat{\mathbf{V}}\| = O_{\mathbb{P}}(1)$, and $\|\widehat{\mathbf{V}}^{-1}\| = O_{\mathbb{P}}(1)$. Note that $\widehat{\mathbf{V}}^{-1} - \mathbf{V}^{-1} = \widehat{\mathbf{V}}^{-1}(\mathbf{V} - \widehat{\mathbf{V}})\mathbf{V}^{-1}$. Then, we have

$$\|\widehat{\mathbf{V}}^{-1} - \mathbf{V}^{-1}\| \le \|\widehat{\mathbf{V}}^{-1}\| \cdot \|\mathbf{V} - \widehat{\mathbf{V}}\| \cdot \|\mathbf{V}^{-1}\| = o_{\mathbb{P}}(1).$$

Note also that

$$\begin{aligned}
&\widehat{\mathbf{V}}^{-1}\widehat{\boldsymbol{\Delta}}\widehat{\mathbf{V}}^{-1} - \mathbf{V}^{-1}\boldsymbol{\Delta}\mathbf{V}^{-1} \\
=&(\widehat{\mathbf{V}}^{-1} - \mathbf{V}^{-1})(\widehat{\boldsymbol{\Delta}} - \boldsymbol{\Delta})(\widehat{\mathbf{V}}^{-1} - \mathbf{V}^{-1}) - \mathbf{V}^{-1}(\boldsymbol{\Delta} - \widehat{\boldsymbol{\Delta}})\widehat{\mathbf{V}}^{-1} - \widehat{\mathbf{V}}^{-1}\boldsymbol{\Delta}(\mathbf{V}^{-1} - \widehat{\mathbf{V}}^{-1}) \\
&- (\mathbf{V}^{-1} - \widehat{\mathbf{V}}^{-1})\widehat{\boldsymbol{\Delta}}\mathbf{V}^{-1}.
\end{aligned}$$

Apply the triangle inequality to the above equation to get that

$$\|\widehat{\mathbf{V}}^{-1}\widehat{\boldsymbol{\Delta}}\widehat{\mathbf{V}}^{-1} - \mathbf{V}^{-1}\boldsymbol{\Delta}\mathbf{V}^{-1}\| = o_{\mathbb{P}}(1).$$

This completes the proof. □

A.4. Proofs in Section 3

For the example in Section 1.2, we define $\mathcal{F}^{\mathsf{H}} = \{f^{\mathsf{H}}(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta^{\mathsf{H}}\}$, where $f^{\mathsf{H}}(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta})$ is defined in the main text, and define $h^{\mathsf{H}}(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta}) = f^{\mathsf{H}}(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta}) - \mathbb{E}f^{\mathsf{H}}(\boldsymbol{z}_1, \cdot; \boldsymbol{\theta}) - \mathbb{E}f^{\mathsf{H}}(\cdot, \boldsymbol{z}_2; \boldsymbol{\theta}) + \Gamma^{\mathsf{H}}(\boldsymbol{\theta})$. For the example in Section 3.2, we define $\mathcal{F}^{\mathsf{C}} = \{f^{\mathsf{C}}(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta^{\mathsf{C}}\}$ and $h^{\mathsf{C}}(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta}) = f^{\mathsf{C}}(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta}) - \mathbb{E}f^{\mathsf{C}}(\boldsymbol{z}_1, \cdot; \boldsymbol{\theta}) - \mathbb{E}f^{\mathsf{C}}(\cdot, \boldsymbol{z}_2; \boldsymbol{\theta}) + \Gamma^{\mathsf{C}}(\boldsymbol{\theta})$. For the example in Section 3.3, we define $\mathcal{F}^{\mathsf{K}} = \{f^{\mathsf{K}}(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta^{\mathsf{K}}\}$ and $h^{\mathsf{K}}(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta}) = f^{\mathsf{K}}(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta}) - \mathbb{E}f^{\mathsf{K}}(\boldsymbol{z}_1, \cdot; \boldsymbol{\theta}) - \mathbb{E}f^{\mathsf{K}}(\cdot, \boldsymbol{z}_2; \boldsymbol{\theta}) + \Gamma^{\mathsf{K}}(\boldsymbol{\theta})$. For the example in Section 3.4, we define $\mathcal{F}^{\mathsf{A}} = \left\{f^{\mathsf{A}}(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta^{\mathsf{A}}\right\}$ and $h^{\mathsf{A}}(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta}) = f^{\mathsf{A}}(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta}) - \mathbb{E}f^{\mathsf{A}}(\boldsymbol{z}_1, \cdot; \boldsymbol{\theta}) - \mathbb{E}f^{\mathsf{A}}(\cdot, \boldsymbol{z}_2; \boldsymbol{\theta}) + \mathbb{E}f^{\mathsf{A}}(\cdot, \cdot; \boldsymbol{\theta})$.

A.4.1. Some additional lemmas

Lemma A.5. *Suppose that Condition 1 in the main text holds. Then $\widetilde{\boldsymbol{X}} - \mathbb{E}(\widetilde{\boldsymbol{X}} \mid \boldsymbol{X}^\top\boldsymbol{\beta}_0)$ is multivariate subgaussian.*

Proof. Fix $\boldsymbol{u} \in \mathbb{S}^{p-1}$. Applying the triangle inequality yields that

$$\|\boldsymbol{u}^\top\{\widetilde{\boldsymbol{X}} - \mathbb{E}(\widetilde{\boldsymbol{X}} \mid \boldsymbol{X}^\top\boldsymbol{\beta}_0)\}\|_r \le \underbrace{\|\boldsymbol{u}^\top\widetilde{\boldsymbol{X}}\|_r}_{B_1} + \underbrace{\|\boldsymbol{u}^\top\mathbb{E}(\widetilde{\boldsymbol{X}} \mid \boldsymbol{X}^\top\boldsymbol{\beta}_0)\|_r}_{B_2}.$$

In what follows, we discuss B_1 and B_2 separately. We first consider B_1:

$$B_1 = \|(0, \boldsymbol{u}^\top)\boldsymbol{X}\|_r \le \sup_{\boldsymbol{v}\in\mathbb{S}^p}\|\boldsymbol{v}^\top\boldsymbol{X}\|_r. \tag{A.37}$$

We then consider B_2:

$$\begin{aligned}
B_2 = \{\mathbb{E}|\mathbb{E}(\boldsymbol{u}^\top\widetilde{\boldsymbol{X}}\big|\boldsymbol{X}^\top\boldsymbol{\beta}_0)|^r\}^{\frac{1}{r}} &\le [\mathbb{E}\{\mathbb{E}(|\boldsymbol{u}^\top\widetilde{\boldsymbol{X}}|\big|\boldsymbol{X}^\top\boldsymbol{\beta}_0)\}^r]^{\frac{1}{r}} \\
&\le \{\mathbb{E}\mathbb{E}(|\boldsymbol{u}^\top\widetilde{\boldsymbol{X}}|^r\big|\boldsymbol{X}^\top\boldsymbol{\beta}_0)\}^{\frac{1}{r}} = \|\boldsymbol{u}^\top\widetilde{\boldsymbol{X}}\|_r \le \sup_{\boldsymbol{v}\in\mathbb{S}^p}\|\boldsymbol{v}^\top\boldsymbol{X}\|_r,
\end{aligned}$$

where the second and third inequalities hold because of the convexity of $|\cdot|^r$ for $r \ge 1$. This, combined with (A.37) and Condition 1, implies that

$$\begin{aligned}
\|\boldsymbol{u}^\top(\widetilde{\boldsymbol{X}} - \mathbb{E}(\widetilde{\boldsymbol{X}} \mid \boldsymbol{X}^\top\boldsymbol{\beta}_0))\|_{\psi_2} &= \sup_{r\ge1} r^{-1/2}\mathbb{E}\|\boldsymbol{u}^\top\{\widetilde{\boldsymbol{X}} - \mathbb{E}(\widetilde{\boldsymbol{X}} \mid \boldsymbol{X}^\top\boldsymbol{\beta}_0)\}\|_r \\
&\le 2\sup_{\boldsymbol{v}\in\mathbb{S}^p}\sup_{r\ge1} r^{-1/2}\mathbb{E}\|\boldsymbol{v}^\top\boldsymbol{X}\|_r = 2\sup_{\boldsymbol{v}\in\mathbb{S}^p}\|\boldsymbol{v}^\top\boldsymbol{X}\|_{\psi_2} \le 2c'',
\end{aligned}$$

which completes the proof. □

Next, we give the following lemma which establishes the upper bound for $\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)}\mathbb{E}\{h^{\mathsf{H}}(\cdot, \cdot; \boldsymbol{\theta})\}^2$.

Lemma A.6. *Under Assumptions 5 and 6 in the main text, then for any small $r > 0$ with $\overline{\mathcal{B}}(\boldsymbol{\theta}_0, r) \subset \Theta^{\mathsf{H}}$, $\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)}\mathbb{E}\{h^{\mathsf{H}}(\cdot, \cdot; \boldsymbol{\theta})\}^2 \lesssim \sqrt{p}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2$.*

Proof. Write $H(\boldsymbol{\theta}) = \mathbb{E}\{h^{\mathsf{H}}(\cdot,\cdot;\boldsymbol{\theta})\}^2$. Substitute the equation for $h^{\mathsf{H}}(\boldsymbol{z}_1,\boldsymbol{z}_2;\boldsymbol{\theta})$ into $H(\boldsymbol{\theta})$ and consolidate terms to get that

$$\begin{aligned} H(\boldsymbol{\theta}) =& \underbrace{\mathbb{E}\{f^{\mathsf{H}}(\cdot,\cdot;\boldsymbol{\theta})\}^2}_{H_1(\boldsymbol{\theta})} - \underbrace{\mathbb{E}\{\mathbb{E}_{\mathbb{P}} f^{\mathsf{H}}(\boldsymbol{Z}_1,\cdot;\boldsymbol{\theta})\}^2}_{H_2(\boldsymbol{\theta})} - \underbrace{\mathbb{E}\{\mathbb{E}_{\mathbb{P}} f^{\mathsf{H}}(\cdot,\boldsymbol{Z}_2;\theta)\}^2}_{H_3(\boldsymbol{\theta})} + 2\underbrace{\mathbb{E}\{\mathbb{E}_{\mathbb{P}} f^{\mathsf{H}}(\boldsymbol{Z}_1,\cdot;\boldsymbol{\theta})\mathbb{E}_{\mathbb{P}} f^{\mathsf{H}}(\cdot,\boldsymbol{Z}_2;\boldsymbol{\theta})\}}_{H_4(\boldsymbol{\theta})} \\ & - \{\Gamma^{\mathsf{H}}(\boldsymbol{\theta})\}^2. \end{aligned}$$

Fix $\boldsymbol{\theta} \in \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)$. Expand $H(\boldsymbol{\theta})$ about $\boldsymbol{\theta}_0$ to get

$$H(\boldsymbol{\theta}) = (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \nabla_1 H(\boldsymbol{\theta}'),$$

where $\boldsymbol{\theta}'$ is between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}$. We wish to bound $\|\nabla_1 H(\boldsymbol{\theta}')\|_\infty$. To that end, we discuss $H_j(\boldsymbol{\theta})$ separately for $j = 1,\ldots,4$. With a little abuse of notation, we still use $\boldsymbol{\theta}$ instead of $\boldsymbol{\theta}'$ below.

We first consider $\nabla_1 H_1(\boldsymbol{\theta})$. By the property of exchangeability between integration and derivation with $\nabla_1 H_1(\boldsymbol{\theta})$, we have

$$\nabla_1 H_1(\boldsymbol{\theta}) = \nabla_1 \mathbb{E}\{f^{\mathsf{H}}(\cdot,\cdot;\boldsymbol{\theta})\}^2 = \mathbb{E}[\nabla_1 \mathbb{E}_{\mathbb{P}}\{f^{\mathsf{H}}(\boldsymbol{Z}_1,\cdot;\boldsymbol{\theta})\}^2] = \mathbb{E}\{\nabla_1 h_1(\boldsymbol{Z}_1;\boldsymbol{\theta})\},$$

where

$$\begin{aligned} h_1(\boldsymbol{z};\boldsymbol{\theta}) = \mathbb{E}_{\mathbb{P}}\{f^{\mathsf{H}}(\boldsymbol{z},\boldsymbol{Z};\boldsymbol{\theta})\}^2 =& \mathbb{E}\{\mathbb{1}(y > Y)\mathbb{1}(\boldsymbol{x}^\top\boldsymbol{\beta} > \boldsymbol{X}^\top\boldsymbol{\beta})\} + \mathbb{E}\{\mathbb{1}(y > Y)\mathbb{1}(\boldsymbol{x}^\top\boldsymbol{\beta}_0 > \boldsymbol{X}^\top\boldsymbol{\beta}_0)\} \\ & - 2\mathbb{E}\{\mathbb{1}(y > Y)\mathbb{1}(\boldsymbol{x}^\top\boldsymbol{\beta} > \boldsymbol{X}^\top\boldsymbol{\beta})\mathbb{1}(\boldsymbol{x}^\top\boldsymbol{\beta}_0 > \boldsymbol{X}^\top\boldsymbol{\beta}_0)\}. \end{aligned}$$

Similarly, we can write $\nabla_1 H_2(\boldsymbol{\theta})$, $\nabla_1 H_3(\boldsymbol{\theta})$ and $\nabla_1 H_4(\boldsymbol{\theta})$ respectively as

$$\nabla_1 H_2(\boldsymbol{\theta}) = \mathbb{E}\{\nabla_1 h_2(\cdot;\boldsymbol{\theta})\},\ \nabla_1 H_3(\boldsymbol{\theta}) = \mathbb{E}\{\nabla_1 h_3(\cdot;\boldsymbol{\theta})\},\ \nabla_1 H_4(\boldsymbol{\theta}) = \mathbb{E}\{\nabla_1 h_4(\cdot,\cdot;\boldsymbol{\theta})\},$$

where

$$\begin{aligned} h_2(\boldsymbol{z};\boldsymbol{\theta}) =& [\mathbb{E}\{\mathbb{1}(y > Y)\mathbb{1}(\boldsymbol{x}^\top\boldsymbol{\beta} > \boldsymbol{X}^\top\boldsymbol{\beta})\}]^2 + [\mathbb{E}\{\mathbb{1}(y > Y)\mathbb{1}(\boldsymbol{x}^\top\boldsymbol{\beta}_0 > \boldsymbol{X}^\top\boldsymbol{\beta}_0)\}]^2 \\ & - 2\mathbb{E}\{\mathbb{1}(y > Y)\mathbb{1}(\boldsymbol{x}^\top\boldsymbol{\beta} > \boldsymbol{X}^\top\boldsymbol{\beta})\} \cdot \mathbb{E}\{\mathbb{1}(y > Y)\mathbb{1}(\boldsymbol{x}^\top\boldsymbol{\beta}_0 > \boldsymbol{X}^\top\boldsymbol{\beta}_0)\}, \\ h_3(\boldsymbol{z};\boldsymbol{\theta}) =& [\mathbb{E}\{\mathbb{1}(Y > y)\mathbb{1}(\boldsymbol{X}^\top\boldsymbol{\beta} > \boldsymbol{x}^\top\boldsymbol{\beta})\}]^2 + [\mathbb{E}\{\mathbb{1}(Y > y)\mathbb{1}(\boldsymbol{X}^\top\boldsymbol{\beta}_0 > \boldsymbol{x}^\top\boldsymbol{\beta}_0)\}]^2 \\ & - 2\mathbb{E}\{\mathbb{1}(Y > y)\mathbb{1}(\boldsymbol{X}^\top\boldsymbol{\beta} > \boldsymbol{x}^\top\boldsymbol{\beta})\} \cdot \mathbb{E}\{\mathbb{1}(Y > y)\mathbb{1}(\boldsymbol{X}^\top\boldsymbol{\beta}_0 > \boldsymbol{x}^\top\boldsymbol{\beta}_0)\}, \end{aligned}$$

and

$$\begin{aligned} h_4(\boldsymbol{z}_1,\boldsymbol{z}_2;\boldsymbol{\theta}) =& \mathbb{E}\{\mathbb{1}(y_1 > Y)I(\boldsymbol{x}_1^\top\boldsymbol{\beta} > \boldsymbol{X}^\top\boldsymbol{\beta})\} \cdot \mathbb{E}\{\mathbb{1}(Y > y_2)\mathbb{1}(\boldsymbol{X}^\top\boldsymbol{\beta} > \boldsymbol{x}_2^\top\boldsymbol{\beta})\} \\ & - \mathbb{E}\{\mathbb{1}(y_1 > Y)\mathbb{1}(\boldsymbol{x}_1^\top\boldsymbol{\beta} > \boldsymbol{X}^\top\boldsymbol{\beta})\} \cdot \mathbb{E}\{\mathbb{1}(Y > y_2)\mathbb{1}(\boldsymbol{X}^\top\boldsymbol{\beta}_0 > \boldsymbol{x}_2^\top\boldsymbol{\beta}_0)\} \\ & - \mathbb{E}\{\mathbb{1}(y_1 > Y)\mathbb{1}(\boldsymbol{x}_1^\top\boldsymbol{\beta}_0 > \boldsymbol{X}^\top\boldsymbol{\beta}_0)\} \cdot \mathbb{E}\{\mathbb{1}(Y > y_2)\mathbb{1}(\boldsymbol{X}^\top\boldsymbol{\beta} > \boldsymbol{x}_2^\top\boldsymbol{\beta})\} \\ & + \mathbb{E}\{\mathbb{1}(y_1 > Y)\mathbb{1}(\boldsymbol{x}_1^\top\boldsymbol{\beta}_0 > \boldsymbol{X}^\top\boldsymbol{\beta}_0)\} \cdot \mathbb{E}\{\mathbb{1}(Y > y_2)\mathbb{1}(\boldsymbol{X}^\top\boldsymbol{\beta}_0 > \boldsymbol{x}_2^\top\boldsymbol{\beta}_0)\}. \end{aligned}$$

Thus, we can rewrite $\nabla_1 H(\boldsymbol{\theta})$ as

$$\nabla_1 H(\theta) = \mathbb{E}\{\nabla_1 h_1(\cdot;\boldsymbol{\theta})\} - \mathbb{E}\{\nabla_1 h_2(\cdot;\boldsymbol{\theta})\} - \mathbb{E}\{\nabla_1 h_3(\cdot;\boldsymbol{\theta})\} + 2\mathbb{E}\{\nabla_1 h_4(\cdot,\cdot;\boldsymbol{\theta})\} - \nabla_1\{\Gamma^{\mathsf{H}}(\boldsymbol{\theta})\}^2.$$

To simplify the expression forms of the functions h_j with $j = 1,\ldots,4$, we introduce the following notations:

$$\begin{aligned} \varphi_1(\boldsymbol{z};\boldsymbol{\theta}) &:= \mathbb{E}\{\mathbb{1}(y > Y)\mathbb{1}(\boldsymbol{x}^\top\boldsymbol{\beta} > \boldsymbol{X}^\top\boldsymbol{\beta})\}, \\ \varphi_2(\boldsymbol{z},\boldsymbol{\theta}) &:= \mathbb{E}\{\mathbb{1}(y > Y)\mathbb{1}(\boldsymbol{x}^\top\boldsymbol{\beta} > \boldsymbol{X}^\top\boldsymbol{\beta})\mathbb{1}(\boldsymbol{x}^\top\boldsymbol{\beta}_0 > \boldsymbol{X}^\top\boldsymbol{\beta}_0)\}, \\ \varphi_3(\boldsymbol{z}) &:= \mathbb{E}\{\mathbb{1}(y > Y)\mathbb{1}(\boldsymbol{x}^\top\boldsymbol{\beta}_0 > \boldsymbol{X}^\top\boldsymbol{\beta}_0)\}, \end{aligned}$$

and

$$\begin{aligned} \omega_1(\boldsymbol{z};\boldsymbol{\theta}) &:= \mathbb{E}\{\mathbb{1}(Y > y)\mathbb{1}(\boldsymbol{X}^\top\boldsymbol{\beta} > \boldsymbol{x}^\top\boldsymbol{\beta})\}, \\ \omega_2(\boldsymbol{z};\boldsymbol{\theta}) &:= \mathbb{E}\{\mathbb{1}(Y > y)\mathbb{1}(\boldsymbol{X}^\top\boldsymbol{\beta} > \boldsymbol{x}^\top\boldsymbol{\beta})\mathbb{1}(\boldsymbol{X}^\top\boldsymbol{\beta}_0 > \boldsymbol{x}^\top\boldsymbol{\beta}_0)\}, \\ \omega_3(\boldsymbol{z}) &:= \mathbb{E}\{\mathbb{1}(Y > y)\mathbb{1}(\boldsymbol{X}^\top\boldsymbol{\beta}_0 > \boldsymbol{x}^\top\boldsymbol{\beta}_0)\}. \end{aligned}$$

This, combined with that $\Gamma^{\mathsf{H}}(\boldsymbol{\theta}) = \mathbb{E}\{\varphi_1(\cdot;\boldsymbol{\theta}) - \varphi_3(\cdot)\}$, allows us to rewrite $\nabla_1 H(\boldsymbol{\theta})$ as

$$\begin{aligned} \nabla_1 H(\boldsymbol{\theta}) =& \mathbb{E}\nabla_1\{\varphi_1(\cdot;\boldsymbol{\theta}) - 2\varphi_2(\cdot;\boldsymbol{\theta}) + \varphi_3(\cdot) - \varphi_1^2(\cdot;\boldsymbol{\theta}) + 2\varphi_1(\cdot;\boldsymbol{\theta})\varphi_3(\cdot) - \varphi_3^2(\cdot) \\ & - \omega_1^2(\cdot;\boldsymbol{\theta}) + 2\omega_2(\cdot;\boldsymbol{\theta})\omega_3(\cdot) - \omega_3^2(\cdot) + 2\varphi_1(\cdot;\boldsymbol{\theta})\omega_1(\cdot;\boldsymbol{\theta}) - 2\varphi_1(\cdot;\boldsymbol{\theta})\omega_3(\cdot) \\ & - 2\varphi_3(\cdot)\omega_1(\cdot;\boldsymbol{\theta}) + 2\varphi_3(\cdot)\omega_3(\cdot)\} - 2\Gamma^{\mathsf{H}}(\boldsymbol{\theta})\mathbb{E}\nabla_1\{\varphi_1(\cdot;\boldsymbol{\theta}) - \varphi_3(\cdot)\}. \end{aligned}$$

Since the functions φ_j, ω_j are all bounded, we just need to bound $\|\mathbb{E}|\nabla_1\varphi_j|\|_\infty$ and $\|\mathbb{E}|\nabla_1\omega_j|\|_\infty$ for $j = 1, 2, 3$.

We first consider $\mathbb{E}|\nabla_1\varphi_1(\cdot;\boldsymbol{\theta})|$ and rewrite $\varphi_1(\boldsymbol{Z};\boldsymbol{\theta})$ as follows:

$$\varphi_1(\boldsymbol{Z};\boldsymbol{\theta}) = \int_{\boldsymbol{x}^\top\boldsymbol{\beta}<\boldsymbol{X}^\top\boldsymbol{\beta}} \rho_1(Y,\boldsymbol{x}^\top\boldsymbol{\beta}_0)G(\,\mathrm{d}\boldsymbol{x}),$$

where $\rho_1(y,t)=\mathbb{E}\{\mathbb{1}(y>Y)\mid \boldsymbol{X}^\top\boldsymbol{\beta}_0=t\}$, and $G(\cdot)$ denotes the probability distribution of $\boldsymbol{X}$.

Let $\boldsymbol{u}_i$ denote the unit vector in $\mathbb{R}^{p+1}$ with the ith component equal to one and let ∇_1^i denote the ith component of ∇_1, where $i=2,\ldots,p+1$. By definition,

$$\nabla_1^i\varphi_1(\boldsymbol{Z};\boldsymbol{\beta}) = \lim_{\varepsilon\to 0}\varepsilon^{-1}\{\varphi_1(\boldsymbol{Z};\boldsymbol{\beta}+\varepsilon\boldsymbol{u}_i)-\varphi_1(\boldsymbol{Z};\boldsymbol{\beta})\}.$$

The term in brackets equals

$$\int_{\boldsymbol{x}^\top\boldsymbol{\beta}<\boldsymbol{X}^\top\boldsymbol{\beta}+\varepsilon(X_i-x_i)}\rho_1(Y,\boldsymbol{x}^\top\boldsymbol{\beta}_0)G(\,\mathrm{d}\boldsymbol{x}) - \int_{\boldsymbol{x}^\top\boldsymbol{\beta}<\boldsymbol{X}^\top\boldsymbol{\beta}}\rho_1(Y,\boldsymbol{x}^\top\boldsymbol{\beta}_0)G(\,\mathrm{d}\boldsymbol{x}).$$

Change variables from $\boldsymbol{x}=(x_1,\widetilde{\boldsymbol{x}})$ to $(\boldsymbol{x}^\top\boldsymbol{\beta}_0,\widetilde{\boldsymbol{x}})$, rearrange the terms in the fields of integration to get that

$$\begin{aligned}&\int_{\boldsymbol{x}^\top\boldsymbol{\beta}_0<\boldsymbol{X}^\top\boldsymbol{\beta}-\widetilde{\boldsymbol{x}}^\top(\boldsymbol{\theta}-\boldsymbol{\theta}_0)+\varepsilon(X_i-x_i)}\rho_1(Y,\boldsymbol{x}^\top\boldsymbol{\beta}_0)G(\,\mathrm{d}\boldsymbol{x}) - \int_{\boldsymbol{x}^\top\boldsymbol{\beta}_0<\boldsymbol{X}^\top\boldsymbol{\beta}-\widetilde{\boldsymbol{x}}^\top(\boldsymbol{\theta}-\boldsymbol{\theta}_0)}\rho_1(Y,\boldsymbol{x}^\top\boldsymbol{\beta}_0)G(\,\mathrm{d}\boldsymbol{x})\\ &=\int\left\{\int_{\boldsymbol{X}^\top\boldsymbol{\beta}-\widetilde{\boldsymbol{x}}^\top(\boldsymbol{\theta}-\boldsymbol{\theta}_0)}^{\boldsymbol{X}^\top\boldsymbol{\beta}-\widetilde{\boldsymbol{x}}^\top(\boldsymbol{\theta}-\boldsymbol{\theta}_0)+\varepsilon(X_i-x_i)}\rho_1(Y,t)g_0(t\mid\widetilde{\boldsymbol{x}})\,\mathrm{d}t\right\}G_{\widetilde{\boldsymbol{X}}}(\,\mathrm{d}\widetilde{\boldsymbol{x}}),\end{aligned}$$

where $G_{\widetilde{\boldsymbol{X}}}(\cdot)$ denotes the distribution of $\widetilde{\boldsymbol{X}}$. The inner integral equals

$$\varepsilon(X_i-x_i)\rho_1\{Y,\boldsymbol{X}^\top\boldsymbol{\beta}-\widetilde{\boldsymbol{x}}^\top(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\}g_0\{\boldsymbol{X}^\top\boldsymbol{\beta}-\widetilde{\boldsymbol{x}}^\top(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\mid\widetilde{\boldsymbol{x}}\}+|X_i-x_i|o(|\varepsilon|)\quad\text{as }\varepsilon\to 0.$$

Integrate, then apply the moment condition $\sup_{i=2,\ldots,p+1}\mathbb{E}|X_i|\le\sqrt{C}$ in Assumption 5 to see that

$$\nabla_1^i\varphi_1(\boldsymbol{Z};\boldsymbol{\beta}) = \int(X_i-x_i)\rho_1\{Y,\boldsymbol{X}^\top\boldsymbol{\beta}-\widetilde{\boldsymbol{x}}^\top(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\}g_0\{\boldsymbol{X}^\top\boldsymbol{\beta}-\widetilde{\boldsymbol{x}}^\top(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\mid\widetilde{\boldsymbol{x}}\}G_{\widetilde{\boldsymbol{X}}}(\,\mathrm{d}\widetilde{\boldsymbol{x}}).$$

Since $|\rho_1(y,t)|\le 1$ and $g_0(\cdot\mid\widetilde{\boldsymbol{x}})\le C_0$ by Assumption 6, it then holds that

$$|\nabla_1^i\varphi_1(\boldsymbol{Z};\boldsymbol{\beta})|\le C_0\int|X_i-x_i|G_{\widetilde{\boldsymbol{X}}}(\,\mathrm{d}\widetilde{\boldsymbol{x}})\le C_0(|X_i|+\mathbb{E}|X_i|).$$

Thus,

$$\sup_{i=2,\ldots,p+1}\mathbb{E}|\nabla_1^i\varphi_1(\boldsymbol{Z};\boldsymbol{\beta})|\le 2C_0\sup_{i=2,\ldots,p+1}\mathbb{E}|X_i|\le 2C_0\sqrt{C}.$$

Similarly,

$$\begin{aligned}&\sup_{i=2,\ldots,p+1}\mathbb{E}|\nabla_1^i\varphi_2(\boldsymbol{Z};\boldsymbol{\beta})|\le 2C_0\sup_{i=2,\ldots,p+1}\mathbb{E}|X_i|\le 2C_0\sqrt{C},\\ &\sup_{i=2,\ldots,p+1}\mathbb{E}|\nabla_1^i\omega_1(\boldsymbol{Z};\boldsymbol{\beta})|\le 2C_0\sup_{i=2,\ldots,p+1}\mathbb{E}|X_i|\le 2C_0\sqrt{C},\\ &\sup_{i=2,\ldots,p+1}\mathbb{E}|\nabla_1^i\omega_2(\boldsymbol{Z};\boldsymbol{\beta})|\le 2C_0\sup_{i=2,\ldots,p+1}\mathbb{E}|X_i|\le 2C_0\sqrt{C}.\end{aligned}$$

Put all results together, and we have that

$$\|\nabla_1 H(\boldsymbol{\theta})\|_\infty\le C_1\sup_{i=2,\ldots,p+1}\mathbb{E}|X_i|\le C_1\sqrt{C}$$

for some constant C_1 depending only on C_0. Then

$$H(\boldsymbol{\theta})=(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top H(\boldsymbol{\theta}')\le\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1\|\nabla H(\boldsymbol{\theta}')\|_\infty\le C_2\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1\le C_2\sqrt{p}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2.$$

That is, $\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r)}\mathbb{E}\{h^{\mathsf{H}}(\cdot,\cdot;\boldsymbol{\theta})\}^2\lesssim\sqrt{p}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2$. This completes the proof. □

The next three lemmas give the upper bound for $\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r)}\mathbb{E}\{h^{\mathsf{C}}(\cdot,\cdot;\boldsymbol{\theta})\}^2$, $\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r)}\mathbb{E}\{h^{\mathsf{K}}(\cdot,\cdot;\boldsymbol{\theta})\}^2$, and $\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r)}\mathbb{E}\{h^{\mathsf{A}}(\cdot,\cdot;\boldsymbol{\theta})\}^2$, respectively. Since the proofs of these lemmas are similar to the proof of Lemma A.6, we omit the proofs for simplicity.

Lemma A.7. *Suppose that Assumptions 5–6 in the main text hold. Then for any small $r>0$ with $\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r)\subset\Theta^{\mathsf{C}}$, $\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r)}\mathbb{E}\{h^{\mathsf{C}}(\cdot,\cdot;\boldsymbol{\theta})\}^2\lesssim\sqrt{p}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2$.*

Lemma A.8. *Suppose that Assumptions 5–6 in the main text hold. Then for any small $r>0$ with $\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r)\subset\Theta^{\mathsf{K}}$, $\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r)}\mathbb{E}\{h^{\mathsf{K}}(\cdot,\cdot;\boldsymbol{\theta})\}^2\lesssim\sqrt{p}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2$.*

Lemma A.9. *Suppose that Assumptions 5 and 10 in the main text hold. Then for any small $r > 0$ with $\overline{\mathcal{B}}(\boldsymbol{\theta}_0, r) \subset \Theta^{\mathsf{A}}$, $\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r)} \mathbb{E}\{h^{\mathsf{A}}(\cdot,\cdot;\boldsymbol{\theta})\}^2 \lesssim \sqrt{p}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2$.*

A.4.2. Proof of Corollary 3.1

Proof. Note that $\mathcal{F}^{\mathsf{H}}$ is uniformly bounded. To prove Corollary 3.1(i) and (ii), it suffices to show that the VC-dimension of $\mathcal{F}^{\mathsf{H}}$ is $\sim p$ by Theorems 2.2 and 2.3.

To see this, define the following function:

$$g(\boldsymbol{z}_1, \boldsymbol{z}_2, t; \gamma, \gamma_1, \gamma_2, \boldsymbol{\delta}_1, \boldsymbol{\delta}_2) = \gamma t + \gamma_1 y_1 + \gamma_2 y_2 + \boldsymbol{\delta}_1^\top \boldsymbol{x} + \boldsymbol{\delta}_2^\top \boldsymbol{x},$$

and the following function class:

$$\mathscr{G} = \{g(\boldsymbol{z}_1, \boldsymbol{z}_2, t; \gamma, \gamma_1, \gamma_2, \boldsymbol{\delta}_1, \boldsymbol{\delta}_2) : \gamma, \gamma_1, \gamma_2 \in \mathbb{R}, \boldsymbol{\delta}_1, \boldsymbol{\delta}_2 \in \mathbb{R}^{p+1}\}.$$

Note that $\mathscr{G}$ is a $(2p+5)$-dimensional vector space of real-valued functions. By Lemma 18 in Pollard (1984) and Lemma 2.4 in Pakes and Pollard (1989), $\{\mathscr{G} \geq s\}$ and $\{\mathscr{G} > s\}$ are VC-classes of VC-dimensions $2p+5$ for any $s \in \mathbb{R}$. We further have, for any $\boldsymbol{\theta} \in \Theta^{\mathsf{H}}$, $\boldsymbol{\beta} = (1, \boldsymbol{\theta}^\top)^\top$, and $\boldsymbol{\beta}_0 = (1, \boldsymbol{\theta}_0^\top)^\top$,

$$\begin{aligned}
\text{subgraph}\{f^{\mathsf{H}}(\cdot,\cdot;\boldsymbol{\theta})\} =& \{(\boldsymbol{z}_1, \boldsymbol{z}_2, t) \in \mathcal{S}\otimes\mathcal{S}\otimes\mathbb{R} : t < f^{\mathsf{H}}(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta})\} \\
=& \Big\{\{y_1 - y_2 > 0\} \cap \{\boldsymbol{x}_1^\top\boldsymbol{\beta} - \boldsymbol{x}_2^\top\boldsymbol{\beta} > 0\} \cap \{\boldsymbol{x}_1^\top\boldsymbol{\beta}_0 - \boldsymbol{x}_2^\top\boldsymbol{\beta}_0 > 0\}^{\mathsf{c}} \cap \{t \geq 1\}^{\mathsf{c}}\Big\} \\
& \cup \Big\{\{y_1 - y_2 > 0\}^{\mathsf{c}} \cap \{t \geq 0\}^{\mathsf{c}}\Big\} \cup \Big\{ \{y_1 - y_2 > 0\} \cap \{\boldsymbol{x}_1^\top\boldsymbol{\beta} - \boldsymbol{x}_2^\top\boldsymbol{\beta} > 0\}^{\mathsf{c}} \cap \\
& \{\boldsymbol{x}_1^\top\boldsymbol{\beta}_0 - \boldsymbol{x}_2^\top\boldsymbol{\beta}_0 > 0\} \cap \{t \geq -1\}^{\mathsf{c}} \Big\} \\
=& \Big\{\{g_1 > 0\} \cap \{g_2 > 0\} \cap \{g_3 > 0\}^{\mathsf{c}} \cap \{g_4 \geq 1\}^{\mathsf{c}}\Big\} \cup \Big\{\{g_1 > 0\}^{\mathsf{c}} \cap \{g_4 \geq 0\}^{\mathsf{c}}\Big\} \\
& \cup \Big\{\{g_1 > 0\} \cap \{g_2 > 0\}^{\mathsf{c}} \cap \{g_3 > 0\} \cap \{g_4 \geq -1\}^{\mathsf{c}}\Big\}
\end{aligned}$$

for $g_1, \ldots, g_4 \in \mathscr{G}$. This, combined with Lemma 9.7 in Kosorok (2007), implies that $\mathcal{F}^{\mathsf{H}}$ is a VC-class of VC-dimension $\sim p$. Then, apply Theorems 2.2 and 2.3 to complete the proof of Corollary 3.1(i) and (ii).

Next, we prove Corollary 3.1(iii). By Lemma A.6, we see that for any $c > 0$, $\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0, c\sqrt{p/n})} \mathbb{E}$ $\{h^{\mathsf{H}}(\cdot,\cdot;\boldsymbol{\theta})\}^2 \lesssim p/\sqrt{n}$ if $p/n \to 0$. Connecting this with Theorem 2.4 implies that $\epsilon_n \sim p/\sqrt{n}$ and $\eta_n \sim p/\sqrt{n}$. Thus, by Theorem 2.4, we conclude that if $\log(n/p^2)p^{3/2}/n^{5/4} \to 0$, we have

$$\|\widehat{\boldsymbol{\theta}}_n^{\mathsf{H}} - \boldsymbol{\theta}_0 + (\mathbf{V}^{\mathsf{H}})^{-1}\mathbb{P}_n\nabla_1\tau^{\mathsf{H}}(\cdot;\boldsymbol{\theta}_0)\|^2 = O_{\mathbb{P}}\big\{\log(n/p^2)p^{3/2}/n^{5/4}\big\}. \tag{A.38}$$

In particular, if $\log(n/p^2)p^{3/2}/n^{1/4} \to 0$, then for any $\boldsymbol{\gamma} \in \mathbb{R}^p$,

$$\sqrt{n}\boldsymbol{\gamma}^\top(\widehat{\boldsymbol{\theta}}_n^{\mathsf{H}} - \boldsymbol{\theta}_0)/\{\boldsymbol{\gamma}^\top(\mathbf{V}^{\mathsf{H}})^{-1}\boldsymbol{\Delta}^{\mathsf{H}}(\mathbf{V}^{\mathsf{H}})^{-1}\boldsymbol{\gamma}\}^{1/2} \Rightarrow N(0,1).$$

This completes the proof.

To prove Corollary 3.1(iv), we only need to evaluate the order of $\widetilde{\nu}^{\mathsf{H}}$, the VC-dimension of $\widetilde{\mathcal{F}}^{\mathsf{H}} := \{f^{\mathsf{H}}(\boldsymbol{z},\cdot;\boldsymbol{\theta}) + f^{\mathsf{H}}(\cdot,\boldsymbol{z};\boldsymbol{\theta}) : \boldsymbol{z} \in \mathbb{R}^{p+1}, \boldsymbol{\theta} \in \Theta^{\mathsf{H}}\}$. Following similar arguments above, we can know that $\widetilde{\nu}^{\mathsf{H}}$ is also of order p. Then, the claim in Corollary 3.1(iv) follows from Theorem 2.6. □

A.4.3. Proof of Corollary 3.2

Proof. Similar to the proof of Corollary 3.1, it can be easy to show that the VC-dimensions of $\mathcal{F}^{\mathsf{C}}$ and $\widetilde{\mathcal{F}}^{\mathsf{C}} := \{f^{\mathsf{C}}(\boldsymbol{z},\cdot;\boldsymbol{\theta}) + f^{\mathsf{C}}(\cdot,\boldsymbol{z};\boldsymbol{\theta}) : \boldsymbol{z} \in \mathbb{R}^{p+1}, \boldsymbol{\theta} \in \Theta^{\mathsf{C}}\}$ are both of order p. This, combined with that $\mathcal{F}^{\mathsf{C}}$ is uniformly bounded, proves Corollary 3.1(i) and (ii) by Theorems 2.2 and 2.3. Corollary 3.1(iii) follows from Lemma A.7 and Theorem 2.4. Corollary 3.1(iv) follows from Theorem 2.6. □

A.4.4. Proof of Corollary 3.3

Proof. Similar to the proof of Corollary 3.1, one could show that the VC-dimension of $\mathcal{F}^{\mathsf{K}}$ and $\widetilde{\mathcal{F}}^{\mathsf{K}} := \{f^{\mathsf{K}}(\boldsymbol{z},\cdot;\boldsymbol{\theta}) + f^{\mathsf{K}}(\cdot,\boldsymbol{z};\boldsymbol{\theta}) : \boldsymbol{z} \in \mathbb{R}^{p+1}, \boldsymbol{\theta} \in \Theta^{\mathsf{K}}\}$ are both of order p. Then, the proofs of Corollary 3.3(i) and (ii) follow directly from the proof of Corollary 3.1. Finally, Lemma A.8, together with Theorem 2.4 implies Corollary 3.3(iii). Corollary 3.3(iv) follows from Theorem 2.6. □

A.4.5. Proof of Corollary 3.4

Proof. (i) Similar to the proof of Theorem 2.2, the proof is twofold. We first show that $\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta})$ converges in probability to $\Gamma^{\mathsf{A}}(\boldsymbol{\theta})$ uniformly in $\boldsymbol{\theta} \in \Theta^{\mathsf{A}}$, and then establish the consistency of $\widehat{\boldsymbol{\theta}}_n^{\mathsf{A}}$.

Step 1. Since $K(\cdot)$ is continuously differential with compact support by Assumption 9(vi), $K(\cdot)$ is bounded and is also a function of bounded variation. Thus, $K(\cdot)$ can be written as $K(\cdot) = K_1(\cdot) - K_2(\cdot)$ with appropriate bounded and monotone functions $K_1(\cdot)$ and $K_2(\cdot)$. Let C_1 and C_2 denote the upper bounds of $|K_1(\cdot)|$ and $|K_2(\cdot)|$ respectively.

Let $\mathcal{F}_1^{\mathsf{A}} = \big\{\mathbb{1}(Y_1 > Y_2)\mathbb{1}(\boldsymbol{X}_1^\top\boldsymbol{\beta} > \boldsymbol{X}_2^\top\boldsymbol{\beta})K_1\{(W_1 - W_2)/b\} : \boldsymbol{\theta} \in \Theta^{\mathsf{A}}\big\}$ and $\mathcal{F}_2^{\mathsf{A}} = \big\{\mathbb{1}(Y_1 > Y_2)\mathbb{1}(\boldsymbol{X}_1^\top\boldsymbol{\beta} > \boldsymbol{X}_2^\top\boldsymbol{\beta})K_2\{(W_1 - W_2)/b\} : \boldsymbol{\theta} \in \Theta^{\mathsf{A}}\big\}$. Then, $\mathcal{F}^{\mathsf{A}} = \mathcal{F}_1^{\mathsf{A}} - \mathcal{F}_2^{\mathsf{A}}$. Similar to the proof of Corollary 3.1, it can be easy to verify that the VC-dimensions of $\mathcal{F}_1^{\mathsf{A}}$ and $\mathcal{F}_2^{\mathsf{A}}$ are both $\sim p$ by considering the class of subgraphs of all functions in $\mathcal{F}_1^{\mathsf{A}}$ and $\mathcal{F}_2^{\mathsf{A}}$ separately. By Lemma 16 in Nolan and Pollard (1987), the covering number of $\mathcal{F}^{\mathsf{A}}$ is bounded through $N_r(\varepsilon, \mathbb{P}\otimes\mathbb{P}, \mathcal{F}^{\mathsf{A}}, C_1 + C_2) \le N_r(\varepsilon/4, \mathbb{P}\otimes\mathbb{P}, \mathcal{F}_1^{\mathsf{A}}, C_1)N_r(\varepsilon/4, \mathbb{P}\otimes\mathbb{P}, \mathcal{F}_2^{\mathsf{A}}, C_2)$. This, combined with Theorem 9.3 in Kosorok (2007), Lemmas A.2, A.4, and Hoeffding decomposition implies that

$$\sup_{\boldsymbol{\theta}\in\Theta^{\mathsf{A}}} \big|\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta}) - \mathbb{E}\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta})\big| = O_{\mathbb{P}}\Big(\frac{\sqrt{p}}{b\sqrt{n}}\Big) = O_{\mathbb{P}}\Big(\sqrt{\frac{p}{n^{1-2\delta}}}\Big) = o_{\mathbb{P}}(1). \tag{A.39}$$

Next, we try to bound $\sup_{\boldsymbol{\theta}\in\Theta^{\mathsf{A}}} \big|\mathbb{E}\{\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta})\} - \Gamma^{\mathsf{A}}(\boldsymbol{\theta})\big|$. Note that

$$\begin{aligned}
\mathbb{E}\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta}) &= \mathbb{E}[\{m(\boldsymbol{Z}_1, \boldsymbol{Z}_2; \boldsymbol{\theta}) - m(\boldsymbol{Z}_1, \boldsymbol{Z}_2; \boldsymbol{\theta}_0)\}K_b(W_1 - W_2)] \\
&= \mathbb{E}[\mathbb{E}\{m(\boldsymbol{Z}_1, \boldsymbol{Z}_2; \boldsymbol{\theta}) - m(\boldsymbol{Z}_1, \boldsymbol{Z}_2; \boldsymbol{\theta}_0) \mid W_1, W_2\}K_b(W_1 - W_2)] \\
&= \frac{1}{b}\iint \psi(w_1, w_2; \boldsymbol{\theta})K\Big(\frac{w_1 - w_2}{b}\Big)\phi(w_1)\phi(w_2)\,\mathrm{d}w_1\,\mathrm{d}w_2 \\
&= \iint \psi(bu + w_2, w_2; \boldsymbol{\theta})\phi(bu + w_2)K(u)\phi(w_2)\,\mathrm{d}u\,\mathrm{d}w_2.
\end{aligned} \tag{A.40}$$

A Jth-order Tylor expansion of $\mathbb{E}\{\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta})\}$ with respect to b at 0 and Assumption 9(vi)–(viii) imply that

$$\sup_{\boldsymbol{\theta}\in\Theta^{\mathsf{A}}} \big|\mathbb{E}\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta}) - \Gamma^{\mathsf{A}}(\boldsymbol{\theta})\big| \lesssim b^J = o(1). \tag{A.41}$$

This, combined with (A.39) and the triangular inequality, implies that

$$\sup_{\boldsymbol{\theta}\in\Theta^{\mathsf{A}}} \big|\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta}) - \Gamma^{\mathsf{A}}(\boldsymbol{\theta})\big| = o_{\mathbb{P}}(1).$$

Thus, the uniform convergence of $\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta})$ is shown.

Step 2. Following *Step 2* in the proof of Theorem 2.2, it can be easy to show that $\|\widehat{\boldsymbol{\theta}}_n^{\mathsf{A}} - \boldsymbol{\theta}_0\| \xrightarrow{\mathbb{P}} 0$. This completes proof of Corollary 3.4(i).

(ii) Similar to the proof of Theorem 2.3, the proof is conducted in four steps. We first define $f_n^{\mathsf{A}}(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta}) = f^{\mathsf{A}}(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta})/b$. Thus, $\mathbb{E}\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta}) = \mathbb{E}f_n^{\mathsf{A}}(\cdot, \cdot; \boldsymbol{\theta})$. By a Hoeffding decomposition of $\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta})$, we have

$$\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta}) = \mathbb{E}\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta}) + \mathbb{P}_n g_n^{\mathsf{A}}(\cdot; \boldsymbol{\theta}) + \mathbb{U}_n h_n^{\mathsf{A}}(\cdot, \cdot; \boldsymbol{\theta}),$$

where

$$g_n^{\mathsf{A}}(\boldsymbol{z}; \boldsymbol{\theta}) = \mathbb{E}f_n^{\mathsf{A}}(\boldsymbol{z}, \cdot; \boldsymbol{\theta}) + \mathbb{E}f_n^{\mathsf{A}}(\cdot, \boldsymbol{z}; \boldsymbol{\theta}) - 2\mathbb{E}\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta}),$$

and

$$h_n^{\mathsf{A}}(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta}) = f_n^{\mathsf{A}}(\boldsymbol{z}_1, \boldsymbol{z}_2; \boldsymbol{\theta}) - \mathbb{E}f_n^{\mathsf{A}}(\boldsymbol{z}_1, \cdot; \boldsymbol{\theta}) - \mathbb{E}f_n^{\mathsf{A}}(\cdot, \boldsymbol{z}_2; \boldsymbol{\theta}) + \mathbb{E}\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta}).$$

The first three steps aim to establish bounds that are similar to (A.10), (A.14) and (A.16), respectively. The last step establishes the rate of convergence of $\widehat{\boldsymbol{\theta}}_n^{\mathsf{A}}$.

Step 1. We first consider $\mathbb{E}\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta})$. By (A.39), there exists a constant $C > 0$ such that

$$\sup_{\boldsymbol{\theta}\in\Theta^{\mathsf{A}}} \big|\mathbb{E}\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta}) - \Gamma^{\mathsf{A}}(\boldsymbol{\theta})\big| \le Cb^J. \tag{A.42}$$

Fix $\boldsymbol{\theta} \in \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r) \subset \Theta^{\mathsf{A}}$. Similar to *Step 1* in the proof of Theorem 2.3, we have

$$\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)} \Big|\Gamma(\boldsymbol{\theta}) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top\mathbf{V}^{\mathsf{A}}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\Big| \le c_{\max}\rho(r)\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2. \tag{A.43}$$

This, combined with (A.42), implies that

$$\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)} \Big|\mathbb{E}\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta}) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top\mathbf{V}^{\mathsf{A}}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\Big| \le c_{\max}\rho(r)\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 + Cb^J. \tag{A.44}$$

Step 2. Similar to (A.40), a change of variables and a Jth-order Tylor expansion imply that

$$|\mathbb{P}_n g_n^{\mathsf{A}}(\cdot;\boldsymbol{\theta}) - \mathbb{P}_n\{\tau^{\mathsf{A}}(\cdot;\boldsymbol{\theta}) - \tau^{\mathsf{A}}(\cdot;\boldsymbol{\theta}_0) - 2\mathbb{E}\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta})\}| \le C' b^J$$

for some constant $C' > 0$. This, combined with (A.42), implies that

$$|\mathbb{P}_n g_n^{\mathsf{A}}(\cdot;\boldsymbol{\theta}) - \mathbb{P}_n\{\tau^{\mathsf{A}}(\cdot;\boldsymbol{\theta}) - \tau^{\mathsf{A}}(\cdot;\boldsymbol{\theta}_0) - 2\Gamma^{\mathsf{A}}(\boldsymbol{\theta})\}| \le (C + C')b^J. \tag{A.45}$$

Following the proof of Theorem 2.3 in *Step 2*, we additionally have

$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r)} \left|\mathbb{P}_n\{\tau^{\mathsf{A}}(\cdot;\boldsymbol{\theta}) - \tau^{\mathsf{A}}(\cdot;\boldsymbol{\theta}_0) - 2\Gamma^{\mathsf{A}}(\boldsymbol{\theta})\} - \frac{1}{\sqrt{n}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top \boldsymbol{W}_n^{\mathsf{A}}\right| > 4c_{\max}\rho(r)\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 + \frac{6\nu_0 r}{\sqrt{n}} d_p(\varepsilon)\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\right\} \le \varepsilon,$$

where $\boldsymbol{W}_n^{\mathsf{A}} = \sqrt{n}\mathbb{P}_n\nabla_1\tau^{\mathsf{A}}(\cdot;\boldsymbol{\theta}_0)$. Combining this with (A.45) implies that

$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r)} \left|\mathbb{P}_n g_n^{\mathsf{A}}(\cdot;\boldsymbol{\theta}) - \frac{1}{\sqrt{n}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top \boldsymbol{W}_n^{\mathsf{A}}\right| > (C + C')b^J + 2c_{\max}\rho(r)\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 + \frac{6\nu_0 r}{\sqrt{n}} d_p(\varepsilon)\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\right\} \le \varepsilon, \tag{A.46}$$

Step 3. Following the proof of Theorem 2.1(i), one can get

$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r)} |\mathbb{U}_n h^{\mathsf{A}}(\cdot,\cdot;\boldsymbol{\theta})| > \delta_n p/n\right\} \le \epsilon,$$

where δ_n is a sequence of nonnegative real numbers converging to zero. Thus,

$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r)} |\mathbb{U}_n h_n^{\mathsf{A}}(\cdot,\cdot;\boldsymbol{\theta})| > \delta_n p/(bn)\right\} \le \epsilon. \tag{A.47}$$

Step 4. By the Hoeffding decomposition of $\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta})$ and the results in (A.44), (A.46) and (A.47), we have

$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r)} \left|\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta}) - \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top \mathbf{V}^{\mathsf{A}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) - \frac{1}{\sqrt{n}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top \boldsymbol{W}_n^{\mathsf{A}}\right| > (2C + C')b^J + 5c_{\max}\rho(r)\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 + \frac{6\nu_0 r}{\sqrt{n}} d_p(\varepsilon)\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| + \delta_n\frac{p}{bn}\right\} \le 2\varepsilon. \tag{A.48}$$

Then, following the proof of Theorem 2.3 in *Step 4*, we conclude that there exists a sufficiently large constant $C'_\varepsilon > 0$ such that

$$\mathbb{P}\left\{\|\widehat{\boldsymbol{\theta}}_n^{\mathsf{A}} - \boldsymbol{\theta}_0\| \le C'_\varepsilon\sqrt{\frac{p}{n^{1-\delta}}}\right\} \ge 1 - 3\varepsilon \tag{A.49}$$

holds for sufficiently large n.

Since $\mathcal{F}^{\mathsf{A}}$ is uniformly bounded and $\sup_{\overline{\mathcal{B}}(\boldsymbol{\theta}_0, C'_\varepsilon\sqrt{p/n^{1-\delta}})} \mathbb{E}\{h^{\mathsf{A}}(\cdot,\cdot;\boldsymbol{\theta})\}^2 \lesssim p/\sqrt{n^{1-\delta}}$ by Lemma A.9, Theorem 2.1(ii) implies that there exists a constant C''_ε such that

$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0, C'_\varepsilon\sqrt{p/n^{1-\delta}})} \left|\mathbb{U}_n h^{\mathsf{A}}(\cdot,\cdot;\boldsymbol{\theta})\right| > C''_\varepsilon \log(n^{1-\delta}/p^2)p^{\frac{3}{2}}/n^{\frac{5-\delta}{4}}\right\} \le \varepsilon$$

holds for sufficiently large n. Thus,

$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0, C'_\varepsilon\sqrt{p/n^{1-\delta}})} \left|\mathbb{U}_n h_n^{\mathsf{A}}(\cdot,\cdot;\boldsymbol{\theta})\right| > C''_\varepsilon \log(n^{1-\delta}/p^2)p^{\frac{3}{2}}/n^{\frac{5(1-\delta)}{4}}\right\} \le \varepsilon. \tag{A.50}$$

In view of $\delta < \frac{1}{5}$, it holds that $\{\log(n^{1-\delta}/p^2)/n^{\frac{5(1-\delta)}{4}}\}/(1/n) \to 0$ as $n \to \infty$. This, combined with (A.51), implies that

$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0, C'_\varepsilon\sqrt{p/n^{1-\delta}})} \left|\mathbb{U}_n h_n^{\mathsf{A}}(\cdot,\cdot;\boldsymbol{\theta})\right| > C''_\varepsilon p^{\frac{3}{2}}/n\right\} \le \varepsilon. \tag{A.51}$$

Based on similar analyses at the beginning of this step, we conclude that, there exists a sufficiently large constant $c'_\varepsilon > 0$ such that

$$\mathbb{P}\left\{\|\widehat{\boldsymbol{\theta}}_n^{\mathsf{A}} - \boldsymbol{\theta}_0\| \le c'_\varepsilon\sqrt{\frac{p^{3/2}}{n}}\right\} \ge 1 - 6\varepsilon$$

holds for sufficiently large n. This, combined with (A.49), implies that there exists a sufficiently large constant $C_\varepsilon > 0$ such that

$$\mathbb{P}\left\{\|\widehat{\boldsymbol{\theta}}_n^{\mathsf{A}} - \boldsymbol{\theta}_0\| \le C_\varepsilon \sqrt{\frac{p}{n^{1-\delta}} \wedge \frac{p^{3/2}}{n}}\right\} \ge 1 - 9\varepsilon.$$

This completes proof of (ii).

(iii) Similar to the proof of Theorem 2.4, we first define $\boldsymbol{t}_n^{*\mathsf{A}} = -(\mathbf{V}^{\mathsf{A}})^{-1}\boldsymbol{W}_n^{\mathsf{A}}$. Similarly, there exists a constant c_ε' such that

$$\mathbb{P}\big(\mathcal{A}_{n,\varepsilon}'\big) \ge 1 - 2\varepsilon \tag{A.52}$$

holds for sufficiently large n, where $\mathcal{A}_{n,\varepsilon}' := \{\boldsymbol{Z} : \widehat{\boldsymbol{\theta}}_n^{\mathsf{A}} \in \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r_n), \boldsymbol{t}_n^{*\mathsf{A}}/\sqrt{n} + \boldsymbol{\theta}_0 \in \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r_n)\}$ and $r_n := c_\varepsilon'\sqrt{p/n^{1-\delta}}$.

Fix $\boldsymbol{\theta} \in \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r_n)$. Then following the proofs of Corollary 3.4(ii) in *Step 1–2*, we have

$$\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)} \left|\mathbb{E}\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta}) - \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top \mathbf{V}^{\mathsf{A}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right| \le c_{\max}\rho(r_n)\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 + Cb^J, \tag{A.53}$$

and

$$\begin{aligned}\mathbb{P}\Bigg\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)} \left|\mathbb{P}_n g_n^{\mathsf{A}}(\cdot;\boldsymbol{\theta}) - \frac{1}{\sqrt{n}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top \boldsymbol{W}_n^{\mathsf{A}}\right| > (C+C')b^J + 4c_{\max}\rho(r_n)\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 + \\ \frac{6\nu_0 r_n}{\sqrt{n}} d_p(\varepsilon)\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \Bigg\} \le \varepsilon.\end{aligned} \tag{A.54}$$

Similar to (A.51), we have

$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)} \left|\mathbb{U}_n h_n^{\mathsf{A}}(\cdot,\cdot;\boldsymbol{\theta})\right| > C_\varepsilon' \log(n^{1-\delta}/p^2)p^{\frac{3}{2}}/n^{\frac{5(1-\delta)}{4}}\right\} \le \varepsilon. \tag{A.55}$$

This, together with (A.53) and (A.54), implies that

$$\begin{aligned}\mathbb{P}\Bigg\{\sup_{\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r_n)} \left|\Gamma_n^{\mathsf{A}}(\boldsymbol{\theta}) - \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top \mathbf{V}^{\mathsf{A}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) - \frac{1}{\sqrt{n}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top \boldsymbol{W}_n^{\mathsf{A}}\right| > (2C+C')b^J + \\ 5c_{\max}\rho(r_n)\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 + \frac{6\nu_0 r_n}{\sqrt{n}} d_p(\varepsilon)\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| + C_\varepsilon' \log(n^{1-\delta}/p^2)p^{\frac{3}{2}}/n^{\frac{5(1-\delta)}{4}} \Bigg\} \le 2\varepsilon.\end{aligned} \tag{A.56}$$

The remaining proofs are straightforward and follow the proof of Theorem 2.4. In conclusion, if $\log(n^{1-\delta}/p^2)p^{3/2}/n^{(5-5\delta)/4} \to 0$, we have

$$\|\widehat{\boldsymbol{\theta}}_n^{\mathsf{A}} - \boldsymbol{\theta}_0 + (\mathbf{V}^{\mathsf{A}})^{-1}\mathbb{P}_n\nabla_1\tau^{\mathsf{A}}(\cdot;\boldsymbol{\theta}_0)\|^2 = O_{\mathbb{P}}\big\{n^{-\delta J} \vee \log(n^{1-\delta}/p^2)p^{3/2}/n^{(5-5\delta)/4}\big\}.$$

In addition, if $\log(n^{1-\delta}/p^2)p^{3/2}/n^{(1-5\delta)/4} \to 0$, then for any $\boldsymbol{\gamma} \in \mathbb{R}^p$,

$$\sqrt{n}\boldsymbol{\gamma}^\top(\widehat{\boldsymbol{\theta}}_n^{\mathsf{A}} - \boldsymbol{\theta}_0)/\{\boldsymbol{\gamma}^\top(\mathbf{V}^{\mathsf{A}})^{-1}\boldsymbol{\Delta}^{\mathsf{A}}(\mathbf{V}^{\mathsf{A}})^{-1}\boldsymbol{\gamma}\}^{1/2} \Rightarrow N(0,1).$$

This completes the proof of (iii).

(iv) Note that $\mathbb{E}\tau_n^{\mathsf{A}}(\boldsymbol{z};\boldsymbol{\theta}) \ne \tau^{\mathsf{A}}(\boldsymbol{z};\boldsymbol{\theta})$. The proof is a little different from that in proving Theorem 2.6. To see this, define $\widetilde{\mathcal{F}}^{\mathsf{A}} = \{f^{\mathsf{A}}(\boldsymbol{z},\cdot;\boldsymbol{\theta}) + \bar{f}^{\mathsf{A}}(\cdot,\boldsymbol{z};\boldsymbol{\theta}) : \boldsymbol{z} \in \mathbb{R}^{p+1}, \boldsymbol{\theta} \in \Theta^{\mathsf{A}}\}$. Following the similar arguments in proof of (i), one can show that the VC-dimension of $\mathcal{F}^{\mathsf{A}}$ is of order p. It then follows from Lemma A.2 that

$$\sup_{\mathbb{R}^m\otimes\Theta^{\mathsf{A}}} |\tau_n^{\mathsf{A}}(\boldsymbol{z};\boldsymbol{\theta}) - \mathbb{E}\tau_n^{\mathsf{A}}(\boldsymbol{z};\boldsymbol{\theta})| = O_{\mathbb{P}}\big\{\sqrt{p}/(b\sqrt{n})\big\} = O_{\mathbb{P}}\big(p/n^{1-2\delta}\big).$$

Similar to the derivations in (A.40) and (A.41), we have

$$\sup_{\mathbb{R}^m\otimes\Theta^{\mathsf{A}}} |\mathbb{E}\tau_n^{\mathsf{A}}(\boldsymbol{z};\boldsymbol{\theta}) - \tau^{\mathsf{A}}(\boldsymbol{z};\boldsymbol{\theta})| = O(b^J) = O(n^{-\delta J}).$$

Then, following from the proof of Theorem 2.6, we get that

$$\begin{aligned}\|\widehat{\boldsymbol{\Delta}}^{\mathsf{A}} - \boldsymbol{\Delta}^{\mathsf{A}}\| &= O_{\mathbb{P}}\big(p/\sqrt{n}\big) + O_{\mathbb{P}}\big[p\{\varepsilon_n + \varepsilon_n^{-1}(r_n^2\sqrt{p/n} + pr_n^4 + \sqrt{p/n^{1-2\delta}} + n^{-\delta J}) + r_n\}^2\big], \\ \|\widehat{\mathbf{V}}^{\mathsf{A}} - \mathbf{V}^{\mathsf{A}}\| &= O_{\mathbb{P}}\big(p/\sqrt{n}\big) + O_{\mathbb{P}}\big[p\{\varepsilon_n + \varepsilon_n^{-2}(r_n^2\sqrt{p/n} + pr_n^4 + \sqrt{p/n^{1-2\delta}} + n^{-\delta J}) + \varepsilon_n^{-1}r_n\}^2\big],\end{aligned}$$

where $r_n = \sqrt{p/n^{1-\delta}} \wedge \sqrt{p^{3/2}/n}$. By assumption, $\log(n^{1-\delta}/p^2)p^{3/2}/n^{(1-5\delta)/4} = o(1)$, $\varepsilon_n\sqrt{p} = o(1)$, and $\varepsilon_n^{-2}p/\sqrt{n^{1-2\delta}} = o(1)$, one can show that

$$\|\widehat{\boldsymbol{\Delta}}^{\mathsf{A}} - \boldsymbol{\Delta}^{\mathsf{A}}\| = o_{\mathbb{P}}(1), \quad \text{and} \quad \|\widehat{\mathbf{V}}^{\mathsf{A}} - \mathbf{V}^{\mathsf{A}}\| = o_{\mathbb{P}}(1).$$

The remaining proof follows exactly from that in the proof of Theorem 2.6. □

A.5. Proof of Theorem 3.1

Proof. We check Assumption 3(iii) and (v) separately under Conditions 1–3.

(iii) The proof proceeds in two steps. We first calculate the third order mixed partial derivatives of $\mathbb{E}\tau^{\mathsf{H}}(\cdot;\cdot)$. Then we establish the bound of $\|\mathbf{V}^{\mathsf{H}}(\boldsymbol{\theta})-\mathbf{V}^{\mathsf{H}}\|$ for any $\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r)$.
Step 1. Fix $\boldsymbol{z}=(\boldsymbol{x},y)^T\in\mathbb{R}^m$ and $\boldsymbol{\theta}\in\overline{\mathcal{B}}(\boldsymbol{\theta}_0,r)$. Note that

$$\tau^{\mathsf{H}}(\boldsymbol{z};\boldsymbol{\theta})=\int_{-\infty}^{\boldsymbol{x}^\top\boldsymbol{\beta}}\int_{-\infty}^{y}g_0(t\mid s;\boldsymbol{\theta})G_Y(\mathrm{d}s)\,\mathrm{d}t+\int_{\boldsymbol{x}^\top\boldsymbol{\beta}}^{\infty}\int_{y}^{\infty}g_0(t\mid s;\boldsymbol{\theta})G_Y(\mathrm{d}s)\,\mathrm{d}t+C(\boldsymbol{\theta}_0),$$

where $G_Y(\cdot)$ denotes the marginal distribution of Y and $g_0(\cdot\mid s;\boldsymbol{\theta})$ denotes the conditional density function of $\boldsymbol{X}^\top\boldsymbol{\beta}$ given $Y=s$, $C(\boldsymbol{\theta}_0)$ is a term that does not depend on $\boldsymbol{\theta}$, and

$$g_0(t\mid s;\boldsymbol{\theta})=\int g_0(t\mid s,\widetilde{\boldsymbol{x}};\boldsymbol{\theta})G_{\widetilde{\boldsymbol{X}}\mid Y=s}(\mathrm{d}\widetilde{\boldsymbol{x}})=\int f_0(t-\widetilde{\boldsymbol{x}}^\top\boldsymbol{\theta}\mid s,\widetilde{\boldsymbol{x}};\boldsymbol{\theta})G_{\widetilde{\boldsymbol{X}}\mid Y=s}(\mathrm{d}\widetilde{\boldsymbol{x}}).$$

For simplicity, we consider only the first part of $\tau^{\mathsf{H}}(\boldsymbol{z};\boldsymbol{\theta})$ and denote

$$\tau_1^{\mathsf{H}}(\boldsymbol{z};\boldsymbol{\theta})=\int_{-\infty}^{\boldsymbol{x}^\top\boldsymbol{\beta}}\int_{-\infty}^{y}g_0(t\mid s;\boldsymbol{\theta})G_Y(\mathrm{d}s)\,\mathrm{d}t.$$

After some simple calculations, we have

$$\frac{\partial\tau_1^{\mathsf{H}}(\boldsymbol{z};\boldsymbol{\theta})}{\partial\theta_i}=\widetilde{x}_i\left\{\int_{-\infty}^{y}g_0(\boldsymbol{x}^\top\boldsymbol{\beta}\mid s;\boldsymbol{\theta})G_Y(\mathrm{d}s)-\int_{-\infty}^{\boldsymbol{x}^\top\boldsymbol{\beta}}\int_{-\infty}^{y}g_{0,1}(t\mid s;\boldsymbol{\theta})G_Y(\mathrm{d}s)\,\mathrm{d}t\right\},$$

where

$$g_{0,1}(t\mid s;\boldsymbol{\theta})=\int f_0^{(1)}(t-\widetilde{\boldsymbol{x}}^\top\boldsymbol{\theta}\mid s,\widetilde{\boldsymbol{x}})G_{\widetilde{\boldsymbol{X}}\mid Y=s}(\mathrm{d}\widetilde{\boldsymbol{x}}).$$

Additionally, we have

$$\frac{\partial^3\tau_1^{\mathsf{H}}(\boldsymbol{z};\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j\partial\theta_k}=\widetilde{x}_i\widetilde{x}_j\widetilde{x}_k\underbrace{\left\{\int_{-\infty}^{y}g_{0,2}(\boldsymbol{x}^\top\boldsymbol{\beta}\mid s;\boldsymbol{\theta})G_Y(\mathrm{d}s)-\int_{-\infty}^{\boldsymbol{x}^\top\boldsymbol{\beta}}\int_{-\infty}^{y}g_{0,3}(t\mid s;\boldsymbol{\theta})G_Y(\mathrm{d}s)\,\mathrm{d}t\right\}}_{A_1(\boldsymbol{x},y;\boldsymbol{\theta})},$$

where

$$g_{0,m}(t\mid s;\boldsymbol{\theta})=\int f_0^{(m)}(t-\widetilde{\boldsymbol{x}}^\top\boldsymbol{\theta}\mid s,\widetilde{\boldsymbol{x}})G_{\widetilde{\boldsymbol{X}}\mid Y=s}(\mathrm{d}\widetilde{\boldsymbol{x}}),\quad m=2,3.$$

According to Condition 2, we know that $A_1(\boldsymbol{x},y;\boldsymbol{\theta})$ is uniformly upper bounded: $|A_1(\boldsymbol{x},y;\boldsymbol{\theta})|\le K$ for some absolute constant $K>0$. We could then similarly define $A_2(\boldsymbol{x},y;\boldsymbol{\theta})$ for the second part and write $A(\boldsymbol{x},y;\boldsymbol{\theta})=A_1(\boldsymbol{x},y;\boldsymbol{\theta})+A_2(\boldsymbol{x},y;\boldsymbol{\theta})$.
Step 2. For any $\boldsymbol{\gamma}\in\mathbb{S}^{p-1}$, we consider $\boldsymbol{\gamma}^\top\{\mathbf{V}^{\mathsf{H}}(\boldsymbol{\theta})-\mathbf{V}^{\mathsf{H}}\}\boldsymbol{\gamma}$. Expand $\boldsymbol{\gamma}^\top\{\mathbf{V}^{\mathsf{H}}(\boldsymbol{\theta})-\mathbf{V}^{\mathsf{H}}\}\boldsymbol{\gamma}$ about $\boldsymbol{\theta}_0$ to get

$$\boldsymbol{\gamma}^\top\{\mathbf{V}^{\mathsf{H}}(\boldsymbol{\theta})-\mathbf{V}^{\mathsf{H}}\}\boldsymbol{\gamma}=\sum_{i,j,k}\gamma_i\gamma_j(\theta_k-\theta_{0,k})\frac{\partial^3\mathbb{E}\tau^{\mathsf{H}}(\cdot;\boldsymbol{\theta}^*)}{\partial\theta_i\partial\theta_j\partial\theta_k}=\boldsymbol{\gamma}^\top\mathbb{E}\left\{A(\boldsymbol{X},Y;\boldsymbol{\theta})\widetilde{\boldsymbol{X}}^\top(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{X}}\right\}\boldsymbol{\gamma}.$$

Then,

$$\sup_{\boldsymbol{\gamma}\in\mathbb{S}^{p-1}}|\boldsymbol{\gamma}^\top\{\mathbf{V}^{\mathsf{H}}(\boldsymbol{\theta})-\mathbf{V}^{\mathsf{H}}\}\boldsymbol{\gamma}|\le 2K\left[\mathbb{E}\{\widetilde{\boldsymbol{X}}^\top(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\}^2\right]^{1/2}\left\{\mathbb{E}(\boldsymbol{\gamma}^\top\widetilde{\boldsymbol{X}})^4\right\}^{1/2}.$$

By Condition 1, we know that there exists an absolute constant C such that

$$\sup_{\boldsymbol{\gamma}\in\mathbb{S}^{p-1}}|\boldsymbol{\gamma}^\top\{\mathbf{V}^{\mathsf{H}}(\boldsymbol{\theta})-\mathbf{V}^{\mathsf{H}}\}\boldsymbol{\gamma}|\le KC\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le KCr_0.$$

Then, we can choose r_0 small enough such that $KCr_0\le c_{\min}/(11c_{\max})$. The first part of Assumption 3(iii) has been verified.
Next, we try to verify the second part of Assumption 3(iii). According to the results in *Step 1*, we expand $\mathbf{V}_{ij}^{\mathsf{H}}(\boldsymbol{\theta})-\mathbf{V}_{ij}^{\mathsf{H}}$ about $\boldsymbol{\theta}_0$ to get that

$$\sup_{i,j}|\mathbf{V}_{ij}^{\mathsf{H}}(\boldsymbol{\theta})-\mathbf{V}_{ij}^{\mathsf{H}}|\le cr,$$

where c depends only on the absolute constants K and C'. Then, by the relationship between different matrix norms, we have that

$$\|\mathbf{V}^{\mathrm{H}}(\boldsymbol{\theta}) - \mathbf{V}^{\mathrm{H}}\| \le \|\mathbf{V}^{\mathrm{H}}(\boldsymbol{\theta}) - \mathbf{V}^{\mathrm{H}}\|_1 \le p \sup_{i,j} |\mathbf{V}^{\mathrm{H}}_{ij}(\boldsymbol{\theta}) - \mathbf{V}^{\mathrm{H}}_{ij}(\boldsymbol{\theta})| \le cpr.$$

Finally,

$$\|\mathbf{I}_p - (\mathbf{V}^{\mathrm{H}})^{-1/2}\mathbf{V}^{\mathrm{H}}(\boldsymbol{\theta})(\mathbf{V}^{\mathrm{H}})^{-1/2}\| \le \|(\mathbf{V}^{\mathrm{H}})^{-1/2}\|\|\mathbf{V}^{\mathrm{H}}(\boldsymbol{\theta}) - \mathbf{V}^{\mathrm{H}}\|\|(\mathbf{V}^{\mathrm{H}})^{-1/2}\| \le \frac{cpr}{c_{\min}}.$$

This completes the verification of Assumption 3(iii).

(v) We first consider $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Since $\boldsymbol{X}$ is multivariate subgaussian by Condition 1, it holds that $\sup_{i=1,\ldots,p+1} \mathbb{E}|X_i|^2 \le c_0$. According to calculations in the proof of Theorem 4 in Sherman (1993), we have

$$\nabla_2 \tau^{\mathrm{H}}(\boldsymbol{Z}, \boldsymbol{\theta}_0) = \{\widetilde{\boldsymbol{X}} - \mathbb{E}(\widetilde{\boldsymbol{X}} \mid \boldsymbol{X}^\top \boldsymbol{\beta}_0)\}\{\widetilde{\boldsymbol{X}} - \mathbb{E}(\widetilde{\boldsymbol{X}} \mid \boldsymbol{X}^\top \boldsymbol{\beta}_0)\}^\top \lambda_2^{\mathrm{H}}(Y, \boldsymbol{X}^\top \boldsymbol{\beta}_0).$$

For any $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in \mathbb{S}^{p-1}$, Lemma A.5 implies that under Conditions 1 and 3, $\boldsymbol{\gamma}_1^\top\{\widetilde{\boldsymbol{X}} - \mathbb{E}(\widetilde{\boldsymbol{X}} \mid \boldsymbol{X}^\top \boldsymbol{\beta}_0)\}$ and $\boldsymbol{\gamma}_2^\top\{\widetilde{\boldsymbol{X}} - \mathbb{E}(\widetilde{\boldsymbol{X}} \mid \boldsymbol{X}^\top \boldsymbol{\beta}_0)\}\lambda_2^{\mathrm{H}}(Y, \boldsymbol{X}^\top \boldsymbol{\beta}_0)$ are both subgaussian with subgaussian norms $2c'$ and $2c'c''$, respectively. Because the product of two subgaussian random variables is subexponential, $\boldsymbol{\gamma}_1^\top \nabla_2 \tau(\boldsymbol{Z}, \boldsymbol{\theta}_0)\boldsymbol{\gamma}_2$ is subexponential with a subexponential norm that depends only on c' and c''. By the definition of subexponential variables and $\zeta^{\mathrm{H}}(\boldsymbol{z}; \boldsymbol{\theta}_0) = \tau^{\mathrm{H}}(\boldsymbol{z}; \boldsymbol{\theta}_0) - \mathbb{E}\{\tau^{\mathrm{H}}(\cdot; \boldsymbol{\theta}_0)\} = \tau^{\mathrm{H}}(\boldsymbol{z}; \boldsymbol{\theta}_0) - \mathbf{V}$, we have

$$\begin{aligned}
\mathbb{E}\exp\{\lambda \boldsymbol{\gamma}_1^\top \nabla_2 \zeta^{\mathrm{H}}(\cdot, \boldsymbol{\theta}_0)\boldsymbol{\gamma}_2\} &= \mathbb{E}\exp[\lambda \boldsymbol{\gamma}_1^\top\{\nabla_2 \tau^{\mathrm{H}}(\cdot, \boldsymbol{\theta}_0) - \mathbf{V}\}\boldsymbol{\gamma}_2] \\
&\le \exp[C_0\lambda^2 \|\boldsymbol{\gamma}_1^\top\{\nabla_2 \tau^{\mathrm{H}}(\boldsymbol{Z}, \boldsymbol{\theta}_0) - \mathbf{V}\}\boldsymbol{\gamma}_2\|_{\psi_1}^2] \\
&\le \exp\{4C_0\lambda^2 \|\boldsymbol{\gamma}_1^\top \nabla_2 \tau^{\mathrm{H}}(\boldsymbol{Z}, \boldsymbol{\theta}_0)\boldsymbol{\gamma}_2\|_{\psi_1}^2\} \\
&\le \exp(\nu_0^2\lambda^2/2), \quad \text{for } |\lambda| \le \ell_0,
\end{aligned} \tag{A.57}$$

where ν_0 and ℓ_0 are constants that depend on constants c_0, c', c''. This shows that Assumption 3(v) holds at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Note that there are several equivalent definitions for a generic zero-mean subexponential variable U. One of them is defined as follows: there is a constant $c_1 > 0$ such that $\mathbb{E}\exp(\lambda U)$ is bounded for all $|\lambda| \le c_1$. This definition implies that, for the subexponential variable $\boldsymbol{\gamma}_1^\top \nabla_2 \zeta^{\mathrm{H}}(\boldsymbol{Z}, \boldsymbol{\theta}_0)\gamma_2$, there is a constant $c_2 > 0$ such that $\mathbb{E}\exp\{\lambda \boldsymbol{\gamma}_1^\top \nabla_2 \zeta^{\mathrm{H}}(\cdot, \boldsymbol{\theta}_0)\boldsymbol{\gamma}_2\}$ is bounded for all $|\lambda| \le c_2$. Because $\mathbb{E}\exp\{\lambda \boldsymbol{\gamma}_1^\top \nabla_2 \zeta^{\mathrm{H}}(\cdot, \boldsymbol{\theta})\boldsymbol{\gamma}_2\}$ is a continuous function in $(\lambda, \boldsymbol{\theta}^\top) \in [-c_2, c_2] \otimes \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)$, and in addition that the domain of this function is a compact set, it then holds

$$\sup_{|\lambda| \le c_2} \sup_{\boldsymbol{\theta} \in \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)} \mathbb{E}\exp\{\lambda \boldsymbol{\gamma}_1^\top \nabla_2 \zeta^{\mathrm{H}}(\cdot, \boldsymbol{\theta})\boldsymbol{\gamma}_2\} < C.$$

Thus, $\boldsymbol{\gamma}_1^\top \nabla_2 \zeta^{\mathrm{H}}(\cdot, \boldsymbol{\theta})\boldsymbol{\gamma}_2$ is subexponential for any $\boldsymbol{\theta} \in \overline{\mathcal{B}}(\boldsymbol{\theta}_0, r)$. Similar to (A.57), we can establish the bound in Assumption 3(v).

This completes the proof. □

References

Abrevaya, J., Shin, Y., 2011. Rank estimation of partially linear index models. Econom. J. 14 (3), 409–437.

Bahadur, R., 1966. A note on quantiles in large samples. Ann. Math. Stat. 37 (3), 577–580.

Belloni, A., Chernozhukov, V., Chetverikov, D., Wei, Y., 2018. Uniformly valid post-regularization confidence regions for many functional parameters in Z-estimation framework. Ann. Statist. 46 (6B), 3643–3675.

Belloni, A., Chernozhukov, V., Kato, K., 2014. Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. Biometrika 102 (1), 77–94.

Caner, M., 2014. Near exogeneity and weak identification in generalized empirical likelihood estimators: Many moment asymptotics. J. Econometrics 182 (2), 247–268.

Cattaneo, M.D., Jansson, M., Newey, W.K., 2018a. Alternative asymptotics and the partially linear model with many regressors. Econom. Theory 34, 277–301.

Cattaneo, M.D., Jansson, M., Newey, W.K., 2018b. Inference in linear regression models with many covariates and heteroskedasticity. J. Amer. Statist. Assoc. 113 (523), 1350–1361.

Cavanagh, C., Sherman, R.P., 1998. Rank estimators for monotonic index models. J. Econometrics 84 (2), 351–381.

Chernozhukov, V., Chetverikov, D., Kato, K., 2017. Central limit theorems and bootstrap in high dimensions. Ann. Probab. 45 (4), 2309–2352.

Chernozhukov, V., Hansen, C., Spindler, M., 2015. Valid post-selection and post-regularization inference: An elementary, general approach. Annu. Rev. Econ. 7, 649–688.

Dudley, R.M., 1999. Uniform Central Limit Theorems. Cambridge University Press.

Fan, J., Liao, Y., Yao, J., 2015. Power enhancement in high-dimensional cross-sectional tests. Econometrica 83 (4), 1497–1541.

Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Statist. 42 (3), 1166–1202.

Han, A.K., 1987. Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. J. Econometrics 35 (2–3), 303–316.

Han, F., Ji, H., Ji, Z., Wang, H., 2017. A provable smoothing approach for high dimensional generalized regression with applications in genomics. Electron. J. Stat. 11 (2), 4347–4403.

Han, C., Phillips, P.C., 2006. GMM with many moment conditions. Econometrica 74 (1), 147–192.

He, X., Shao, Q.-M., 1996. A general bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. Ann. Statist. 24 (6), 2608–2630.
He, X., Shao, Q.-M., 2000. On parameters of increasing dimensions. J. Multivariate Anal. 73 (1), 120–135.
Hoeffding, W., 1948. A class of statistics with asymptotically normal distribution. Ann. Math. Stat. 19 (3), 293–325.
Honoré, B.E., Powell, J., 2005. Pairwise difference estimators for nonlinear models. In: Andrews, D.W.K., Stock, J.H. (Eds.), Identification and Inference in Econometric Models. Essays in Honor of Thomas Rothenberg. Cambridge University Press, pp. 520–553.
Huber, P.J., 1967. The behavior of maximum likelihood estimates under nonstandard conditions. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, CA, pp. 221–233.
Huber, P.J., 1973. Robust regression: Asymptotics, conjectures and Monte Carlo. Ann. Statist. 1 (5), 799–821.
Javanmard, A., Montanari, A., 2018. De-biasing the lasso: Optimal sample size for Gaussian designs. Ann. Statist. 46 (6A), 2593–2622.
Jurečková, J., Sen, P.K., Picek, J., 2012. Methodology in Robust and Nonparametric Statistics. CRC Press.
Khan, S., Tamer, E., 2007. Partial rank estimation of duration models with general forms of censoring. J. Econometrics 136 (1), 251–280.
Kiefer, J., 1967. On Bahadur's representation of sample quantiles. Ann. Math. Stat. 38 (5), 1323–1342.
Kosorok, M.R., 2007. Introduction to Empirical Processes and Semiparametric Inference. Springer.
Lee, J.D., Sun, D.L., Sun, Y., Taylor, J.E., 2016. Exact post-selection inference, with application to the lasso. Ann. Statist. 44 (3), 907–927.
Lei, L., Bickel, P.J., Karoui, N.E., 2018. Asymptotics for high dimensional regression m-estimates: Fixed design results. Probab. Theory Related Fields 172 (3–4), 983—1079.
Mammen, E., 1989. Asymptotics with increasing dimension for robust regression with applications to the bootstrap. Ann. Statist. 17 (1), 382–400.
Mammen, E., 1993. Bootstrap and wild bootstrap for high dimensional linear models. Ann. Statist. 21 (1), 255–285.
Negahban, S.N., Ravikumar, P., Wainwright, M.J., Yu, B., 2012. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. Statist. Sci. 27 (4), 538–557.
Newey, W.K., Windmeijer, F., 2009. Generalized method of moments with many weak moment conditions. Econometrica 77 (3), 687–719.
Nolan, D., Pollard, D., 1987. U-processes: rates of convergence. Ann. Statist. 15 (2), 780–799.
Pakes, A., Pollard, D., 1989. Simulation and the asymptotics of optimization estimators. Econometrica 57 (5), 1027–1057.
de la Pena, V., Giné, E., 2012. Decoupling: From Dependence to Independence. Springer, New York.
Pollard, D., 1984. Convergence of Stochastic Processes. Springer.
Portnoy, S., 1984. Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. Ann. Statist. 12 (4), 1298–1309.
Portnoy, S., 1985. Asymptotic behavior of M estimators of p regression parameters when p^2/n is large; II. Normal approximation. Ann. Statist. 13 (4), 1403–1417.
Portnoy, S., 1988. Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. Ann. Statist. 16 (1), 356–366.
Sherman, R.P., 1993. The limiting distribution of the maximum rank correlation estimator. Econometrica 61 (1), 123–137.
Sherman, R.P., 1994. Maximal inequalities for degenerate U-processes with applications to optimization estimators. Ann. Statist. 22 (1), 439–459.
Spokoiny, V., 2012a. Parametric estimation. Finite sample theory. Ann. Statist. 40 (6), 2877–2909.
Spokoiny, V., 2012b. Supplement to "parametric estimation. Finite sample theory". Ann. Statist..
Spokoiny, V., 2013. Bernstein-von Mises Theorem for growing parameter dimension. arXiv preprint arXiv:1302.3430.
Subbotin, V.Y., 2008. Essays on the Econometric Theory of Rank Regressions (PhD thesis). Northwestern University.
van der Vaart, A., Wellner, J., 1996. Weak Convergence and Empirical Processes. Springer.
Wang, H., 2007. A note on iterative marginal optimization: a simple algorithm for maximum rank correlation estimation. Comput. Statist. Data Anal. 51 (6), 2803–2812.
Yu, B., 1997. Assouad, Fano, and Le Cam. In: Festschrift for Lucien Le Cam. Springer, New York, pp. 423–435.
Zhang, C.-H., Zhang, S.S., 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. J. R. Stat. Soc. Ser. B Stat. Methodol. 76 (1), 217–242.