UNIFORMLY VALID POST-REGULARIZATION CONFIDENCE REGIONS FOR MANY FUNCTIONAL PARAMETERS IN Z-ESTIMATION FRAMEWORK

A. BELLONI, V. CHERNOZHUKOV, D. CHETVERIKOV, AND Y. WEI

Abstract. In this paper we develop procedures to construct simultaneous confidence bands for $\tilde{p}$ potentially infinite-dimensional parameters after model selection for general moment condition models where $\tilde{p}$ is potentially much larger than the sample size of available data, n. This allows us to cover settings with functional response data where each of the $\tilde{p}$ parameters is a function. The procedure is based on the construction of score functions that satisfy certain orthogonality condition. The proposed simultaneous confidence bands rely on uniform central limit theorems for high-dimensional vectors (and not on Donsker arguments as we allow for $\tilde{p} \gg n$). To construct the bands, we employ a multiplier bootstrap procedure which is computationally efficient as it only involves resampling the estimated score functions (and does not require resolving the high-dimensional optimization problems). We formally apply the general theory to inference on regression coefficient process in the distribution regression model with a logistic link, where two implementations are analyzed in detail. Simulations and an application to real data are provided to help illustrate the applicability of the results.

1. Introduction

High-dimensional models have become increasingly popular in the last two decades. Much research has been conducted on estimation of these models. However, inference about parameters in these models is much less understood, although the literature on inference is growing quickly; see the list of references below. In particular, despite its practical relevance, there is almost no research on the problem of construction of simultaneous confidence bands on many target parameters in these models (with one exception being [9]). In this paper we provide a solution to this problem by constructing simultaneous confidence bands for parameters in a very general framework of moment condition models, allowing for many functional parameters, where each parameter itself can be an infinite-dimensional object, and the number of parameters can be much larger than the sample size of available data. As a substantive application, we apply the results to provide simultaneous confidence bands for parameters in a functional logistic regression model, which includes the so called distributional regression and conditional transformation models as special cases. (In particular, this contribution goes much beyond [9], which considers the special case of many scalar parameters).

Date: December 23, 2015.
Key words and phrases. inference after model selection, moment condition models with a continuum of target parameters, Lasso and Post-Lasso with functional response data.

Specifically, we consider the problem of estimating the set of parameters $(\theta_{uj})_{u\in\mathcal{U},j\in[\tilde{p}]}$ in the moment condition model

$$\mathrm{E}_P[m_{uj}(W,\theta_{uj},\eta_{uj})]=0,\quad u\in\mathcal{U},\ j\in[\tilde{p}],\tag{1.1}$$

where W is a random element that takes values in a measurable space $(\mathcal{W},\mathcal{A}_{\mathcal{W}})$ according to the probability measure P, and $\mathcal{U}\subset\mathbb{R}^{d_u}$ and $[\tilde{p}]:=\{1,\dots,\tilde{p}\}$ are sets of indices. For each $u\in\mathcal{U}$ and $j\in[\tilde{p}]$, m_{uj} is a known score function, θ_{uj} is a scalar parameter of interest, and η_{uj} is a potentially high-dimensional nuisance parameter. Assuming that a random sample of size n, $(W_i)_{i=1}^n$, from the distribution of W is available together with suitable estimators $\widehat{\eta}_{uj}$ of η_{uj} for $u\in\mathcal{U}$ and $j\in[\tilde{p}]$, we aim to construct simultaneous confidence bands for $(\theta_{uj})_{u\in\mathcal{U},j\in[\tilde{p}]}$ that are valid uniformly over a large class of probability measures P, say $\mathcal{P}$. Specifically, for each $u\in\mathcal{U}$ and $j\in[\tilde{p}]$, we construct an appropriate estimator $\check{\theta}_{uj}$ of θ_{uj} along with an estimator of the standard deviation of $\sqrt{n}(\check{\theta}_{uj}-\theta_{uj})$, $\widehat{\sigma}_{uj}$, such that

$$\sup_{P\in\mathcal{P}}\left|\mathrm{P}_P\left(\check{\theta}_{uj}-\frac{c_\alpha\widehat{\sigma}_{uj}}{\sqrt{n}}\leqslant\theta_{uj}\leqslant\check{\theta}_{uj}+\frac{c_\alpha\widehat{\sigma}_{uj}}{\sqrt{n}},\ \text{for all}\ \ u\in\mathcal{U}\ \text{and}\ j\in[\tilde{p}]\right)-(1-\alpha)\right|=o(1),\tag{1.2}$$

where $\alpha\in(0,1)$ and c_α is an appropriate critical value, which we choose to construct using a multiplier bootstrap method. The left- and the right-hand sides of the inequalities inside the probability statement (1.2) then can be used as bounds in simultaneous confidence bands for θ_{uj}'s. In this paper, we are particularly interested in the case when $\tilde{p}$ is potentially much larger than n and $\mathcal{U}$ is an uncountable subset of $\mathbb{R}^{d_u}$, so that for each $j\in[\tilde{p}]$, $(\theta_{uj})_{u\in\mathcal{U}}$ is an infinite-dimensional (that is, functional) parameter.

This general framework covers a broad variety of applications. For example, consider a finite-dimensional generalized linear model for a response variable Y and covariates D and X given by

$$\mathrm{E}_P[Y\mid D,X]=\Lambda(D\theta_0+X'\beta_0),\tag{1.3}$$

where D is a scalar, X is a p-vector (p is potentially large), $\Lambda\colon\mathbb{R}\to\mathbb{R}$ is a known link function, θ_0 is a parameter of interest, and β_0 is a nuisance parameter. This model is a particular case of our general framework with the moment condition

$$\mathrm{E}_P\Big[\{Y-\Lambda(D\theta_0+X'\beta_0)\}D\Big]=0,\tag{1.4}$$

where $\mathcal{U}$ is a singleton, $\tilde{p}=1$, $\theta_{uj}=\theta_0$ and $\eta_{uj}=\beta_0$ for $u\in\mathcal{U}$ and $j\in[\tilde{p}]$, and $W=(Y,D,X)$. Another example fitting into our framework is a logistic regression model with functional response data

$$\mathrm{E}_P[Y_u\mid D,X]=\Lambda(D'\theta_u+X'\beta_u),\quad u\in\mathcal{U},\tag{1.5}$$

where D is now a $\tilde{p}$-vector with $\tilde{p}$ being potentially much larger than n, $\theta_u=(\theta_{u1},\dots,\theta_{u\tilde{p}})'$ is a $\tilde{p}$-vector of parameters of interest, $\mathcal{U}=[0,1]$ is a set of indices, $Y_u=1\{Y\leqslant(1-u)\underline{y}+u\bar{y}\}$ for some constants $\underline{y}\leqslant\bar{y}$ and all $u\in\mathcal{U}$, and $\Lambda\colon\mathbb{R}\to\mathbb{R}$ is the logistic link function defined by $\Lambda(t)=\exp(t)/\{1+\exp(t)\}$ for all $t\in\mathbb{R}$. Here we have $\widetilde{p}$ infinite-dimensional parameters $(\theta_{uj})_{u\in\mathcal{U}}$,

$j = 1, \ldots, \tilde{p}$. This example is important because it demonstrates that our methods can be used for inference about the whole distribution of the response variable Y given D and X in a high-dimensional setting, and not only about some particular features of it such as mean or median. This model is called a distribution regression model in [21] and a conditional transformation model in [23], who argue that the model provides a rich class of models for conditional distributions, and offers a useful generalization of traditional proportional hazard models as well as a useful alternative to quantile regression. We develop inference methods for many functional parameters of this model in detail in Section 3.

In the presence of high-dimensional nuisance parameters, construction of valid confidence bands is delicate. High-dimensionality requires relying upon regularization that leads to lack of asymptotic linearization of the estimators. This lack of asymptotic linearization in turn typically translates into severe distortions in coverage probability of the confidence bands constructed by traditional techniques that are based on perfect model selection; see [29], [30], [31], [39].

To deal with this problem, we consider moment conditions

$$\mathrm{E}_P[\psi_{uj}(W, \theta_{uj}, \eta_{uj})] = 0, \quad u \in \mathcal{U}, \ j \in [\tilde{p}] \tag{1.6}$$

based on score functions ψ_{uj} with an additional "near orthogonality" property that makes them immune to first-order changes in the value of the nuisance parameter, namely

$$\partial_r \Big\{ \mathrm{E}_P \Big[\psi_{uj}(W, \theta_{uj}, \eta_{uj} + r\tilde{\eta}\}) \Big] \Big\} \Big|_{r=0} \approx 0, \quad u \in \mathcal{U}, \ j \in [\tilde{p}], \tag{1.7}$$

for all $\tilde{\eta}$ in an appropriate set where ∂_r denotes the derivative with respect to r. For example, in the finite-dimensional generalized linear model (1.3), a score function with such a near orthogonality property is

$$\psi(W, \theta, \eta_0) = \Big\{ Y - \Lambda(D\theta + X'\beta_0) \Big\} (D - X'\gamma_0),$$

where the nuisance parameter is $\eta_0 = (\beta_0', \gamma_0')'$, and $\gamma_0 \in \arg\min_\gamma \mathrm{E}_P[f_0^2\{D - X'\gamma\}^2]$ for $f_0^2 = \Lambda'(D\theta_0 + X'\beta_0)$. It satisfies the moment condition (1.6) and also satisfies the near orthogonality condition (1.7) since

$$\begin{aligned} \partial_\beta \Big\{ \mathrm{E}_P[\psi(W, \theta_0, \beta, \gamma_0)] \Big\} \Big|_{\beta=\beta_0} &= -\mathrm{E}_P \Big[f_0^2 \{D - X'\gamma_0\} X \Big] = 0, \\ \partial_\gamma \Big\{ \mathrm{E}_P[\psi(W, \theta_0, \beta_0, \gamma)] \Big\} \Big|_{\gamma=\gamma_0} &= -\mathrm{E}_P \Big[\{Y - \Lambda(D\theta_0 + X'\beta_0)\} X \Big] = 0, \end{aligned}$$

where the first line holds by the definition of γ_0, and the second by (1.3). Because of this orthogonality property, we can exploit the moment conditions based on these new score functions to construct a regular, $\sqrt{n}$-consistent, estimator of θ_0 even if non-regular, regularized or post-regularized, estimators of β_0 and γ_0 are used to cope with high-dimensionality. Then we can construct confidence bands for θ_0 based on this regular estimator.

Our general approach, which is developed in Section 2, can be described as follows. First, we transform the moment conditions (1.1) into those based on the score functions (1.6) with the near orthogonality property (1.7), and use these new moment conditions to construct an estimator $\check{\theta}_{uj}$ of θ_{uj} for all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$. Second, under appropriate regularity conditions, we establish a Bahadur representation for $\check{\theta}_{uj}$'s. Third, employing the Bahadur representation, we are able to derive a suitable Gaussian approximation for the distribution of $\check{\theta}_{uj}$'s. Importantly, the Gaussian approximation is possible even if both $\tilde{p}$ and the dimension of the index set $\mathcal{U}$, d_u, are allowed to grow with n, and $\tilde{p}$ asymptotically remains much larger than n. Finally, from the Gaussian approximation, we construct simultaneous confidence bands using a multiplier bootstrap method. This approach makes use of the results on high-dimensional central limit and bootstrap theorems established in [15], [17], [18], [19], and [20].

Although regularity conditions underlying our approach can be verified for many models defined by moment conditions, for illustration purposes, we explicitly verify these conditions for the logistic regression model with functional response data (1.5) in Section 3. We also examine the performance of the proposed procedures in a Monte Carlo simulation study and provide an example based on real data in Section 5. In addition, in the Supplementary Material, we discuss the construction of simultaneous confidence bands based on a double-selection estimator. This estimator does not require to explicitly construct the new score functions but nonetheless is first-order equivalent to the estimator based on such functions.

We also develop new results for ℓ_1-penalized M-estimators in Section 4 to handle functional data and criterion functions that depend on nuisance functions for which only estimates are available (for brevity of the paper, generic results are deferred to Appendix I of the Supplementary Material, and Section 4 only contains results that are relevant for the logistic regression model studied in Section 3). Specifically, we develop a method to select penalty parameters for these estimators and extend the existing theory to cover functional data to achieve rates of convergence and sparsity guarantees that hold uniformly over $u \in \mathcal{U}$. The ability to allow both for functional data and for nuisance functions is crucial in the implementation and in theoretical analysis of the methods proposed in this paper.

Orthogonality conditions like that in (1.7) have played an important role in statistics and econometrics. In low-dimensional settings, a similar condition was used by Neyman in [36] while in semiparametric models the orthogonality conditions were used in [34], [1], [35], [40] and [32]. In high-dimensional settings, [5] and [2] were the first to use the orthogonality condition (1.7) in a linear instrumental variables model with many instruments. Related ideas have also been used in the literature to construct confidence bands in high-dimensional linear models, generalized linear models, and other non-linear models; see [6], [45], [7], [42], [12], [25], [24], [9], [8], [11], [46], [37]. We contribute to this quickly growing literature by providing procedures to construct *simultaneous* confidence bands for *many infinite-dimensional* parameters identified by moment conditions.

2. Confidence Regions for Function-Valued Parameters Based on Moment Conditions

In this section, we formally introduce the model and state our results under high-level conditions. In the next section, we will apply these results to construct simultaneous confidence bands for many infinite-dimensional parameters in the logistic regression model with functional response data.

We are interested in a set of parameters $(\theta_{uj})_{u\in\mathcal{U},j\in[\tilde{p}]}$ where for each $u\in\mathcal{U}\subset\mathbb{R}^{d_u}$ and $j\in[\tilde{p}]=\{1,\ldots,\tilde{p}\}$, we have $\theta_{uj}\in\Theta_{uj}$, a convex subset of $\mathbb{R}$. Here $\mathcal{U}$ is possibly an uncountable set of indices, and $\tilde{p}$ is potentially large. We assume that for each $u\in\mathcal{U}$ and $j\in[\tilde{p}]$, the parameter θ_{uj} satisfies the moment condition

$$\mathrm{E}_P[\psi_{uj}(W,\theta_{uj},\eta_{uj})]=0, \tag{2.1}$$

where W is a random element that takes values in a measurable space $(\mathcal{W},\mathcal{A}_{\mathcal{W}})$, with law determined by a probability measure $P\in\mathcal{P}_n$, η_{uj} is a nuisance parameter with $\eta_{uj}\in T_{uj}$, a convex set equipped with a norm $\|\cdot\|_e$, and the score function $\psi_{uj}\colon\mathcal{W}\times\Theta_{uj}\times T_{uj}\to\mathbb{R}$ is a measurable map (where we equip Θ_{uj} and T_{uj} with their Borel σ-fields). Here $\mathcal{P}_n$ is some set of probability measures on $(\mathcal{W},\mathcal{A}_{\mathcal{W}})$.

We focus on the estimation of $(\theta_{uj})_{u\in\mathcal{U},j\in[\tilde{p}]}$ using a random sample $(W_i)_{i=1}^n$ from the distribution of W. We assume that for each $u\in\mathcal{U}$ and $j\in[\tilde{p}]$, the nuisance parameter η_{uj} can be estimated by $\widehat{\eta}_{uj}$ using the same data $(W_i)_{i=1}^n$. In the next section, we discuss examples where $\widehat{\eta}_{uj}$'s are based on Lasso or Post-Lasso methods (although other modern regularization and post-regularization methods can be applied). For each $u\in\mathcal{U}$ and $j\in[\tilde{p}]$, we construct the estimator $\check{\theta}_{uj}$ of θ_{uj} as an approximate ϵ_n-solution in Θ_{uj} to a sample analog of the moment condition (2.1), that is,

$$\sup_{u\in\mathcal{U},j\in[\tilde{p}]}\left\{\left|\mathbb{E}_n[\psi_{uj}(W,\check{\theta}_{uj},\widehat{\eta}_{uj})]\right|-\inf_{\theta\in\Theta_{uj}}\left|\mathbb{E}_n[\psi_{uj}(W,\theta,\widehat{\eta}_{uj})]\right|\right\}\leqslant\epsilon_n=o(\delta_n n^{-1/2}), \tag{2.2}$$

where $(\delta_n)_{n\geqslant 1}$ is some sequence of positive constants converging to zero.

Let C_0 be a strictly positive (and finite) constant, and for each $u\in\mathcal{U}$ and $j\in[\tilde{p}]$, let $\mathcal{T}_{uj}$ be some subset of T_{uj}, whose properties are specified below in assumptions. As discussed before, we rely on the following near orthogonality condition:

Definition 2.1 (Near orthogonality condition)**.** *For each $u\in\mathcal{U}$ and $j\in[\tilde{p}]$, we say that ψ_{uj} obeys the near orthogonality condition with respect to $\mathcal{T}_{uj}\subset T_{uj}$ if the following conditions hold: The Gateaux derivative map*

$$\mathrm{D}_{u,j,\bar{r}}[\eta-\eta_{uj}]:=\partial_r\left\{\mathrm{E}_P\Big[\psi_{uj}(W,\theta_{uj},\eta_{uj}+r(\eta-\eta_{uj}))\Big]\right\}\Big|_{r=\bar{r}}$$

exists for all $\bar{r}\in[0,1)$ and $\eta\in\mathcal{T}_{uj}$ and (nearly) vanishes at $\bar{r}=0$, namely,

$$\left|\mathrm{D}_{u,j,0}[\eta-\eta_{uj}]\right|\leqslant C_0\delta_n n^{-1/2}, \tag{2.3}$$

for all $\eta\in\mathcal{T}_{uj}$. ∎

If the original score functions m_{uj} do not satisfy this near orthogonality condition, we have to transform them into score functions ψ_{uj} that satisfy this condition. At the end of this section, we describe two methods to obtain score functions ψ_{uj}. Together these methods cover a wide variety of applications.

Let ω and c_0 be some strictly positive (and finite) constants, and let $n_0 \geqslant 3$ be some positive integer. Also, let $(B_{1n})_{n\geqslant 1}$ and $(B_{2n})_{n\geqslant 1}$ be some sequences of positive constants, possibly growing to infinity, where $B_{1n} \geqslant 1$ for all $n \geqslant 1$. Denote

$$\mathrm{u_n} := \mathrm{E}_P\left[\sup_{u\in\mathcal{U},j\in[\tilde{p}]}\left|\sqrt{n}\mathbb{E}_n[\psi_{uj}(W,\theta_{uj},\eta_{uj})]\right|\right], \quad J_{uj} := \partial_\theta\Big\{\mathrm{E}_P[\psi_{uj}(W,\theta,\eta_{uj})]\Big\}\Big|_{\theta=\theta_{uj}}. \tag{2.4}$$

The quantity $\mathrm{u_n}$ measures how rich the process $\{\psi_{uj}(\cdot,\theta_{uj},\eta_{uj})\colon u\in\mathcal{U}, j\in[\tilde{p}]\}$ is. In many applications, it satisfies $\mathrm{u_n} \leqslant C(1+d_u+\log\tilde{p})^{1/2}$ for some constant C. The quantity J_{uj} measures the degree of identifiability of θ_{uj} by the moment condition (2.1). In many applications, it is bounded in absolute value from above and away from zero.

We are now ready to state our main regularity conditions.

Assumption 2.1 (Moment condition problem)**.** *For all $n \geqslant n_0$, $P \in \mathcal{P}_n$, $u \in \mathcal{U}$, and $j \in [\tilde{p}]$, the following conditions hold: (i) The true parameter value θ_{uj} obeys (2.1), and Θ_{uj} contains a ball of radius $C_0 n^{-1/2}\mathrm{u_n}\log n$ centered at θ_{uj}. (ii) The map $(\theta,\eta)\mapsto \mathrm{E}_P[\psi_{uj}(W,\theta,\eta)]$ is twice continuously Gateaux-differentiable on $\Theta_{uj}\times\mathcal{T}_{uj}$. (iii) The moment function ψ_{uj} obeys the near orthogonality condition given in Definition 2.1 for the set $\mathcal{T}_{uj}\subset T_{uj}$. (iv) For all $\theta\in\Theta_{uj}$, $|\mathrm{E}_P[\psi_{uj}(W,\theta,\eta_{uj})]| \geqslant 2^{-1}|J_{uj}(\theta-\theta_{uj})|\wedge c_0$, where J_{uj} satisfies $c_0 \leqslant |J_{uj}| \leqslant C_0$. (v) For all $r\in[0,1)$, $\theta\in\Theta_{uj}$, and $\eta\in\mathcal{T}_{uj}$,*

(a) $\mathrm{E}_P[(\psi_{uj}(W,\theta,\eta)-\psi_{uj}(W,\theta_{uj},\eta_{uj}))^2] \leqslant C_0(|\theta-\theta_{uj}|\vee\|\eta-\eta_{uj}\|_e)^\omega$,
(b) $|\partial_r \mathrm{E}_P\left[\psi_{uj}(W,\theta,\eta_{uj}+r(\eta-\eta_{uj}))\right]| \leqslant B_{1n}\|\eta-\eta_{uj}\|_e$,
(c) $|\partial_r^2 \mathrm{E}_P[\psi_{uj}(W,\theta_{uj}+r(\theta-\theta_{uj}),\eta_{uj}+r(\eta-\eta_{uj}))]| \leqslant B_{2n}(|\theta-\theta_{uj}|^2\vee\|\eta-\eta_{uj}\|_e^2)$.

Assumption 2.1 is mild and standard in moment condition problems. Assumption 2.1(i) requires θ_{uj} to be sufficiently separated from the boundary of Θ_{uj}. Assumption 2.1(iii) is discussed above. Assumption 2.1(iv) implies sufficient identifiability of θ_{uj}. Assumptions 2.1(ii,v) are smoothness conditions. Assumption 2.1(ii) is rather weak because it only requires differentiability of the function $(\theta,\eta)\mapsto \mathrm{E}_P[\psi_{uj}(W,\theta,\eta)]$ and does not require differentiability of the function $(\theta,\eta)\mapsto\psi_{uj}(W,\theta,\eta)$.

Next, we state conditions related to the estimators $\widehat{\eta}_{uj}$, $u\in\mathcal{U}$ and $j\in[\tilde{p}]$. Let $(\Delta_n)_{n\geqslant 1}$ and $(\tau_n)_{n\geqslant 1}$ be some sequences of positive constants converging to zero. Also, let $(a_n)_{n\geqslant 1}$, $(v_n)_{n\geqslant 1}$, and $(K_n)_{n\geqslant 1}$ be some sequences of positive constants, possibly growing to infinity, where $a_n \geqslant n\vee K_n$ and $v_n\geqslant 1$ for all $n\geqslant 1$. Finally, let $q\geqslant 2$ be some constant.

Assumption 2.2 (Estimation of nuisance parameters)**.** *For all $n\geqslant n_0$ and $P\in\mathcal{P}_n$, the following conditions hold: (i) With probability at least $1-\Delta_n$, we have $\widehat{\eta}_{uj}\in\mathcal{T}_{uj}$ for all $u\in\mathcal{U}$ and $j\in[\tilde{p}]$. (ii) For all $u\in\mathcal{U}$, $j\in[\tilde{p}]$, and $\eta\in\mathcal{T}_{uj}$, $\|\eta-\eta_{uj}\|_e\leqslant\tau_n$. (iii) For all $u\in\mathcal{U}$ and $j\in[\tilde{p}]$, we have*

$\eta_{uj} \in \mathcal{T}_{uj}$. *(iv) The function class* $\mathcal{F}_1 = \{\psi_{uj}(\cdot,\theta,\eta) \colon u \in \mathcal{U}, j \in [\tilde p], \theta \in \Theta_{uj}, \eta \in \mathcal{T}_{uj}\}$ *is suitably measurable and its uniform entropy numbers obey*

$$\sup_Q \log N(\epsilon \|F_1\|_{Q,2}, \mathcal{F}_1, \|\cdot\|_{Q,2}) \leqslant v_n \log(a_n/\epsilon), \quad \text{for all } 0 < \epsilon \leqslant 1 \tag{2.5}$$

where F_1 *is a measurable envelope for* $\mathcal{F}_1$ *that satisfies* $\|F_1\|_{P,q} \leqslant K_n$. *(v) For all* $f \in \mathcal{F}_1$, *we have* $c_0 \leqslant \|f\|_{P,2} \leqslant C_0$. *(vi) The complexity characteristics* a_n *and* v_n *satisfy*

(a) $(v_n \log a_n/n)^{1/2} \leqslant C_0 \tau_n$,
(b) $(B_{1n}\tau_n + \mathrm{u_n} \log n/\sqrt{n})^{\omega/2}(v_n \log a_n)^{1/2} + n^{-1/2+1/q} v_n K_n \log a_n \leqslant C_0 \delta_n$,
(c) $n^{1/2} B_{1n}^2 B_{2n} \tau_n^2 \leqslant C_0 \delta_n$.

Assumption 2.2 provides sufficient conditions for the estimation of the nuisance parameters $(\eta_{uj})_{u\in\mathcal{U}, j\in[\tilde p]}$. It shows that the choice of the sets $\mathcal{T}_{uj}$ is delicate: setting $\mathcal{T}_{uj}$ large, on the one hand, makes it easy to satisfy Assumption 2.2(i) but, on the other hand, yields large values of a_n and v_n making it difficult to satisfy Assumption 2.2(vi). Suitable measurability of $\mathcal{F}_1$, required in Assumption 2.2(iv), is a mild condition that is satisfied in most practical cases; see Appendix A for clarifications. The index v_n captures the complexity of the class of functions $\mathcal{F}_1$ and typically grows with d_u, $\tilde p$, and $\mathcal{T}_{uj}$'s. In particular, in the case of approximately sparse models, like a logistic regression model with functional response data studied in Section 3, v_n can be typically set up-to a constant as the sum of the dimension of the approximating model, the dimension of the selected model, and the dimension of $\mathcal{U}$. However, we note that our conditions potentially cover other frameworks, where assumptions other than approximate sparsity are used to make the estimation problem manageable.

We stress that the class $\mathcal{F}_1$ does not need to be Donsker because its uniform entropy numbers are allowed to increase with n. This is important because allowing for non-Donsker classes is necessary to deal with high-dimensional nuisance parameters. Note also that our conditions are very different from the conditions imposed in various settings with nonparametrically estimated nuisance functions; see, e.g., [44], [43], and [27].

The following theorem is our first main result in this paper:

Theorem 2.1 (Uniform Bahadur representation)**.** *Under Assumptions 2.1 and 2.2, for an estimator* $(\check\theta_{uj})_{u\in\mathcal{U}, j\in[\tilde p]}$ *that obeys equation (2.2), we have*

$$\sqrt{n}\sigma_{uj}^{-1}(\check\theta_{uj} - \theta_{uj}) = \mathbb{G}_n \bar\psi_{uj} + O_P(\delta_n) \text{ in } \ell^\infty(\mathcal{U} \times [\tilde p])$$

uniformly over $P \in \mathcal{P}_n$, *where* $\bar\psi_{uj}(\cdot) := -\sigma_{uj}^{-1} J_{uj}^{-1} \psi_{uj}(\cdot, \theta_{uj}, \eta_{uj})$ *and* $\sigma_{uj}^2 := J_{uj}^{-2} \mathrm{E}_P[\psi_{uj}^2(W, \theta_{uj}, \eta_{uj})]$.

The uniform Bahadur representation derived in Theorem 2.1 is useful for the construction of simultaneous confidence bands for $(\theta_{uj})_{u\in\mathcal{U}, j\in[\tilde p]}$ as in (1.2). For this purpose, we apply new high-dimensional central limit and bootstrap theorems that have been recently developed in a sequence of papers [15], [17], [18], [19], and [20]. To apply these theorems, we make use of the following regularity condition.

Let $(\bar{\delta}_n)_{n\geqslant 1}$ be a sequence of positive constants converging to zero. Also, let $(\varrho_n)_{n\geqslant 1}$, $(\bar{\varrho}_n)_{n\geqslant 1}$, $(A_n)_{n\geqslant 1}$, $(\bar{A}_n)_{n\geqslant 1}$, and $(L_n)_{n\geqslant 1}$ be some sequences of positive constants, possibly growing to infinity, where $\varrho_n \geqslant 1$, $A_n \geqslant n$, and $\bar{A}_n \geqslant n$ for all $n \geqslant 1$. In addition, from now on, we assume that $q > 4$. Denote by $\widehat{\psi}_{uj}(\cdot) := -\widehat{\sigma}_{uj}^{-1}\widehat{J}_{uj}^{-1}\psi_{uj}(\cdot, \check{\theta}_{uj}, \widehat{\eta}_{uj})$ an estimator of $\bar{\psi}_{uj}(\cdot)$, with $\widehat{J}_{uj}$ and $\widehat{\sigma}_{uj}$ being suitable estimators of J_{uj} and σ_{uj}.

Assumption 2.3 (Score regularity)**.** *For all $n \geqslant n_0$ and $P \in \mathcal{P}_n$, the following conditions hold: (i) The function class $\mathcal{F}_0 = \{\bar{\psi}_{uj}(\cdot)\colon u \in \mathcal{U}, j \in [\tilde{p}]\}$ is suitably measurable and its uniform entropy numbers obey*

$$\sup_Q \log N(\epsilon\|F_0\|_{Q,2}, \mathcal{F}_0, \|\cdot\|_{Q,2}) \leqslant \varrho_n \log(A_n/\epsilon), \quad \textit{for all } 0 < \epsilon \leqslant 1,$$

where F_0 is a measurable envelope for $\mathcal{F}_0$ that satisfies $\|F_0\|_{P,q} \leqslant L_n$. (ii) For all $f \in \mathcal{F}_0$ and $k = 3, 4$, we have $\mathrm{E}_P[|f(W)|^k] \leqslant C_0 L_n^{k-2}$. (iii) The function class $\widehat{\mathcal{F}}_0 = \{\bar{\psi}_{uj}(\cdot) - \widehat{\psi}_{uj}(\cdot)\colon u \in \mathcal{U}, j \in [\tilde{p}]\}$ satisfies with probability $1 - \Delta_n$: $\log N(\epsilon, \widehat{\mathcal{F}}_0, \|\cdot\|_{\mathbb{P}_n,2}) \leqslant \bar{\varrho}_n \log(\bar{A}_n/\epsilon)$ for all $0 < \epsilon \leqslant 1$ and $\|f\|_{\mathbb{P}_n,2} \leqslant \bar{\delta}_n$ for all $f \in \widehat{\mathcal{F}}_0$.

This assumption is technical, and its verification in applications is rather standard. For the Gaussian approximation result below, we actually only need the first and the second part of this assumption. The third part will be needed for establishing validity of the simultaneous confidence bands based on the multiplier bootstrap procedure.

Next, let $(\mathcal{N}_{uj})_{u\in\mathcal{U}, j\in[\tilde{p}]}$ denote a tight zero-mean Gaussian process indexed by $\mathcal{U} \times [\tilde{p}]$ with covariance operator given by $\mathrm{E}_P[\bar{\psi}_{uj}(W)\bar{\psi}_{u'j'}(W)]$ for $u, u' \in \mathcal{U}$ and $j, j' \in [\tilde{p}]$. We have the following corollary of Theorem 2.1, which is our second main result in this paper.

Corollary 2.1 (Gaussian approximation)**.** *Suppose that Assumptions 2.1, 2.2, and 2.3(i,ii) hold. In addition, suppose that the following growth conditions hold: $\delta_n^2 \varrho_n \log A_n = o(1)$, $L_n^{2/7} \varrho_n \log A_n = o(n^{1/7})$, and $L_n^{2/3} \varrho_n \log A_n = o(n^{1/3-2/(3q)})$. Then*

$$\sup_{t\in\mathbb{R}} \left| \mathrm{P}_P\left(\sup_{u\in\mathcal{U}, j\in[\tilde{p}]} |\sqrt{n}\sigma_{uj}^{-1}(\check{\theta}_{uj} - \theta_{uj})| \leqslant t \right) - \mathrm{P}_P\left(\sup_{u\in\mathcal{U}, j\in[\tilde{p}]} |\mathcal{N}_{uj}| \leqslant t \right) \right| = o(1)$$

uniformly over $P \in \mathcal{P}_n$.

Based on Corollary 2.1, we are now able to construct simultaneous confidence bands for θ_{uj}'s as in (1.2). In particular, we will use the Gaussian multiplier bootstrap method employing the estimates $\widehat{\psi}_{uj}$ of $\bar{\psi}_{uj}$. To describe the method, define the process

$$\widehat{\mathcal{G}} = (\widehat{\mathcal{G}}_{uj})_{u\in\mathcal{U}, j\in[\tilde{p}]} = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \widehat{\psi}_{uj}(W_i) \right)_{u\in\mathcal{U}, j\in[\tilde{p}]} \tag{2.6}$$

where $(\xi_i)_{i=1}^n$ are independent standard normal random variables which are independent from the data $(W_i)_{i=1}^n$. Then the multiplier bootstrap critical value c_α is defined as the $(1-\alpha)$ quantile of the conditional distribution of $\sup_{u\in\mathcal{U}, j\in[\tilde{p}]} |\widehat{\mathcal{G}}_{uj}|$ given the data $(W_i)_{i=1}^n$. To prove validity of this critical value for the construction of simultaneous confidence bands of the form (1.2), we will impose the following additional assumption. Let $(\varepsilon_n)_{n\geqslant 1}$ be a sequence of positive constants converging to zero.

Assumption 2.4 (Variation estimation)**.** *For all $n \geqslant n_0$ and $P \in \mathcal{P}_n$,*

$$\mathrm{P}_P\left(\sup_{u\in\mathcal{U},j\in[\tilde{p}]}\left|\frac{\widehat{\sigma}_{uj}}{\sigma_{uj}}-1\right| > \varepsilon_n\right) \leqslant \Delta_n.$$

The following corollary establishing validity of the multiplier bootstrap critical value c_α for the simultaneous confidence bands construction is our third main result in this paper.

Corollary 2.2 (Simultaneous confidence bands)**.** *Suppose that Assumptions 2.1 – 2.4 hold. In addition, suppose that the growth conditions of Corollary 2.1 hold. Finally, suppose that $\varepsilon_n\varrho_n\log A_n = o(1)$, and $\bar{\delta}_n^2\bar{\varrho}_n\varrho_n(\log\bar{A}_n)\cdot(\log A_n) = o(1)$. Then*

$$\mathrm{P}_P\left(\check{\theta}_{uj} - \frac{c_\alpha\widehat{\sigma}_{uj}}{\sqrt{n}} \leqslant \theta_{uj} \leqslant \check{\theta}_{uj} + \frac{c_\alpha\widehat{\sigma}_{uj}}{\sqrt{n}},\ \textit{for all } u\in\mathcal{U} \textit{ and } j\in[\tilde{p}]\right) = 1-\alpha-o(1)$$

uniformly over $P \in \mathcal{P}_n$.

2.1. **Construction of score functions satisfying near orthogonality condition.** We conclude this section with a short description of two methods that allow to construct score functions ψ_{uj} that satisfy the required near orthogonality condition. Together these methods cover a wide variety of applications.

First, suppose that the original functions m_{uj} are score functions of the model of the form $(\theta,\eta)\mapsto P_{\theta,\eta}$, where $\theta = (\theta_{uj})_{u\in\mathcal{U},j\in[\tilde{p}]}$ and $\eta = (\eta_{uj})_{u\in\mathcal{U},j\in[\tilde{p}]}$ are sets of parameters and $P_{\theta,\eta}$ is the distribution of W. In this case, one can transform m_{uj}'s into efficient score functions ψ_{uj} that obey the near orthogonality condition (2.3) with $C_0 = 0$ by projecting m_{uj}'s onto the orthocomplement of the tangent space induced by the nuisance parameter η; see Chapter 25 of [43] for a detailed description of this construction. Other relevant references include [44], [27], [7], and [9].

Second, suppose that the original moment conditions take the form

$$\mathrm{E}_P[m_{uj}(W,\theta_{uj},\beta_{uj}) \mid X] = 0$$

for some random variable X and some vector of nuisance parameters β_{uj}, as, for example, in the logistic regression model (1.5). In this case, one can define the score functions ψ_{uj} for $\eta = (\beta,\gamma)$ as

$$\psi_{uj}(W,\theta,\eta) = m_{uj}(W,\theta,\beta)\cdot\Big(\widetilde{m}_{uj,\theta}(X) - \gamma'\widetilde{m}_{uj,\beta}(X)\Big)$$

where

$$\widetilde{m}_{uj,\theta}(X) = \partial_\theta\Big\{\mathrm{E}_P[m_{uj}(W,\theta,\beta_{uj}) \mid X]\Big\}\Big|_{\theta=\theta_{uj}} \text{ and } \widetilde{m}_{uj,\beta}(X) = \partial_\beta\Big\{\mathrm{E}_P[m_{uj}(W,\theta_{uj},\beta) \mid X]\Big\}\Big|_{\beta=\beta_{uj}}.$$

These score functions ψ_{uj} satisfy (2.1) with $\eta_{uj} = (\beta_{uj},\gamma_{uj})$ where

$$\gamma_{uj} = \left(\partial_\beta\Big\{\mathrm{E}_P[m_{uj}(W,\theta_{uj},\beta)\widetilde{m}'_{uj,\beta}(X)]\Big\}\Big|_{\beta=\beta_{uj}}\right)^{-1}\cdot\partial_\beta\Big\{\mathrm{E}_P[m_{uj}(W,\theta_{uj},\beta)\widetilde{m}_{uj,\theta}(X)]\Big\}\Big|_{\theta=\theta_{uj}},$$

and also obey the near orthogonality condition (2.3) with $C_0 = 0$.

3. Application to Logistic Regression Model with Functional Response Data

In this section we apply our main results to a logistic regression model with functional response data. We consider a response variable $Y \in \mathbb{R}$ that induces a functional response $(Y_u)_{u\in\mathcal{U}}$ by $Y_u = 1\{Y \leqslant (1-u)\underline{y}+u\bar{y}\}$ for a set of indices $\mathcal{U} = [0,1]$ and some constants $\underline{y} \leqslant \bar{y}$. We are interested in the dependence of this functional response on a $\tilde{p}$-vector of covariates, $D = (D_1, \dots, D_{\tilde{p}})' \in \mathbb{R}^{\tilde{p}}$, controlling for a p-vector of additional covariates $X = (X_1, \dots, X_p)' \in \mathbb{R}^p$. We allow both $\tilde{p}$ and p to be (much) larger than the sample size of available data, n.

For each $u \in \mathcal{U}$, we assume that Y_u satisfies the generalized linear model with the logistic link function

$$\mathrm{E}_P[Y_u \mid D, X] = \Lambda(D'\theta_u + X'\beta_u) + r_u \tag{3.1}$$

where $\theta_u = (\theta_{u1}, \dots, \theta_{u\tilde{p}})'$ is a vector of parameters of interest, $\beta_u = (\beta_{u1}, \dots, \beta_{up})'$ is a vector of nuisance parameters, $r_u = r_u(D, X)$ is an approximation error, $\Lambda\colon \mathbb{R} \to \mathbb{R}$ is a logistic link function defined by $\Lambda(t) = \exp(t)/\{1 + \exp(t)\}$ for all $t \in \mathbb{R}$, and $P \in \mathcal{P}_n$ is the distribution of the triple (Y, D, X). As in the previous section, we construct simultaneous confidence bands for the parameters $(\theta_{uj})_{u\in\mathcal{U},j\in[\tilde{p}]}$ based on a random sample $(Y_i, D_i, X_i)_{i=1}^n$ from the distribution of (Y, D, X).

Observe that for each $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, the standard score function for estimating the coefficient θ_{uj} based on (3.1) is given by

$$m_{uj}(W, \theta) = \Big\{Y_u - \Lambda\Big(D_j\theta + X^j(\theta'_{u[\tilde{p}]\setminus j}, \beta'_u)'\Big) - r_u\Big\}D_j,$$

where $W = (Y, D, X)$ and $X^j = (D'_{[\tilde{p}]\setminus j}, X')$, so that $\mathrm{E}_P[m_{uj}(W, \theta_{uj})] = 0$. However, this score function does not satisfy the desired near orthogonality condition in general. We therefore proceed to construct an appropriate score function using an approach from Section 2.1. Specifically, for each $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, define the coefficients

$$\gamma_u^j \in \arg\min_{\gamma} \mathrm{E}_P[f_u^2(D_j - X^j\gamma)^2] \tag{3.2}$$

where $f_u^2 = f_u^2(D, X) = \mathrm{Var}(Y_u \mid D, X) = \mathrm{E}_P[Y_u \mid D, X](1 - \mathrm{E}_P[Y_u \mid D, X])$, so that

$$f_u D_j = f_u X^j\gamma_u^j + v_u^j, \quad \mathrm{E}_P[f_u X^j v_u^j] = 0. \tag{3.3}$$

Also, denote $\beta_u^j = \theta_{uj}\gamma_u^j + (\theta'_{u[\tilde{p}]\setminus j}, \beta'_u)'$. Then a score function ψ_{uj} is

$$\psi_{uj}(W, \theta, \eta_{uj}) = \Big\{Y_u - \Lambda\Big((D_j - X^j\gamma_u^j)\theta + X^j\beta_u^j\Big) - r_u\Big\}(D_j - X^j\gamma_u^j)$$

where the nuisance parameter is $\eta_{uj} = (r_u, \beta_u^j, \gamma_u^j)$. As we demonstrate in the proof of Theorem 3.1 below, this function satisfies the desired near orthogonality condition. (Observe that when $r_u = 0$ almost surely, we have $\mathrm{Var}(Y_u \mid D, X) = \Lambda'(D'\theta_u + X'\beta_u)$, so that we could define the weights $f_u^2 = f_u^2(D, X)$ using the alternative formula $f_u^2(D, X) = \Lambda'(D'\theta_u + X'\beta_u)$, which was used, in particular, in the Introduction. When $r_u \neq 0$ with positive probability, however, two expressions are not the same. We find it more convenient to work with the formula $f_u^2(D,X) = \mathrm{Var}(Y_u \mid D, X)$.)

Next, we discuss possible estimators of η_{uj}. First, if D and X are chosen appropriately, $r_u = r_u(D,X)$ is asymptotically negligible, so it can be estimated by $\mathcal{O} = \mathcal{O}(D,X)$, the identically zero function of D and X. Second, for γ_u^j, we consider an estimator $\widetilde{\gamma}_u^j$ defined as a post-regularization weighted least squares estimator corresponding to the problem (3.2). Third, for β_u^j, we consider a plug-in estimator

$$\widehat{\beta}_u^j = \widetilde{\theta}_{uj}\widetilde{\gamma}_u^j + (\widetilde{\theta}'_{u[\tilde{p}]\setminus j}, \widetilde{\beta}'_u)' \tag{3.4}$$

where $\widetilde{\theta}_u$ and $\widetilde{\beta}_u$ are suitable estimators of θ_u and β_u. In particular, we assume that $\widetilde{\theta}_u$ and $\widetilde{\beta}_u$ are post-regularization maximum likelihood estimators corresponding to the log-likelihood function $(\theta,\beta) \mapsto -M_u(W,\theta,\beta)$ where

$$M_u(W,\theta,\beta) = -\Big(1\{Y_u = 1\}\log\Lambda(D'\theta + X'\beta) + 1\{Y_u = 0\}\log(1 - \Lambda(D'\theta + X'\beta))\Big). \tag{3.5}$$

The details of the estimators $\widetilde{\theta}_u$, $\widetilde{\beta}_u$, and $\widetilde{\gamma}_u^j$ are given in Algorithm 1 below. The results in this paper can also be easily extended to the case where $\widetilde{\theta}_u$, $\widetilde{\beta}_u$, and $\widetilde{\gamma}_u^j$ are replaced by penalized maximum likelihood estimators $\widehat{\theta}_u$ and $\widehat{\beta}_u$ and penalized weighted least squares estimator $\widehat{\gamma}_u^j$, respectively.

To sum up, our estimator of η_{uj} is $\widehat{\eta}_{uj} = (\mathcal{O}, \widehat{\beta}_u^j, \widetilde{\gamma}_u^j)$. Substituting this estimator into the score function ψ_{uj} gives

$$\psi_{uj}(W,\theta,\widehat{\eta}_{uj}) = \Big\{Y_u - \Lambda\Big((D_j - X^j\widetilde{\gamma}_u^j)\theta + X^j\widehat{\beta}_u^j\Big)\Big\}(D_j - X^j\widetilde{\gamma}_u^j) \tag{3.6}$$

and the sample analog (2.2) of the moment condition (2.1) can be implemented as

$$\check{\theta}_{uj} \in \arg\inf_{\theta\in\Theta_{uj}} \Big|\mathbb{E}_n\Big[\psi_{uj}(W,\theta,\widehat{\eta}_{uj})\Big]\Big|. \tag{3.7}$$

The algorithm is summarized as follows.

Algorithm 1. (Based on the score function.) For each $u \in \mathcal{U}$ and $j \in [\tilde{p}]$:
Step 1. Run post-ℓ_1-penalized logistic estimator (4.2) of Y_u on D and X to compute $(\widetilde{\theta}_u, \widetilde{\beta}_u)$.
Step 2. Define the weights $\widehat{f}_u^2 = \widehat{f}_u^2(D,X) = \Lambda'(D'\widetilde{\theta}_u + X'\widetilde{\beta}_u)$.
Step 3. Run the post-lasso estimator (4.5) of $\widehat{f}_u D_j$ on $\widehat{f}_u X^j$ to compute $\widetilde{\gamma}_u^j$.
Step 4. Compute $\widehat{\beta}_u^j$ in (3.4).
Step 5. Solve (3.7) with $\psi_{uj}(W,\theta,\widehat{\eta}_{uj})$ defined in (3.6) to compute $\check{\theta}_{uj}$.

Next, we specify our regularity conditions. For all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, denote $Z_u^j = D_j - X^j\gamma_u^j$. Also, denote $a_n = p \vee \tilde{p} \vee n$. Let q, c_1, and C_1 be some strictly positive (and finite) constants where $q > 4$. Also, let $(\delta_n)_{n\geqslant 1}$ and $(\bar{\Delta}_n)_{n\geqslant 1}$ be some sequences of positive constants converging to zero. Finally, let $(M_{n,1})_{n\geqslant 1}$ and $(M_{n,2})_{n\geqslant 1}$ be some sequences of positive constants, possibly growing to infinity, where $M_{n,1} \geqslant 1$ and $M_{n,2} \geqslant 1$ for all n.

Assumption 3.1 (Parameters)**.** *For all $u \in \mathcal{U}$, we have $\|\theta_u\| + \|\beta_u\| + \max_{j\in[\tilde{p}]}\|\gamma_u^j\| \leqslant C_1$ and $\max_{j\in[\tilde{p}]}\sup_{\theta\in\Theta_{uj}}|\theta| \leqslant C_1$. In addition, for all $u_1, u_2 \in \mathcal{U}$, we have $(\|\theta_{u_2} - \theta_{u_1}\| + \|\beta_{u_2} - \beta_{u_1}\|) \leqslant$*

$C_1|u_2-u_1|$. *Finally, for all* $u \in \mathcal{U}$ *and* $j \in [\tilde{p}]$, Θ_{uj} *contains a ball of radius* $(\log\log n)(\log a_n)^{3/2}/n^{1/2}$ *centered at* θ_{uj}.

Assumption 3.2 (Sparsity). *There exist* $s = s_n$ *and* $\bar{\gamma}_u^j$, $u \in \mathcal{U}$ *and* $j \in [\tilde{p}]$, *such that for all* $u \in \mathcal{U}$, $\|\beta_u\|_0 + \|\theta_u\|_0 + \max_{j\in[\tilde{p}]} \|\bar{\gamma}_u^j\|_0 \leqslant s_n$ *and* $\max_{j\in[\tilde{p}]}(\|\bar{\gamma}_u^j - \gamma_u^j\| + s_n^{-1/2}\|\bar{\gamma}_u^j - \gamma_u^j\|_1) \leqslant C_1(s_n \log a_n/n)^{1/2}$.

Assumption 3.3 (Distribution of Y). *The conditional pdf of* Y *given* (D, X) *is bounded by* C_1.

Assumptions 3.1-3.3 are mild and standard in the literature. In particular, Assumption 3.1 requires the parameter spaces Θ_{uj} to be bounded, and also requires that for each $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, the parameter θ_{uj} to be sufficiently separated from the boundaries of the parameter space Θ_{uj}. Assumption 3.2 requires approximate sparsity of the model (3.1). Note that in Assumption 3.2, given that $\bar{\gamma}_u^j$'s exist, we can and will assume without loss of generality that $\bar{\gamma}_u^j = \gamma_{uT}^j$ for some $T \subset \{1, \ldots, p + \tilde{p} - 1\}$, where $T = T_u^j$ is allowed to depend on u and j. Assumption 3.3 can be relaxed at the expense of more technicalities.

Assumption 3.4 (Covariates). *For all* $u \in \mathcal{U}$, *we have (i)* $\inf_{\|\xi\|=1} \mathrm{E}_P[f_u^2\{(D', X')\xi\}^2] \geqslant c_1$, *(ii)* $\min_{j,k}(\mathrm{E}_P[|f_u^2 Z_u^j X_k^j|^2] \wedge \mathrm{E}_P[|f_u^2 D_j X_k^j|^2]) \geqslant c_1$, *and (iii)* $\max_{j,k} \mathrm{E}_P[|Z_u^j X_k^j|^3]^{1/3} \log^{1/2} a_n \leqslant \delta_n n^{1/6}$. *In addition, (iv)* $\sup_{\|\xi\|=1} \mathrm{E}_P[\{(D', X')\xi\}^4] \leqslant C_1$, *(v)* $M_{n,1} \geqslant \mathrm{E}_P[\sup_{u\in\mathcal{U}, j\in[\tilde{p}]} |Z_u^j|^{2q}]^{1/(2q)}$, *(vi)* $M_{n,1}^2 s_n \log a_n \leqslant \delta_n n^{1/2-1/q}$, *(vii)* $M_{n,2} \geqslant \{\mathrm{E}_P[(\|D\|_\infty \vee \|X\|_\infty)^{2q}]\}^{1/(2q)}$, *(viii)* $M_{n,2}^2 s_n \log^{1/2} a_n \leqslant \delta_n n^{1/2-1/q}$, *and (ix)* $M_{n,1}^2 M_{n,2}^4 s_n \leqslant \delta_n n^{1-3/q}$.

This assumption requires that there is no multicollinearity between covariates in vectors D and X. In addition, it imposes constraints on various moments of covariates. Since these constraints might be difficult to grasp, at the end of this section, in Corollary 3.3, we provide an example for which these constraints simplify into easily interpretable conditions.

Assumption 3.5 (Approximation error). *For all* $u \in \mathcal{U}$, *we have (i)* $\sup_{\|\xi\|=1} \mathrm{E}_P[r_u^2\{(D', X')\xi\}^2] \leqslant C_1 \mathrm{E}_P[r_u^2]$, *(ii)* $\mathrm{E}_P[r_u^2] \leqslant C_1 s_n \log a_n/n$, *(iii)* $\max_{j\in[\tilde{p}]} |\mathrm{E}_P[r_u Z_u^j]| \leqslant \delta_n n^{-1/2}$, *and (iv)* $|r_u(D, X)| \leqslant f_u^2(D, X)/4$ *almost surely. In addition, with probability* $1 - \bar{\Delta}_n$, *(v)* $\sup_{u\in\mathcal{U}, j\in[\tilde{p}]}(\mathbb{E}_n[(r_u Z_u^j/f_u)^2] + \mathbb{E}_n[r_u^2/f_u^6]) \leqslant C_1 s_n \log a_n/n$.

This assumption requires the approximation error $r_u = r_u(D, X)$ to be sufficiently small. Under Assumption 3.4, the first condition of Assumption 3.5 holds if the approximation error is such that $r_u^2 \leqslant C \mathrm{E}_P[r_u^2]$ almost surely for some constant C.

Under specified assumptions, our estimator $(\check{\theta}_{uj})_{u\in\mathcal{U}, j\in[\tilde{p}]}$ satisfies the following uniform Bahadur representation theorem.

Theorem 3.1 (Uniform Bahadur representation for logistic model). *Suppose that Assumptions 3.1 – 3.5 hold for all* $P \in \mathcal{P}_n$. *In addition, suppose that the following growth condition holds:* $\delta_n^2 \log a_n = o(1)$. *Then for the estimator* $(\check{\theta}_{uj})_{u\in\mathcal{U}, j\in[\tilde{p}]}$ *satisfying* (3.7), *we have*

$$\sqrt{n}\sigma_{uj}^{-1}(\check{\theta}_{uj} - \theta_{uj}) = \mathbb{G}_n \bar{\psi}_{uj} + O_P(\delta_n) \text{ in } \ell^\infty(\mathcal{U} \times [\tilde{p}])$$

uniformly over $P \in \mathcal{P}_n$*, where* $\bar{\psi}_{uj}(W) := -\sigma_{uj}^{-1} J_{uj}^{-1} \psi_{uj}(W, \theta_{uj}, \eta_{uj})$*,* $\sigma_{uj}^2 := \mathrm{E}_P[J_{uj}^{-2} \psi_{uj}^2(W, \theta_{uj}, \eta_{uj})]$*, and* J_{uj} *is defined in* (2.4)*.*

This theorem allows us to establish a Gaussian approximation result for the supremum of the process $\{\sqrt{n}\sigma_{uj}^{-1}(\check{\theta}_{uj} - \theta_{uj}) \colon u \in \mathcal{U}, j \in [\tilde{p}]\}$:

Corollary 3.1 (Gaussian approximation for logistic model)**.** *Suppose that Assumptions 3.1 – 3.5 hold for all* $P \in \mathcal{P}_n$*. In addition, suppose that the following growth conditions hold:* $\delta_n^2 \log a_n = o(1)$*,* $M_{n,1}^{2/7} \log a_n = o(n^{1/7})$*, and* $M_{n,1}^{2/3} \log a_n = o(n^{1/3 - 2/(3q)})$*. Then*

$$\sup_{t \in \mathbb{R}} \left| \mathrm{P}_P \left(\sup_{u \in \mathcal{U}, j \in [\tilde{p}]} |\sqrt{n}\sigma_{uj}^{-1}(\check{\theta}_{uj} - \theta_{uj})| \leqslant t \right) - \mathrm{P}_P \left(\sup_{u \in \mathcal{U}, j \in [\tilde{p}]} |\mathcal{N}_{uj}| \leqslant t \right) \right| = o(1)$$

uniformly over $P \in \mathcal{P}_n$*, where* $(\mathcal{N}_{uj})_{u \in \mathcal{U}, j \in [\tilde{p}]}$ *is a tight zero-mean Gaussian process indexed by* $\mathcal{U} \times [\tilde{p}]$ *with the covariance operator given by* $\mathrm{E}_P[\bar{\psi}_{uj}(W)\bar{\psi}_{u'j'}(W)]$ *for* $u, u' \in \mathcal{U}$ *and* $j, j' \in [\tilde{p}]$*.*

Based on this corollary, we are now able to construct simultaneous confidence bands for the parameters θ_{uj}. Observe that

$$J_{uj} = -\mathrm{E}_P\Big[\Lambda'\Big((D_j - X^j\gamma_u^j)\theta_{uj} + X^j\beta_u^j\Big)(D_j - X^j\gamma_u^j)^2\Big],$$

and so it can be estimated by

$$\widehat{J}_{uj} = -\mathbb{E}_n\Big[\Lambda'\Big((D_j - X^j\widetilde{\gamma}_u^j)\widetilde{\theta}_{uj} + X^j\widehat{\beta}_u^j\Big)(D_j - X^j\widetilde{\gamma}_u^j)^2\Big]$$

for all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$. In addition, $\sigma_{uj}^2 = \mathrm{E}_P[J_{uj}^{-2}\psi_{uj}^2(W, \theta_{uj}, \eta_{uj})]$, and so it can be estimated by

$$\widehat{\sigma}_{uj}^2 = \mathbb{E}_n[\widehat{J}_{uj}^{-2}\psi_{uj}^2(W, \widetilde{\theta}_{uj}, \widehat{\eta}_{uj})]$$

for all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$. Moreover, as in Section 2, for all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, define $\widehat{\psi}_{uj}(W) = -\widehat{\sigma}_{uj}\widehat{J}_{uj}\psi_{uj}(W, \check{\theta}_{uj}, \widehat{\eta}_{uj})$, and let c_α be the $(1-\alpha)$ quantile of the conditional distribution of $\sup_{u \in \mathcal{U}, j \in [\tilde{p}]} |\widehat{\mathcal{G}}_{uj}|$ given the data $(W_i)_{i=1}^n$ where the process $\widehat{\mathcal{G}} = (\widehat{\mathcal{G}}_{uj})_{u \in \mathcal{U}, j \in [\tilde{p}]}$ is defined in (2.6). Then we have

Corollary 3.2 (Uniform confidence bands for logistic model)**.** *Suppose that Assumptions 3.1 – 3.5 hold for all* $P \in \mathcal{P}_n$*. In addition, suppose that the following growth conditions hold:* $\delta_n^2 \log a_n = o(1)$*,* $M_{n,1}^{2/7} \log a_n = o(n^{1/7})$*,* $M_{n,1}^{2/3} \log a_n = o(n^{1/3 - 2/(3q)})$*, and* $s_n \log^3 a_n = o(n)$*. Then*

$$\mathrm{P}_P\left(\check{\theta}_{uj} - \frac{c_\alpha \widehat{\sigma}_{uj}}{\sqrt{n}} \leqslant \theta_{uj} \leqslant \check{\theta}_{uj} + \frac{c_\alpha \widehat{\sigma}_{uj}}{\sqrt{n}}, \text{ for all } u \in \mathcal{U} \text{ and } j \in [\tilde{p}]\right) = 1 - \alpha - o(1) \tag{3.8}$$

uniformly over $P \in \mathcal{P}_n$*.*

To conclude this section, we provide an example for which conditions of Corollary 3.2 are easy to interpret. Recall that $a_n = n \vee p \vee \tilde{p}$.

Corollary 3.3 (Uniform confidence bands for logistic model under simple conditions)**.** *Suppose that Assumptions 3.1 – 3.3, 3.4(i,ii,iv), and 3.5(i,ii,iv,v) hold for* $q > 4$ *for all* $P \in \mathcal{P}_n$*. In addition, suppose that* $\{\mathrm{E}_P[(\|D\|_\infty \vee \|X\|_\infty)^{2q}]\}^{1/(2q)} \leqslant C_1$ *and* $\sup_{u\in\mathcal{U},j\in[\tilde{p}]} \|\gamma_u^j\|_1 \leqslant C_1$*. Finally, suppose that* $\log^7 a_n/n = o(1)$*,* $s_n^2 \log^3 a_n/n^{1-2/q} = o(1)$*, and* $\sup_{u\in\mathcal{U},j\in[\tilde{p}]} |\mathrm{E}_P[r_u Z_u^j]| = o((n\log a_n)^{-1/2})$*. Then* (3.8) *holds uniformly over* $P \in \mathcal{P}_n$*.*

Comment 3.1 (Estimation of variance)**.** When constructing the confidence bands based on (3.8), we find in simulations that it is beneficial to replace the estimators $\widehat{\sigma}_{uj}^2$ of σ_{uj}^2 by $\max\{\widehat{\sigma}_{uj}^2, \widehat{\Sigma}_{uj}^2\}$ where $\widehat{\Sigma}_{uj}^2 = \mathbb{E}_n[\widehat{f}_u^2(D - X^j\widetilde{\gamma}_u^j)^2]$ is an alternative consistent estimator of σ_{uj}^2. ■

Comment 3.2 (Alternative implementations, double selection)**.** We note that the theory developed here is applicable for different estimators that construct the new score function with the desired orthogonality condition implicitly. For example, the double selection idea yields an implementation of an estimator that is first-order equivalent to the estimator based on the score function. The algorithm yielding the double selection estimator is as follows.

Algorithm 2. (Based on double selection) For each $u \in \mathcal{U}$ and $j \in [\tilde{p}]$:
Step $1'$. Run post-ℓ_1-penalized logistic estimator (4.2) of Y_u on D and X to compute $(\widetilde{\theta}_u, \widetilde{\beta}_u)$.
Step $2'$. Define the weights $\widehat{f}_u^2 = \widehat{f}_u^2(D, X) = \Lambda'(D_i'\widetilde{\theta}_u + X_i'\widetilde{\beta}_u)$.
Step $3'$. Run the lasso estimator (4.4) of $\widehat{f}_u D_j$ on $\widehat{f}_u X$ to compute $\widehat{\gamma}_u^j$.
Step $4'$. Run logistic regression of Y_u on D_j and all the selected variables in Steps $1'$ and $3'$ to compute $\check{\theta}_{uj}$.

As mentioned by a referee, it is surprising that the double selection procedure has uniform validity. The use of the additional variables selected in Step 3', through the first order conditions of the optimization problem, induces the necessary near-orthogonality condition. We refer to the Supplementary Material for a more detailed discussion. ■

Comment 3.3 (Alternative implementations, one-step correction)**.** Another implementation for which the theory developed here applies is to replace Step 5 in Algorithm 1 with a one-step procedure. This relates to the debiasing procedure proposed in [42] to the case when the set $\mathcal{U}$ is a singleton. In this case instead of minimizing the criterion (3.7) in Step 5, the method makes a full Newton step from the initial estimate,

Step $5''$. Compute $\bar{\theta}_{uj} = \widehat{\theta}_{uj} - \widehat{J}_{uj}^{-1}\mathbb{E}_n[\psi_{uj}(W, \widehat{\theta}_{uj}, \widehat{\eta}_{uj})]$.

The theory developed here directly apply to those estimators as well. ■

4. ℓ_1-Penalized M-Estimators: Nuisance Functions and Functional Data

In this section, we define the estimators $\widetilde{\theta}_u$, $\widetilde{\beta}_u$, and $\widetilde{\gamma}_u^j$, which were used in the previous section, and study their properties. We consider the same setting as that in the previous section. The results in this section rely upon a set of new results for ℓ_1-penalized M-estimators with functional data presented in Appendix I of the Supplementary Material.

4.1. **ℓ_1-Penalized Logistic Regression for Functional Response Data: Asymptotic Properties.** Here we consider the generalized linear model with the logistic link function and functional response data (3.1). As explained in the previous section, we assume that $\widetilde{\theta}_u$ and $\widetilde{\beta}_u$ are post-regularization maximum likelihood estimators of θ_u and β_u corresponding to the log-likelihood function $M_u(W, \theta, \beta) = M_u(Y_u, D, X, \theta, \beta)$ defined in (3.5). To define these estimators, let $\widehat{\theta}_u$ and $\widehat{\beta}_u$ be ℓ_1-penalized maximum likelihood (logistic regression) estimators

$$(\widehat{\theta}_u, \widehat{\beta}_u) \in \arg\min_{\theta,\beta} \left(\mathbb{E}_n[M_u(Y_u, D, X, \theta, \beta)] + \frac{\lambda}{n} \|\widehat{\Psi}_u(\theta', \beta')'\|_1 \right) \quad (4.1)$$

where λ is a penalty level and $\widehat{\Psi}_u$ a diagonal matrix of penalty loadings. We choose parameters λ and $\widehat{\Psi}_u$ according to Algorithm 3 described below. Using the ℓ_1-penalized estimators $\widehat{\theta}_u$ and $\widehat{\beta}_u$, we then define post-regularization estimators $\widetilde{\theta}_u$ and $\widetilde{\beta}_u$ by

$$(\widetilde{\theta}_u, \widetilde{\beta}_u) \in \arg\min_{\theta} \mathbb{E}_n[M_u(Y_u, D, X, \theta, \beta)] \quad : \quad \mathrm{supp}(\theta, \beta) \subseteq \mathrm{supp}(\widehat{\theta}_u, \widehat{\beta}_u). \quad (4.2)$$

We derive the rate of convergence and sparsity properties of $\widetilde{\theta}_u$ and $\widetilde{\beta}_u$ as well as of $\widehat{\theta}_u$ and $\widehat{\beta}_u$ in Theorem 4.1 below. Recall that $a_n = n \vee p \vee \tilde{p}$.

Algorithm 3 (Penalty Level and Loadings for Logistic Regression)**.** Choose $\gamma \in [1/n, 1/\log n]$ and $c > 1$ (in practice, we set $c = 1.1$ and $\gamma = .1/\log n$). Define $\lambda = c\sqrt{n}\Phi^{-1}(1 - \gamma/(2(p + \tilde{p})N_n))$ with $N_n = n$. To select $\widehat{\Psi}_u$, choose a constant $\bar{m} \geqslant 0$ as an upper bound on the number of loops and proceed as follows: (0) Let $\widetilde{X} = (D', X')'$, $m = 0$, and initialize $\widehat{l}_{uk,0} = \frac{1}{2}\{\mathbb{E}_n[\widetilde{X}_k^2]\}^{1/2}$ for $k \in [p + \tilde{p}]$. (1) Compute $(\widehat{\theta}_u, \widehat{\beta}_u)$ and $(\widetilde{\theta}_u, \widetilde{\beta}_u)$ based on $\widehat{\Psi}_u = \mathrm{diag}(\{\widehat{l}_{uk,m}, k \in [p + \tilde{p}]\})$. (2) Set $\widehat{l}_{uk,m+1} := \{\mathbb{E}_n[\widetilde{X}_k^2(Y_u - \Lambda(D'\widetilde{\theta}_u + X'\widetilde{\beta}_u))^2]\}^{1/2}$. (3) If $m \geqslant \bar{m}$, report the current value of $\widehat{\Psi}_u$ and stop; otherwise set $m \leftarrow m + 1$ and go to step (1).

Theorem 4.1 (Rates and Sparsity for Functional Response under Logistic Link)**.** *Suppose that Assumptions 3.1 – 3.5 hold for all $P \in \mathcal{P}_n$. In addition, suppose that the penalty level λ and the matrices of penalty loadings $\widehat{\Psi}_u$ are chosen according to Algorithm 3. Moreover, suppose that the following growth condition holds: $\delta_n^2 \log a_n = o(1)$. Then there exists a constant $\bar{C}$ such that uniformly over all $P \in \mathcal{P}_n$ with probability $1 - o(1)$,*

$$\sup_{u \in \mathcal{U}} \left(\|\widehat{\theta}_u - \theta_u\| + \|\widehat{\beta}_u - \beta_u\| \right) \leqslant \bar{C} \sqrt{\frac{s_n \log a_n}{n}}, \quad \sup_{u \in \mathcal{U}} \left(\|\widehat{\theta}_u - \theta_u\|_1 + \|\widehat{\beta}_u - \beta_u\|_1 \right) \leqslant \bar{C} \sqrt{\frac{s_n^2 \log a_n}{n}},$$

and the estimators $\widehat{\theta}_u$ and $\widehat{\beta}_u$ are uniformly sparse: $\sup_{u \in \mathcal{U}} \|\widehat{\theta}_u\|_0 + \|\widehat{\beta}_u\|_0 \leqslant \bar{C} s_n$. Also, uniformly over all $P \in \mathcal{P}_n$, with probability $1 - o(1)$,

$$\sup_{u \in \mathcal{U}} \left(\|\widetilde{\theta}_u - \theta_u\| + \|\widetilde{\beta}_u - \beta_u\| \right) \leqslant \bar{C} \sqrt{\frac{s_n \log a_n}{n}}, \ \sup_{u \in \mathcal{U}} \left(\|\widetilde{\theta}_u - \theta_u\|_1 + \|\widetilde{\beta}_u - \beta_u\|_1 \right) \leqslant \bar{C} \sqrt{\frac{s_n^2 \log a_n}{n}}.$$

4.2. **Lasso with Estimated Weights: Asymptotic Properties.** Here we consider the weighted linear model (3.3) for $u \in \mathcal{U}$ and $j \in [\tilde{p}]$. Using the parameter $\bar{\gamma}_u^j$ appearing in Assumption 3.2, it will be convenient to rewrite this model as

$$f_u D_j = f_u X^j \bar{\gamma}_u^j + f_u \bar{r}_{uj} + v_u^j, \quad \mathrm{E}_P[f_u X^j v_u^j] = 0 \tag{4.3}$$

where $\bar{r}_{uj} = X^j(\gamma_u^j - \bar{\gamma}_u^j)$ is an approximation error, which is asymptotically negligible under Assumption 3.2. As explained in the previous section, we assume that $\widetilde{\gamma}_u^j$ is a post-regularization weighted least squares estimator of γ_u^j (or $\bar{\gamma}_u^j$). To define this estimator, let $\widehat{\gamma}_u^j$ be an ℓ_1-penalized (weighted Lasso) estimator

$$\widehat{\gamma}_u^j \in \arg\min_{\gamma} \left(\tfrac{1}{2}\mathbb{E}_n[\widehat{f}_u^2 (D_j - X^j\gamma)^2] + \frac{\lambda}{n}\|\widehat{\Psi}_{uj}\gamma\|_1 \right) \tag{4.4}$$

where λ and $\widehat{\Psi}_{uj}$ are the associated penalty level and the diagonal matrix of penalty loadings specified below in Algorithm 4 and where $\widehat{f}_u^2$'s are estimated weights. As in Algorithm 1 in the previous section, we set $\widehat{f}_u^2 = \widehat{f}_u^2(D,X) = \Lambda'(D'\widetilde{\theta}_u + X'\widetilde{\beta}_u)$. Using $\widehat{\gamma}_u^j$, we define a post-regularized weighted least squares estimator

$$\widetilde{\gamma}_u^j \in \arg\min_{\gamma} \tfrac{1}{2}\mathbb{E}_n[\widehat{f}_u^2 (D_j - X^j\gamma)^2] \quad : \; \mathrm{supp}(\gamma) \subseteq \mathrm{supp}(\widehat{\gamma}_u^j). \tag{4.5}$$

We derive the rate of convergence and sparsity properties of $\widetilde{\gamma}_u^j$ as well as of $\widehat{\gamma}_u^j$ in Theorem 4.2 below.

Algorithm 4 (Penalty Level and Loadings for Weighted Lasso)**.** Choose $\gamma \in [1/n, 1/\log n]$ and $c > 1$ (in practice, we set $c = 1.1$ and $\gamma = .1/\log n$). Define $\lambda = c\sqrt{n}\Phi^{-1}(1 - \gamma/(2(p+\tilde{p})N_n))$ with $N_n = p\tilde{p}^2 n^2$. To select $\widehat{\Psi}_{uj}$, choose a constant $\bar{m} \geqslant 1$ as an upper bound on the number of loops and proceed as follows: (0) Set $m = 0$ and $\widehat{l}_{ujk,0} = \max_{1\leqslant i\leqslant n} \|\widehat{f}_{ui} X_i^j\|_\infty \{\mathbb{E}_n[\widehat{f}_u^2 D_j^2]\}^{1/2}$. (1) Compute $\widehat{\gamma}_u^j$ and $\widetilde{\gamma}_u^j$ based on $\widehat{\Psi}_{uj} = \mathrm{diag}(\{\widehat{l}_{ujk,m}, k \in [p+\tilde{p}-1]\})$. (2) Set $\widehat{l}_{ujk,m+1} := \{\mathbb{E}_n[\widehat{f}_u^4 (D_j - X^j\widetilde{\gamma}_u^j)^2 (X_k^j)^2]\}^{1/2}$. (3) If $m \geqslant \bar{m}$, report the current value of $\widehat{\Psi}_u^j$ and stop; otherwise set $m \leftarrow m+1$ and go to step (1).

Theorem 4.2 (Rates and Sparsity for Lasso with Estimated Weights)**.** *Suppose that Assumptions 3.1 – 3.5 hold for all $P \in \mathcal{P}_n$. In addition, suppose that the penalty level λ and the matrices of penalty loadings $\widehat{\Psi}_{uj}$ are chosen according to Algorithm 4. Moreover, suppose that the following growth condition holds: $\delta_n^2 \log a_n = o(1)$. Then there exists a constant $\bar{C}$ such that uniformly over all $P \in \mathcal{P}_n$ with probability $1 - o(1)$,*

$$\max_{j\in[\tilde{p}]} \sup_{u\in\mathcal{U}} \|\widehat{\gamma}_u^j - \bar{\gamma}_u^j\| \leqslant \bar{C}\sqrt{\frac{s_n \log a_n}{n}} \;\; \textit{and} \;\; \max_{j\in[\tilde{p}]} \sup_{u\in\mathcal{U}} \|\widehat{\gamma}_u^j - \bar{\gamma}_u^j\|_1 \leqslant \bar{C}\sqrt{\frac{s_n^2 \log a_n}{n}},$$

and the estimator $\widehat{\gamma}_u^j$ is uniformly sparse, $\max_{j\in[\tilde{p}]}\sup_{u\in\mathcal{U}} \|\widehat{\gamma}_u^j\|_0 \leqslant \bar{C}s_n$. Also, uniformly over all $P \in \mathcal{P}_n$, with probability $1 - o(1)$,

$$\max_{j\in[\tilde{p}]} \sup_{u\in\mathcal{U}} \|\widetilde{\gamma}_u^j - \bar{\gamma}_u^j\| \leqslant \bar{C}\sqrt{\frac{s_n \log a_n}{n}}, \;\; \textit{and} \;\; \max_{j\in[\tilde{p}]} \sup_{u\in\mathcal{U}} \|\widetilde{\gamma}_u^j - \bar{\gamma}_u^j\|_1 \leqslant \bar{C}\sqrt{\frac{s_n^2 \log a_n}{n}}.$$

5. Simulations and Illustrative Application

5.1. **Monte Carlo Simulations.** In this section we provide a simulation study to investigate the finite sample properties of the proposed estimators and the associated confidence regions. We report only the performance of the estimator based on the double selection procedure due to space constraints and note that it has very similar to the performance of the estimator based on score functions with near orthogonality property. We will compare the proposed procedure with the traditional estimator that refits the model selected by the corresponding ℓ_1-penalized M-estimator (naive post-selection estimator).

We consider a logistic regression model with functional response data where the response $Y_u = 1\{y \leqslant u\}$ for $u \in \mathcal{U}$ a compact set. We specify two different designs: (1) a location model where $y = x'\beta_0 + \xi$ where ξ is distributed as a logistic random variable, the first component of x is the intercept and the other $p-1$ components are distributed as $N(0,\Sigma)$ with $\Sigma_{k,j} = |0.5|^{|k-j|}$; (2) a location-shift model where $y = \{(x'\beta_0 + \xi)/x'\vartheta_0\}^3$ where ξ is distributed as a logistic random variable, $x_j = |w_j|$ where w is a p-vector distributed as $N(0,\Sigma)$ with $\Sigma_{k,j} = |0.5|^{|k-j|}$, and ϑ_0 has non-negative components. Such specification implies that for each $u \in \mathcal{U}$

$$\text{Design 1: } \theta_u = u(1,0,\ldots,0)' - \beta_0 \quad \text{and} \quad \text{Design 2: } \theta_u = u^{1/3}\vartheta_0 - \beta_0.$$

In our simulations we will consider $n = 500$ and $p = 2000$. For the location model (Design 1) we will consider two different choices for β_0: (i) $\beta_{0j}^{(i)} = 2/j^2$ for $j = 1,\ldots,p$, and (ii) $\beta_{0j}^{(ii)} = (1/2)/\{j-3.5\}^2$ for $j > 1$ with the intercept coefficient $\beta_{01}^{(ii)} = -10$. (These choices ensure $\max_{j>1}|\beta_{0j}| = 2$ and that y is around zero in Design 2(ii).) We set $\vartheta_0 = \frac{1}{8}(1,1,1,1,0,0,\ldots,0,0,1,1,1,1)'$. For Design 1 we have $\mathcal{U} = [1, 2.5]$ and for Design 2 we have $\mathcal{U} = [-.5, .5]$. The results are based on 500 replications (the bootstrap procedure is performed 5000 times for each replication).

We report the (empirical) rejection frequencies for confidence regions with 95% nominal coverage. That is, the fraction of simulations the confidence regions of a particular method did not cover the true value (thus .05 rejection frequency is the ideal performance). We report the rejection frequencies for the proposed estimator and the post-naive selection estimator.

Table 1 presents the performance of the methods when applied to construct a confidence interval for a single parameter ($\tilde{p} = 1$ and $\mathcal{U}$ is a singleton). Since the setting is not symmetric we investigate the performance for different components. Specifically, we consider $\{u\} \times \{j\}$ for $j = 1,\ldots,10$. First consider the location model (Design 1). The difference between the performance of the naive estimator for Design 1(i) and 1(ii) highlights its fragile performance which is highly dependent on the unknown parameters. In Design 1(i) the Naive method achieve (pointwise) rejection frequencies between .032 and .162 when the nominal level is .05. However, in the Design 1(ii) the range goes from 0.018 to 0.904. We also note that it is important to look at the performance of each component and avoid averaging across components (large j components are essentially not in the model, indeed for $j > 50$ we obtain rejection frequencies very close to .05 regardless of the model selection procedure). In contrast the proposed estimator exhibits a more robust behavior. For Design 1(i) the rejection

$p = 2000, n = 500$		Pointwise Rejection Frequencies for each $j \in \{1, \ldots, 10\}$									
Design	Method	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$j=8$	$j=9$	$j=10$
1(i)	Proposed	0.042	0.040	0.062	0.050	0.044	0.028	0.044	0.048	0.042	0.036
	Naive	0.100	0.098	0.108	0.108	0.162	0.110	0.078	0.072	0.068	0.032
1(ii)	Proposed	0.044	0.040	0.054	0.056	0.056	0.052	0.044	0.048	0.054	0.050
	Naive	0.038	0.030	0.070	0.886	0.698	0.172	0.062	0.030	0.052	0.018
2(i)	Proposed	0.046	0.054	0.044	0.052	0.054	0.040	0.038	0.052	0.036	0.044
	Naive	0.046	0.050	0.038	0.070	0.054	0.040	0.064	0.038	0.038	0.044
2(ii)	Proposed	0.092	0.074	0.034	0.088	0.082	0.046	0.068	0.058	0.068	0.070
	Naive	0.034	0.972	0.182	0.312	0.916	0.096	0.028	0.018	0.016	0.024

TABLE 1. We report the rejection frequencies of each method for (pointwise) confidence intervals for each $j \in \{1, \ldots, 10\}$. For Design 1 we used $\mathcal{U} = \{1\}$ and for Design 2 we used $\mathcal{U} = \{.5\}$. The results are based on 500 replications.

frequencies are between .028 and .062 while for Design 1(ii) the rejection frequencies of the proposed estimator were between 0.044 and 0.056.

Table 2 presents the performance for simultaneous confidence bands of the form $\{[\tilde{\theta}_{uj} - \text{cv}\tilde{\sigma}_{uj}, \tilde{\theta}_{uj} + \text{cv}\tilde{\sigma}_{uj}]$ for $u \in \mathcal{U} \times [\tilde{p}]$ where $\tilde{\theta}_{uj}$ is a point estimate, $\tilde{\sigma}_{uj}$ is an estimate of the pointwise standard deviation, and cv is a critical value that accounts for the uniform estimation. For the point estimate we consider the proposed estimator and the post-naive selection estimator which have estimates of standard deviation. We consider two critical values: from the multiplier bootstrap (MB) procedure and the Bonferroni (BF) correction (which we expect to be conservative). For each of the four different designs (1(i), 1(ii), 2(i) and 2(ii) described above), we consider four different choices of $\mathcal{U} \times [\tilde{p}]$. Table 2 displays rejection frequencies for confidence regions with 95% nominal coverage (and again .05 would be the ideal performance). The simulation results confirms the differences between the performance of the methods and overall the proposed procedure is closer to the nominal value of .05 . The proposed estimator performed within a factor of two to the nominal value in 10 out of the 16 designs considered (and 13 out 16 within a factor of three). The post-naive selection estimator performed within a factor of two only in 3 out of the 16 designs when using the multiplier bootstrap as critical value (7 out of 16 within a factor of three) and similarly with the Bonferroni correction as the critical value.

5.2. **Application to US Presidential Approval Ratings.** In this section we illustrate the applicability of the tools proposed in this work with data on US presidential approval ratings used in [13]. There several economic and political factors that impact presidential approval ratings. In this illustration, we are interested on the impact of unemployment rates and on the impact of time in office on the approval rate of a sitting president. However, the impact of such factors might not be homogeneous and in fact depend on current ratings. For example, a sitting president is likely to have a fraction of voters who would support him regardless of economic factors. Thus, a low unemployment rate might not have an effect when approval ratings are low and have a significant effect when approval ratings are high. This would imply a different effect on different parts of the conditional distribution of the approval rating.

$p = 2000, n = 500$		Uniform over $\mathcal{U} \times [\tilde{p}]$			
Design	Method	$[1, 2.5] \times \{1\}$	$\{1\} \times [10]$	$[1, 2.5] \times [10]$	$\{1\} \times [1000]$
1(i)	Proposed	0.054	0.036	0.048	0.040
	Naive (MB)	0.126	0.136	0.172	0.032
	Naive (BF)	0.014	0.124	0.026	0.032
1(ii)	Propose	0.270	0.036	0.032	0.142
	Naive (MB)	0.014	0.802	0.934	0.404
	Naive (BF)	0.000	0.802	0.718	0.376
Design	Method	$[-.5, .5] \times \{1\}$	$\{.5\} \times [10]$	$[-.5, .5] \times [10]$	$\{.5\} \times [1000]$
2(i)	Proposed	0.364	0.038	0.052	0.062
	Naive (MB)	0.116	0.040	0.022	0.048
	Naive (BF)	0.018	0.038	0.000	0.046
2(ii)	Proposed	0.140	0.090	0.408	0.084
	Naive (MB)	0.002	0.946	0.996	0.362
	Naive (BF)	0.000	0.946	0.944	0.298

TABLE 2. We report the rejection frequencies of each method for the (uniform) confidence bands for $\mathcal{U} \times [\tilde{p}]$. The proposed estimator computes the critical value based on the multiplier bootstrap procedure. For the naive post-selection estimator we report the results for two choices of critical values, one choice based on the multiplier bootstrap (MB), and another based on Bonferroni (BF) correction. The results are based on 500 replications.

To study the distributional effect of these factors we use a logistic regression model with functional data as described in Section 3. Letting Y denote the approval rating, we define $Y_u = 1\{Y \leqslant u\}$ to be the binary variable that indicates if the approval rating is below the threshold $u \in \mathcal{U} = [.45, .65]$. For each level of approval rating $u \in \mathcal{U}$ we estimate the model

$$\mathrm{E}[Y_u \mid D_{unemp}, D_{time}, X] = \Lambda(\theta_{u,unemp} D_{unemp} + \theta_{u,time} D_{time} + X'\beta_u) + r_u$$

where D_{unemp} denotes the unemployment rate, D_{time} the number of months the president has been in office, r_u a (small) approximation error, and X denotes several additional control variables. In addition to the linear terms for the variables[1] used in [13], we also consider interactions among controls. Therefore, for each $u \in \mathcal{U}$ we have a generalized linear model using logistic link function with 160 variables and 603 observations.

We construct simultaneous confidence bands for both coefficients ($\tilde{p} = 2$) uniformly over $u \in \mathcal{U}$. Although $p < n$ for every $u \in \mathcal{U}$, the full model (applying logistic regression with all regressors) led to numerical instabilities and other numerical failures. We proceed to construct (asymptotically) valid confidence regions based on the double selection procedure for functional logistic regression.

Figure 1 displays the estimation results. Specifically, the figure displays point estimates of each coefficient for every $u \in \mathcal{U}$ (solid line), pointwise confidence intervals (dotted line), and uniform confidence bands (dot-dash lines). Point estimates account for model selection mistakes and are computed based the double selection procedure. For 95% coverage, the pointwise critical value is taken to be $\Phi^{-1}(.975) \approx 1.96$ from the normal approximation and the critical value for uniform

[1]Those include dummy variables for each president, Watergate scandal, causalities in different wars, political shocks, and other variables, see [13] for a complete description.

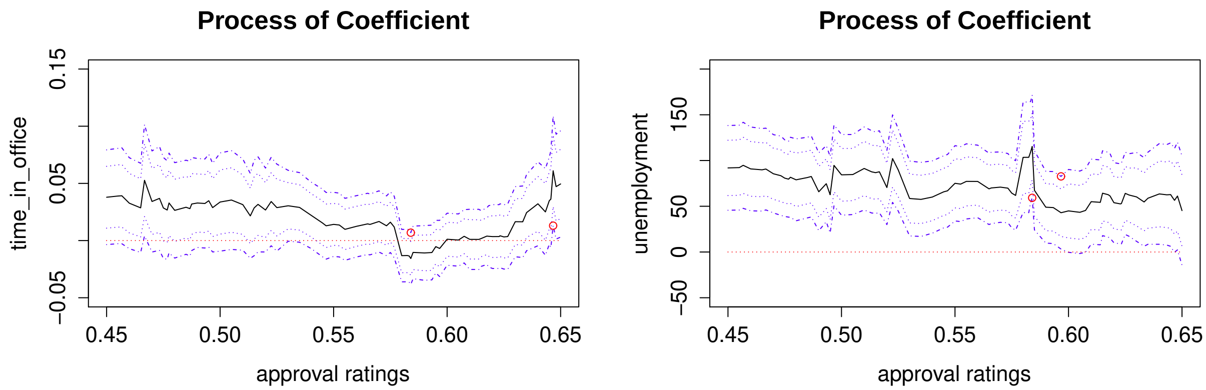

FIGURE 1. The panels display pointwise (dotted) and uniform (dotdash) confidence bands for the unemployment coefficient and the time-in-office coefficient. These confidence regions are set to have 95% coverage and were constructed based on the double selection algorithm. The critical value for the uniform confidence bands was set to 2.93 based on the multiplier bootstrap procedure with 5000 repetitions. For both coefficients, the lowest value of the upper confidence band is smaller than the largest value of the lower confidence band.

confidence bands (uniformly over both coefficients and over $u \in \mathcal{U}$) was calculated to be 3.003 based on 5000 repetitions of the multiplier bootstrap.

At 95% confidence level, the uniform confidence band rule out "no effect" over $\mathcal{U}$ for both variables. Indeed, the straight line at 0 is not contained in the uniform confidence bands for either variable. Next consider the process of the unemployment coefficient. The analysis suggests that the unemployment rate has an overall negative effect on the approval rate of a sitting president (as the coefficient is positive increasing the probability to be below a threshold). Regarding time-in-office, the effect also seems to be predominantly negative. However, the impact is not homogeneous across $u \in \mathcal{U}$. Indeed, the lowest value of the upper confidence band (0.007, $u = 0.584$) is smaller than the largest value of the lower confidence band (0.013, $u = 0.647$), see the circles in the plot of the process of the time-in-office coefficient. The impact seems to be greater for the lower and higher values of $u \in \mathcal{U}$ while the effect of seems negligible in the range of $u \in [.575, .625]$. In particular, at 95% level, no effect is ruled out for large values of u.

Appendix A. Notation

A.1. **Overall Notation.** Throughout the paper, the symbols P and E denote probability and expectation operators with respect to a generic probability measure. If we need to signify the dependence on a probability measure P, we use P as a subscript in P_P and E_P. In the proofs, we sometimes also use P as a subscript for random variables as in W_P. Note also that we use capital letters such as W to denote random elements and use the corresponding lower case letters such as w to denote fixed values that these random elements can take. For a positive integer k, $[k]$ denotes the set $\{1, \ldots, k\}$.

We denote by $\mathbb{P}_n$ the (random) empirical probability measure that assigns probability n^{-1} to each $W_i \in (W_i)_{i=1}^n$. $\mathbb{E}_n$ denotes the expectation with respect to the empirical measure, and $\mathbb{G}_n = \mathbb{G}_{n,P}$ denotes the empirical process $\sqrt{n}(\mathbb{E}_n - \mathrm{E}_P)$, that is,

$$\mathbb{G}_{n,P}(f) = \mathbb{G}_{n,P}(f(W)) = n^{-1/2}\sum_{i=1}^n \{f(W_i) - \mathrm{E}_P[f(W)]\}, \quad \mathrm{E}_P[f(W)] := \int f(w)dP(w),$$

indexed by a class of measurable functions $\mathcal{F}\colon \mathcal{W} \to \mathbb{R}$; see [44, chap. 2.3]. In what follows, we use $\|\cdot\|_{P,q}$ to denote the $L^q(P)$ norm; for example, we use $\|f(W)\|_{P,q} = (\int |f(w)|^q dP(w))^{1/q}$ and $\|f(W)\|_{\mathbb{P}_n,q} = (n^{-1}\sum_{i=1}^n |f(W_i)|^q)^{1/q}$. For a vector $v = (v_1, \dots, v_p)' \in \mathbb{R}^p$, $\|v\|_0$ denotes the ℓ_0-"norm" of v, that is, the number of non-zero components of v, $\|v\|_1$ denotes the ℓ_1-norm of v, that is, $\|v\|_1 = |v_1| + \dots + |v_p|$, and $\|v\|$ denotes the Euclidean norm of v, that is, $\|v\| = \sqrt{v'v}$.

We say that a class of functions $\mathcal{F} = \{f(\cdot, t)\colon t \in T\}$, where $f\colon \mathcal{W} \times T \to \mathbb{R}$, is suitably measurable if it is an image admissible Suslin class, as defined in [22], p 186. In particular, $\mathcal{F}$ is suitably measurable if $f\colon \mathcal{W} \times T \to \mathbb{R}$ is measurable and T is a Polish space equipped with its Borel σ-field, see [22], p 186.

Appendix B. A Bound on Sparse Eigenvalues for Many Random Matrices

The following lemma is a generalization of the main result in [41] to many matrices. The proof of the lemma is given in the Supplementary Material.

Lemma B.1. *Let $\mathcal{U}$ denote a finite set and $(X_{ui})_{u\in\mathcal{U}}$, $i = 1, \dots, n$, be independent (across i) random vectors such that $X_{ui} \in \mathbb{R}^p$ with $p \geqslant 2$ and $(\mathrm{E}[\max_{1\leqslant i\leqslant n}\max_{u\in\mathcal{U}}\|X_{ui}\|_\infty^2])^{1/2} \leqslant K$. Furthermore, for $k \geqslant 1$, define*

$$\delta_n := \frac{K\sqrt{k}}{\sqrt{n}}\left(\log^{1/2}|\mathcal{U}| + \log^{1/2} p + (\log k)(\log^{1/2} p)(\log^{1/2} n)\right),$$

Then,

$$\mathrm{E}\left[\sup_{\|\theta\|_0\leqslant k, \|\theta\|=1}\max_{u\in\mathcal{U}}\left|\mathbb{E}_n\left[(\theta' X_u)^2 - \mathrm{E}[(\theta' X_u)^2]\right]\right|\right] \lesssim \delta_n^2 + \delta_n \sup_{\|\theta\|_0\leqslant k, \|\theta\|=1, u\in\mathcal{U}}\sqrt{\mathbb{E}_n\mathrm{E}[(\theta' X_u)^2]}$$

up-to a universal constant.

Appendix C. Proofs for Section 2

In this appendix, we use C to denote a strictly positive constant that is independent of n and $P \in \mathcal{P}_n$. The value of C may change at each appearance. Also, the notation $a_n \lesssim b_n$ means that $a_n \leqslant C b_n$ for all n and some C. The notation $a_n \gtrsim b_n$ means that $b_n \lesssim a_n$. Moreover, the notation $a_n = o(1)$ means that there exists a sequence $(b_n)_{n\geqslant 1}$ of positive numbers such that (i) $|a_n| \leqslant b_n$ for all n, (ii) b_n is independent of $P \in \mathcal{P}_n$ for all n, and (iii) $b_n \to 0$ as $n \to \infty$. Finally, the notation $a_n = O_P(b_n)$ means that for all $\epsilon > 0$, there exists C such that $\mathrm{P}_P(a_n > C b_n) \leqslant 1 - \epsilon$ for all n. Using this notation allows us to avoid repeating "uniformly over $P \in \mathcal{P}_n$" many times in the proofs of Theorem 2.1 and Corollaries 2.1 and 2.2. Throughout this appendix, we assume that $n \geqslant n_0$.

Proof of Theorem 2.1. We split the proof into five steps.

Step 1. (Preliminary Rate Result). We claim that with probability $1-o(1)$,

$$\sup_{u\in\mathcal{U},j\in[\tilde{p}]}|\check{\theta}_{uj}-\theta_{uj}|\lesssim B_{1n}\tau_n.$$

By definition of $\check{\theta}_{uj}$, we have for each $u\in\mathcal{U}$ and $j\in[\tilde{p}]$,

$$\Big|\mathbb{E}_n[\psi_{uj}(W,\check{\theta}_{uj},\widehat{\eta}_{uj})]\Big|\leqslant\inf_{\theta\in\Theta_{uj}}\Big|\mathbb{E}_n[\psi_{uj}(W,\theta,\widehat{\eta}_{uj})]\Big|+\epsilon_n,$$

which implies via the triangle inequality that uniformly over $u\in\mathcal{U}$ and $j\in[\tilde{p}]$, with probability $1-o(1)$,

$$\Big|\ \mathrm{E}_P[\psi_{uj}(W,\theta,\eta_{uj})]|_{\theta=\check{\theta}_{uj}}\ \Big|\leqslant\epsilon_n+2I_1+2I_2\lesssim B_{1n}\tau_n,\quad\text{where}\tag{C.1}$$

$$I_1:=\sup_{u\in\mathcal{U},j\in[\tilde{p}],\theta\in\Theta_{uj}}\Big|\mathbb{E}_n[\psi_{uj}(W,\theta,\widehat{\eta}_{uj})]-\mathbb{E}_n[\psi_{uj}(W,\theta,\eta_{uj})]\Big|\lesssim B_{1n}\tau_n,$$
$$I_2:=\sup_{u\in\mathcal{U},j\in[\tilde{p}],\theta\in\Theta_{uj}}\Big|\mathbb{E}_n[\psi_{uj}(W,\theta,\eta_{uj})]-\mathrm{E}_P[\psi_{uj}(W,\theta,\eta_{uj})]\Big|\lesssim\tau_n.$$

and the bounds on I_1 and I_2 are derived in Step 2 (note also that $\epsilon_n=o(\tau_n)$ by construction of the estimator and Assumption 2.2(vi)). Since by Assumption 2.1(iv), $2^{-1}|J_{uj}(\check{\theta}_{uj}-\theta_{uj})|\wedge c_0$ does not exceed the left-hand side of (C.1), $\inf_{u\in\mathcal{U},j\in[\tilde{p}]}|J_{uj}|\gtrsim 1$, and by Assumption 2.2(vi), $B_{1n}\tau_n=o(1)$, we conclude that

$$\sup_{u\in\mathcal{U},j\in[\tilde{p}]}|\check{\theta}_{uj}-\theta_{uj}|\lesssim\left(\inf_{u\in\mathcal{U},j\in[\tilde{p}]}|J_{uj}|\right)^{-1}B_{1n}\tau_n\lesssim B_{1n}\tau_n,\tag{C.2}$$

with probability $1-o(1)$ yielding the claim of this step.

Step 2. (Bounds on I_1 and I_2) We claim that with probability $1-o(1)$,

$$I_1\lesssim B_{1n}\tau_n\ \text{ and }\ I_2\lesssim\tau_n.$$

To show these relations, observe that with probability $1-o(1)$, we have $I_1\leqslant 2I_{1a}+I_{1b}$ and $I_2\leqslant I_{1a}$, where

$$I_{1a}:=\sup_{u\in\mathcal{U},j\in[\tilde{p}],\theta\in\Theta_{uj},\eta\in\mathcal{T}_{uj}}\Big|\mathbb{E}_n[\psi_{uj}(W,\theta,\eta)]-\mathrm{E}_P[\psi_{uj}(W,\theta,\eta)]\Big|,$$
$$I_{1b}:=\sup_{u\in\mathcal{U},j\in[\tilde{p}],\theta\in\Theta_{uj},\eta\in\mathcal{T}_{uj}}\Big|\mathrm{E}_P[\psi_{uj}(W,\theta,\eta)]-\mathrm{E}_P[\psi_{uj}(W,\theta,\eta_{uj})]\Big|.$$

To bound I_{1b}, we employ Taylor's expansion:

$$\begin{aligned}I_{1b}&\leqslant\sup_{u\in\mathcal{U},j\in[\tilde{p}],\theta\in\Theta_{uj},\eta\in\mathcal{T}_{uj},r\in[0,1)}\partial_r\mathrm{E}_P\Big[\psi_{uj}(W,\theta,\eta_{uj}+r(\eta-\eta_{uj}))\Big]\\&\leqslant B_{1n}\sup_{u\in\mathcal{U},j\in[\tilde{p}],\eta\in\mathcal{T}_{uj}}\|\eta-\eta_{uj}\|_e\leqslant B_{1n}\tau_n,\end{aligned}$$

by Assumptions 2.1(v) and 2.2(ii).

To bound I_{1a}, we apply the maximal inequality of Lemma L.2 to the class $\mathcal{F}_1$ defined in Assumption 2.2 to conclude that with probability $1-o(1)$,

$$I_{1a} \lesssim n^{-1/2}\Big(\sqrt{v_n \log a_n} + n^{-1/2+1/q} v_n K_n \log a_n\Big). \tag{C.3}$$

Here we used: $\log \sup_Q N(\epsilon\|F_1\|_{Q,2}, \mathcal{F}_1, \|\cdot\|_{Q,2}) \leqslant v_n \log(a_n/\epsilon)$ for all $0<\epsilon\leqslant 1$ with $\|F_1\|_{P,q}\leqslant K_n$ by Assumption 2.2(iv); $\sup_{f\in\mathcal{F}_1}\|f\|^2_{P,2}\leqslant C_0$ by Assumption 2.2(v); $a_n \geqslant n \vee K_n$ and $v_n\geqslant 1$ by the choice of a_n and v_n. In turn, the right-hand side of (C.3) is bounded from above by $O(\tau_n)$ by Assumption 2.2(vi) since $(v_n \log a_n/n)^{1/2}\lesssim \tau_n$ and

$$n^{-1+1/q} v_n K_n \log a_n = n^{-1/2} n^{-1/2+1/q} v_n K_n \log a_n \lesssim n^{-1/2}\delta_n \lesssim n^{-1/2} \lesssim \tau_n.$$

Combining presented bounds gives the claim of this step.

Step 3. (Linearization) Here we prove the claim of the theorem. Fix $u\in\mathcal{U}$ and $j\in[\tilde p]$. By definition of $\check\theta_{uj}$, we have

$$\sqrt{n}\Big|\mathbb{E}_n[\psi_{uj}(W,\check\theta_{uj},\widehat\eta_{uj})]\Big| \leqslant \inf_{\theta\in\Theta_{uj}} \sqrt{n}\Big|\mathbb{E}_n[\psi_{uj}(W,\theta,\widehat\eta_{uj})]\Big| + \epsilon_n\sqrt{n}. \tag{C.4}$$

Also, for any $\theta\in\Theta_{uj}$ and $\eta\in\mathcal{T}_{uj}$, we have

$$\begin{aligned}\sqrt{n}\mathbb{E}_n[\psi_{uj}(W,\theta,\eta)] &= \sqrt{n}\mathbb{E}_n[\psi_{uj}(W,\theta_{uj},\eta_{uj})] - \mathbb{G}_n\psi_{uj}(W,\theta_{uj},\eta_{uj}) \\ &\quad - \sqrt{n}\Big(\mathrm{E}_P[\psi_{uj}(W,\theta_{uj},\eta_{uj})] - \mathrm{E}_P[\psi_{uj}(W,\theta,\eta)]\Big) + \mathbb{G}_n\psi_{uj}(W,\theta,\eta).\end{aligned} \tag{C.5}$$

Moreover, by Taylor's expansion of the function $r\mapsto \mathrm{E}_P[\psi_{uj}(W,\theta_{uj}+r(\theta-\theta_{uj}),\eta_{uj}+r(\eta-\eta_{uj}))]$,

$$\begin{aligned}&\mathrm{E}_P[\psi_{uj}(W,\theta,\eta)] - \mathrm{E}_P[\psi_{uj}(W,\theta_{uj},\eta_{uj})] \\ &\quad = J_{uj}(\theta-\theta_{uj}) + \mathrm{D}_{u,j,0}[\eta-\eta_{uj}] + \partial_r^2 \mathrm{E}_P[W,\theta_{uj}+r(\theta-\theta_{uj}),\eta_{uj}+r(\eta-\eta_{uj})]\big|_{r=\bar r}\end{aligned} \tag{C.6}$$

for some $\bar r\in(0,1)$. Substituting this equality into (C.5), taking $\theta=\check\theta_{uj}$ and $\eta=\eta_{uj}$, and using (C.4) gives

$$\begin{aligned}&\sqrt{n}\Big|\mathbb{E}_n[\psi_{uj}(W,\theta_{uj},\eta_{uj})] + J_{uj}(\check\theta_{uj}-\theta_{uj}) + \mathrm{D}_{u,j,0}[\widehat\eta_{uj}-\eta_{uj}]\Big| \\ &\quad \leqslant \epsilon_n\sqrt{n} + \inf_{\theta\in\Theta_{uj}}\sqrt{n}|\mathbb{E}_n[\psi_{uj}(W,\theta,\widehat\eta_{uj})]| + |II_1(u,j)| + |II_2(u,j)|,\end{aligned} \tag{C.7}$$

where

$$II_1(u,j) := \sqrt{n}\sup_{r\in[0,1)}\left|\partial_r^2\mathrm{E}_P\Big[\psi_{uj}(W,\theta_{uj}+r(\theta-\theta_{uj}),\eta_{uj}+r\{\eta-\eta_{uj}\})\Big]\Big|_{\theta=\check\theta_{uj},\eta=\widehat\eta_{uj}}\right|,$$

$$II_2(u,j) := \mathbb{G}_n\Big(\psi_{uj}(W,\theta,\eta)-\psi_{uj}(W,\theta_{uj},\eta_{uj})\Big)\Big|_{\theta=\check\theta_{uj},\eta=\widehat\eta_{uj}}.$$

It will be shown in Step 4 that

$$\sup_{u\in\mathcal{U},j\in[\tilde p]}\Big(|II_1(u,j)|+|II_2(u,j)|\Big) = O_P(\delta_n). \tag{C.8}$$

In addition, it will be shown in Step 5 that

$$\sup_{u\in\mathcal{U},j\in[\tilde p]}\inf_{\theta\in\Theta_{uj}}\sqrt{n}|\mathbb{E}_n[\psi_{uj}(W,\theta,\widehat\eta_{uj})]| = O_P(\delta_n). \tag{C.9}$$

Moreover, $\epsilon_n\sqrt{n} = o(\delta_n)$ by construction of the estimator. Therefore, the expression in (C.7) is $O_P(\delta_n)$. Further,

$$\sup_{u\in\mathcal{U},j\in[\tilde p]} \left|\mathrm{D}_{u,j,0}[\widehat\eta_{uj} - \eta_{uj}]\right| = O_P(\delta_n n^{-1/2})$$

by the near orthogonality condition since $\widehat\eta_{uj} \in \mathcal{T}_{uj}$ for all $u \in \mathcal{U}$ and $j \in [\tilde p]$ with probability $1 - o(1)$ by Assumption 2.2(i). Therefore, Assumption 2.1(iv) gives

$$\sup_{u\in\mathcal{U},j\in[\tilde p]} \left|J_{uj}^{-1}\sqrt{n}\mathbb{E}_n[\psi_{uj}(W,\theta_{uj},\eta_{uj})] + \sqrt{n}(\check\theta_{uj} - \theta_{uj})\right| = O_P(\delta_n).$$

The asserted claim now follows by dividing both parts of the display above by σ_{uj} (under the supremum on the left-hand side) and noting that σ_{uj} is bounded below from zero uniformly over $u \in \mathcal{U}$ and $j \in [\tilde p]$ by Assumptions 2.2(iii) and 2.2(v).

Step 4. (Bounds on $II_1(u,j)$ and $II_2(u,j)$). Here we prove (C.8). First, with probability $1 - o(1)$,

$$\sup_{u\in\mathcal{U},j\in[\tilde p]} |II_1(u,j)| \leqslant \sqrt{n}B_{2n} \sup_{u\in\mathcal{U},j\in[\tilde p]} |\check\theta_{uj} - \theta_{uj}|^2 \vee \|\widehat\eta_{uj} - \eta_{uj}\|_e^2 \lesssim \sqrt{n}B_{1n}^2B_{2n}\tau_n^2 \lesssim \delta_n,$$

where the first inequality follows from Assumptions 2.1(v) and 2.2(i), the second from Step 1 and Assumptions 2.2(ii) and 2.2(vi), and the third from Assumption 2.2(vi).

Second, with probability $1 - o(1)$,

$$\sup_{u\in\mathcal{U},j\in[\tilde p]} |II_2(u,j)| \lesssim \sup_{f\in\mathcal{F}_2} |\mathbb{G}_n(f)|$$

where

$$\mathcal{F}_2 = \left\{\psi_{uj}(\cdot,\theta,\eta) - \psi_{uj}(\cdot,\theta_{uj},\eta_{uj})\colon u\in\mathcal{U}, j\in[\tilde p], \eta\in\mathcal{T}_{uj}, |\theta - \theta_{uj}| \leqslant CB_{1n}\tau_n\right\}$$

for sufficiently large constant C. To bound $\sup_{f\in\mathcal{F}_2}|\mathbb{G}_n(f)|$, we apply Lemma L.2. Observe that

$$\begin{aligned}\sup_{f\in\mathcal{F}_2}\|f\|_{P,2}^2 &\leqslant \sup_{u\in\mathcal{U},j\in[\tilde p],|\theta-\theta_{uj}|\leqslant CB_{1n}\tau_n,\eta\in\mathcal{T}_{uj}} \mathrm{E}_P\left[(\psi_{uj}(W,\theta,\eta) - \psi_{uj}(W,\theta_{uj},\eta_{uj}))^2\right]\\ &\leqslant \sup_{u\in\mathcal{U},j\in[\tilde p],|\theta-\theta_{uj}|\leqslant CB_{1n}\tau_n,\eta\in\mathcal{T}_{uj}} C_0(|\theta-\theta_{uj}| \vee \|\eta-\eta_{uj}\|_e)^\omega \lesssim (B_{1n}\tau_n)^\omega,\end{aligned}$$

where we used Assumption 2.1(v) and Assumption 2.2(ii). Also, observe that $(B_{1n}\tau_n)^{\omega/2} \geqslant n^{-\omega/4}$ by Assumption 2.2(vi) since $B_{1n} \geqslant 1$. Therefore, an application of Lemma L.2 with an envelope $F_2 = 2F_1$ and $\sigma = (CB_{1n}\tau_n)^{\omega/2}$ for sufficiently large constant C gives with probability $1 - o(1)$,

$$\sup_{f\in\mathcal{F}_2}|\mathbb{G}_n(f)| \lesssim (B_{1n}\tau_n)^{\omega/2}\sqrt{v_n\log a_n} + n^{-1/2+1/q}v_nK_n\log a_n, \tag{C.10}$$

since $\sup_{f\in\mathcal{F}_2}|f| \leqslant 2\sup_{f\in\mathcal{F}_1}|f| \leqslant 2F_1$ and $\|F_1\|_{P,q} \leqslant K_n$ by Assumption 2.2(iv) and

$$\log\sup_Q N(\epsilon\|F_2\|_{Q,2},\mathcal{F}_2,\|\cdot\|_{Q,2}) \lesssim v_n\log(a_n/\epsilon), \quad \text{for all } 0 < \epsilon \leqslant 1$$

by Lemma K.1 because $\mathcal{F}_2 \subset \mathcal{F}_1 - \mathcal{F}_1$ for $\mathcal{F}_1$ defined in Assumption 2.2(iv). The claim of this step now follows from an application of Assumption 2.2(vi) to bound the right-hand side of (C.10).

Step 5. Here we prove (C.9). For all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, let $\bar{\theta}_{uj} = \theta_{uj} - J_{uj}^{-1}\mathbb{E}_n[\psi_{uj}(W, \theta_{uj}, \eta_{uj})]$. Then $\sup_{u\in\mathcal{U}, j\in[\tilde{p}]} |\bar{\theta}_{uj} - \theta_{uj}| = O_P(\mathrm{u_n}/\sqrt{n})$ since $\mathrm{u_n} = \mathrm{E}_P[\sup_{u\in\mathcal{U},j\in[\tilde{p}]} |\sqrt{n}\mathbb{E}_n[\psi_{uj}(W_{uj}, \theta_{uj}, \eta_{uj})]|]$ and J_{uj} is bounded in absolute value below from zero uniformly over $u \in \mathcal{U}$ and $j \in [\tilde{p}]$ by Assumption 2.1(iv). Therefore, $\check{\theta}_{uj} \in \Theta_{uj}$ for all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$ with probability $1 - o(1)$ by Assumption 2.1(i). Hence, with the same probability, for all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$,

$$\inf_{\theta\in\Theta_{uj}} \sqrt{n}\Big|\mathbb{E}_n[\psi_{uj}(W, \theta, \widehat{\eta}_{uj})]\Big| \leqslant \sqrt{n}\Big|\mathbb{E}_n[\psi_{uj}(W, \bar{\theta}_{uj}, \widehat{\eta}_{uj})]\Big|,$$

and so it suffices to show that

$$\sup_{u\in\mathcal{U},j\in[\tilde{p}]} \sqrt{n}\Big|\mathbb{E}_n[\psi_{uj}(W, \bar{\theta}_{uj}, \widehat{\eta}_{uj})]\Big| = O_P(\delta_n). \tag{C.11}$$

To prove (C.11), for given $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, substitute $\theta = \bar{\theta}_{uj}$ and $\eta = \widehat{\eta}_{uj}$ into (C.5) and use Taylor's expansion in (C.6). This gives

$$\begin{aligned}\sqrt{n}\Big|\mathbb{E}_n[\psi_{uj}(W, \bar{\theta}_{uj}, \widehat{\eta}_{uj})]\Big| \leqslant \sqrt{n}\Big|\mathbb{E}_n[\psi_{uj}(W, \theta_{uj}, \eta_{uj})] + J_{uj}(\bar{\theta}_{uj} - \theta_{uj}) + \mathrm{D}_{u,j,0}[\widehat{\eta}_{uj} - \eta_{uj}]\Big| \\ + |\widetilde{II}_1(u, j)| + |\widetilde{II}_2(u, j)|\end{aligned}$$

where $\widetilde{II}_1(u, j)$ and $\widetilde{II}_2(u, j)$ are defined as $II_1(u, j)$ and $II_2(u, j)$ in Step 3 but with $\check{\theta}_{uj}$ replaced by $\bar{\theta}_{uj}$. Then, given that $\sup_{u\in\mathcal{U},j\in[\tilde{p}]} |\bar{\theta}_{uj} - \theta_{uj}| \lesssim \mathrm{u_n} \log n/\sqrt{n}$ with probability $1 - o(1)$, the argument in Step 4 shows that

$$\sup_{u\in\mathcal{U},j\in[\tilde{p}]} \Big(|\widetilde{II}_1(u, j)| + |\widetilde{II}_2(u, j)|\Big) = O_P(\delta_n).$$

In addition,

$$\mathbb{E}_n[\psi_{uj}(W, \theta_{uj}, \eta_{uj})] + J_{uj}(\bar{\theta}_{uj} - \theta_{uj}) = 0$$

by the definition of $\bar{\theta}_{uj}$ and $\sup_{u\in\mathcal{U},j\in[\tilde{p}]} |\mathrm{D}_{u,j,0}[\widehat{\eta}_{uj} - \eta_{uj}]| = O_P(\delta_n n^{-1/2})$ by the near orthogonality condition. Combining these bounds gives (C.11), so that the claim of this step follows, and completes the proof of the theorem. ■

Proof of Corollary 2.1. To prove the asserted claim, we will apply Lemma 2.4 in [18]. Denote

$$Z_n = \sup_{u\in\mathcal{U},j\in[\tilde{p}]} \Big|n^{1/2}\sigma_{uj}^{-1}(\check{\theta}_{uj} - \theta_{uj})\Big|.$$

Under our assumptions, $L_n^{2/7}\varrho_n \log A_n = o(n^{1/7})$, and so given that $\varrho_n \geqslant 1$ and $A_n \geqslant n$, it follows that $\log L_n \lesssim \log n$. Hence, since $\mathrm{E}_P[\bar{\psi}_{uj}^2(W)] = 1$, Assumption 2.3(i) and Corollary 2.2.8 in [44] imply that

$$\mathrm{E}_P\left[\sup_{u\in\mathcal{U},j\in[\tilde{p}]} |\mathcal{N}_{uj}|\right] \lesssim \sqrt{\varrho_n \log(A_n L_n)} \lesssim \sqrt{\varrho_n \log A_n}. \tag{C.12}$$

Further, Theorem 2.1 shows that

$$\left|Z_n - \sup_{u\in\mathcal{U},j\in[\tilde{p}]} |\mathbb{G}_n\bar{\psi}_{uj}|\right| = O_P(\delta_n), \tag{C.13}$$

and Theorem 2.1 in [20], together with Assumptions 2.3(i,ii), shows that one can construct a version $\widetilde{Z}_n$ of $\sup_{u\in\mathcal{U},j\in[\tilde{p}]}|\mathcal{N}_{uj}|$ such that

$$\left|\sup_{u\in\mathcal{U},j\in[\tilde{p}]}|\mathbb{G}_n\bar{\psi}_{uj}|-\widetilde{Z}_n\right|=O_P\left(\frac{L_n\varrho_n\log A_n}{n^{1/2-1/q}}+\frac{L_n^{1/3}(\varrho_n\log A_n)^{2/3}}{n^{1/6}}\right). \tag{C.14}$$

Combining (C.13) and (C.14) gives

$$|Z_n-\widetilde{Z}_n|=O_P\left(\delta_n+\frac{L_n\varrho_n\log A_n}{n^{1/2-1/q}}+\frac{L_n^{1/3}(\varrho_n\log A_n)^{2/3}}{n^{1/6}}\right). \tag{C.15}$$

Therefore, it follows from Lemma 2.4 in [18] that (C.12) and (C.15) imply

$$\sup_{t\in\mathbb{R}}\left|\mathrm{P}_P(Z_n\leqslant t)-\mathrm{P}_P(\widetilde{Z}_n\leqslant t)\right|=o(1) \tag{C.16}$$

under our growth conditions $\delta_n^2\rho_n\log A_n=o(1)$, $L_n^{2/7}\rho_n\log A_n=o(n^{1/7})$, and $L_n^{2/3}\rho_n\log A_n=o(n^{1/3-2/(3q)})$; note that formally their Lemma 2.4 requires Z_n to be the supremum of an empirical process but this requirement is not used in the proof. The asserted claim now follows by substituting the definitions of Z_n and $\widetilde{Z}_n$. ■

Proof of Corollary 2.2. Denote $\widetilde{Z}_n^*=\sup_{u\in\mathcal{U},j\in[\tilde{p}]}|\widehat{\mathcal{G}}_{uj}|$. For all $\vartheta\in(0,1)$, let c_ϑ^0 be the $(1-\vartheta)$ quantile of $\sup_{u\in\mathcal{U},j\in[\tilde{p}]}|\mathcal{N}_{uj}|$. We proceed in several steps.

Step 1. Here we show that c_ϑ^0 satisfies the bound

$$c_\vartheta^0\leqslant \mathrm{E}_P\left[\sup_{u\in\mathcal{U},j\in[\tilde{p}]}|\mathcal{N}_{uj}|\right]+\sqrt{2\log(1/\vartheta)}$$

for all $\vartheta\in(0,1)$. Indeed, recall that $\mathrm{E}_P[\mathcal{N}_{uj}^2]=1$ for all $u\in\mathcal{U}$ and $j\in[\tilde{p}]$. Therefore, this bound follows from Borell's inequality; see Proposition A.2.1 in [44].

Step 2. Here we show that for any $\vartheta\in(0,1)$ and $\beta\in(0,\vartheta)$,

$$c_{\vartheta-\beta}^0-c_\vartheta^0\geqslant c\beta\Big/\mathrm{E}_P\left[\sup_{u\in\mathcal{U},j\in[\tilde{p}]}|\mathcal{N}_{uj}|\right]$$

for some absolute constant $c>0$. Indeed, given that $\mathrm{E}_P[\mathcal{N}_{uj}^2]=1$ for all $u\in\mathcal{U}$ and $j\in[\tilde{p}]$, this bound follows from Corollary 2.1 in [16].

Step 3. Here we show that

$$\left|\widetilde{Z}_n^*-\sup_{u\in\mathcal{U},j\in[\tilde{p}]}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\xi_i\bar{\psi}_{uj}(W_i)\right|\right|=O_P\left(\bar{\delta}_n\sqrt{\bar{\varrho}_n\log\bar{A}_n}\right). \tag{C.17}$$

Indeed, the left-hand side of (C.17) is bounded from above by

$$\sup_{u\in\mathcal{U},j\in[\tilde{p}]}\left|\widehat{\mathcal{G}}_{uj}-\frac{1}{\sqrt{n}}\sum_{i=1}^n\xi_i\bar{\psi}_{uj}(W_i)\right|=\sup_{u\in\mathcal{U},j\in[\tilde{p}]}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\xi_i(\widehat{\psi}_{uj}(W_i)-\bar{\psi}_{uj}(W_i))\right|.$$

Conditional on $(W_i)_{i=1}^n$, $n^{-1/2}\sum_{i=1}^n \xi_i(\widehat{\psi}_{uj}(W_i) - \bar{\psi}_{uj}(W_i))$ is zero-mean Gaussian with variance $\mathbb{E}_n[(\widehat{\psi}_{uj}(W_i) - \bar{\psi}_{uj}(W_i))^2] \leqslant \bar{\delta}_n^2$ with probability at least $1-\Delta_n$ by Assumption 2.3(iii). Thus, with the same probability,

$$\mathrm{E}_P\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i(\widehat{\psi}_{uj}(W_i) - \bar{\psi}_{uj}(W_i)) \mid (W_i)_{i=1}^n\right] \lesssim \bar{\delta}_n\sqrt{\bar{\varrho}_n \log \bar{A}_n}$$

by Assumption 2.3(iii) and Corollary 2.2.8 in [44]. Since $\Delta_n \to 0$, (C.17) follows.

Step 4. Here we show that one can construct a version $\widetilde{Z}_n$ of $\sup_{u\in\mathcal{U},j\in[\tilde{p}]}|\mathcal{N}_{uj}|$ such that

$$\left|\sup_{u\in\mathcal{U},j\in[\tilde{p}]}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i\bar{\psi}_{uj}(W_i)\right| - \widetilde{Z}_n\right| = O_P\left(\frac{L_n\varrho_n\log A_n}{n^{1/2-1/q}} + \frac{L_n^{1/2}(\varrho_n\log A_n)^{3/4}}{n^{1/4}}\right).$$

Indeed, as was discussed in the proof of Corollary 2.1, we have $\log L_n \lesssim \log n$ and $\mathrm{E}_P[\bar{\psi}_{uj}^2(W_i)] = 1$. Therefore, the claim follows from Theorem 2.2 in [20] combined with Assumption 2.3(i,ii).

Step 5. Here we show that there exists a sequence of positive constants $(\vartheta_n)_{n\geqslant 1}$ such that $\vartheta_n \to 0$ and $\mathrm{P}_P\Big(c_\alpha(1+\varepsilon_n) > c^0_{\alpha-\vartheta_n}\Big) \to 0$. Indeed, Steps 3 and 4 imply that $|\widetilde{Z}_n^* - \widetilde{Z}_n| = O_P(r_n)$ where

$$r_n \lesssim \bar{\delta}_n\sqrt{\bar{\varrho}_n\log\bar{A}} + \frac{L_n\varrho_n\log A_n}{n^{1/2-1/q}} + \frac{L_n^{1/2}(\varrho_n\log A_n)^{3/4}}{n^{1/4}}.$$

Also, as in the proof of Corollary 2.1, $\mathrm{E}_P[\widetilde{Z}_n] \lesssim (\varrho_n\log A_n)^{1/2}$. Hence, $r_n\mathrm{E}_P[\widetilde{Z}_n] = o(1)$ under our conditions, and so there exists a sequence of positive constants $(\chi_n)_{n\geqslant 1}$ such that $\chi_n \to \infty$ but $\chi_n r_n \mathrm{E}_P[\widetilde{Z}_n] = o(1)$. Further, let $\beta_n = \{\mathrm{P}_P(|\widetilde{Z}_n^* - \widetilde{Z}_n| > \chi_n r_n)\}^{1/2}$. Observe that $\beta_n = o(1)$ and

$$\mathrm{P}_P\left(\mathrm{P}_P\Big(|\widetilde{Z}_n^* - \widetilde{Z}_n| > \chi_n r_n \mid (W_i)_{i=1}^n\Big) > \beta_n\right) \leqslant \beta_n.$$

The last display implies that with probability at least $1-\beta_n$, $c_\alpha \leqslant c^0_{\alpha-\beta_n} + \chi_n r_n$. Hence, with the same probability, using the bounds in Steps 1 and 2, we obtain for some sequence of positive constants $(\vartheta_n)_{n\geqslant 1}$ such that $\vartheta_n = o(1)$,

$$c_\alpha(1+\varepsilon_n) \leqslant (c^0_{\alpha-\beta_n} + \chi_n r_n)(1+\varepsilon_n) \leqslant c^0_{\alpha-\vartheta_n}$$

since $\chi_n r_n \mathrm{E}_P[\widetilde{Z}_n] = o(1)$ and $\varepsilon_n(\mathrm{E}_P[\widetilde{Z}_n])^2 = o(1)$ by assumption. The claim of this step follows.

Step 6. Here we complete the proof. We have

$$\begin{aligned}
\mathrm{P}_P\Big(\sup_{u\in\mathcal{U},j\in[\tilde{p}]}|\sqrt{n}\widehat{\sigma}_{uj}^{-1}(\check{\theta}_{uj} - \theta_{uj})| \leqslant c_\alpha\Big) &= \mathrm{P}_P\Big(|\sqrt{n}\sigma_{uj}^{-1}(\check{\theta}_{uj} - \theta_{uj})| \leqslant c_\alpha\widehat{\sigma}_{uj}/\sigma_{uj},\ \forall u\in\mathcal{U}, j\in[\tilde{p}]\Big)\\
&\leqslant \mathrm{P}_P\Big(\sup_{u\in\mathcal{U},j\in[\tilde{p}]}|\sqrt{n}\sigma_{uj}^{-1}(\check{\theta}_{uj} - \theta_{uj})| \leqslant c_\alpha(1+\varepsilon_n)\Big) + o(1)\\
&\leqslant \mathrm{P}_P\Big(\sup_{u\in\mathcal{U},j\in[\tilde{p}]}|\sqrt{n}\sigma_{uj}^{-1}(\check{\theta}_{uj} - \theta_{uj})| \leqslant c^0_{\alpha-\vartheta_n}\Big) + o(1)\\
&= 1-\alpha+\vartheta_n + o(1) = 1-\alpha+o(1)
\end{aligned}$$

where the third line follows by Assumption 2.4, the fourth by Step 5, and the fifth by Corollary 2.1. Similar arguments also give the same bound from the other side. Therefore,

$$\mathrm{P}_P\left(\sup_{u\in\mathcal{U},j\in[\tilde{p}]}|\sqrt{n}\widehat{\sigma}_{uj}^{-1}(\check{\theta}_{uj}-\theta_{uj})|\leqslant c_\alpha\right)=1-\alpha+o(1). \tag{C.18}$$

This completes the proof. ■

Appendix D. Proofs for Sections 3 and 4

See Supplementary Material.

References

[1] Andrews, D.W.K. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica*, 62(1):43–72.

[2] Belloni, A., Chen, D, Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80:2369–2429. ArXiv, 2010.

[3] Belloni, A. and Chernozhukov, V. (2011). ℓ_1-penalized quantile regression for high dimensional sparse models. *Annals of Statistics*, 39(1):82–130. ArXiv, 2009.

[4] Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547. ArXiv, 2009.

[5] Belloni, A., Chernozhukov, V., and Hansen, C. (2010). Lasso methods for gaussian instrumental variables models. ArXiv, 2010.

[6] Belloni, A., Chernozhukov, V., and Hansen, C. (2013). Inference for high-dimensional sparse econometric models. *Advances in Economics and Econometrics. 10th World Congress of Econometric Society. August 2010*, III:245–295. ArXiv, 2011.

[7] Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection amongst high-dimensional controls. *Review of Economic Studies*, 81:608–650. ArXiv, 2011.

[8] Belloni, A., Chernozhukov, V., and Kato, K. (2013). Robust inference in high-dimensional approximately sparse quantile regression models. ArXiv, 2013.

[9] Belloni, A., Chernozhukov, V., and Kato, K. (2015). Uniform post selection inference for LAD regression models and other Z-estimators. *Biometrika*, (102):77–94. ArXiv, 2013.

[10] Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root-lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806. Arxiv, 2010.

[11] Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2013). Program evaluation with high-dimensional data. ArXiv, 2013.

[12] Belloni, A., Chernozhukov, V., and Wang, L. (2014). Pivotal estimation via square-root lasso in nonparametric regression. *The Annals of Statistics*, 42(2):757–788. ArXiv, 2013.

[13] Berlemann, M., Enkelmann, S., and Kuhlenkasper, T. (2014). Unraveling the relationship between presidential approval and the economy: a multidimensional semiparametric approach. *Journal of Applied Econometrics*.

[14] Bickel, P., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732. ArXiv, 2008.

[15] Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819. ArXiv, 2012.

[16] Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 42(5):1787–1818. ArXiv, 2013.

[17] Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Central limit theorems and bootstrap in high dimensions. ArXiv, 2014.

[18] Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564–1597. ArXiv, 2012.

[19] Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Comparison and anti-concentration bounds for maxima of gaussian random vectors. *Probability Theory and Related Fields*, 162:47–70.

[20] Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related gaussian couplings. ArXiv, 2015.

[21] Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013). Inference on counterfactual distributions *Econometrica*, 81:2205–2268.

[22] Dudley, R. (1999). *Uniform central limit theorems*, volume 63 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge.

[23] Hothorn, T., Kneib, T., and Bühlmann, P. (2014). Conditional transformation models. *J. R. Statist. Soc. B*, 76:3–27.

[24] Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, 15:2869–2909. ArXiv, 2013.

[25] Javanmard, A. and Montanari, A. (2014). Hypothesis testing in high-dimensional regression under the gaussian random design model: asymptotic theory. *IEEE Transactions on Information Theory*, 60:6522–6554. ArXiv, 2013.

[26] Jing, B.-Y., Shao, Q.-M., and Wang, Q. (2003). Self-normalized Cramer-type large deviations for independent random variables. *Ann. Probab.*, 31(4):2167–2215.

[27] Kosorok, M. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Series in Statistics. Springer, Berlin.

[28] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces (Isoperimetry and processes)*. Ergebnisse der Mathematik undihrer Grenzgebiete, Springer-Verlag.

[29] Leeb, H. and Pötscher, B. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(2):338–376.

[30] Leeb, H. and Pötscher, B. (2008). Recent developments in model selection and related areas. *Econometric Theory*, 24(2):319–322.

[31] Leeb, H. and Pötscher, B. (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *J. Econometrics*, 142(1):201–211.

[32] Linton, O. (1996). Edgeworth approximation for MINPIN estimators in semiparametric regression models. *Econometric Theory*, 12(1):30–60.

[33] Negahban, S., Ravikumar, P., Wainwright, P., and Yu, B. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557. ArXiv, 2010.

[34] Newey, W. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2):99–135.

[35] Newey, W. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6):1349–1382.

[36] Neyman, J. (1979). $c(\alpha)$ tests and their use. *Sankhya*, 41:1–21.

[37] Ning, Y. and Liu, H. (2014). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. ArXiv, 2014.

[38] Pisier, G. (1999). *The volume of convex bodies and Banach space geometry*, volume 94. Cambridge University Press.

[39] Pötscher, B. and Leeb, H. (2009). On the distribution of penalized maximum likelihood estimators: the LASSO, SCAD, and thresholding. *J. Multivariate Anal.*, 100(9):2065–2082.

[40] Robins, J. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.*, 90(429):122–129.

[41] Rudelson, M. and Vershynin, R. (2008). On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61:1025–1045.

[42] van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42:1166–1202. ArXiv, 2013.

[43] van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

[44] van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes*. Springer Series in Statistics.

[45] Zhang, C.-H, and Zhang, S. (2014). Confidence intervals for low-dimensional parameters with high-dimensional data. *J. R. Statist. Soc. B*, 76:217–242. ArXiv, 2012.

[46] Zhao, T., Kolar, M., and Liu, H. (2014). A general framework for robust testing and confidence regions in high-dimensional quantile regression. ArXiv, 2014.

Supplementary Material for "Uniformly Valid Post-Regularization Confidence Regions for Many Functional Parameters in Z-Estimation Framework"

Appendix E. Proofs for Section 3

In this appendix, we use c and C to denote strictly positive constants that depend only on c_1 and C_1 (but do not depend on n, u, j, or $P \in \mathcal{P}_n$). The values of c and C may change at each appearance. Also, the notation $a_n \lesssim b_n$ means that $a_n \leqslant Cb_n$ for all n and some C. The notation $a_n \gtrsim b_n$ means that $b_n \lesssim a_n$. Moreover, the notation $a_n = o(1)$ means that there exists a sequence $(b_n)_{n\geqslant 1}$ of positive numbers such that (i) $|a_n| \leqslant b_n$ for all n, (ii) b_n is independent of $P \in \mathcal{P}_n$ for all n, and (iii) $b_n \to 0$ as $n \to \infty$. Finally, the notation $a_n = o_P(b_n)$ means that for any C, we have $\mathrm{P}_P(a_n > Cb_n) = o(1)$. Using this notation allows us to avoid repeating "uniformly over $P \in \mathcal{P}_n$" and "uniformly over $u \in \mathcal{U}$ and $j \in [\tilde{p}]$" many times in the proofs of Theorem 3.1 and Corollaries 3.1 – 3.3.

Proof of Theorem 3.1. Observe that for all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, we have $\mathrm{E}_P[|Z_u^j|^2] \lesssim 1$ by Assumptions 3.1 and 3.4. We use this fact several times in the proof without further notice.

For $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, define

$$T_{uj} = \Big\{\eta = (\eta^{(1)}, \eta^{(2)}, \eta^{(3)}) \colon \eta^{(1)} \in \ell^\infty(\mathbb{R}^{\tilde{p}+p}), \eta^{(2)} \in \mathbb{R}^{\tilde{p}-1+p}, \eta^{(3)} \in \mathbb{R}^{\tilde{p}-1+p}\Big\},$$

so that $\eta_{uj} = (r_u, \beta_u^j, \gamma_u^j) \in T_{uj}$, and T_{uj} is convex. Endow T_{uj} with a norm $\|\cdot\|_e$ defined by

$$\|\eta\|_e = \sqrt{\mathrm{E}_P[\eta^{(1)}(D,X)^2]} \vee \|\eta^{(2)}\| \vee \|\eta^{(3)}\|, \quad \eta = (\eta^{(1)}, \eta^{(2)}, \eta^{(3)}) \in T_{uj}.$$

Further, recall that $a_n = p \vee \tilde{p} \vee n$ and define $\tau_n = C(s_n \log a_n/n)^{1/2}$ and

$$\mathcal{T}_{uj} = \{\eta_{uj}\} \cup \Big\{\eta = (\eta^{(1)}, \eta^{(2)}, \eta^{(3)}) \in T_{uj} \colon \eta^{(1)} = \mathcal{O}, \|\eta^{(2)}\|_0 \vee \|\eta^{(3)}\|_0 \leqslant Cs_n,$$
$$\|\eta^{(2)} - \beta_u^j\| \vee \|\eta^{(3)} - \gamma_u^j\| \leqslant \tau_n, \|\eta^{(3)} - \gamma_u^j\|_1 \leqslant C\sqrt{s_n}\tau_n\Big\}$$

for sufficiently large C.

First, we verify Assumption 2.1(i). To bound $\mathrm{u_n}$, below we establish the following inequality:

$$\|\gamma_{u_2}^j - \gamma_{u_1}^j\|_1 \lesssim \sqrt{p+\tilde{p}}|u_2 - u_1|. \tag{E.1}$$

Recall that

$$f_u^2 = f_u^2(D,X) = \mathrm{Var}(Y_u \mid D, X) = \mathrm{E}_P[Y_u \mid D, X](1 - \mathrm{E}_P[Y_u \mid D, X]),$$

and so

$$|f_{u_2}^2 - f_{u_1}^2| \leqslant \Big|\mathrm{E}_P[Y_{u_2} \mid D, X] - \mathrm{E}_P[Y_{u_1} \mid D, X]\Big| \lesssim |u_2 - u_1| \tag{E.2}$$

by Assumption 3.3. In addition,

$$\begin{aligned}
\mathrm{E}_P\Big[f_{u_2}^2\Big(X^j(\gamma_{u_1}^j-\gamma_{u_2}^j)\Big)^2\Big] &= \mathrm{E}_P\Big[f_{u_2}^2\Big(X^j(\gamma_{u_1}^j-\gamma_{u_2}^j)\Big)\Big(D_j-X^j\gamma_{u_2}^j-(D_j-X^j\gamma_{u_1}^j)\Big)\Big]\\
&= -\mathrm{E}_P\Big[f_{u_2}^2\Big(X^j(\gamma_{u_1}^j-\gamma_{u_2}^j)\Big)(D_j-X^j\gamma_{u_1}^j)\Big]\\
&= -\mathrm{E}_P\Big[(f_{u_2}^2-f_{u_1}^2)\Big(X^j(\gamma_{u_1}^j-\gamma_{u_2}^j)\Big)(D_j-X^j\gamma_{u_1}^j)\Big]\\
&= -\mathrm{E}_P\Big[(f_{u_2}^2-f_{u_1}^2)\Big(X^j(\gamma_{u_1}^j-\gamma_{u_2}^j)\Big)Z_{u_1}^j\Big] \qquad \text{(E.3)}
\end{aligned}$$

where the first line follows from adding and subtracting D_j, the second from the equality $\mathrm{E}_P[f_{u_2}^2(D_j-X^j\gamma_{u_2}^j)X^j]=0$, the third from the equality $\mathrm{E}_P[f_{u_1}^2(D_j-X^j\gamma_{u_1}^j)X^j]=0$, and the fourth from $D_j-X^j\gamma_{u_1}^j=Z_{u_1}^j$. Now, by the Cauchy-Schwarz inequality, the expression in (E.3) is bounded in absolute value by

$$\begin{aligned}
&\Big(\mathrm{E}_P\Big[\Big(X^j(\gamma_{u_1}^j-\gamma_{u_2}^j)\Big)^2\Big]\cdot\mathrm{E}_P\Big[(f_{u_1}^2-f_{u_2}^2)^2(Z_{u_1}^j)^2\Big]\Big)^{1/2}\\
&\quad\lesssim\Big(\mathrm{E}_P\Big[f_{u_2}^2\Big(X^j(\gamma_{u_1}^j-\gamma_{u_2}^j)\Big)^2\Big]\cdot\mathrm{E}_P\Big[(f_{u_1}^2-f_{u_2}^2)^2(Z_{u_1}^j)^2\Big]\Big)^{1/2}\\
&\quad\lesssim|u_2-u_1|\Big(\mathrm{E}_P\Big[f_{u_2}^2\Big(X^j(\gamma_{u_1}^j-\gamma_{u_2}^j)\Big)^2\Big]\Big)^{1/2}
\end{aligned}$$

where the second line follows from $\mathrm{E}_P[(X^j(\gamma_{u_1}^j-\gamma_{u_2}^j))^2]\lesssim\|\gamma_{u_1}^j-\gamma_{u_2}^j\|^2\lesssim\mathrm{E}_P[f_{u_2}^2(X^j(\gamma_{u_1}^j-\gamma_{u_2}^j))^2]$, which holds by Assumption 3.4, and the third line follows from (E.2). Hence, $(\mathrm{E}_P[f_{u_2}^2(X^j(\gamma_{u_1}^j-\gamma_{u_2}^j))^2])^{1/2}\lesssim|u_2-u_1|$, and so

$$\|\gamma_{u_2}^j-\gamma_{u_1}^j\|\lesssim\Big(\mathrm{E}_P\Big[f_{u_2}^2\Big(X^j(\gamma_{u_1}^j-\gamma_{u_2}^j)\Big)^2\Big]\Big)^{1/2}\lesssim|u_2-u_1|. \qquad \text{(E.4)}$$

Therefore,

$$\|\gamma_{u_2}^j-\gamma_{u_1}^j\|_1\leqslant\sqrt{p+\tilde p}\|\gamma_{u_2}^j-\gamma_{u_1}^j\|\lesssim\sqrt{p+\tilde p}|u_2-u_1|,$$

and so (E.1) follows.

Next, let

$$\begin{aligned}
\mathcal{G}_1&=\Big\{(Y,D,X)\mapsto 1\{Y\leqslant u\bar y+(1-u)\underline{y}\}\colon u\in\mathcal{U}\Big\},\\
\mathcal{G}_2&=\Big\{(Y,D,X)\mapsto \mathrm{E}_P[g(Y,D,X)\mid D,X]\colon g\in\mathcal{G}_1\Big\},\\
\mathcal{G}_{3,j}&=\Big\{(Y,D,X)\mapsto D_j-X^j\gamma_u^j\colon u\in\mathcal{U}\Big\},\quad j\in[\tilde p].
\end{aligned}$$

Then the function class $\widetilde{\mathcal{F}}=\{\psi_{uj}(\cdot,\theta_{uj},\eta_{uj})\colon u\in\mathcal{U},j\in[\tilde p]\}$ satisfies

$$\widetilde{\mathcal{F}}\subset(\mathcal{G}_1-\mathcal{G}_2)\cdot(\cup_{j\in[\tilde p]}\mathcal{G}_{3j}).$$

Observe that $\mathcal{G}_1$ is a VC-subgraph class with index bounded by C, and so by Theorem 2.6.7 in [44], its uniform entropy numbers obey

$$\sup_Q\log N(\epsilon\|\widetilde F_1\|_{Q,2},\mathcal{G}_1,\|\cdot\|_{Q,2})\leqslant C\log(C/\epsilon),\quad\text{for all }0<\epsilon\leqslant 1, \qquad \text{(E.5)}$$

where $\widetilde{F}_1 \equiv 1$ is its envelope. In addition, Lemma K.2 implies that the uniform entropy numbers of $\mathcal{G}_2$ obey the same inequalities with the same envelope $\widetilde{F}_1$ (but possibly different constant C). Moreover, for any $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, by Assumptions 3.1 and 3.2 and the triangle inequality,

$$\|\gamma_u^j\|_1 \lesssim \|\bar{\gamma}_u^j\|_1 + s_n\sqrt{\log a_n/n} \lesssim \sqrt{s_n}\|\bar{\gamma}_u^j\| + s_n\sqrt{\log a_n/n} \lesssim \sqrt{s_n}\|\gamma_u^j\| + s_n\sqrt{\log a_n/n} \lesssim \sqrt{s_n}, \quad \text{(E.6)}$$

because by Assumption 3.4, $s_n \log a_n/n = o(1)$. Therefore, (E.1) and Lemma K.3 with $k = 1$ imply that for all $j \in [\tilde{p}]$, the uniform entropy numbers of $\mathcal{G}_{3,j}$ obey

$$\sup_Q \log N(\epsilon\|\widetilde{F}_3\|_{Q,2}, \mathcal{G}_{3,j}, \|\cdot\|_{Q,2}) \leqslant C\log(a_n/\epsilon), \quad \text{for all } 0 < \epsilon \leqslant 1,$$

where $\widetilde{F}_3(Y, D, X) = \sup_{u\in\mathcal{U}} |Z_u^j| + M_{n,2}^{-1}(\|D\|_\infty \vee \|X\|_\infty)$ is its envelope, and so Lemma K.1 gives that the uniform entropy numbers of $\cup_{j\in[\tilde{p}]}\mathcal{G}_{3,j}$ obey the same inequalities with the same envelope $\widetilde{F}_3$. Hence, Lemma K.1 also shows that the uniform entropy numbers of $\widetilde{\mathcal{F}}$ obey

$$\sup_Q \log N(\epsilon\|\widetilde{F}\|_{Q,2}, \widetilde{\mathcal{F}}, \|\cdot\|_{Q,2}) \leqslant C\log(a_n/\epsilon), \quad \text{for all } 0 < \epsilon \leqslant 1, \quad \text{(E.7)}$$

where $\widetilde{F}(Y, D, X) = C\{\sup_{u\in\mathcal{U}} |Z_u^j| + M_{n,2}^{-1}(\|D\|_\infty \vee \|X\|_\infty\})$ is its envelope. Now observe that $\|\widetilde{F}\|_{P,q} \lesssim M_{n,1}$ and that $\|f\|_{P,2} \lesssim 1$ uniformly over $f \in \widetilde{\mathcal{F}}$ by Assumption 3.4. Therefore, it follows from Lemma L.2 that

$$\begin{aligned}
\mathrm{u_n} &= \mathrm{E}_P\Big[\sup_{u\in\mathcal{U}, j\in[\tilde{p}]} \Big|\sqrt{n}\mathbb{E}_n[\psi_{uj}(W, \theta_{uj}, \eta_{uj})]\Big|\Big] \\
&\lesssim \log^{1/2}(a_n M_{n,1}) + n^{-1/2+1/q} M_{n,1}\log(a_n M_{n,1}) \lesssim \log^{1/2}(a_n M_{n,1})(1 + \delta_n) \lesssim \sqrt{\log a_n}
\end{aligned}$$

where the last two inequalities follow from Assumption 3.4 and the facts that $\delta_n = o(1)$ and that $\log M_{n,1} \lesssim \log n$, which is another consequence of Assumption 3.4. Hence, Assumption 3.1 implies that for all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, Θ_{uj} contains a ball of radius $C_0 n^{-1/2}\mathrm{u_n}\log n$ centered at θ_{uj} for all sufficiently large n for any constant C_0. Therefore, Assumption 2.1(i) holds.

Next, Assumption 2.1(ii) follows from the observation that for all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, the map $(\theta, \eta) \mapsto \psi_{uj}(W, \theta, \eta)$ is twice continuously Gateaux-differentiable on $\Theta_{uj} \times T_{uj}$, and so is the map $(\theta, \eta) \mapsto \mathrm{E}_P[\psi_{uj}(W, \theta, \eta)]$.

To verify the near orthogonality condition in Assumption 2.1(iii), note that for all $u \in \mathcal{U}$, $j \in [\tilde{p}]$, and $\eta = (\eta^{(1)}, \eta^{(2)}, \eta^{(3)}) \in \mathcal{T}_{uj}$ with $\eta \neq \eta_{uj}$, we have

$$\mathrm{D}_{u,j,0}[\eta - \eta_{uj}] = \mathrm{E}_P\Big[r_u Z_u^j - \Lambda'(Z_u^j\theta_{uj} + X^j\beta_u^j)Z_u^j X^j\Big(\eta^{(2)} - \beta_u^j - \theta_{uj}(\eta^{(3)} - \gamma_u^j)\Big)\Big]$$

where we used the equality $\mathrm{E}_P[\{Y_u - \Lambda(Z_u^j\theta_{uj} + X^j\beta_u^j) - r_u\}X^j] = 0$. In addition, $|\mathrm{E}_P[r_u Z_u^j]| \leqslant \delta_n n^{-1/2}$ by Assumption 3.5. Further, recall that $\mathrm{E}_P[f_u^2 Z_u^j X^j] = 0$ and observe that

$$\begin{aligned}
f_u^2 = f_u^2(D, X) &= \mathrm{Var}(Y_u \mid D, X) = \mathrm{E}_P[Y_u \mid D, X](1 - \mathrm{E}_P[Y_u \mid D, X]) \\
&= (\Lambda(D'\theta_u + X'\beta_u) + r_u)(1 - \Lambda(D'\theta_u + X'\beta_u) - r_u) \\
&= \Lambda'(D'\theta_u + X'\beta_u) + r_u - r_u^2 - 2r_u\Lambda(D'\theta_u + X'\beta_u)
\end{aligned}$$

where we used the equality $\Lambda'(t) = \Lambda(t) - \Lambda^2(t)$, which holds for all $t \in \mathbb{R}$. Hence,

$$\begin{aligned}
&\Big|\mathrm{E}_P\Big[\Lambda'(Z_u^j\theta_{uj} + X^j\beta_u^j)Z_u^jX^j\Big(\eta^{(2)} - \beta_u^j - \theta_{uj}(\eta^{(3)} - \gamma_u^j)\Big)\Big]\Big| \\
&\quad\lesssim \Big(\mathrm{E}_P\Big[(r_uZ_u^j)^2\Big]\cdot\mathrm{E}_P\Big[\Big(X^j(\eta^{(2)} - \beta_u^j - \theta_{uj}(\eta^{(3)} - \gamma_u^j))\Big)^2\Big]\Big)^{1/2} \\
&\quad\lesssim \Big(\mathrm{E}_P[r_u^2]\Big)^{1/2}\Big(\|\eta^{(2)} - \beta_u^j\| + \|\eta^{(3)} - \gamma_u^j\|\Big) \lesssim s_n\log a_n/n \lesssim \delta_n n^{-1/2}
\end{aligned}$$

where the second line follows from the Cauchy-Schwarz inequality and the observations that $|r_u| \leqslant 1$ and that $|\Lambda(t)| \leqslant 1$ for all $t \in \mathbb{R}$, and the third line from Assumptions 3.4 and 3.5 (the last inequality holds because $s_n^2\log^2 a_n \leqslant \delta_n^2 n$ by Assumption 3.4). Also, when $\eta = \eta_{uj}$, we have $|\mathrm{D}_{u,j,0}[\eta - \eta_{uj}]| = 0$, and so Assumption 2.1(iii) holds.

Next, we verify Assumption 2.1(iv). Fix $u \in \mathcal{U}$ and $j \in [\tilde{p}]$. Observe that

$$\begin{aligned}
J_{uj} &= -\mathrm{E}_P\Big[\Lambda'(Z_u^j\theta_{uj} + X^j\beta_u^j)|Z_u^j|^2\Big] \\
&= -\mathrm{E}_P\Big[f_u^2|Z_u^j|^2\Big] + \mathrm{E}_P\Big[(r_u - r_u^2 - 2r_u\Lambda(D'\theta_u + X'\beta_u))|Z_u^j|^2\Big].
\end{aligned}$$

Hence, by Assumptions 3.1, 3.4 and 3.5 and the Cauchy-Schwarz inequality,

$$|J_{uj}| \geqslant c_1 - 4\Big(\mathrm{E}_P[r_u^2]\mathrm{E}_P[|Z_u^j|^4]\Big)^{1/2} = c_1 + o(1),$$

and also $|J_{uj}| \lesssim 1$ uniformly over $u \in \mathcal{U}$ and $j \in [\tilde{p}]$. In addition,

$$\mathrm{E}_P[\psi_{uj}(W,\theta,\eta_{uj})] = J_{uj}(\theta - \theta_{uj}) + \frac{1}{2}\partial_\theta^2\Big\{\mathrm{E}_P[\psi_{uj}(W,\theta,\eta_{uj})]\Big\}\Big|_{\theta=\bar{\theta}}(\theta - \theta_{uj})^2$$

for some $\bar{\theta} \in \Theta_{uj}$. Moreover, for all $\theta \in \Theta_{uj}$, we have $|\partial_\theta^2\mathrm{E}_P[\psi_{uj}(W,\theta,\eta_{uj})]| \leqslant \mathrm{E}_P[|Z_u^j|^3] \lesssim 1$ by Assumptions 3.1 and 3.4 since $|\Lambda''(t)| \leqslant 1$ for all $t \in \mathbb{R}$. These inequalities together imply Assumption 2.1(iv).

Next, we verify Assumption 2.1(v) with $\omega = 2$ and $B_{1n} = B_{2n} = C$ for sufficiently large C. Fix $u \in \mathcal{U}$, $j \in [\tilde{p}]$, $r \in (0,1]$, $\theta \in \Theta_{uj}$, and $\eta = (\eta^{(1)}, \eta^{(2)}, \eta^{(3)}) \in \mathcal{T}_{uj}$. We consider the case $\eta \neq \eta_{uj}$, and the other case is similar. Denote

$$I_{1,1} = 2|X^j(\eta^{(3)} - \gamma_u^j)| + |r_uZ_u^j|, \quad I_{1,2} = \Big|(D_j - X^j\eta^{(3)})\theta + X^j\eta^{(2)} - Z_u^j\theta_{uj} - X^j\beta_u^j\Big|\cdot|Z_u^j|.$$

Then

$$\Big|\psi_{uj}(W,\theta,\eta) - \psi_{uj}(W,\theta_{uj},\eta_{uj})\Big| \leqslant I_{1,1} + I_{1,2}$$

since $|\Lambda'(t)| \leqslant 1$ for all $t \in \mathbb{R}$. In addition,

$$\mathrm{E}_P[(r_uZ_u^j)^2] \lesssim \mathrm{E}_P[r_u^2] \leqslant \|\eta - \eta_{uj}\|_e^2, \quad \mathrm{E}_P[(X^j(\eta^{(3)} - \gamma_u^j))^2] \lesssim \|\eta^{(3)} - \gamma_u^j\|^2 \leqslant \|\eta - \eta_{uj}\|_e^2$$

by Assumptions 3.5 and 3.4, respectively. Thus, $\mathrm{E}_P[I_{1,1}^2] \lesssim \|\eta - \eta_{uj}\|_e^2$. Also,

$$\begin{aligned}
\mathrm{E}_P[I_{1,2}^2] &\leqslant \mathrm{E}_P\Big[(Z_u^j)^2\Big] \cdot \mathrm{E}_P\Big[((D_j - X^j\eta^{(3)})\theta + X^j\eta^{(2)} - Z_u^j\theta_{uj} - X^j\beta_u^j)^2\Big] \\
&\lesssim \mathrm{E}_P\Big[((D_j - X^j\eta^{(3)})\theta + X^j\eta^{(2)} - Z_u^j\theta_{uj} - X^j\beta_u^j)^2\Big] \\
&\lesssim \mathrm{E}_P\Big[(X^j(\eta^{(2)} - \beta_u^j))^2\Big] + \mathrm{E}_P\Big[(D_j(\theta - \theta_{uj}))^2\Big] + \mathrm{E}_P\Big[(X^j(\eta^{(3)}\theta - \gamma_u^j\theta_{uj}))^2\Big] \\
&\lesssim \|\eta^{(2)} - \beta_u^j\|^2 + |\theta - \theta_{uj}|^2 + \mathrm{E}_P\Big[(X^j\eta^{(3)}(\theta - \theta_{uj}))^2\Big] + \mathrm{E}_P\Big[(X^j(\eta^{(3)} - \gamma_u^j)\theta_{uj})^2\Big] \\
&\lesssim \|\eta^{(2)} - \beta_u^j\|^2 + |\theta - \theta_{uj}|^2 + \|\eta^{(3)} - \gamma_u^j\|^2 \lesssim \|\eta - \eta_{uj}\|_e^2 + |\theta - \theta_{uj}|^2
\end{aligned}$$

where the first line follows from the Cauchy-Schwarz inequality, the second from $\mathrm{E}_P[(Z_u^j)^2] \lesssim 1$, the third from the triangle inequality, the fourth from Assumption 3.4 and the triangle inequality, and the fifth from Assumptions 3.1 and 3.4 and the fact that $\|\eta^{(3)}\| \lesssim \|\gamma_u^j\| + (s_n \log a_n/n)^{1/2} \lesssim 1$. Therefore, Assumption 2.1(v-a) holds.

To verify Assumption 2.1(v-b), observe that under our conditions,

$$\partial_r \mathrm{E}_P\Big[\psi_{uj}(W, \theta, \eta_{uj} + r(\eta - \eta_{uj}))\Big] = \mathrm{E}_P\Big[\partial_r\psi_{uj}(W, \theta, \eta_{uj} + r(\eta - \eta_{uj}))\Big].$$

Further, denote

$$\begin{aligned}
x_r &= Z_u^j\theta - r\theta X^j(\eta^{(3)} - \gamma_u^j) + X^j\beta_u^j + rX^j(\eta^{(2)} - \beta_u^j), \\
I_{2,1} &= -X^j(\eta^{(3)} - \gamma_u^j)(Y_u - \Lambda(x_r) - (1-r)r_u), \\
I_{2,2} &= r_u(Z_u^j - rX^j(\eta^{(3)} - \gamma_u^j)), \\
I_{2,3} &= -\Lambda'(x_r)(Z_u^j - rX^j(\eta^{(3)} - \gamma_u^j))(X^j(\eta^{(2)} - \beta_u^j) - \theta X^j(\eta^{(3)} - \gamma_u^j)).
\end{aligned}$$

Then $\partial_r\psi_{uj}(W, \theta, \eta_{uj} + r(\eta - \eta_{uj})) = I_{2,1} + I_{2,2} + I_{2,3}$, and so

$$\mathrm{E}_P\Big[\partial_r\psi_{uj}(W, \theta, \eta_{uj} + r(\eta - \eta_{uj}))\Big] = \mathrm{E}_P[I_{2,1}] + \mathrm{E}_P[I_{2,2}] + \mathrm{E}_P[I_{2,3}].$$

Now, observe that

$$\mathrm{E}_P[|I_{2,1}|] \lesssim \mathrm{E}_P\Big[|X^j(\eta^{(3)} - \gamma_u^j)|\Big] \lesssim \Big(\mathrm{E}_P\Big[|X^j(\eta^{(3)} - \gamma_u^j)|^2\Big]\Big)^{1/2} \lesssim \|\eta - \eta_{uj}\|_e$$

where the first inequality holds since $|r_u| \leqslant 1$, the second by Jensen's inequality, and the third by Assumption 3.4. Also, by the Cauchy-Schwarz inequality,

$$\mathrm{E}_P[|I_{2,2}|] \leqslant \Big(\mathrm{E}_P[r_u^2] \cdot \mathrm{E}_P\Big[(Z_{uj} - rX^j(\eta^{(3)} - \gamma_u^j))^2\Big]\Big)^{1/2} \lesssim \|\eta - \eta_{uj}\|_e$$

where the second inequality follows from $(\mathrm{E}_P[r_u^2])^{1/2} \leqslant \|\eta - \eta_{uj}\|_e$. Moreover, since $|\Lambda'(t)| \leqslant 1$ for all $t \in \mathbb{R}$, the Cauchy-Schwarz inequality gives

$$\mathrm{E}_P[|I_{2,3}|] \lesssim \Big(\mathrm{E}_P\Big[(Z_u^j - rX^j(\eta^{(3)} - \gamma_u^j))^2\Big]\mathrm{E}_P\Big[(X^j(\eta^{(2)} - \beta_u^j) - \theta X^j(\eta^{(3)} - \gamma_u^j))^2\Big]\Big)^{1/2} \lesssim \|\eta - \eta_{uj}\|_e.$$

Therefore, Assumption 2.1(v-b) holds.

To verify Assumption 2.1(v-c), denote

$$
\begin{aligned}
I_{3,1} &= -r_u X^j(\eta^{(3)} - \gamma_u^j) + \Lambda'(x_r) X^j(\eta^{(3)} - \gamma_u^j)(X^j(\eta^{(2)} - \beta_u^j) - \theta X^j(\eta^{(3)} - \gamma_u^j)),\\
I_{3,2} &= -r_u X^j(\eta^{(3)} - \gamma_u^j),\\
I_{3,3} &= -\Lambda''(x_r)(Z_u^j - r X^j(\eta^{(3)} - \gamma_u^j))(X^j(\eta^{(2)} - \beta_u^j) - \theta X^j(\eta^{(3)} - \gamma_u^j))^2\\
&\quad + \Lambda'(x_r) X^j(\eta^{(3)} - \gamma_u^j)(X^j(\eta^{(2)} - \beta_u^j) - \theta X^j(\eta^{(3)} - \gamma_u^j)),
\end{aligned}
$$

so that $\partial_r I_{2,1} = I_{3,1}$, $\partial_r I_{2,2} = I_{3,2}$, and $\partial_r I_{2,3} = I_{3,3}$. Now, observe that since $|\Lambda'(t)| \leqslant 1$ for all $t \in \mathbb{R}$,

$$
\mathrm{E}_P[|I_{3,1}|] \lesssim \sqrt{\mathrm{E}_P[r_u^2]}\|\eta^{(3)} - \gamma_u^j\| + \|\eta^{(3)} - \gamma_u^j\|(\|\eta^{(2)} - \beta_u^j\| + \|\eta^{(3)} - \gamma_u^j\|) \lesssim \|\eta - \eta_{uj}\|_e^2
$$

by the Cauchy-Schwarz inequality, the triangle inequality, and Assumptions 3.1 and 3.4. Similarly, $\mathrm{E}_P[|I_{3,2}|] \lesssim \|\eta - \eta_{uj}\|_e^2$. In addition, since $|\Lambda''(t)| \leqslant 1$ for all $t \in \mathbb{R}$,

$$
\begin{aligned}
\mathrm{E}_P[|I_{3,3}|] &\lesssim \|\eta - \eta_{uj}\|_e^2 + \Big(\mathrm{E}_P\Big[(Z_u^j - r X^j(\eta^{(3)} - \gamma_u^j))^2\Big]\mathrm{E}_P\Big[(X^j(\eta^{(2)} - \beta_u^j) - \theta X^j(\eta^{(3)} - \gamma_u^j))^4\Big]\Big)^{1/2}\\
&\lesssim \|\eta - \eta_{uj}\|_e^2 + \|\eta - \eta_{uj}\|_e^2 \lesssim \|\eta - \eta_{uj}\|_e^2
\end{aligned}
$$

by the arguments used above, the Cauchy-Schwarz inequality, and Assumption 3.4. Also, the terms in $\mathrm{E}_P[\partial_r^2 \psi_{uj}(W, \theta_{uj} + r(\theta - \theta_{uj}), \eta_{uj} + r(\eta - \eta_{uj}))]$ arising from differentiation of $\theta_{uj} + r(\theta - \theta_{uj})$ can be bounded similarly. Therefore, Assumption 2.1(v-c) holds.

Next, we verify Assumption 2.2(i). Observe that by Theorems 4.1 and 4.2, with probability $1 - o(1)$,

$$
\sup_{u\in\mathcal{U}}\Big(\|\widetilde{\theta}_u - \theta_u\| + \|\widetilde{\beta}_u - \beta_u\|\Big) \lesssim \sqrt{s_n \log a_n / n}, \quad \sup_{u\in\mathcal{U}}\Big(\|\widetilde{\theta}_u\|_0 + \|\widetilde{\beta}_u\|_0\Big) \lesssim s_n,
$$

$$
\sup_{u\in\mathcal{U}, j\in[\tilde{p}]} \|\widetilde{\gamma}_u^j - \gamma_u^j\| \lesssim \sqrt{s_n \log a_n / n}, \quad \text{and} \quad \sup_{u\in\mathcal{U}, j\in[\tilde{p}]} \|\widetilde{\gamma}_u^j\|_0 \lesssim s_n.
$$

In addition, $\widehat{\beta}_u^j = \widetilde{\theta}_{uj}\widetilde{\gamma}_u^j + (\widetilde{\theta}'_{u[\tilde{p}]\setminus j}, \widetilde{\beta}'_u)'$, and so uniformly over $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, with probability $1 - o(1)$,

$$
\|\widehat{\beta}_u^j - \beta_u^j\| \leqslant |\widetilde{\theta}_{uj} - \theta_{uj}|\|\gamma_u^j\| + |\widetilde{\theta}_{uj}|\|\widetilde{\gamma}_u^j - \gamma_u^j\| + \|\widetilde{\theta}_u - \theta_u\| + \|\widetilde{\beta}_u - \beta_u\| \lesssim \sqrt{s_n \log a_n / n}
$$

and $\|\widehat{\beta}_u^j\|_0 \lesssim s_n$. Moreover, uniformly over $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, with probability $1 - o(1)$,

$$
\begin{aligned}
\|\widetilde{\gamma}_u^j - \gamma_u^j\|_1 &\leqslant \|\widetilde{\gamma}_u^j - \bar{\gamma}_u^j\|_1 + \|\bar{\gamma}_u^j - \gamma_u^j\|_1 \lesssim \|\widetilde{\gamma}_u^j - \bar{\gamma}_u^j\|_1 + s_n\sqrt{\log a_n / n}\\
&\lesssim \sqrt{s_n}\|\widetilde{\gamma}_u^j - \bar{\gamma}_u^j\| + s_n\sqrt{\log a_n / n} \lesssim s_n\sqrt{\log a_n / n}
\end{aligned}
$$

by Assumption 3.2 and the triangle and the Cauchy-Schwarz inequalities. Therefore, Assumption 2.2(i) holds. In addition, Assumption 2.2(ii) holds by construction of $\mathcal{T}_{uj}$ and since $\mathrm{E}_P[r_u^2] \leqslant C_1 s_n \log a_n / n$, which in turn follows from Assumption 3.5. Also, Assumption 2.2(iii) holds by construction of $\mathcal{T}_{uj}$.

Next, we establish the entropy bound of Assumption 2.2(iv) with $v_n = Cs_n$ and $K_n = CM_{n,1}$ for sufficiently large constant $C > 0$ (recall that $a_n = p \vee \tilde{p} \vee n$). Let

$$\mathcal{G}_4 = \Big\{(Y, D, X) \mapsto (D', X')\xi \colon \xi \in \mathbb{R}^{\tilde{p}+p}, \|\xi\|_0 \leqslant Cs_n, \|\xi\| \leqslant C\Big\},$$
$$\mathcal{G}_{5,j} = \Big\{(Y, D, X) \mapsto \xi(D_j - X^j\gamma_u^j) + X^j\beta_u^j \colon u \in \mathcal{U}, |\xi| \leqslant C\Big\}, \quad j \in [\tilde{p}]$$

for sufficiently large C. Moreover, recall that $W = (Y, D, X)$ and let

$$\mathcal{F}_{1,1} = \Big\{W \mapsto \psi_{uj}(W, \theta, \eta) \colon u \in \mathcal{U}, j \in [\tilde{p}], \theta \in \Theta_{uj}, \eta \in \mathcal{T}_{uj} \setminus \eta_{uj}\Big\},$$
$$\mathcal{F}_{1,2} = \Big\{W \mapsto \psi_{uj}(W, \theta, \eta_{uj}) \colon u \in \mathcal{U}, j \in [\tilde{p}], \theta \in \Theta_{uj}\Big\}.$$

Then $\mathcal{F}_1 = \mathcal{F}_{1,1} \cup \mathcal{F}_{1,2}$ and

$$\mathcal{F}_{1,1} \subset (\mathcal{G}_1 - \Lambda(\mathcal{G}_4)) \cdot \mathcal{G}_4,$$
$$\mathcal{F}_{1,2} \subset (\mathcal{G}_1 - \mathcal{G}_2 + \Lambda(\cup_{j\in[\tilde{p}]}\mathcal{G}_{5,j}) - \Lambda(\cup_{j\in[\tilde{p}]}\mathcal{G}_{5,j})) \cdot (\cup_{j\in[\tilde{p}]}\mathcal{G}_{3,j})$$

where $\mathcal{G}_1$, $\mathcal{G}_2$, and $\mathcal{G}_{3,j}$, $j \in [\tilde{p}]$, are defined above, because $\psi_{uj}(W, \theta, \eta_{uj}) = \{1\{Y \leqslant u\bar{y} + (1-u)\underline{y}\} - r_u - \Lambda(Z_u^j\theta + X^j\beta_u^j)\}Z_u^j$. A bound for the uniform entropy numbers of $\mathcal{G}_1$ is established above in (E.5). Also, $\mathcal{G}_4$ is a union over $\binom{p+\tilde{p}}{Cs_n}$ VC-subgraph classes with indices $O(s_n)$, and so is $\Lambda(\mathcal{G}_4)$. Hence, by Lemma K.1,

$$\sup_Q \log N(\epsilon\|\widetilde{F}_{1,1}\|_{Q,2}, \mathcal{F}_{1,1}, \|\cdot\|_{Q,2}) \leqslant Cs_n \log(a_n/\epsilon), \quad \text{for all } 0 < \epsilon \leqslant 1,$$

where

$$\widetilde{F}_{1,1}(W) = \sup_{u\in\mathcal{U}, j\in[\tilde{p}]} \sup_{\gamma\in\mathbb{R}^{\tilde{p}-1+p} \colon \|\gamma-\gamma_u^j\|_1 \leqslant C\sqrt{s_n}\tau_n} \Big(2|D_j - X^j\gamma|\Big)$$

is an envelope of $\mathcal{F}_{1,1}$. Observe that $|D_j - X^j\gamma| \leqslant |D_j - X^j\gamma_u^j| + |X^j(\gamma - \gamma_u^j)| \leqslant \sup_{u\in\mathcal{U}, j\in[\tilde{p}]} |Z_u^j| + \|X\|_\infty C\sqrt{s_n}\tau_n$, and so

$$\|\widetilde{F}_{1,1}\|_{P,q} \lesssim M_{n,1} + \sqrt{s_n}\tau_n M_{n,2} \lesssim M_{n,1}$$

by Assumption 3.4 (observe that $\sqrt{s_n}\tau_n M_{n,2} \lesssim 1$ and $M_{n,1} \geqslant 1$).

Next we turn to $\mathcal{F}_{1,2}$. Bounds for the uniform entropy numbers of $\mathcal{G}_2$ and $\cup_{j\in[\tilde{p}]}\mathcal{G}_{3,j}$ are established above. Consider $\mathcal{G}_{5,j}$ for $j \in [\tilde{p}]$. Note that for all $u_1, u_2 \in \mathcal{U}$,

$$\|\gamma_{u_2}^j - \gamma_{u_1}^j\|_1 \lesssim \sqrt{p + \tilde{p}}|u_2 - u_1|, \quad \|\gamma_{u_1}^j\|_1 \lesssim \sqrt{s_n},$$
$$\|\beta_{u_2}^j - \beta_{u_1}^j\|_1 \lesssim \Big(\|\theta_{u_2} - \theta_{u_1}\|_1 + \|\beta_{u_2} - \beta_{u_1}\|_1 + \|\gamma_{u_2} - \gamma_{u_1}\|_1\Big) \lesssim \sqrt{p + \tilde{p}}|u_2 - u_1|$$

by (E.1), (E.6), and Assumption 3.1. Therefore, for all $\xi_1, \xi_2 \in \mathbb{R}$ such that $|\xi_1| \leqslant C$ and $|\xi_2| \leqslant C$, and all $u_1, u_2 \in \mathcal{U}$,

$$\Big\|(\xi_2, \beta_{u_2}^j - \xi_2\gamma_{u_2}^j) - (\xi_1, \beta_{u_1}^j - \xi_1\gamma_{u_1}^j)\Big\|_1 \leqslant |\xi_2 - \xi_1|(1 + \|\gamma_{u_1}^j\|) + \|\beta_{u_2}^j - \beta_{u_1}^j\|_1 + C\|\gamma_{u_2}^j - \gamma_{u_1}^j\|_1$$
$$\lesssim \sqrt{s_n}|\xi_2 - \xi_1| + \sqrt{p + \tilde{p}}|u_2 - u_1|.$$

Hence, Lemma K.3 implies that for all $j \in [\tilde{p}]$, the uniform entropy numbers of $\Lambda(\mathcal{G}_{5,j})$ obey

$$\sup_Q \log N(\epsilon \|\widetilde{F}_5\|_{Q,2}, \Lambda(\mathcal{G}_{5,j}), \|\cdot\|_{Q,2}) \leqslant C \log(a_n/\epsilon), \quad \text{for all } 0 < \epsilon \leqslant 1,$$

where $\widetilde{F}_5(Y,D,X) = 1 + M_{n,2}^{-1}(\|D\|_\infty \vee \|X\|_\infty)$ is its envelope. Hence, by Lemma K.1,

$$\sup_Q \log N(\epsilon \|\widetilde{F}_{1,2}\|_{Q,2}, \mathcal{F}_{1,2}, \|\cdot\|_{Q,2}) \leqslant C \log(a_n/\epsilon), \quad \text{for all } 0 < \epsilon \leqslant 1,$$

where $\widetilde{F}_{1,2}(W) = C(1 + M_{n,2}^{-1}(\|D\|_\infty \vee \|X\|_\infty)) \cdot (\sup_{u \in \mathcal{U} j \in [\tilde{p}]} |Z_u^j| + M_{n,2}^{-1}(\|D\|_\infty \vee \|X\|_\infty))$ is an envelope of $\mathcal{F}_{1,2}$ that satisfies $\|\widetilde{F}_{1,2}\|_{P,q} \lesssim M_{n,1}$ by Assumption 3.4 and the Cauchy-Schwarz inequality. Applying Lemma K.1 one more time finally shows that the uniform entropy numbers of $\mathcal{F}_1$ obey (2.5) with constants specified above and with an envelope $F_1 = \widetilde{F}_{1,1} \vee \widetilde{F}_{1,2}$ satisfying $\|F_1\|_{P,q} \lesssim M_{n,1}$.

Next, we verify Assumption 2.2(v). Fix $u \in \mathcal{U}$, $j \in [\tilde{p}]$, $\theta \in \Theta_{uj}$, and $\eta = (\eta^{(1)}, \eta^{(2)}, \eta^{(3)}) \in \mathcal{T}_{uj}$. Then

$$\begin{aligned} \mathrm{E}_P[\psi_{uj}(W,\theta,\eta)^2] &= \mathrm{E}_P\Big[\mathrm{E}_P\Big[\Big(Y_u - \Lambda\Big((D_j - X^j\eta^{(3)})\theta + X^j\eta^{(2)}\Big) - \eta^{(1)}\Big)^2 \mid D, X\Big](D_j - X^j\eta^{(3)})^2\Big] \\ &\geqslant \mathrm{E}_P[f_u^2(D_j - X^j\eta^{(3)})^2] \geqslant c \end{aligned}$$

where the first inequality follows from the fact that for any random variable ξ, the function $x \mapsto \mathrm{E}[(\xi - x)^2]$ is minimized at $x = \mathrm{E}[\xi]$ and in this case $f_u^2 = \mathrm{Var}(Y_u \mid D, X)$, and the second inequality follows from Assumption 3.4. In addition,

$$\mathrm{E}_P[\psi_{uj}(W,\theta,\eta)^2] \leqslant \mathrm{E}_P[(D_j - X^j\eta^{(3)})^2] \lesssim 1$$

by Assumptions 3.1 and 3.4. Therefore, Assumption 2.2(v) holds.

Finally, we verify Assumption 2.2(vi). The condition (a) holds by construction of τ_n and v_n. To verify the condition (b) observe that

$$(B_{1n}\tau_n)^{\omega/2}(v_n \log a_n)^{1/2} + n^{-1/2+1/q} v_n K_n \log a_n \lesssim n^{-1/2} s_n \log a_n + n^{-1/2+1/q} s_n M_{n,1} \log a_n \lesssim \delta_n$$

by Assumption 3.4. In addition,

$$(\mathrm{u_n} \log n/\sqrt{n})^{\omega/2}(v_n \log a_n)^{1/2} \lesssim \sqrt{s_n}(\log a_n) \cdot (\log n)/\sqrt{n} \lesssim \delta_n$$

because $\mathrm{u_n} \lesssim (\log a_n)^{1/2}$, which is established above, and $s_n \log a_n \leqslant \delta_n n^{1/2-1/q}$, which holds by Assumption 3.4. The condition (b) follows. The condition (c) holds because

$$n^{1/2} B_{1n}^2 B_{2n} \tau_n^2 \lesssim n^{-1/2} s_n \log a_n \lesssim \delta_n$$

as in the verification of the condition (b). This completes the verification of Assumptions 2.1 and 2.2 and thus completes the proof of the theorem. ■

Proof of Corollary 3.1. The asserted claim will follow from Corollary 2.1 as long as we can verify its conditions. Assumptions 2.1 and 2.2 were verified in the proof of Theorem 3.1. Therefore, it suffices to verify Assumption 2.3(i,ii) and the growth conditions of Corollary 2.1.

First, we verify Assumption 2.3(i). Recall the function class $\widetilde{\mathcal{F}} = \{\psi_{uj}(\cdot, \theta_{uj}, \eta_{uj}) \colon u \in \mathcal{U}, j \in [\tilde{p}]\}$ defined in the proof of Theorem 3.1, where it is also proven that its uniform entropy numbers obey (E.7) with an envelope $\widetilde{F}$ satisfying $\|\widetilde{F}\|_{P,q} \lesssim M_{n,1}$. Also, note that Assumption 2.1(iv) gives $1 \lesssim |J_{uj}| \lesssim 1$ for all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, and that Assumption 2.2(v) gives $1 \lesssim \mathrm{E}_P[\psi_{uj}^2(W, \theta_{uj}, \eta_{uj})] \lesssim 1$ for all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$. Hence,

$$\mathcal{F}_0 \subset \Big\{\xi \cdot f \colon f \in \widetilde{\mathcal{F}}, \xi \in \mathbb{R}, c \leqslant |\xi| \leqslant C\Big\},$$

and so Lemma K.1 implies that the uniform entropy numbers of $\mathcal{F}_0$ obey

$$\sup_Q \log N(\epsilon \|F_0\|_{Q,2}, \mathcal{F}_0, \|\cdot\|_{Q,2}) \leqslant C \log(a_n/\epsilon), \quad \text{for all } 0 < \epsilon \leqslant 1, \tag{E.8}$$

where its envelope F_0 satisfies $\|F_0\|_{P,q} \lesssim M_{n,1}$. Thus, Assumption 2.3(i) holds with $\varrho_n = C$, $A_n = a_n = p \vee \tilde{p} \vee n$, and $L_n = CM_{n,1}$.

Next, we verify Assumption 2.3(ii). For $k = 3, 4$, $u \in \mathcal{U}$, and $j \in [\tilde{p}]$, we have

$$\mathrm{E}_P\Big[|\bar{\psi}_{uj}(W, \theta_{uj}, \eta_{uj})|^k\Big] \lesssim \mathrm{E}_P\Big[|D_j - X^j \gamma_u^j|^k\Big] \lesssim 1$$

by Assumptions 3.1 and 3.4 since $|Y_u| \leqslant 1$, $|\Lambda(t)| \leqslant 1$ for all $t \in \mathbb{R}$, and $|r_u| \leqslant 1$. Thus, Assumption 2.3(ii) holds with the same $L_n = CM_{n,1}$ since $M_{n,1} \geqslant 1$.

Finally, with our choice of A_n and ϱ_n, the growth conditions of Corollary 2.1 hold by assumption. This completes the proof. ∎

Proof of Corollary 3.2. The asserted claim will follow from Corollary 2.2 as long as we can verify its conditions. Assumptions 2.1 and 2.2 were verified in the proof of Theorem 3.1. Assumption 2.3(i,ii) was verified in the proof of Corollary 3.1. Therefore, it suffices to verify Assumptions 2.3(iii) and 2.4 and the growth conditions of Corollary 2.2.

We split the proof into six steps. In Steps 1-3, we verify Assumption 2.4. In Steps 4 and 5, we verify Assumption 2.3(iii). In Step 6, we verify the growth conditions of Corollary 2.2.

Step 1. Here we show that

$$\widehat{J}_{uj} - J_{uj} = o_P(\log^{-1} a_n)$$

uniformly over $u \in \mathcal{U}$ and $j \in [\tilde{p}]$. For $\theta \in \mathbb{R}$, $\beta \in \mathbb{R}^{\tilde{p}-1+p}$, $\gamma \in \mathbb{R}^{\tilde{p}-1+p}$, and $j \in [\tilde{p}]$, define

$$\widetilde{\psi}_j(W, \theta, \beta, \gamma) = -\Lambda'\Big((D_j - X^j\gamma)\theta + X^j\beta\Big)(D_j - X^j\gamma)^2,$$
$$\widetilde{m}_j(\theta, \beta, \gamma) = \mathrm{E}_P[\widetilde{\psi}_j(W, \theta, \beta, \gamma)].$$

Then $\widehat{J}_{uj} = \mathbb{E}_n[\widetilde{\psi}_j(W, \widetilde{\theta}_{uj}, \widehat{\beta}_u^j, \widetilde{\gamma}_u^j)]$ and $J_{uj} = \widetilde{m}_j(\theta_{uj}, \beta_u^j, \gamma_u^j)$ for all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$. Therefore, by the triangle inequality,

$$|\widehat{J}_{uj} - J_{uj}| \leqslant \Big|\widehat{J}_{uj} - \widetilde{m}_j(\widetilde{\theta}_{uj}, \widehat{\beta}_u^j, \widetilde{\gamma}_u^j)\Big| + \Big|\widetilde{m}_j(\widetilde{\theta}_{uj}, \widehat{\beta}_u^j, \widetilde{\gamma}_u^j) - \widetilde{m}_j(\theta_{uj}, \beta_u^j, \gamma_u^j)\Big|.$$

Define

$$\mathcal{G}_6 = \Big\{(Y,D,X) \mapsto -\Lambda'\Big((D_j - X^j\eta^{(3)})\theta + X^j\eta^{(2)}\Big)(D_j - X^j\eta^{(3)})^2 \colon \\ u \in \mathcal{U}, j \in [\tilde p], \theta \in \Theta_{uj}, \eta = (\eta^{(1)},\eta^{(2)},\eta^{(3)}) \in \mathcal{T}_{uj} \setminus \eta_{uj}\Big\}.$$

Then by Assumption 2.2(i), with probability $1 - o(1)$,

$$\Big|\widehat J_{uj} - \widetilde m_j(\widetilde\theta_{uj}, \widehat\beta^j_u, \widetilde\gamma^j_u)\Big| \leqslant \sup_{f\in\mathcal{G}_6}\Big|\mathbb{E}_n[f(W)] - \mathrm{E}_P[f(W)]\Big|$$

for all $u \in \mathcal{U}$ and $j \in [\tilde p]$. In addition, $\mathcal{G}_6 \subset -\Lambda'(\mathcal{G}_4)\cdot\mathcal{G}_4^2$, where the function class $\mathcal{G}_4$ is introduced in the proof of Theorem 3.1. Moreover,

$$\begin{aligned}\Big|(D_j - X^j\eta^{(3)})\theta + X^j\eta^{(2)}\Big| &\lesssim \Big(\|D\|_\infty \vee \|X\|_\infty\Big)\Big(\|\eta^{(2)}\|_1 + \|\eta^{(3)}\|_1\Big)\\ &\lesssim \sqrt{p+\tilde p}\Big(\|D\|_\infty \vee \|X\|_\infty\Big)\Big(\|\eta^{(2)}\| + \|\eta^{(3)}\|\Big)\\ &\lesssim \sqrt{p+\tilde p}\Big(\|D\|_\infty \vee \|X\|_\infty\Big)\end{aligned}$$

uniformly over $u \in \mathcal{U}$, $j \in [\tilde p]$, $\theta \in \Theta_{uj}$, and $\eta = (\eta^{(1)},\eta^{(2)},\eta^{(3)}) = \mathcal{T}_{uj}\setminus\eta_{uj}$ by Assumptions 3.1. Hence, applying Lemmas K.1 and K.4 with $K = M_{n,2}n^{2/q}(p+\tilde p)^{1/2}$ shows that the uniform entropy numbers of $\mathcal{G}_6$ obey

$$\sup_Q \log N(\epsilon\|F_6\|_{Q,2}, \mathcal{G}_6, \|\cdot\|_{Q,2}) \leqslant C s_n \log(a_n/\epsilon), \quad \text{for all } 0 < \epsilon \leqslant 1$$

where $F_6(W) = (1 + (\|D\|_\infty \vee \|X\|_\infty)/(M_{n,2}n^{2/q}))\widetilde F_{1,1}^2(W)$ is its envelope, and $\widetilde F_{1,1}$ is defined in the proof of Theorem 3.1. Also, recall that $\|\widetilde F_{1,1}\|_{P,q} \lesssim M_{n,1}$, which is established in the proof of Theorem 3.1, and so

$$\begin{aligned}&\Big(\mathrm{E}_P\Big[\max_{1\leqslant i\leqslant n}|F_6(W_i)|^{q/4}\Big]\Big)^{4/q}\\ &\leqslant \Big(\mathrm{E}_P\Big[\max_{1\leqslant i\leqslant n}|\widetilde F_{1,1}(W_i)|^{q/2}\Big]\Big)^{4/q} + \Big(\mathrm{E}_P\Big[\max_{1\leqslant i\leqslant n}\Big(\frac{\|D_i\|_\infty \vee \|X_i\|_\infty}{M_{n,2}n^{2/q}}\Big)^{q/4}|\widetilde F_{1,1}(W_i)|^{q/2}\Big]\Big)^{4/q}\\ &\lesssim n^{2/q}M_{n,1}^2 + n^{4/q}\Big(\mathrm{E}_P\Big[\Big(\frac{\|D\|_\infty \vee \|X\|_\infty}{M_{n,2}n^{2/q}}\Big)^{q/2}\Big]\cdot\mathrm{E}_P\Big[|\widetilde F_{1,1}(W)|^q\Big]\Big)^{2/q} \lesssim n^{2/q}M_{n,1}^2\end{aligned}$$

where the first line follows from the triangle inequality, and the second from the Cauchy-Schwarz inequality and Assumption 3.4. In addition, $\|f\|_{P,2} \lesssim 1$ for all $f \in \mathcal{G}_6$. Hence, Lemma L.2 implies that

$$\sup_{f\in\mathcal{G}_6}\Big|\mathbb{E}_n[f(W)] - \mathrm{E}_P[f(W)]\Big| \lesssim \sqrt{\frac{s_n\log a_n}{n}} + \frac{M_{n,1}^2 s_n \log a_n}{n^{1-2/q}} = o(\log^{-1} a_n)$$

with probability $1 - o(1)$ by Assumption 3.4 and the growth condition $s_n \log^3 a_n/n = o(1)$.

Next, using the same arguments as those used to verify Assumption 2.1(v-a) in Theorem 3.1 shows that

$$\Big|\widetilde m_j(\widetilde\theta_{uj}, \widehat\beta^j_u, \widetilde\gamma^j_u) - \widetilde m_j(\theta_{uj}, \beta^j_u, \gamma^j_u)\Big| \lesssim \|\widetilde\theta_{uj} - \theta_{uj}\| + \|\widehat\beta^j_u - \beta^j_u\| + \|\widetilde\gamma^j_u - \gamma^j_u\|$$

and the right-hand of this inequality is bounded from above by $(Cs_n \log a_n/n)^{1/2}$ uniformly over $u \in \mathcal{U}$ and $j \in [\tilde{p}]$ with probability $1 - o(1)$, as demostrated in the proof of Theorem 3.1. In turns, $(Cs_n \log a_n/n)^{1/2} = o(\log^{-1} a_n)$ by assumption. Combining presented bounds gives the claim of this step.

Step 2. Here we show that

$$\mathbb{E}_n[\psi^2_{uj}(W, \widetilde{\theta}_{uj}, \widehat{\eta}_{uj})] - \mathrm{E}_P[\psi^2_{uj}(W, \theta_{uj}, \eta_{uj})] = o_P(\log^{-1} a_n)$$

uniformly over $u \in \mathcal{U}$ and $j \in [\tilde{p}]$. The proof of this claim is similar to that in Step 1, where the main difference is that instead of $-\Lambda'(\cdot)$ in the function class $\mathcal{G}_6$, we set $Y_u^2 - 2Y_u\Lambda(\cdot) + \Lambda^2(\cdot)$, with the resulting function class having the same envelope and its uniform entropy numbers obeying the same bounds as those derived fo $\mathcal{G}_6$ (up-to a possibly different constants).

Step 3. Here we finish the verification of Assumption 2.4. Observe that $1 \lesssim J_{uj} \lesssim 1$ and $1 \lesssim \mathrm{E}_P[\psi^2_{uj}(W, \theta_{uj}, \eta_{uj})] \lesssim 1$ for all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$ by Assumptions 2.1(iv) and 2.2(v). Hence, $1 \lesssim \sigma^2_{uj} \lesssim 1$, and so

$$\begin{aligned}\left|\frac{\widehat{\sigma}_{uj}}{\sigma_{uj}} - 1\right| &\leqslant \left|\frac{\widehat{\sigma}^2_{uj}}{\sigma^2_{uj}} - 1\right| \lesssim \left|\widehat{\sigma}^2_{uj} - \sigma^2_{uj}\right| \\ &\leqslant \left|\widehat{J}^{-2}_{uj} - J^{-2}_{uj}\right| \mathbb{E}_n[\psi^2_{uj}(W, \widetilde{\theta}_{uj}, \widehat{\eta}_{uj})] + J^{-2}_{uj}\left|\mathbb{E}_n[\psi^2_{uj}(W, \widetilde{\theta}_{uj}, \widehat{\eta}_{uj})] - \mathrm{E}_P[\psi^2_{uj}(W, \theta_{uj}, \eta_{uj})]\right| \\ &\lesssim \left|\widehat{J}_{uj} - J_{uj}\right| + \left|\mathbb{E}_n[\psi^2_{uj}(W, \widetilde{\theta}_{uj}, \widehat{\eta}_{uj})] - \mathrm{E}_P[\psi^2_{uj}(W, \theta_{uj}, \eta_{uj})]\right| = o_P(\log^{-1} a_n)\end{aligned}$$

uniformly over $u \in \mathcal{U}$ and $j \in [\tilde{p}]$ by Steps 1 and 2. Therefore, Assumption 2.4 holds for some ε_n and Δ_n satisfying $\varepsilon_n \log a_n = o(1)$ and $\Delta_n = o(1)$.

Step 4. Here we show that the inequality concerning the entropy numbers of $\widehat{\mathcal{F}}_0$ in Assumption 2.3(iii) holds with $\bar{\varrho}_n = C$, $\bar{A}_n = a_n = p \vee \tilde{p} \vee n$, and $\Delta_n = o(1)$. By construction of $\widehat{\psi}_{uj}$, the function class $\{\widehat{\psi}_{uj}(\cdot) \colon u \in \mathcal{U}, j \in [\tilde{p}]\}$ contains at most $n\tilde{p}$ functions (as u varies, new functions appear only as u crosses one of the observations $(Y_i)_{i=1}^n$). Also, it follows from (E.8) in the proof of Corollary 3.1 that the entropy numbers of $\mathcal{F}_0 = \{\bar{\psi}_{uj}(\cdot) \colon u \in \mathcal{U}, j \in [\tilde{p}]\}$ obey

$$N(\epsilon, \mathcal{F}_0, \|\cdot\|_{\mathbb{P}_{n,2}}) \leqslant C\log(a_n\|F_0\|_{\mathbb{P}_{n,2}}/\epsilon) \leqslant C\log(a_n/\epsilon), \quad \text{for all } 0 < \epsilon \leqslant 1,$$

with probability $1 - o(1)$. Hence,

$$\log N(\epsilon, \widehat{\mathcal{F}}_0, \|\cdot\|_{\mathbb{P}_{n,2}}) \leqslant C\log(a_n/\epsilon), \quad \text{for all } 0 < \epsilon \leqslant 1$$

with probability $1 - o(1)$. The claim of this step follows.

Step 5. Here we show that the second part of Assumption 2.3(iii), that is, that with probability $1 - \Delta_n$, we have $\|f\|_{\mathbb{P}_{n,2}} \leqslant \bar{\delta}_n$ for all $f \in \widehat{\mathcal{F}}_0$, holds for some $\bar{\delta}_n$ and Δ_n satisfying $\bar{\delta}_n = o(\log^{-1} a_n)$ and $\Delta_n = o(1)$. By the triangle inequality,

$$\begin{aligned}&\left\|\widehat{\sigma}^{-1}_{uj}\widehat{J}^{-1}_{uj}\psi_{uj}(W, \check{\theta}_{uj}, \widehat{\eta}_{uj}) - \sigma^{-1}_{uj}J^{-1}_{uj}\psi_{uj}(W, \theta_{uj}, \eta_{uj})\right\|_{\mathbb{P}_{n,2}} \\ &\quad\leqslant |\widehat{\sigma}^{-1}_{uj}\widehat{J}^{-1}_{uj} - \sigma^{-1}_{uj}J^{-1}_{uj}| \cdot \left\|\psi_{uj}(W, \theta_{uj}, \eta_{uj})\right\|_{\mathbb{P}_{n,2}} + \widehat{\sigma}^{-1}_{uj}\widehat{J}^{-1}_{uj}\left\|\psi_{uj}(W, \check{\theta}_{uj}, \widehat{\eta}_{uj}) - \psi_{uj}(W, \theta_{uj}, \eta_{uj})\right\|_{\mathbb{P}_{n,2}}.\end{aligned}$$

We bound two terms on the right-hand side of this inequality in turn. To bound the first term, observe that

$$|\widehat{\sigma}_{uj}^{-1}\widehat{J}_{uj}^{-1}-\sigma_{uj}^{-1}J_{uj}^{-1}|=o_P(\log^{-1}a_n)$$

uniformly over $u\in\mathcal{U}$ and $j\in[\tilde{p}]$ by Steps 1 and 3 and since $1\lesssim J_{uj}\lesssim 1$ and $1\lesssim\sigma_{uj}\lesssim 1$, which is discussed in Step 3. Also, as established in the proof of Theorem 3.1, the uniform entropy numbers of the function class $\widetilde{\mathcal{F}}=\{\psi_{uj}(\cdot,\theta_{uj},\eta_{uj})\colon u\in\mathcal{U},j\in[\tilde{p}]\}$ obey (E.7) with an envelope $\widetilde{F}$ satisfying $\|\widetilde{F}\|_{P,q}\lesssim M_{n,1}$. Moveover, $\mathrm{E}_P[f^2(W)]\lesssim 1$ uniformly over $f\in\widetilde{\mathcal{F}}$ by Assumption 2.2(v). Therefore, Lemma L.3 shows that

$$\begin{aligned}\mathrm{E}_P\Big[\sup_{u\in\mathcal{U},j\in[\tilde{p}]}\mathbb{E}_n[\psi_{uj}^2(W,\theta_{uj},\eta_{uj})]\Big]&\lesssim 1+n^{-1/2+1/q}M_{n,1}\Big(\sqrt{\log a_n}+n^{-1/2+1/q}M_{n,1}\log a_n\Big)\\&\lesssim 1+n^{-1+2/q}M_{n,1}^2\log a_n\lesssim 1,\end{aligned}$$

where the second inequality follows from Assumption 3.4. Hence,

$$|\widehat{\sigma}_{uj}^{-1}\widehat{J}_{uj}^{-1}-\sigma_{uj}^{-1}J_{uj}^{-1}|\cdot\|\psi_{uj}(W,\theta_{uj},\eta_{uj})\|_{\mathbb{P}_n,2}=o_P(\log^{-1}a_n)$$

uniformly over $u\in\mathcal{U}$ and $j\in[\tilde{p}]$.

To bound the second term, define

$$\mathcal{G}_7=\Big\{\psi_{uj}(\cdot,\theta,\eta)-\psi_{uj}(\cdot,\theta_{uj},\eta_{uj})\colon u\in\mathcal{U},j\in[\tilde{p}],\theta\in\Theta_{uj},|\theta-\theta_{uj}|\leqslant\sqrt{Cs_n\log a_n/n},\eta\in\mathcal{T}_{uj}\Big\}$$

for sufficiently large constant $C>0$ and $\mathcal{T}_{uj}$ appearing in Assumption 2.2. Then $\mathcal{G}_7\subset\mathcal{F}_1-\mathcal{F}_1$, and so Lemma K.1 together with the bound for the uniform entropy numbers of $\mathcal{F}_1$ established in the proof of Theorem 3.1 imply that the uniform entropy numbers of $\mathcal{G}_7$ obey

$$\sup_Q\log N(\epsilon\|F\|_{Q,2},\mathcal{G}_7,\|\cdot\|_{Q,2})\leqslant Cs_n\log(a_n/\epsilon),\quad\text{for all }0<\epsilon\leqslant 1,$$

where F_7 is its envelope satisfying $\|F_7\|_{P,2}\lesssim M_{n,1}$. In addition, Assumption 2.2(i) together with Step 1 in the proof of Theorem 2.1 imply that with probability $1-o(1)$,

$$\psi_{uj}(\cdot,\check{\theta}_{uj},\widehat{\eta}_{uj})-\psi_{uj}(\cdot,\theta_{uj},\eta_{uj})\in\mathcal{G}_7$$

for all $u\in\mathcal{U}$ and $j\in[\tilde{p}]$ (recall that in the proof of Theorem 3.1, we set $B_{1n}=C$ and $\tau_n=(Cs_n\log a_n/n)^{1/2}$). Also, Assumptions 2.1(v-a) and 2.2(ii) show that $\mathrm{E}_P[f^2(W)]\lesssim s_n\log a_n/n$ uniformly over $f\in\mathcal{G}_7$. Hence, it follows from Lemma L.3 that

$$\begin{aligned}\mathrm{E}_P\Big[\sup_{f\in\mathcal{G}_7}\mathbb{E}_n[f^2(W)]\Big]&\lesssim s_n\log a_n/n+n^{-1/2+1/q}M_{n,1}\Big(n^{-1/2}s_n\log a_n+n^{-1/2+1/q}M_{n,1}s_n\log a_n\Big)\\&\lesssim n^{-1+2/q}M_{n,1}^2s_n\log a_n.\end{aligned}$$

Hence,

$$\widehat{\sigma}_{uj}^{-1}\widehat{J}_{uj}^{-1}\|\psi_{uj}(W,\check{\theta}_{uj},\widehat{\eta}_{uj})-\psi_{uj}(W,\theta_{uj},\eta_{uj})\|_{\mathbb{P}_n,2}=o_P(\log^{-1}a_n)$$

uniformly over $u\in\mathcal{U}$ and $j\in[\tilde{p}]$ by Assumption 3.4. Combining presented bounds gives the asserted claim and completes the verification of Assumption 2.3.

Step 6. Recall that the growth conditions of Corollary 2.1 were verified in the proof of Corollary 3.1, where we set $\varrho_n = C$ and $A_n = a_n$. The other growth conditions of Corollary 2.2, $\varepsilon_n \varrho_n \log A_n = o(1)$ and $\bar{\delta}_n^2 \bar{\varrho}_n \varrho_n (\log \bar{A}_n) \cdot (\log A_n) = o(1)$ hold because we have $\varepsilon_n = o(\log^{-1} a_n)$, $\bar{\varrho}_n = C$, $\bar{A}_n = a_n$, and $\bar{\delta}_n = o(\log^{-1} a_n)$. This completes the proof of the corollary. ■

Proof of Corollary 3.3. To prove the asserted claim, we will apply Corollary 3.2. Below we will verify Assumptions 3.4(iii,v,vi,vii,viii,ix), 3.5(iii), and the growth conditions of Corollary 3.2.

Set

$$\bar{M}_n = \Big(\mathrm{E}_P\Big[(\|D\|_\infty \vee \|X\|_\infty)^{2q}\Big]\Big)^{1/(2q)} \quad \text{and} \quad \bar{C}_n := 1 + \sup_{u \in \mathcal{U}, j \in [\tilde{p}]} \|\gamma_u^j\|_1$$

so that $\bar{M}_n \leqslant C_1$ and $\bar{C}_n \leqslant 1 + C_1$ by assumption. Observe that for all $u \in \mathcal{U}$ and $j \in [\tilde{p}]$,

$$|Z_u^j| = |D_j - X^j \gamma_u^j| \leqslant (\|D\|_\infty \vee \|X\|_\infty) \cdot (1 + \|\gamma_u^j\|_1) \lesssim \bar{C}_n (\|D\|_\infty \vee \|X\|_\infty)$$

Then

$$\max_{j,k} \Big(\mathrm{E}_P[|Z_u^j X_k^j|^3]\Big)^{1/3} \lesssim \bar{C}_n \Big(\mathrm{E}_P\Big[(\|D\|_\infty \vee \|X\|_\infty)^6\Big]\Big)^{1/3} \lesssim \bar{C}_n \bar{M}_n^2 \lesssim 1.$$

Therefore, given that $\log^6 a_n = o(n)$ by assumption, it follows that Assumption 3.4(iii) holds for some δ_n satisfying $\delta_n^2 \log a_n = o(1)$.

Also,

$$\left(\mathrm{E}_P\left[\sup_{u \in \mathcal{U}, j \in [\tilde{p}]} |Z_u^j|^{2q}\right]\right)^{1/(2q)} \lesssim \bar{C}_n \bar{M}_n \lesssim 1,$$

and so Assumption 3.4(v) holds with $M_{n,1} = C$ for sufficiently large constant C. In addition, since $s_n^2 \log^3 a_n = o(n^{1-2/q})$, Assumption 3.4(vi) holds for some δ_n satisfying $\delta_n^2 \log a_n = o(1)$.

Further, Assumption 3.4(vii) holds with $M_{n,2} = \bar{M}_n$ by definition of $\bar{M}_n$. In addition, since $s_n^2 \log^2 a_n = o(n^{1-2/q})$, Assumption 3.4(viii) holds for some δ_n satisfying $\delta_n^2 \log a_n = o(1)$. Also, Assumption 3.4(ix) holds for some δ_n satisfying $\delta_n^2 \log a_n = o(1)$ since $M_{n,2} = \bar{M}_n \leqslant C_1$, $M_{n,1} \leqslant C$, $s_n \leqslant \delta_n n^{1/2-1/q}$, and $q > 4$.

Moreover, since $\sup_{u \in \mathcal{U}, j \in [\tilde{p}]} |\mathrm{E}_P[r_u Z_u^j]| = o((n \log a_n)^{-1/2})$, Assumption 3.5(iii) holds for some δ_n satisfying $\delta_n^2 \log a_n = o(1)$. Finally, the growth conditions $M_{n,1}^{2/7} \log a_n = o(n^{1/7})$ and $M_{n,1}^{2/3} \log a_n = o(n^{1/3-2/(3q)})$ hold because $M_{n,1} \lesssim \bar{C}_n \bar{M}_n \lesssim C$, $\log^7 a_n = o(n)$, and $\log^3 a_n = o(n^{1-2/q})$.

Thus, there exists δ_n such that Assumptions 3.4(iii,v,vi,vii,viii,ix) and 3.5(iii) as well as all growth conditions of Corollary 3.2 are satisfied. Since all other conditions of Corollary 3.2 are assumed, the asserted claim follows from that in Corollary 3.2. ■

Appendix F. Proofs for Section 4

In this appendix, we use C to denote a strictly positive constant that is independent of n and $P \in \mathcal{P}_n$. The value of C may change at each appearance. Also, the notation $a_n \lesssim b_n$ means that $a_n \leqslant C b_n$ for all n and some C. The notation $a_n \gtrsim b_n$ means that $b_n \lesssim a_n$. Moreover, the notation

$a_n = o(1)$ means that there exists a sequence $(b_n)_{n\geqslant 1}$ of positive numbers such that (i) $|a_n| \leqslant b_n$ for all n, (ii) b_n is independent of $P \in \mathcal{P}_n$ for all n, and (iii) $b_n \to 0$ as $n \to \infty$. Finally, the notation $a_n \lesssim_P b_n$ means that for any $\epsilon > 0$, there exists C such that $\mathrm{P}_P(a_n > Cb_n) \leqslant \epsilon$ for all n, and the notation $a_n \gtrsim_P b_n$ means that $b_n \lesssim_P a_n$. Using this notation allows us to avoid repeating "uniformly over $P \in \mathcal{P}_n$" many times in the proofs of Theorems 4.1 and 4.2.

Proof of Theorem 4.1. In this proof, we will rely upon results in Appendix I. In particular, the asserted claims will follow from an application of Lemmas I.1, I.2, and I.3 (with some extra work). To follow the notation in Appendix I, define $X_u = (D', X')'$ and $w_u = f_u^2$ and redefine $\theta_u = (\theta_u', \beta_u')'$, $\widehat{\theta}_u = (\widehat{\theta}_u', \widehat{\beta}_u')'$, and $p = \tilde{p} + p$. Also, define $a_u = a_u(X_u)$ as a solution to the following equation:

$$\Lambda(X_u'\theta_u) + r_u = \Lambda(X_u'\theta_u + a_u). \tag{F.1}$$

Since Λ is increasing, for each value of X_u, a_u is uniquely defined. Then θ_u satisfies (I.1) with

$$M_u(Y_u, X_u, \theta, a) = -\Big(1\{Y_u = 1\}\log\Big(\Lambda(X_u'\theta + a(X_u))\Big) + 1\{Y_u = 0\}\log\Big(1 - \Lambda(X_u'\theta + a(X_u))\Big)\Big)$$

for θ being a vector in $\mathbb{R}^p$ and a being a function of X_u. Similarly, $\widehat{\theta}_u$ satisfies (I.2) where $M_u(Y_u, X_u, \theta) = M_u(Y_u, X_u, \theta, \mathcal{O})$ and $\mathcal{O} = \mathcal{O}(X_u)$, the identically zero function of X_u.

To apply Lemmas I.1, I.2, and I.3, we need to verify Assumption I.1. In addition, one of the conditions in these lemmas is that (I.5) holds with probability $1 - o(1)$. Verification of this condition will be done with the help of Lemma I.4, which in turn relies upon Condition WL. Therefore, below we also verify this condition.

We first verify Condition WL with $\epsilon_n = 1/n$ and $N_n = n$. Observe that since $\mathcal{U} = [0, 1]$, we have for any $\epsilon \in (0, 1]$ that $N(\epsilon, \mathcal{U}, d_{\mathcal{U}}) \leqslant 1/\epsilon$, and so ϵ_n and N_n satisfy the inequality $N_n \geqslant N(\epsilon_n, \mathcal{U}, d_{\mathcal{U}})$, which is the first requirement of Condition WL. Further, as in front of Condition WL in Appendix I, let

$$\begin{aligned}
S_u &= \partial_\theta M_u(Y_u, X_u, \theta, a_u)|_{\theta=\theta_u} \\
&= -\Big(1\{Y_u = 1\}\frac{\Lambda'(X_u'\theta_u + a_u(X_u))}{\Lambda(X_u'\theta_u + a_u(X_u))} - 1\{Y_u = 0\}\frac{\Lambda'(X_u'\theta_u + a_u(X_u))}{1 - \Lambda(X_u'\theta_u + a_u(X_u))}\Big) \cdot X_u \\
&= -\Big(1\{Y_u = 1\}(1 - \Lambda(X_u'\theta_u + a_u(X_u)) - 1\{Y_u = 0\}(\Lambda(X_u'\theta_u + a_u(X_u))\Big) \cdot X_u \\
&= -\Big(Y_u - \mathrm{E}_P[Y_u \mid X_u]\Big) \cdot X_u.
\end{aligned}$$

Then $|S_{uk}| \leqslant |X_{uk}|$. In addition, since $\gamma \geqslant 1/n$, we have

$$\Phi^{-1}(1 - \gamma/(2pN_n)) \lesssim \sqrt{\log(pn)} \lesssim \sqrt{\log a_n}.$$

Also, since $f_u \leqslant 1$, Assumption 3.4(ii,iii) yields $\log^{1/2} a_n \lesssim \delta_n n^{1/6}$. Moreover, Assumption 3.4(iv) yields $(\mathrm{E}_P[|S_{uk}|^3])^{1/3} \lesssim 1$ uniformly over $u \in \mathcal{U}$ and $k \in [p]$. Therefore, Condition WL(i) holds for some φ_n satisfying $\varphi_n \lesssim \delta_n$. Assumption 3.4(i,iv) also implies that uniformly over $u \in \mathcal{U}$ and $k \in [p]$,

$$\mathrm{E}_P[|S_{uk}|^2] \leqslant \mathrm{E}_P[|X_{uk}|^2] \lesssim 1 \text{ and } \mathrm{E}_P[|S_{uk}|^2] = \mathrm{E}_P[|f_u X_{uk}|^2] \gtrsim 1, \tag{F.2}$$

and so Condition WL(ii) holds for some $\underline{C}$ and $\bar{C}$ depending only on the constants in Assumption 3.4.

To verify Condition WL(iii), we apply Lemma I.5. Observe that $Y_u = 1\{Y \leqslant (1-u)\underline{y} + u\bar{y}\}$ and the class of functions $\{H(\cdot, u)\colon u \in \mathcal{U}\}$ with $H(y, u) = 1\{y \leqslant (1-u)\underline{y} + u\bar{y}\}$ is VC-subgraph with index bounded by some C. Also, X_u does not depend on u, and by Assumption 3.4(iv,vii,viii), $\mathrm{E}_P[|X_{uk}|^4] \lesssim 1$ uniformly over $k \in [p]$ and $(\mathrm{E}_P[\|X_u\|_\infty^{2q}])^{1/(2q)} \lesssim (\delta_n n^{1/2-1/q})^{1/2}$. Moreover, by Assumption 3.3, $\mathrm{E}_P[|Y_u - Y_{u'}|^4] = |u - u'|$ uniformly over $u, u' \in \mathcal{U}$. Therefore, Lemma I.5 with $2q$ replacing q implies that Condition WL(iii) holds with $\Delta_n = (\log n)^{-1}$ and some φ_n satisfying

$$\varphi_n \lesssim \frac{\delta_n^{1/2} \log a_n}{n^{1/4}} \vee \frac{\log^{1/2} a_n}{n^{1/4}} = o(1),$$

where the last assertion follows from $\log^{1/2} a_n \lesssim \delta_n n^{1/6}$, established above, and $\delta_n^2 \log^4 a_n = o(n)$, which holds by $\delta_n^2 \log a_n = o(1)$.

Next we verify Assumption I.1. It is well-known that the function $\theta \mapsto M_u(Y_u, X_u, \theta)$ is convex almost surely, which is the first requirement of Assumption I.1. Further, let us verify Assumption I.1(b). By Condition WL(iii), which was verified above, we have with probability $1 - o(1)$ that $|(\mathbb{E}_n - \mathrm{E}_P)[S_{uk}^2]| = o(1)$ uniformly over $u \in \mathcal{U}$ and $k \in [p]$. So, it follows from (F.2) that with the same probability we have $\mathbb{E}_n[S_{uk}^2] = (1 - o(1))\mathrm{E}_P[S_{uk}^2]$ uniformly over $u \in \mathcal{U}$ and $k \in [p]$, and so Assumption I.1(b) holds for some Δ_n, ℓ, and L satisfying $\Delta_n = o(1)$, $\ell = 1 - o(1)$, and $L \lesssim 1$ for any $\widehat{\Psi}_u$ such that

$$(1 - o(1))\mathrm{E}_P[S_{uk}^2] \leqslant \widehat{\Psi}_{ukk}^2 \lesssim 1 \text{ with probability } 1 - o(1) \text{ uniformly over } u \in \mathcal{U} \text{ and } k \in [p]. \tag{F.3}$$

Thus, it suffices to verify (F.3). In the case $\bar{m} = 0$, we have by Lemma L.2 and Assumption 3.4(i,iv,vii,viii) that $\mathbb{E}_n[X_{uk}^2] = (1 - o(1))\mathrm{E}[X_{uk}^2]$ with probability $1 - o(1)$ uniformly over $u \in \mathcal{U}$ and $k \in [p]$. Thus, (F.3) holds since in this case,

$$\widehat{\Psi}_{ukk}^2 = \frac{1}{4}\mathbb{E}_n[X_{uk}^2] = \frac{1 - o(1)}{4}\mathrm{E}_P[X_{uk}^2] \lesssim 1$$

and $4^{-1}\mathrm{E}[X_{uk}^2] \geqslant \mathrm{E}[f_u^2 X_{uk}^2] = \mathrm{E}[S_{uk}^2]$ (recall that $f_u^2 \leqslant 1/4$) with probability $1 - o(1)$ uniformly over $u \in \mathcal{U}$ and $k \in [p]$.

To establish (F.3) for $\bar{m} \geqslant 1$, we proceed by induction. Assuming that (F.3) holds when the number of loops in Algorithm 3 is $\bar{m} - 1$, we can complete the proof of the theorem to show that $\|X_u'(\widetilde{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2} \lesssim (s_n \log a_n/n)^{1/2}$ with probability $1 - o(1)$ uniformly over $u \in \mathcal{U}$ for $m = \bar{m} - 1$.

Then for $m = \bar{m}$, we have by the triangle inequality that

$$|\widehat{l}_{uk,m} - l_{u0k}| \leqslant \Big(\mathbb{E}_n[X_{uk}^2\{\Lambda(X_u'\theta_u) + r_u - \Lambda(X_u'\widetilde{\theta}_u)\}^2]\Big)^{1/2}$$
$$\leqslant \Big(\|\Lambda(X_u'\theta_u) - \Lambda(X_u'\widetilde{\theta}_u)\|_{\mathbb{P}_n,2} + \|r_u\|_{\mathbb{P}_n,2}\Big) \cdot \max_{1\leqslant i\leqslant n}\|X_{ui}\|_\infty \lesssim_P \delta_n$$

uniformly over $u \in \mathcal{U}$ and $k \in [p]$ since $\max_{1\leqslant i\leqslant n}\|X_{ui}\|_\infty \lesssim_P n^{1/(2q)}M_{n,2}$ by Assumption 3.4(vii), $\|r_u\|_{\mathbb{P}_n,2} \lesssim_P (s_n \log a_n/n)^{1/2}$ by Assumption 3.5(v), $n^{1/(2q)}M_{n,2}(s_n\log a_n/n)^{1/2} \leqslant \delta_n$ by Assumption 3.4(viii), and the fact that Λ is 1-Lipschitz (observe that $M_{n,2} \geqslant 1$, and so $M_{n,2} \leqslant M_{n,2}^2$). Thus, (F.3) holds with the number of loops in Algorithm 3 being $\bar{m}$. This completes verification of Assumption I.1(b).

To verify Assumption I.1(a), note that for any $\delta \in \mathbb{R}^p$,

$$\{\partial_\theta M_u(Y_u, X_u, \theta_u) - \partial_\theta M_u(Y_u, X_u, \theta_u, a_u)\}'\delta = \{\Lambda(X_u'\theta_u) - \Lambda(X_u'\theta_u + a_u(X_u))\}X_u'\delta = -r_u X_u'\delta,$$

and so

$$\mathbb{E}_n[\partial_\theta M_u(Y_u, X_u, \theta_u) - \partial_\theta M_u(Y_u, X_u, \theta_u, a_u)]'\delta \leqslant \|r_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}X_u'\delta\|_{\mathbb{P}_n,2}$$
$$\leqslant C_n\|\sqrt{w_u}X_u'\delta\|_{\mathbb{P}_n,2}$$

where the first line follows from the Cauchy-Schwarz inequality, and the second holds with probability $1 - \bar{\Delta}_n$ for some C_n satisfying $C_n \lesssim (s_n \log a_n/n)^{1/2}$ by Assumption 3.5. Thus, Assumption I.1(a) follows for given C_n and $\Delta_n = \bar{\Delta}_n = o(1)$.

To verify Assumption I.1(c), note that Lemma O.2 in [11] imply that for any $u \in \mathcal{U}$, $A_u \subset \mathbb{R}^p$, and $\delta \in A_u$,

$$\mathbb{E}_n[M_u(Y_u, X_u, \theta_u + \delta)] - \mathbb{E}_n[M_u(Y_u, X_u, \theta_u)] - \mathbb{E}_n[\partial_\theta M_u(Y_u, X_u, \theta_u)]'\delta$$
$$+ 2\|a_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}\|\sqrt{w_u}X_u'\delta\|_{\mathbb{P}_n,2} \geqslant \Big(\|\sqrt{w_u}X_u'\delta\|_{\mathbb{P}_n,2}^2\Big) \wedge \Big(\bar{q}_{A_u}\|\sqrt{w_u}X_u'\delta\|_{\mathbb{P}_n,2}\Big)$$

where

$$\bar{q}_{A_u} = \inf_{\delta\in A_u} \frac{(\mathbb{E}_n[w_u|X_u'\delta|^2])^{3/2}}{\mathbb{E}_n[w_u|X_u'\delta|^3]}. \tag{F.4}$$

Next, we bound $\|a_u/\sqrt{w_u}\|_{\mathbb{P}_n,2}$. Fix some arbitrary value of X_u. Consider the case that $r_u = r_u(X_u) \geqslant 0$. Then $a_u = a_u(X_u) \geqslant 0$, and so combining the mean-value theorem and (F.1) shows that for some $t \in (0, a_u)$,

$$r_u = a_u\Lambda'(X_u'\theta_u + t).$$

Now, since the function Λ' is unimodal,

$$\Lambda'(X_u'\theta_u + t) \geqslant \Lambda'(X_u'\theta_u) \wedge \Lambda'(X_u'\theta_u + a_u).$$

Further, observe that $\Lambda'(X_u'\theta_u + a_u) = f_u^2$ and

$$\Lambda'(X_u'\theta_u) = \Lambda(X_u'\theta_u)\cdot(1 - \Lambda(X_u'\theta_u)) = (\Lambda(X_u'\theta_u) + r_u - r_u)\cdot(1 - \Lambda(X_u'\theta_u) - r_u + r_u)$$
$$= f_u^2 - r_u(1 - \Lambda(X_u'\theta_u)) + r_u(\Lambda(X_u'\theta_u) + r_u) \geqslant f_u^2 - r_u + 2r_u\Lambda(X_u'\theta_u) \geqslant f_u^2 - r_u.$$

In addition, by Assumption 3.5, $|r_u| \leqslant f_u^2/4$, so that $\Lambda'(X_u'\theta_u) \geqslant 3f_u^2/4$. Thus,

$$|a_u| \leqslant 4r_u/(3f_u^2).$$

Similarly, the same inequality can be obtained in the case that $r_u = r_u(X_u) < 0$. Conclude that

$$\|a_u/\sqrt{w_u}\|_{\mathbb{P}_n,2} \lesssim \|r_u/f_u^3\|_{\mathbb{P}_n,2} \lesssim \sqrt{s_n \log a_n/n}$$

with probability at least $1-\bar{\Delta}_n$ uniformly over $u \in \mathcal{U}$. Therefore, Assumption I.1(c) holds for any $A_u \subset \mathbb{R}^p$ with $\Delta_n = \bar{\Delta}_n$, $C_n \lesssim (s_n \log a_n/n)^{1/2}$ and $\bar{q}_{A_u}$ defined in (F.4).

Next, we apply Lemma I.1. We have to verify the condition on $\bar{q}_{A_u}$ required in the lemma. To do so, recall that $A_u = A_{u,1} \cup A_{u,2}$ where

$$A_{u,1} = \{\delta\colon \|\delta_{T_u}^c\|_1 \leqslant 2\tilde{c}\|\delta_{T_u}\|_1\},$$
$$A_{u,2} = \Big\{\delta\colon \|\delta\|_1 \leqslant \frac{3n}{\lambda}\frac{c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1} C_n \|\sqrt{w_u}X_u'\delta\|_{\mathbb{P}_n,2}\Big\}.$$

Then $\bar{q}_{A_u}$ defined in (F.4) equals $\bar{q}_{A_{u,1}} \wedge \bar{q}_{A_{u,2}}$ where $\bar{q}_{A_{u,1}}$ and $\bar{q}_{A_{u,2}}$ are defined similarly. To bound $\bar{q}_{A_{u,1}}$, we have

$$\begin{aligned}\bar{q}_{A_{u,1}} &\geqslant \inf_{\delta\in A_{u,1}} \frac{\mathbb{E}_n\left[w_u|X_u'\delta|^2\right]^{1/2}}{\max_{1\leqslant i\leqslant n}\|X_{ui}\|_\infty\|\delta\|_1} \gtrsim_P \inf_{\delta\in A_{u,1}} \frac{\mathbb{E}_n\left[w_u|X_u'\delta|^2\right]^{1/2}}{n^{1/(2q)}M_{n,2}\|\delta\|_1}\\ &\geqslant \inf_{\delta\in A_{u,1}} \frac{\mathbb{E}_n\left[w_u|X_u'\delta|^2\right]^{1/2}}{n^{1/(2q)}M_{n,2}(1+2\tilde{c})\sqrt{s_n}\|\delta_{T_u}\|} \geqslant \frac{\bar{\kappa}_{2\tilde{c}}}{n^{1/(2q)}M_{n,2}(1+2\tilde{c})\sqrt{s_n}}\end{aligned}$$

uniformly over $u \in \mathcal{U}$ by Assumption 3.4(viii) and definition of $\bar{\kappa}_{2\tilde{c}}$. By Lemma F.3, sparse eigenvalues of order $\ell_n s_n$, for some sequence $\ell_n \to \infty$, are bounded away from zero and from above so that $\bar{\kappa}_{2\tilde{c}}$ is bounded away from zero with probability $1-o(1)$. Conclude that

$$\bar{q}_{A_{u,1}} \gtrsim_P \frac{1}{n^{1/(2q)}M_{n,2}(1+2\tilde{c})\sqrt{s_n}} \geqslant \frac{1}{\delta_n^{1/2}n^{1/4}} \gtrsim \left(\frac{s_n\log a_n}{\delta_n n}\right)^{1/2}$$

uniformly over $u \in \mathcal{U}$ where the second inequality holds by Assumption 3.4(viii) and the third by Assumption 3.4(vi) (when we apply Assumption 3.4(vi), we use the fact that $M_{n,1} \gtrsim 1$, which in turn follows from Assumption 3.4(i)). Next, to bound $\bar{q}_{A_{u,2}}$, we have

$$\begin{aligned}\bar{q}_{A_{u,2}} &\geqslant \inf_{\delta\in A_{u,2}} \frac{\mathbb{E}_n\left[w_u|X_u'\delta|^2\right]^{1/2}}{\max_{1\leqslant i\leqslant n}\|X_{ui}\|_\infty\|\delta\|_1} \gtrsim_P \inf_{\delta\in A_{u,2}} \frac{\mathbb{E}_n\left[w_u|X_u'\delta|^2\right]^{1/2}}{n^{1/(2q)}M_{n,2}\|\delta\|_1}\\ &\geqslant \frac{\lambda}{3nC_n}\frac{\ell c-1}{c}\frac{\|\widehat{\Psi}_{u0}^{-1}\|_\infty^{-1}}{n^{1/(2q)}M_{n,2}} \gtrsim_P \frac{\lambda}{C_n n^{1+1/(2q)}M_{n,2}}\end{aligned}$$

uniformly over $u \in \mathcal{U}$ since $\sup_{u\in\mathcal{U}}\|\widehat{\Psi}_{u0}^{-1}\|_\infty \lesssim 1$ with probability $1-o(1)$. Substituting $\lambda = c\sqrt{n}\Phi^{-1}(1-\gamma/(2pN_n))$ and $C_n \lesssim (s_n\log a_n/n)^{1/2}$ gives

$$\bar{q}_{A_{u,2}} \gtrsim_P \frac{1}{n^{1/(2q)}M_{n,2}\sqrt{s_n}} \geqslant \frac{1}{\delta_n^{1/2}n^{1/4}} \gtrsim \left(\frac{s_n\log a_n}{\delta_n n}\right)^{1/2}$$

uniformly over $u \in \mathcal{U}$. Moreover,

$$\left(L + \frac{1}{c}\right) \|\widehat{\Psi}_{u0}\|_\infty \frac{\lambda\sqrt{s_n}}{n\bar{\kappa}_{2\tilde{c}}} + 6\tilde{c}C_n \lesssim \left(\frac{s_n \log a_n}{n}\right)^{1/2}$$

since $\sup_{u\in\mathcal{U}} \|\widehat{\Psi}_{u0}\|_\infty \lesssim 1$ with probability $1 - o(1)$. Hence, since $\delta_n = o(1)$, the condition on $\bar{q}_{A_u}$ required in Lemma I.1 is satisfied with probability $1 - o(1)$. In addition, note that (I.5) holds with probability $1 - o(1)$ by Lemma I.4. Therefore, applying Lemma I.1 gives

$$\|\sqrt{w_u}X_u'(\widehat{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2} \lesssim (s_n \log a_n/n)^{1/2} \quad \text{and} \quad \|\widehat{\theta}_u - \theta_u\|_1 \lesssim (s_n^2 \log a_n/n)^{1/2} \tag{F.5}$$

with probability $1 - o(1)$ uniformly over $u \in \mathcal{U}$.

The second inequality in (F.5) gives the second inequality in the first asserted claim of the theorem. To transform the first inequality in (F.5) into the first inequality in the first asserted claim of the theorem (and also to prove other claims), we apply Lemma I.2. We have to verify (I.6). To do so, note that

$$\sup_{u\in\mathcal{U}} \max_{1\leqslant i\leqslant n} |X_{ui}'(\widehat{\theta}_u - \theta_u)| \lesssim_P n^{1/(2q)} M_{n,2}\|\widehat{\theta}_u - \theta_u\|_1 \lesssim \{n^{-1+1/q}M_{n,2}^2 s_n^2 \log a_n\}^{1/2} \lesssim \delta_n = o(1)$$

by Assumption 3.4(vii, viii) and since $M_{n,2} \geqslant 1$. Also, note that uniformly over t and Δt in $\mathbb{R}$ with $|\Delta t| \leqslant 1$, we have

$$\begin{aligned} |\Lambda(t + \Delta t) - \Lambda(t)| &= \left|\frac{e^{t+\Delta t}}{1 + e^{t+\Delta t}} - \frac{e^t}{1 + e^t}\right| = \frac{|e^{t+\Delta t} - e^t|}{(1 + e^{t+\Delta t})(1 + e^t)} \\ &= \frac{e^t|e^{\Delta t} - 1|}{(1 + e^{t+\Delta t})(1 + e^t)} \lesssim \frac{e^t|e^{\Delta t} - 1|}{(1 + e^t)^2} \lesssim \Lambda'(t)\Delta t. \end{aligned} \tag{F.6}$$

Thus, with probability $1 - o(1)$,

$$\begin{aligned} |[\partial_\theta M_u(Y_{ui}, X_{ui}, \widehat{\theta}_u) - \partial_\theta M_u(Y_{ui}, X_{ui}, \theta_u)]'\delta| &= |\Lambda(X_{ui}'\widehat{\theta}_u) - \Lambda(X_{ui}'\theta_u)| \cdot |X_{ui}'\delta| \\ &\lesssim \Lambda'(X_{ui}'\theta_u) \cdot |X_{ui}'(\widehat{\theta}_u - \theta_u)| \cdot |X_{ui}'\delta| \end{aligned}$$

uniformly over $i = 1, \ldots, n$ and $u \in \mathcal{U}$. Also, since $|r_u| \leqslant w_u/4$ by Assumption 3.5 and $w_u = f_u^2 \leqslant 1$,

$$\Lambda'(X_{ui}'\theta_u) = f_{ui}^2 - r_{ui} + 2r_{ui}\Lambda(X_{ui}'\theta_u) + r_{ui}^2 \leqslant w_{ui} + 3|r_{ui}| + w_{ui}^2 \leqslant 3w_{ui} \leqslant 3\sqrt{w_{ui}};$$

see the expression for $\Lambda'(X_{ui}'\theta_u)$ above in this proof. Therefore, for some constant C,

$$\begin{aligned} |\mathbb{E}_n[\partial_\theta M_u(Y_u, X_u, \widehat{\theta}_u) - \partial_\theta M_u(Y_u, X_u, \theta_u)]'\delta| &\leqslant C\|\sqrt{w_u}X_u'(\widehat{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2}\|X_u'\delta\|_{\mathbb{P}_n,2} \\ &\leqslant L_n\|X_u'\delta\|_{\mathbb{P}_n,2} \end{aligned}$$

with probability $1 - o(1)$ uniformly over $u \in \mathcal{U}$ for some L_n satisfying $L_n \lesssim (s_n \log a_n/n)^{1/2}$. Hence, since $\sup_{u\in\mathcal{U}} \phi_{\max}(\ell_n s_n, u) \lesssim 1$ with probability $1 - o(1)$ for some $\ell_n \to \infty$ sufficiently slowly by Lemma F.3, Lemma I.2 implies that $\sup_{u\in\mathcal{U}} \|\widehat{\theta}_u\|_0 \lesssim s_n$ with probability $1 - o(1)$, which is the second asserted claim of the theorem.

In turn, since $\sup_{u\in\mathcal{U}} \|\widehat{\theta}_u\|_0 \lesssim s_n$ with probability $1 - o(1)$, Lemma F.3 also establishes that

$$\|\sqrt{w_u}X_u'(\widehat{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2} \gtrsim \|\sqrt{w_u}X_u'(\widehat{\theta}_u - \theta_u)\|_{\mathrm{P}_P,2} \gtrsim \|\widehat{\theta}_u - \theta_u\|$$

with probability $1 - o(1)$ uniformly over $u \in \mathcal{U}$, where the inequality follows from Assumption 3.4(i). Combining these inequalities with (F.5) gives the first inequality in the first asserted claim of the theorem.

It remains to prove the claim about the estimators $\widetilde{\theta}_u$. We apply Lemma I.3. We have to verify the condition (I.7) on $\bar{q}_{A_u}$ required in the lemma. To do so, we first bound $\bar{q}_{A_u}$ from below for $A_u = \{\delta \in \mathbb{R}^p : \|\delta\|_0 \leqslant Cs\}$ where C is a constant such that $\widehat{s}_u + s_n \leqslant Cs_n$ with probability $1 - o(1)$ uniformly over $u \in \mathcal{U}$. We have

$$\bar{q}_{A_u} = \inf_{\delta \in A_u} \frac{\mathbb{E}_n\left[w_u |X_u'\delta|^2\right]^{3/2}}{\mathbb{E}_n\left[w_u |X_u'\delta|^3\right]} \geqslant \inf_{\delta \in A_u} \frac{\mathbb{E}_n\left[w_u |X_u'\delta|^2\right]^{1/2}}{\max_{1\leqslant i\leqslant n} \|X_{ui}\|_\infty \|\delta\|_1}$$

$$\geqslant \inf_{\|\delta\|_0 \leqslant Cs_n} \frac{\mathbb{E}_n\left[w_u |X_u'\delta|^2\right]^{1/2}}{\max_{1\leqslant i\leqslant n} \|X_{ui}\|_\infty \sqrt{Cs_n}\|\delta\|_2} \gtrsim_P \frac{\sqrt{\phi_{\min}(Cs_n, u)}}{\sqrt{s_n} n^{1/(2q)} M_{n,2}} \gtrsim \frac{\log^{1/4} a_n}{\delta_n^{1/2} n^{1/4}}$$

uniformly over $u \in \mathcal{U}$, where the inequality preceding the last one follows from Assumption 3.4(vii) and the definition of $\phi_{\min}(Cs_n, u)$, and the last one follows from Assumption 3.4(viii) and the observation that by Lemma F.3, $\inf_{u\in\mathcal{U}} \phi_{\min}(Cs_n, u)$ is bounded away from zero with probability $1 - o(1)$ uniformly over $u \in \mathcal{U}$.

Next we bound from above the right-hand side of (I.7). It follows by (I.8) that uniformly over $u \in \mathcal{U}$ with probability $1 - o(1)$,

$$\mathbb{E}_n[M_u(Y_u, X_u, \widetilde{\theta}_u)] - \mathbb{E}_n[M_u(Y_u, X_u, \theta_u)] \lesssim s_n \log a_n / n$$

since $\lambda/n \lesssim (\log a_n/n)^{1/2}$, $\|\widehat{\theta}_u - \theta_u\|_1 \lesssim (s_n^2 \log a_n/n)^{1/2}$, and $\sup_{u\in\mathcal{U}} \|\widehat{\Psi}_{u0}\|_\infty \lesssim 1$ with probability $1 - o(1)$. Furthermore, $C_n \lesssim (s_n \log a_n/n)^{1/2}$ and

$$\sup_{u\in\mathcal{U}} \|\mathbb{E}_n[S_u]\|_\infty \leqslant \sup_{u\in\mathcal{U}} \|\widehat{\Psi}_{u0}\|_\infty \|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[S_u]\|_\infty \lesssim \lambda/n$$

with probability $1 - o(1)$ by the choice of λ; see Lemma I.4. Hence, it follows that the right-hand side of (I.7) is bounded up-to a constant by $(s_n \log a_n/n)^{1/2}$ with probability $1 - o(1)$ uniformly over $u \in \mathcal{U}$. Since $s_n^2 \log a_n/n \lesssim 1$ (see Assumption 3.4(viii) and recall that $M_{n,2} \geqslant 1$) and $\delta_n = o(1)$, the condition (I.7) on $\bar{q}_{A_u}$ required in Lemma I.3 holds with probability $1 - o(1)$ uniformly over $u \in \mathcal{U}$. Hence, Lemma I.3 implies that

$$\|\sqrt{w_u} X_u(\widetilde{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2} \lesssim (s_n \log a_n/n)^{1/2}$$

with probability $1 - o(1)$ uniformly over $u \in \mathcal{U}$. Finally, as in the case of $\widehat{\theta}_u$'s, we also have $\|\sqrt{w_u} X_u(\widetilde{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2} \gtrsim \|\widetilde{\theta}_u - \theta_u\|$ with probability $1 - o(1)$ uniformly over $u \in \mathcal{U}$, which gives the last asserted claim and completes the proof of the theorem. ∎

Proof of Theorem 4.2. The strategy of this proof is similar to that of Theorem 4.1. In particular, we will rely upon results in Appendix I with $\mathcal{U}$ and p replaced by $\widetilde{\mathcal{U}} = \mathcal{U} \times [\tilde{p}]$ and $\bar{p} = p + \tilde{p} - 1$, respectively, where for $\tilde{u} = (u, j) \in \widetilde{\mathcal{U}}$, we set $Y_{\tilde{u}} = D_j$, $X_{\tilde{u}} = (X^j)' = (D_{[\tilde{p}]\backslash j}', X')'$, $\theta_{\tilde{u}} = \bar{\gamma}_u^j$,

$\widehat{\theta}_{\tilde u} = \widehat{\gamma}^j_u$, $\widetilde{\theta}_{\tilde u} = \widetilde{\gamma}^j_u$, $a_{\tilde u} = (f_u, \bar r_{uj})$, $\bar r_{\tilde u} = \bar r_{uj} = X^j(\gamma^j_u - \bar\gamma^j_u)$, and $w_{\tilde u} = \widehat{f}^2_u$. Note that for all $\tilde u \in \widetilde{\mathcal{U}}$, we have that $\theta_{\tilde u}$ satisfies (I.1) where

$$M_{\tilde u}(Y_{\tilde u}, X_{\tilde u}, \theta, a) = 2^{-1} f^2 (D_j - X^j\theta - r)^2$$

for θ being a vector in $\mathbb{R}^{\bar p}$ and a being a pair (f, r) of functions of D and X. Similarly, $\widehat{\theta}_{\tilde u}$ satisfies (I.2) where $M_{\tilde u}(Y_{\tilde u}, X_{\tilde u}, \theta) = M_{\tilde u}(Y_{\tilde u}, X_{\tilde u}, \theta, (\widehat{f}_u, \mathcal{O}))$ and $\mathcal{O} = \mathcal{O}(D, X)$, the identically zero function of D and X.

We first verify Condition WL with

$$\epsilon_n = \frac{\delta_n^2}{n^{1/2+1/q}(p+\tilde p)^{1/2}(M^2_{n,1} \vee M^2_{n,2})},$$

$N_n = p\tilde p^2 n^2$, and the following semi-metric $d_{\widetilde{\mathcal{U}}}$ on $\widetilde{\mathcal{U}}$: for all $\tilde u = (u, j)$ and $\tilde u' = (u', j')$ in $\widetilde{\mathcal{U}}$, $d_{\widetilde{\mathcal{U}}}(\tilde u, \tilde u') = |u - u'|$ if $j = j'$ and $d_{\widetilde{\mathcal{U}}}(\tilde u, \tilde u') = 1$ otherwise. Observe that $N(\epsilon, \widetilde{\mathcal{U}}, d_{\widetilde{\mathcal{U}}}) \leqslant \tilde p/\epsilon$ for all $\epsilon > 0$. Also, note that $M^2_{n,1} \vee M^2_{n,2} \leqslant \delta_n n^{1/2-1/q}$ by Assumption 3.4(vi,viii), and so

$$1/\epsilon_n \leqslant n(p+\tilde p)^{1/2}/\delta_n \leqslant n^2(p+\tilde p)^{1/2}$$

since $\delta_n \geqslant 1/n$ by Assumption 3.4(i,vii,viii). Thus, ϵ_n and N_n satisfy the inequality $N_n \geqslant N(\epsilon_n, \widetilde{\mathcal{U}}, d_{\widetilde{\mathcal{U}}})$, which is the first requirement of Condition WL.

Next, we verify Condition WL(i). As in front of Condition WL in Appendix I, for $\tilde u = (u, j)$, let

$$S_{\tilde u} = S_{uj} = \partial_\theta M_{\tilde u}(Y_{\tilde u}, X_{\tilde u}, \theta, a_{\tilde u})|_{\theta=\theta_{\tilde u}} = -f^2_u(D_j - X^j\gamma^j_u)(X^j)' = -f^2_u Z^j_u (X^j)'.$$

Then the inequality $\Phi^{-1}(1-t) \lesssim \sqrt{\log(1/t)}$, which holds uniformly over $t \in (0, 1/2)$, implies that

$$(\mathrm{E}_P[|S_{\tilde u k}|^3])^{1/3}\Phi^{-1}(1-\gamma/2pN_n) \lesssim (\mathrm{E}_P[|Z^j_u X^j_k|^3])^{1/3}\log^{1/2} a_n \leqslant \delta_n n^{1/6}$$

uniformly over $\tilde u \in \widetilde{\mathcal{U}}$ and $k \in [\bar p]$, where the second inequality holds by Assumption 3.4(iii). Hence, Condition WL(i) holds for some φ_n satisfying $\varphi_n \lesssim \delta_n$.

To verify Condition WL(ii), note that by Assumption 3.4(ii), we have $\mathrm{E}_P[S^2_{\tilde u k}] \gtrsim 1$ and

$$\mathrm{E}_P[S^2_{\tilde u k}] \leqslant \mathrm{E}_P[|Z^j_u X^j_k|^2] \leqslant \mathrm{E}_P[|Z^j_u|^4 + |X^j_k|^4] \lesssim 1$$

uniformly over $\tilde u = (u, j) \in \widetilde{\mathcal{U}}$ and $k \in [\bar p]$ by Assumption 3.4(iv).

To verify Condition WL(iii), we use the decomposition

$$S_{\tilde u k} - S_{\tilde u' k} = -(f^2_u - f^2_{u'})Z^j_u X^j_k + f^2_{u'}X^j(\gamma^j_u - \gamma^j_{u'})X^j_k$$

for $\tilde u = (u, j)$ and $\tilde u' = (u', j)$ in $\widetilde{\mathcal{U}}$. By (E.2) and (E.1) we have

$$|f^2_u - f^2_{u'}| \lesssim |u - u'| \quad \text{and} \quad \|\gamma^j_u - \gamma^j_{u'}\|_1 \lesssim \sqrt{p+\tilde p}|u - u'| \tag{F.7}$$

uniformly over $u, u' \in \mathcal{U}$ and $j \in [\tilde{p}]$. Therefore, uniformly over $\tilde{u} = (u, j)$ and $\tilde{u}' = (u', j)$ in $\widetilde{\mathcal{U}}$ such that $|u - u'| \leqslant \epsilon_n$, we have

$$
\begin{aligned}
|S_{\tilde{u}k} - S_{\tilde{u}'k}| &\lesssim \Big(|Z_u^j| \|X^j\|_\infty + \|X^j\|_\infty^2 \sqrt{p + \tilde{p}}\Big) \cdot |u - u'| \\
&\lesssim \Big(|Z_u^j|^2 + \|X^j\|_\infty^2\Big) \cdot \sqrt{p + \tilde{p}} \epsilon_n. \qquad \text{(F.8)}
\end{aligned}
$$

Since $\mathrm{E}_P[\max_{1 \leqslant i \leqslant n, j \in [\tilde{p}]} \sup_{u \in \mathcal{U}} |Z_{ui}^j|^2] \leqslant n^{1/q} M_{n,1}^2$ and $\mathrm{E}_P[\max_{1 \leqslant i \leqslant n, j \in [\tilde{p}]} \|X_i^j\|_\infty^2] \leqslant n^{1/q} M_{n,2}^2$ by Assumption 3.4(v,vii), we have by Markov's inequality that with probability $1 - o(1)$,

$$
\sup_{d_{\widetilde{\mathcal{U}}}(\tilde{u}, \tilde{u}') \leqslant \epsilon_n} \max_{k \in [\bar{p}]} |\mathbb{E}_n[S_{\tilde{u}k} - S_{\tilde{u}'k}]| \lesssim \delta_n n^{-1/2}.
$$

In addition, uniformly over $\tilde{u}, \tilde{u}' \in \widetilde{\mathcal{U}}$ with $d_{\widetilde{\mathcal{U}}}(\tilde{u}, \tilde{u}') \leqslant \epsilon_n$ and $k \in [\bar{p}]$,

$$
\begin{aligned}
|\mathrm{E}_P[S_{\tilde{u}k}^2 - S_{\tilde{u}'k}^2]| &\leqslant \Big(\mathrm{E}_P[(S_{\tilde{u}k} - S_{\tilde{u}'k})^2]\Big)^{1/2} \cdot \Big(\mathrm{E}_P[(S_{\tilde{u}k} + S_{\tilde{u}'k})^2]\Big)^{1/2} \\
&\lesssim \Big((M_{n,1}^2 + M_{n,2}^2)(p + \tilde{p})^{1/2} \epsilon_n\Big)^{1/2} \lesssim \delta_n.
\end{aligned}
$$

Further, let $\mathcal{U}^\epsilon$ denote a minimal ϵ_n-net for $\mathcal{U}$. Using (F.7) and (F.8), we obtain that with probability $1 - o(1)$,

$$
\sup_{u \in \mathcal{U}} \max_{j \in [\tilde{p}], k \in [\bar{p}]} |(\mathbb{E}_n - \mathrm{E}_P)[S_{ujk}^2]| \lesssim \sup_{u \in \mathcal{U}^\epsilon} \max_{j \in [\tilde{p}], k \in [\bar{p}]} |(\mathbb{E}_n - \mathrm{E}_P)[S_{ujk}^2]| + \delta_n.
$$

To bound the first term on the right-hand side of this inequality, we apply Lemma B.1 with X_{ui} and p replaced by $S_{\tilde{u}i}$ and $\bar{p}$, respectively, and $k = 1$, where $S_{\tilde{u}i} = (S_{\tilde{u}1i}, \ldots, S_{\tilde{u}[\bar{p}]i})' = -f_{ui}^2 Z_{ui}^j (X_i^j)'$ for $\tilde{u} = (u, j) \in \widetilde{\mathcal{U}}$ and $i = 1, \ldots, n$. With

$$
K^2 = \mathrm{E}_P\left[\max_{1 \leqslant i \leqslant n} \sup_{u \in \mathcal{U}} \max_{j \in [\tilde{p}], k \in [\bar{p}]} S_{ujki}^2\right] \leqslant \mathrm{E}_P\left[\max_{1 \leqslant i \leqslant n} \sup_{u \in \mathcal{U}, j \in [\tilde{p}]} |Z_{ui}^j|^2 \|X_i^j\|_\infty^2\right] \lesssim n^{2/q}(M_{n,1}^4 + M_{n,2}^4),
$$

which holds by Assumption 3.4(v,vii), the lemma yields

$$
\max_{u \in \mathcal{U}^\epsilon} \max_{j \in [\tilde{p}], k \in [\bar{p}]} \Big|(\mathbb{E}_n - \mathrm{E}_P)[S_{ujk}^2]\Big| \lesssim_P n^{-1/2} n^{1/q} (M_{n,1}^2 + M_{n,2}^2) \log a_n \lesssim \delta_n \log^{1/2} a_n = o(1)
$$

by Assumption 3.4(vi) and (viii). Thus, Condition WL(iii) holds with some Δ_n and φ_n satisfying $\Delta_n = o(1)$ and $\varphi_n = o(1)$.

Next, we verify Assumption I.1. The function $\theta \mapsto M_{\tilde{u}}(Y_{\tilde{u}}, X_{\tilde{u}}, \theta)$ is convex almost surely, which is the first requirement of Assumption I.1. Further, to verify Assumption I.1(a), note that

$$
[\partial_\theta M_u(Y_u, X, \theta_u) - \partial_\theta M_u(Y_u, X, \theta_u, a_u)]'\delta = -\Big(\widehat{f}_u^2 \bar{r}_{uj} + (\widehat{f}_u^2 - f_u^2) Z_u^j\Big) \cdot X^j \delta,
$$

so that by the Cauchy-Schwarz and triangle inequalities, since $w_u = \widehat{f}_u^2$, we have

$$
\begin{aligned}
&|\mathbb{E}_n[\partial_\theta M_u(Y_u, X, \theta_u) - \partial_\theta M_u(Y_u, X, \theta_u, a_u)]'\delta| \\
&\qquad \leqslant \Big(\|\widehat{f}_u \bar{r}_{uj}\|_{\mathbb{P}_n,2} + \|(\widehat{f}_u^2 - f_u^2) Z_u^j / \widehat{f}_u\|_{\mathbb{P}_n,2}\Big) \cdot \|\sqrt{w_u} X^j \delta\|_{\mathbb{P}_n,2}.
\end{aligned}
$$

To bound $\sup_{u\in\mathcal{U},j\in[\tilde{p}]}\|\widehat{f}_u\bar{r}_{uj}\|_{\mathbb{P}_n,2}$, note that $\widehat{f}_u\leqslant 1$, and so Lemma F.1 shows that with probability $1-o(1)$,

$$\sup_{u\in\mathcal{U},j\in[\tilde{p}]}\|\widehat{f}_u\bar{r}_{uj}\|_{\mathbb{P}_n,2}\leqslant\sup_{u\in\mathcal{U},j\in[\tilde{p}]}\|\bar{r}_{uj}\|_{\mathbb{P}_n,2}\lesssim(s_n\log a_n/n)^{1/2}.$$

Also, by Lemma F.2, with probability $1-o(1)$,

$$\sup_{u\in\mathcal{U},j\in[\tilde{p}]}\|(\widehat{f}_u^2-f_u^2)Z_u^j/\widehat{f}_u\|_{\mathbb{P}_n,2}\lesssim(s_n\log a_n/n)^{1/2}. \tag{F.9}$$

Hence, Assumption I.1(a) holds for some Δ_n and C_n satisfying $\Delta_n=o(1)$ and $C_n\lesssim(s_n\log a_n/n)^{1/2}$.

To prove Assumption I.1(b), as in Appendix I, for $\tilde{u}=(u,j)\in\widetilde{\mathcal{U}}$, let $\widehat{\Psi}_{\tilde{u}0}=\widehat{\Psi}_{uj0}=\operatorname{diag}(\{l_{\tilde{u}0k},k\in[\bar{p}]\})$ where $l_{\tilde{u}0k}=l_{uj0k}=(\mathbb{E}_n[S_{\tilde{u}k}^2])^{1/2}$, $k\in[\bar{p}]$. Note that by Condition WL(ii,iii), which is verified above,

$$1\lesssim\widehat{\Psi}_{\tilde{u}0k}\lesssim 1$$

with probability $1-o(1)$ uniformly over $\tilde{u}\in\widetilde{\mathcal{U}}$ and $k\in[\bar{p}]$. Now, suppose that $\bar{m}=0$ (even though Algorithm 4 requires $\bar{m}\geqslant 1$). Then uniformly over $u\in\mathcal{U}$, $j\in[\tilde{p}]$, and $k\in[\bar{p}]$ with probability $1-o(1)$,

$$\widehat{l}_{ujk,0}\gtrsim\Big(\mathbb{E}_n[\widehat{f}_u^4(D_jX_k^j)^2]\Big)^{1/2}\gtrsim\Big(\mathbb{E}_n[f_u^4(D_jX_k^j)^2]\Big)^{1/2}\gtrsim 1$$

where the second inequality follows from the observation that $|\widehat{f}_{ui}^2-f_{ui}^2|\leqslant f_{ui}^2$ with probability $1-o(1)$ uniformly over $i=1,\dots,n$ and $u\in\mathcal{U}$ (see (F.12) in the proof of Lemma F.2), and the third from the same derivations as those used to obtain Condition WL(ii,iii). Also, uniformly over $u\in\mathcal{U}$, $j\in[\tilde{p}]$, and $k\in[\bar{p}]$,

$$\widehat{l}_{ujk,0}\leqslant\max_{1\leqslant i\leqslant n}\|X_i^j\|_\infty(\mathbb{E}_n[D_j^2])^{1/2}\lesssim_P n^{1/(2q)}M_{n,2}$$

by Assumption 3.4(vii) since $\widehat{f}_u\leqslant 1$. Therefore, Assumption I.1(b) holds with some Δ_n, ℓ, and L satisfying $\Delta_n=o(1)$, $\ell\gtrsim 1$, and $L\lesssim n^{1/(2q)}M_{n,2}\log^{1/2}a_n$.

To establish Assumption I.1(b) for $\bar{m}\geqslant 1$, which is required by Algorithm 4, we proceed by induction. Assuming that Assumption I.1(b) holds with some Δ_n, ℓ, and L satisfying $\Delta_n=o(1)$, $\ell\gtrsim 1$, and $L\lesssim n^{1/(2q)}M_{n,2}\log^{1/2}a_n$ when the number of loops in Algorithm 4 is $\bar{m}-1$, we can complete the proof of the theorem to show that $\|\widehat{f}_uX^j(\tilde{\gamma}_u^j-\bar{\gamma}_u^j)\|_{\mathbb{P}_n,2}\lesssim(L+1)(s_n\log a_n/n)^{1/2}$ with probability $1-o(1)$ uniformly over $u\in\mathcal{U}$ and $j\in[\tilde{p}]$ for $m=\bar{m}-1$. Thus, by the triangle inequality,

$$\|\widehat{f}_uX^j(\tilde{\gamma}_u^j-\gamma_u^j)\|_{\mathbb{P}_n,2}\lesssim(L+1)(s_n^2\log a_n/n)^{1/2} \tag{F.10}$$

with probability $1-o(1)$ uniformly over $u\in\mathcal{U}$ and $j\in[\tilde{p}]$ since

$$\begin{aligned}\|\widehat{f}_uX^j(\bar{\gamma}_u^j-\gamma_u^j)\|_{\mathbb{P}_n,2}&\leqslant\|X^j(\bar{\gamma}_u^j-\gamma_u^j)\|_{\mathbb{P}_n,2}\\&\leqslant\max_{1\leqslant i\leqslant n}\|X_i^j\|\cdot\|\bar{\gamma}_u^j-\gamma_u^j\|_1\lesssim_P n^{1/(2q)}M_{n,2}(s_n^2\log a_n/n)^{1/2}\end{aligned}$$

uniformly over $u \in \mathcal{U}$ and $j \in [\tilde{p}]$. Then for $m = \bar{m}$, we have uniformly over $u \in \mathcal{U}$, $j \in [\tilde{p}]$, and $k \in [\bar{p}]$,

$$
\begin{aligned}
|\widehat{l}_{ujk,m} - l_{uj0k}| &= \left| \left(\mathbb{E}_n[\widehat{f}_u^4 (X_k^j)^2 (D^j - X^j \widetilde{\gamma}_u^j)^2]\right)^{1/2} - \left(\mathbb{E}_n[f_u^4 (X_k^j)^2 (D^j - X^j \gamma_u^j)^2]\right)^{1/2} \right| \\
&\leqslant \left| \left(\mathbb{E}_n[\widehat{f}_u^4 (X_k^j)^2 (X^j (\widetilde{\gamma}_u^j - \gamma_u^j))^2]\right)^{1/2} - \left(\mathbb{E}_n[(\widehat{f}_u^2 - f_u^2)^2 (X_k^j)^2 (D^j - X^j \gamma_u^j)^2]\right)^{1/2} \right| \\
&\lesssim \left(\|\widehat{f}_u X^j (\widetilde{\gamma}_u^j - \gamma_u^j)\|_{\mathbb{P}_n,2} + \|(\widehat{f}_u^2 - f_u^2) Z_u^j / \widehat{f}_u \|_{\mathbb{P}_n,2} \right) \cdot \max_{1 \leqslant i \leqslant n} \|X^j\|_\infty \\
&\lesssim_P n^{1/(2q)} M_{n,2} (s_n^2 \log a_n^2 / n)^{1/2} n^{1/(2q)} M_{n,2} = n^{-1/2 + 1/q} M_{n,2}^2 s_n \log a_n \\
&\lesssim \delta_n \log^{1/2} a_n = o(1)
\end{aligned}
$$

where the second line follows from the triangle inequality and the observation that $(\widehat{f}_u^2 - f_u^2)^2 \leqslant f_u^4 - \widehat{f}_u^4$, the third from $\widehat{f}_i \leqslant 1$ and $f_u \leqslant 1$, the fourth from (F.9) and (F.10), and the fifth from Assumption 3.4(viii). Thus, for $\bar{m} \geqslant 1$, Assumption I.1(b) holds for some Δ_n, ℓ, and L satisfying $\Delta_n = o(1)$, $\ell \gtrsim 1$, and $L \lesssim 1$.

Further, Assumption I.1(c) holds with $\Delta_n = 0$ and $\bar{q}_{A_{\tilde{u}}} = \infty$ for any $A_{\tilde{u}}$ since for any $\tilde{u} = (u, j) \in \widetilde{\mathcal{U}}$ and $\delta \in \mathbb{R}^{\bar{p}}$, we have

$$
\mathbb{E}_n[M_{\tilde{u}}(Y_{\tilde{u}}, X_{\tilde{u}}, \theta_{\tilde{u}} + \delta)] - \mathbb{E}_n[M_{\tilde{u}}(Y_{\tilde{u}}, X_{\tilde{u}}, \theta_{\tilde{u}})] = -\mathbb{E}_n[\widehat{f}_u^2 (D_j - X^j \bar{\gamma}_u^j) X^j \delta] + 2^{-1} \mathbb{E}_n[(\widehat{f}_u X^j \delta)^2]
$$

and

$$
2^{-1} \mathbb{E}_n[(\widehat{f}_u X^j \delta)^2] \geqslant \mathbb{E}_n[\widehat{f}_u (X^j \delta)^2] = \|\sqrt{w_u} X_{\tilde{u}}' \delta\|_{\mathbb{P}_n,2}
$$

where we used $\widehat{f}_u \leqslant 1/2$.

We are now ready to apply Lemma I.1. Observe that by Lemma I.4, λ satisfies (I.5) with probability $1 - o(1)$. Also, as established in (F.12) in the proof of Lemma F.2, $|\widehat{f}_{ui}^2 - f_{ui}^2| \leqslant f_{ui}^2/2$ with probability $1 - o(1)$ uniformly over $i = 1, \dots, n$ and $u \in \mathcal{U}$. Therefore, since Lemma F.3 implies that for some ℓ_n satisfying $\ell_n \to \infty$,

$$
1 \lesssim \min_{\|\delta\|_0 \leqslant \ell_n s_n} \frac{\|f_u X_{\tilde{u}}' \delta\|_{\mathbb{P}_n,2}^2}{\|\delta\|^2} \leqslant \max_{\|\delta\|_0 \leqslant \ell_n s_n} \frac{\|X_{\tilde{u}}' \delta\|_{\mathbb{P}_n,2}^2}{\|\delta\|^2} \lesssim 1
$$

with probability $1 - o(1)$ uniformly over $\tilde{u} \in \widetilde{\mathcal{U}}$, it follows that $\bar{\kappa}_{2\tilde{c}} \gtrsim 1$ with probability $1 - o(1)$. In addition, $\sup_{\tilde{u} \in \widetilde{\mathcal{U}}} \|\widehat{\Psi}_{\tilde{u}0}\|_\infty \lesssim 1$ and $\sup_{\tilde{u} \in \widetilde{\mathcal{U}}} \|\widehat{\Psi}_{\tilde{u}0}^{-1}\|_\infty \lesssim 1$ with probability $1 - o(1)$. Therefore, applying Lemma I.1 gives

$$
\|\widehat{f}_u X^j (\widehat{\gamma}_u^j - \bar{\gamma}_u^j)\|_{\mathbb{P}_n,2} \lesssim (s_n \log a_n / n)^{1/2} \quad \text{and} \quad \|\widehat{\gamma}_u^j - \bar{\gamma}_u^j\|_1 \lesssim (s_n^2 \log a_n / n)^{1/2} \tag{F.11}
$$

with probability $1 - o(1)$ uniformly over $u \in \mathcal{U}$ and $j \in [\tilde{p}]$.

The second inequality in (F.11) gives the second inequality in the first asserted claim of the theorem. To transform the first inequality in (F.11) into the first inequality in the first asserted claim of the theorem (and also to prove other claims), we apply Lemma I.2. We have to verify (I.6). To do so, note that for $\tilde{u} = (u, j) \in \widetilde{\mathcal{U}}$, we have

$$
|[\partial_\theta M_{\tilde{u}}(Y_{\tilde{u}}, X_{\tilde{u}}, \widehat{\theta}_{\tilde{u}}) - \partial_\theta M_{\tilde{u}}(Y_{\tilde{u}}, X_{\tilde{u}}, \theta_{\tilde{u}})]'\delta| \leqslant |\widehat{f}_u X^j (\widehat{\theta}_{\tilde{u}} - \theta_{\tilde{u}})| \cdot |\widehat{f}_u X^j \delta|.
$$

Therefore, by the Cauchy-Schwarz inequality and since $\widehat{f}_u \leqslant 1$,

$$|\mathbb{E}_n[\partial_\theta M_{\tilde{u}}(Y_{\tilde{u}}, X_{\tilde{u}}, \widehat{\theta}_{\tilde{u}}) - \partial_\theta M_{\tilde{u}}(Y_{\tilde{u}}, X_{\tilde{u}}, \theta_{\tilde{u}})]'\delta| \leqslant \|\widehat{f}_u X^j(\widehat{\gamma}_u^j - \bar{\gamma}_u^j)\|_{\mathbb{P}_n,2}\|\widehat{f}_u X^j\delta\|_{\mathbb{P}_n,2}$$
$$\leqslant L_n\|X^j\delta\|_{\mathbb{P}_n,2}$$

with probability $1-o(1)$ uniformly over $\tilde{u} = (u,j) \in \widetilde{\mathcal{U}}$ for some L_n satisfying $L_n \lesssim (s_n \log a_n/n)^{1/2}$. Thus, since $\sup_{\tilde{u}\in\widetilde{\mathcal{U}}} \phi_{\max}(\ell_n s_n, \tilde{u}) \lesssim 1$ for some $\ell_n \to \infty$ with probability $1 - o(1)$ by Lemma F.3, it follows from Lemma I.2 that $\sup_{u\in\mathcal{U}} \|\widehat{\gamma}_u^j\|_0 \lesssim s_n$ with probability $1 - o(1)$ uniformly over $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, which is the second asserted claim of the theorem.

In turn, with probability $1 - o(1)$, uniformly over $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, we have

$$\|\widehat{f}_u X^j(\widehat{\gamma}_u^j - \bar{\gamma}_u^j)\|_{\mathbb{P}_n,2} \gtrsim \|f_u X^j(\widehat{\gamma}_u^j - \bar{\gamma}_u^j)\|_{\mathbb{P}_n,2} \gtrsim \|\widehat{\gamma}_u^j - \bar{\gamma}_u^j\|.$$

Combining these inequalities with (F.11) gives the first inequality in the first asserted claim of the theorem.

It remains to prove the claim about the estimators $\widetilde{\gamma}_u^j$. We apply Lemma I.3. The condition (I.7) on $\bar{q}_{A_{\tilde{u}}}$ required in the lemma holds almost surely since $\bar{q}_{A_{\tilde{u}}} = \infty$. Also, it follows from (I.8) that uniformly over $u \in \mathcal{U}$ and $j \in [\tilde{p}]$ with probability $1 - o(1)$,

$$\mathbb{E}_n[M_{\tilde{u}}(Y_{\tilde{u}}, X_{\tilde{u}}, \widetilde{\theta}_{\tilde{u}})] - \mathbb{E}_n[M_{\tilde{u}}(Y_{\tilde{u}}, X_{\tilde{u}}, \theta_{\tilde{u}})] \lesssim s_n \log a_n/n$$

since $\lambda/n \lesssim (\log a_n/n)^{1/2}$, $\sup_{u\in\mathcal{U},j\in[\tilde{p}]} \|\widehat{\gamma}_u^j - \bar{\gamma}_u^j\|_1 \lesssim (s_n^2 \log a_n/n)^{1/2}$, and $\sup_{u\in\mathcal{U},j\in[\tilde{p}]} \|\widehat{\Psi}_{uj0}\|_\infty \lesssim 1$ with probability $1 - o(1)$. Furthermore, $C_n \lesssim (s_n \log a_n/n)^{1/2}$ and

$$\sup_{\tilde{u}\in\widetilde{\mathcal{U}}} \|\mathbb{E}_n[S_{\tilde{u}}]\|_\infty \leqslant \sup_{\tilde{u}\in\widetilde{\mathcal{U}}} \|\widehat{\Psi}_{\tilde{u}0}\|_\infty \|\widehat{\Psi}_{\tilde{u}0}^{-1}\mathbb{E}_n[S_{\tilde{u}}]\|_\infty \lesssim \lambda/n$$

with probability $1-o(1)$ by the choice of λ; see Lemma I.4. In addition, uniformly over $\tilde{u} \in \widetilde{\mathcal{U}}$ with probability $1 - o(1)$, we have $\widehat{s}_{\tilde{u}} + s_n \lesssim s_n$ and $\phi_{\min}(Cs_n, \tilde{u}) \gtrsim 1$ for arbitrarily large C. Hence, by Lemma I.3,

$$\|\sqrt{w_{\tilde{u}}}X^j(\widetilde{\gamma}_u^j - \bar{\gamma}_u^j)\|_{\mathbb{P}_n,2} \lesssim (s_n \log a_n/n)^{1/2}$$

with probability $1 - o(1)$ uniformly over $u \in \mathcal{U}$ and $j \in [\tilde{p}]$. Finally, as in the case of $\widehat{\gamma}_u^j$'s, we also have $\|\sqrt{w_{\tilde{u}}}X^j(\widetilde{\gamma}_u^j - \bar{\gamma}_u^j)\|_{\mathbb{P}_n,2} \gtrsim \|\widetilde{\gamma}_u^j - \bar{\gamma}_u^j\|$ with probability $1-o(1)$ uniformly over $u \in \mathcal{U}$ and $j \in [\tilde{p}]$, which gives the last asserted claim and completes the proof of the theorem. ■

Auxiliary Lemmas for Proofs of Theorems 4.1 and 4.2

Lemma F.1 (Control of Approximation Error)**.** *Suppose that Assumptions 3.1 – 3.5 hold for all $P \in \mathcal{P}_n$. Then for $\bar{r}_{uj} = X^j(\gamma_u^j - \bar{\gamma}_u^j)$, we have with probability $1 - o(1)$ that*

$$\sup_{u\in\mathcal{U},j\in[\tilde{p}]} \mathbb{E}_n[\bar{r}_{uj}^2] \lesssim s_n \log a_n/n$$

uniformly over $P \in \mathcal{P}_n$.

Lemma F.2 (Control of Estimated Weights and Score)**.** *Suppose that Assumptions 3.1 – 3.5 hold for all $P \in \mathcal{P}_n$. Then with probability $1 - o(1)$, we have*

$$\sup_{u \in \mathcal{U}, j \in [\tilde{p}]} \|(\widehat{f}_u^2 - f_u^2) Z_u^j / \widehat{f}_u\|_{\mathbb{P}_n,2} \lesssim (s_n \log a_n / n)^{1/2}$$

uniformly over $P \in \mathcal{P}_n$.

Lemma F.3 (Functional Sparse Eigenvalues)**.** *Suppose that Assumptions 3.1 – 3.5 hold for all $P \in \mathcal{P}_n$. Then for $\ell_n \to \infty$ slowly enough, we have*

$$\sup_{u \in \mathcal{U}} \sup_{\|\delta\|_0 \leqslant \ell_n s_n, \|\delta\|=1} |1 - \|f_u(D', X')\delta\|_{\mathbb{P}_n,2} / \|f_u(D', X')\delta\|_{\mathrm{P},2}| = o_P(1) \quad \textit{and}$$

$$\sup_{\|\delta\|_0 \leqslant \ell_n s_n, \|\delta\|=1} |1 - \|(D', X')\delta\|_{\mathbb{P}_n,2} / \|(D', X')\delta\|_{\mathrm{P},2}| = o_P(1)$$

uniformly over $P \in \mathcal{P}_n$.

Proof of Lemma F.1. By Assumption 3.2, we have that $\sup_{u \in \mathcal{U}} \max_{j \in [\tilde{p}]} \|\bar{\gamma}_u^j\|_0 \leqslant s_n$ and

$$\sup_{u \in \mathcal{U}} \max_{j \in [\tilde{p}]} \left(\|\bar{\gamma}_u^j - \gamma_u^j\| + s_n^{-1/2} \|\bar{\gamma}_u^j - \gamma_u^j\|_1 \right) \leqslant C_1 (s_n \log a_n / n)^{1/2}.$$

Also, by (E.4), we have $\|\gamma_u^j - \gamma_{u'}^j\| \leqslant L_\gamma |u - u'|$ for some constant L_γ uniformly over $u, u' \in \mathcal{U}$. By the triangle inequality,

$$\sup_{u \in \mathcal{U}, j \in [\tilde{p}]} \mathbb{E}_n[\bar{r}_{uj}^2] \leqslant \sup_{u \in \mathcal{U}, j \in [\tilde{p}]} |(\mathbb{E}_n - \mathrm{E}_P)[\bar{r}_{uj}^2]| + \sup_{u \in \mathcal{U}, j \in [\tilde{p}]} \mathrm{E}_P[\bar{r}_{uj}^2].$$

Consider the class of functions $\mathcal{G} = \{(D, X) \mapsto X^j(\gamma_u^j - \bar{\gamma}_u^j) \colon u \in \mathcal{U}, j \in [\tilde{p}]\}$ and $\mathcal{G}_{j,T} = \{(D, X) \mapsto X^j(\gamma_u^j - \gamma_{uT}^j) \colon u \in \mathcal{U}\}$ for $j \in [\tilde{p}]$ and $T \subset [p + \tilde{p} - 1]$ being a subset of the components of X^j with $|T| \leqslant s_n$. Since $\bar{\gamma}_u^j = \gamma_{uT}^j$ for some $T = T_u^j$, it follows that $\mathcal{G} \subset \cup_{j \in [\tilde{p}], |T| \leqslant s_n} \mathcal{G}_{j,T}$. Also, we have $\|\gamma_{uT}^j - \gamma_{u'T}^j\| \leqslant \|\gamma_u^j - \gamma_{u'}^j\|$ for all $u, u' \in \mathcal{U}$, $j \in [\tilde{p}]$, and $T \subset [p + \tilde{p} - 1]$. Therefore, for fixed j and T, we have

$$\begin{aligned} &\left| (X^j(\gamma_{uT}^j - \gamma_u^j))^2 - (X^j(\gamma_{u'T}^j - \gamma_{u'}^j))^2 \right| \\ &\quad = \left| X^j(\gamma_{uT}^j - \gamma_u^j + \gamma_{u'T}^j - \gamma_{u'}^j) X^j(\gamma_{uT}^j - \gamma_{u'T}^j + \gamma_{u'}^j - \gamma_u^j) \right| \\ &\quad \leqslant \|(D', X')'\|_\infty^2 \|\gamma_{uT}^j - \gamma_u^j + \gamma_{u'T}^j - \gamma_{u'}^j\|_1 \|\gamma_{uT}^j - \gamma_{u'T}^j + \gamma_{u'}^j - \gamma_u^j\|_1 \\ &\quad \leqslant 8 \|(D', X')'\|_\infty^2 \sup_{u \in \mathcal{U}} \|\gamma_u^j\|_1 \{p + \tilde{p}\}^{1/2} \|\gamma_u^j - \gamma_{u'}^j\| \\ &\quad \leqslant \left(M^{-1} L'_\gamma \|(D', X')'\|_\infty^2 \right) M |u - u'|. \end{aligned}$$

where $L'_\gamma = 8 \sup_{u \in \mathcal{U}, j \in [\tilde{p}]} \|\gamma_u^j\|_1 \{p + \tilde{p}\}^{1/2} L_\gamma \lesssim a_n$ and we will set $M = a_n^2$. Therefore, we have for the envelope $G(D, X) = \|(D', X')'\|_\infty^2 (M^{-1} L'_\gamma + \sup_{u \in \mathcal{U}, j \in [\tilde{p}]} \|\bar{\gamma}_u^j - \gamma_u^j\|_1^2)$ that for all $0 < \epsilon \leqslant 1$ and all finitely-discrete probability measures Q,

$$\begin{aligned}\log N(\epsilon\|G\|_{Q,2},\mathcal{G}^2,\|\cdot\|_{Q,2}) &\lesssim s_n\log a_n+\max_{j\in[\tilde p],|T|\leqslant s}\log\left(\epsilon\|G\|_{Q,2},\mathcal{G}^2_{j,T},\|\cdot\|_{Q,2}\right)\\ &\lesssim s_n\log a_n+\log N\left(\epsilon/M,\mathcal{U},d_{\mathcal{U}}\right)\\ &\lesssim s_n\log a_n+\log(a_n/\epsilon)\lesssim s_n\log(a_n/\epsilon).\end{aligned}$$

By Lemma L.2, since

$$\left\|\max_{1\leqslant i\leqslant n}G(D_i,X_i)\right\|_{P,2}\lesssim n^{1/q}M_{n,2}\left(a_n^{-2}L_\gamma'+\sup_{u\in\mathcal{U},j\in[\tilde p]}\|\bar\gamma_u^j-\gamma_u^j\|_1^2\right)\lesssim n^{-1+1/q}M_{n,2}s_n^2\log a_n,$$

we have with probability $1-o(1)$ that

$$\begin{aligned}\sup_{u\in\mathcal{U},j\in[\tilde p]}|(\mathbb{E}_n-\mathrm{E}_P)[\bar r_{uj}^2]| &\lesssim\sqrt{\frac{s_n\log a_n\sup_{u\in\mathcal{U},j\in[\tilde p]}\mathrm{E}_P[\bar r_{uj}^4]}{n}}+\frac{s_nn^{-1+1/q}M_{n,2}s_n^2\log^2 a_n}{n}\\ &\lesssim\sqrt{\frac{s_n\log a_n}{n}}\frac{s_n\log a_n}{n}+\delta_n\frac{s_n\log a_n}{n}\lesssim\frac{s_n\log a_n}{n}\end{aligned}$$

where we used that $s_n\geqslant 1$, $\bar r_{uj}=X^j(\gamma_u^j-\bar\gamma_u^j)$, $\mathrm{E}_P[\{(D',X')\xi\}^4]\leqslant C_1\|\xi\|^4$ by Assumption 3.4(iv), $\|\bar\gamma_u^j-\gamma_u^j\|^2\leqslant C_1^2s_n\log a_n/n$ by Assumption 3.2, $M_{n,2}s_n^2\log a_n\leqslant\delta_nn^{1-1/q}$ implied by Assumption 3.4(viii). Finally, the result follows since $\mathrm{E}_P[\bar r_{uj}^2]=\mathrm{E}_P[\{X^j(\gamma_u^j-\bar\gamma_u^j)\}^2]\lesssim\|\gamma_u^j-\bar\gamma_u^j\|^2\lesssim s_n\log a_n/n$ by Assumption 3.2. ■

Proof of Lemma F.2. Recall that $\widehat f_u^2=\widehat f_u^2(D,X)=\Lambda'(D'\widetilde\theta_u+X'\widetilde\beta_u)$ and that by Theorem 4.1, we have $\sup_{u\in\mathcal{U}}(\|\widetilde\theta_u-\theta_u\|+\|\widetilde\beta_u-\beta_u\|)\lesssim(s_n\log a_n/n)^{1/2}$ and $\sup_{u\in\mathcal{U}}\|(\widetilde\theta_u',\widetilde\beta_u')'\|_0\lesssim s_n$ with probability $1-o(1)$. Also,

$$\begin{aligned}\max_{1\leqslant i\leqslant n}|(D_i',X_i')\{(\widetilde\theta_u',\widetilde\beta_u')'-(\theta_u',\beta_u')'\}| &\leqslant\max_{1\leqslant i\leqslant n}\|(D_i',X_i')'\|_\infty\{\|\widetilde\theta_u-\theta_u\|_1+\|\widetilde\beta_u-\beta_u\|_1\}\\ &\lesssim_P n^{1/(2q)}M_{n,2}(s_n^2\log a_n/n)^{1/2}\leqslant\delta_n\end{aligned}$$

by Assumption 3.4(vii, viii) since $M_{n,2}\geqslant 1$. Thus, for $\tilde t_{ui}=D_i'\widetilde\theta_u+X_i'\widetilde\beta_u$ and $t_{ui}=D_i'\theta_u+X_i'\beta_u$, we have with probability $1-o(1)$ that $\sup_{u\in\mathcal{U},i\in[n]}|\tilde t_{ui}-t_{ui}|\leqslant\delta_n^{1/2}=o(1)$, and so $|\Lambda(\tilde t_{ui})-\Lambda(t_{ui})|\lesssim\Lambda'(t_{ui})|\tilde t_{ui}-t_{ui}|$ uniformly over $u\in\mathcal{U}$ and $i=1,\dots,n$ as in (F.6). Hence, the inequality $|x(1-x)-y(1-y)|\leqslant|x-y|$, which holds for all $x,y\in[0,1]$, implies that with probability $1-o(1)$,

$$|\widehat f_{ui}^2-f_{ui}^2|\leqslant|\Lambda(\tilde t_{ui})-\Lambda(t_{ui})-r_{ui}|\lesssim\Lambda'(t_{ui})|\tilde t_{ui}-t_{ui}|+|r_{ui}|\leqslant f_{ui}^2/2 \tag{F.12}$$

since $|r_u|\leqslant f_u^2/4$ by Assumption 3.5 and

$$\Lambda'(t_{ui})=f_{ui}^2+2\Lambda(t_{ui})r_{ui}-r_{ui}+r_{ui}^2\leqslant f_{ui}^2+3|r_{ui}|+r_{ui}^2\leqslant f_{ui}^2+4|r_{ui}|\leqslant 2f_{ui}^2 \tag{F.13}$$

by the definition of f_u^2 and since $|r_u|\leqslant 1$. Therefore, with probability $1-o(1)$,

$$\|(\widehat f_u^2-f_u^2)Z_u^j/\widehat f_u\|_{\mathbb{P}_n,2}\lesssim\|(\widehat f_u^2-f_u^2)Z_u^j/f_u\|_{\mathbb{P}_n,2}$$

uniformly over $u \in \mathcal{U}$ and $j \in [\tilde{p}]$. Hence, it suffices to show that with probability $1 - o(1)$,

$$\sup_{u \in \mathcal{U}, j \in [\tilde{p}]} \|(\widehat{f}_u^2 - f_u^2) Z_u^j / f_u\|_{\mathbb{P}_n, 2} \lesssim (s_n \log a_n / n)^{1/2}.$$

Next, as in (E.1), we have uniformly over $u, u' \in \mathcal{U}$ that $\|\gamma_u^j - \gamma_{u'}^j\|_1 \lesssim (p + \tilde{p})^{1/2} |u - u'|$, and so, given that $Z_u^j - Z_{u'}^j = X^j(\gamma_{u'}^j - \gamma_u^j)$, we have by Assumption 3.4(vii) that

$$\max_{1 \leqslant i \leqslant n} |Z_{ui}^j - Z_{u'i}^j| \leqslant \max_{1 \leqslant i \leqslant n} \|(D_i', X_i')'\|_\infty \|\gamma_u^j - \gamma_{u'}^j\|_1 \lesssim_P n^{1/(2q)} M_{n,2} (p + \tilde{p})^{1/2} |u - u'|.$$

Moreover, as in (E.2), we have uniformly over $u, u' \in \mathcal{U}$ that $|f_u^2 - f_{u'}^2| \lesssim |u - u'|$.

Further, observe that for $a > 1$, the inequality $|x| \leqslant \log(\sqrt{a} - 1)$ implies that

$$\Lambda'(x) = \frac{e^x}{(1 + e^x)^2} = \frac{1}{e^{-x} + 2 + e^x} \geqslant \frac{1}{2(1 + e^{|x|})} \geqslant \frac{1}{2\sqrt{a}}.$$

Also, by Assumptions 3.1, 3.2, and 3.4(vii),

$$|t_{ui}| \leqslant \|(D_i', X_i')'\|_\infty (\|\theta_u\|_1 + \|\beta\|_1) \leqslant n^{1/(2q)} M_{n,2} \sqrt{s_n \log n}$$

with probability $1 - o(1)$ uniformly over $u \in \mathcal{U}$ and $i = 1, \dots, n$. Thus, applying the inequality above with $\sqrt{a} - 1 = \exp(n^{1/(2q)} M_{n,2} \sqrt{s_n \log n})$ gives

$$f_{ui}^2 \geqslant \Lambda'(t_{ui})/2 \gtrsim \exp(-n^{1/(2q)} M_{n,2} \sqrt{s_n \log n})$$

with probability $1 - o(1)$ uniformly over $u \in \mathcal{U}$ and $i = 1, \dots, n$. So,

$$\mathbb{E}_n[f_u^{-2}] \lesssim \exp(n^{1/(2q)} M_{n,2} \sqrt{s_n \log n})$$

with probability $1 - o(1)$ uniformly over $u \in \mathcal{U}$.

In addition, let

$$\epsilon = \epsilon_n = \Big(n^{1 + 1/q} (M_{n,1}^2 \vee M_{n,2}^2)(p + \tilde{p})^{1/2} \exp(n^{1/(2q)} M_{n,2} \sqrt{s_n \log n}) \Big)^{-1},$$

and let $\mathcal{U}^\epsilon$ be an ϵ-net of $\mathcal{U}$ with $|\mathcal{U}^\epsilon| \leqslant 1/\epsilon$. For all $i = 1, \dots, n$, let U_i be a value of $u \in \mathbb{R}$ such that $Y_i = (1 - u)\underline{y} + u\bar{y}$. Note that $\widehat{f}_u$ does not vary with u on any interval $[\underline{u}, \bar{u}] \subset \mathcal{U}$ as long as $U_i \notin [\underline{u}, \bar{u}]$ for all $i = 1, \dots, n$. Also, since $\epsilon \lesssim n^{-3}$, with probability $1 - o(1)$, each interval $[u - 2\epsilon, u + 2\epsilon]$ with $u \in \mathcal{U}^\epsilon$ contains at most one value of U_i's by Assumption 3.3. Now,

$$\begin{aligned} \sup_{u \in \mathcal{U}} \max_{j \in [\tilde{p}]} \mathbb{E}_n[(\widehat{f}_u^2 - f_u^2)^2 (Z_u^j / f_u)^2] &\lesssim \sup_{u \in \mathcal{U}^\epsilon} \max_{j \in [\tilde{p}]} \mathbb{E}_n[(\widehat{f}_u^2 - f_u^2)^2 (Z_u^j / f_u)^2] \\ &+ \sup_{u \in \mathcal{U}, j \in [\tilde{p}]} \inf_{u' \in \mathcal{U}^\epsilon} \Big| \mathbb{E}_n[(\widehat{f}_u^2 - f_u^2)^2 (Z_u^j / f_u)^2 - (\widehat{f}_{u'}^2 - f_{u'}^2)^2 (Z_{u'}^j / f_{u'})^2] \Big|, \end{aligned}$$

and uniformly over $j \in [\tilde p]$ and $u, u' \in \mathcal{U}$ such that $\widehat{f}_u = \widehat{f}_{u'}$, we have with probability $1 - o(1)$ that

$$
\begin{aligned}
&\left|\mathbb{E}_n[(\widehat{f}_u^2 - f_u^2)^2 (Z_u^j/f_u)^2 - (\widehat{f}_{u'}^2 - f_{u'}^2)^2 (Z_{u'}^j/f_{u'})^2]\right| \\
&\quad \leqslant \left|\mathbb{E}_n[\{(\widehat{f}_u^2 - f_u^2)^2 - (\widehat{f}_{u'}^2 - f_{u'}^2)^2\}(Z_u^j/f_u)^2]\right| + \left|\mathbb{E}_n[(\widehat{f}_{u'}^2 - f_{u'}^2)^2\{(Z_u^j/f_u)^2 - (Z_{u'}^j/f_{u'})^2\}]\right| \\
&\quad \lesssim \left|\mathbb{E}_n[\{(f_{u'}^2 - f_u^2)(Z_u^j/f_u)^2]\right| + \left|\mathbb{E}_n[f_{u'}^4\{(Z_u^j/f_u)^2 - (Z_{u'}^j/f_{u'})^2\}]\right| \\
&\quad \lesssim |u' - u| \mathbb{E}_n[(Z_u^j/f_u)^2] + \left|\mathbb{E}_n[(f_{u'}^4/f_u^2)|\{Z_u^j - Z_{u'}^j\}\{Z_u^j + Z_{u'}^j\}|]\right| + \left|\mathbb{E}_n[(f_{u'}^2/f_u^2)(Z_{u'}^j)^2(f_{u'}^2 - f_u^2)]\right| \\
&\quad \lesssim_P |u' - u| \cdot \left(n^{1/q} M_{n,1}^2 + n^{1/(2q)} M_{n,2}(p + \tilde p)^{1/2} n^{1/(2q)} M_{n,1}\right) \cdot \mathbb{E}_n[f_u^{-2}].
\end{aligned}
$$

Thus, by the choice of ϵ, and since with probability $1 - o(1)$ each interval $[u - 2\epsilon, u + 2\epsilon]$ with $u \in \mathcal{U}^\epsilon$ contains at most one value of U_i's, we have with probability $1 - o(1)$ that

$$
\sup_{u \in \mathcal{U}, j \in [\tilde p]} \inf_{u' \in \mathcal{U}^\epsilon} \left|\mathbb{E}_n[(\widehat{f}_u^2 - f_u^2)^2 (Z_u^j/f_u)^2 - (\widehat{f}_{u'}^2 - f_{u'}^2)^2 (Z_{u'}^j/f_{u'})^2]\right| \lesssim s_n \log a_n / n.
$$

Further by (F.12) and (F.13), with probability $1 - o(1)$,

$$
\begin{aligned}
\sup_{u \in \mathcal{U}^\epsilon, j \in [\tilde p]} \mathbb{E}_n[(\widehat{f}_u^2 - f_u^2)^2 (Z_u^j/f_u)^2] &\lesssim \sup_{u \in \mathcal{U}^\epsilon, j \in [\tilde p]} \mathbb{E}_n[\Lambda'(t_{ui})^2 |\tilde t_{ui} - t_{ui}|^2 (Z_u^j/f_u)^2] \\
&\quad + \sup_{u \in \mathcal{U}^\epsilon, j \in [\tilde p]} \mathbb{E}_n[r_u^2 (Z_u^j/f_u)^2] \\
&\lesssim \max_{j \in [\tilde p]} \sup_{u \in \mathcal{U}^\epsilon, \|\delta\|_0 \leqslant C s_n, \|\delta\| = 1} \mathbb{E}_n[\{(D', X')\delta\}^2 (Z_u^j)^2] s_n \log a_n / n \\
&\quad + s_n \log a_n / n
\end{aligned}
$$

for C large enough, where we used that $\sup_{u \in \mathcal{U}} \mathbb{E}_n[r_u^2 (Z_u^j/f_u)^2] \lesssim s_n \log a_n / n$ with probability $1 - o(1)$ by Assumption 3.5, and $\sup_{u \in \mathcal{U}}(\|\widetilde\theta_u - \theta_u\| + \|\widetilde\beta_u - \beta_u\|) \lesssim (s_n \log a_n / n)^{1/2}$ and $\sup_{u \in \mathcal{U}}(\|(\widetilde\theta_u', \widetilde\beta_u')'\|_0 + \|(\theta_u', \beta_u')'\|_0) \lesssim s_n$ with probability $1 - o(1)$. Therefore, since

$$
\mathrm{E}_P[\{(D', X')\delta\}^2 (Z_u^j)^2] \leqslant \mathrm{E}_P[\{(D', X')\delta\}^4]^{1/2} \mathrm{E}_P[(Z_u^j)^4]^{1/2} \lesssim 1,
$$

to establish the statement of the lemma it suffices to show that with probability $1 - o(1)$,

$$
\max_{j \in [\tilde p]} \sup_{u \in \mathcal{U}^\epsilon, \|\delta\|_0 \leqslant C s_n, \|\delta\| = 1} |(\mathbb{E}_n - \mathrm{E}_P)[\{(D', X')\delta\}^2 (Z_u^j)^2]| \lesssim 1.
$$

To do so, we will apply Lemma B.1 with $\mathcal{U}$ replaced by $\mathcal{U}^\epsilon \times [\tilde p]$ and X_u replaced by $Z_u^j (D', X')'$. We have

$$
\begin{aligned}
K &= \left(\mathrm{E}_P\left[\max_{1 \leqslant i \leqslant n, u \in \mathcal{U}^\epsilon} \|Z_{ui}^j (D', X')'\|_\infty^2\right]\right)^{1/2} \leqslant n^{1/q} \left(\mathrm{E}_P\left[\max_{u \in \mathcal{U}^\epsilon} \|Z_u^j (D', X')'\|_\infty^q\right]\right)^{1/q} \\
&\leqslant n^{1/q} \left(\mathrm{E}_P[\|(D', X')'\|_\infty^{2q}] \mathrm{E}_P[\|Z_u^j\|_\infty^{2q}]\right)^{1/(2q)} \leqslant n^{1/q} M_{n,2} M_{n,1}
\end{aligned}
$$

by Assumption 3.4(v,vii). Also,

$$
\sup_{\|\delta\|_0 \leqslant C s_n, \|\delta\| = 1} \max_{u \in \mathcal{U}^\epsilon, j \in [\tilde p]} \mathrm{E}_P[(Z_u^j (D', X')\delta)^2] \lesssim 1
$$

by Assumption 3.4(iv). Then, by Lemma B.1 we have for

$$\tilde{\delta}_n = n^{-1/2} K s_n^{1/2} \Big(\log^{1/2}(\tilde{p}|\mathcal{U}_0^\epsilon|) + (\log s_n)(\log^{1/2} n)(\log^{1/2} a_n) \Big)$$

that

$$\sup_{\|\delta\|_0 \leqslant C s_n, \|\delta\|=1} \max_{u \in \mathcal{U}^\epsilon, j \in [\tilde{p}]} |(\mathbb{E}_n - \mathrm{E}_P)[(Z_u^j(D', X')\delta)^2]| \lesssim_P \tilde{\delta}_n^2 + \tilde{\delta}_n.$$

Now,

$$\tilde{p}|\mathcal{U}^\epsilon| \leqslant \tilde{p}/\epsilon \leqslant n^{1+1/q}(M_{n,1}^2 \vee M_{n,2}^2)(p+\tilde{p})^{3/2} \exp(n^{1/(2q)} M_{n,2}\sqrt{s_n \log n}),$$

so that

$$\log(\tilde{p}|\mathcal{U}^\epsilon|) \lesssim \log a_n + n^{1/(2q)} M_{n,2}\sqrt{s_n \log n}.$$

Using Assumption 3.4(vi,viii,ix) and since $\delta_n^2 \log a_n = o(1)$, we have

$$\frac{(M_{n,1} \vee M_{n,2})^2 s_n \log a_n}{n^{1/2-1/q}} \leqslant \delta_n \log^{1/2} a_n = o(1) \text{ and } \frac{M_{n,1}^2 M_{n,2}^4 s_n}{n^{1-3/q}} = o(1),$$

and so

$$\begin{aligned}
\tilde{\delta}_n &\lesssim \frac{s_n^{1/2} n^{1/q} M_{n,2} M_{n,1} (\log s_n)(\log^{1/2} n)(\log^{1/2} a_n)}{n^{1/2}} + \frac{s_n^{1/2} n^{1/q} M_{n,2} M_{n,1} n^{1/(4q)} M_{n,2}^{1/2} s_n^{1/4} \log^{1/4} n}{n^{1/2}} \\
&\lesssim \Big(\frac{M_{n,1}^2 s_n \log a_n}{n^{1/2-1/q}}\Big)^{1/2}\Big(\frac{M_{n,2}^2 s_n \log a_n}{n^{1/2-1/q}}\Big)^{1/2} + \Big(\frac{M_{n,1}^2 s_n \log n}{n^{1/2-1/q}}\Big)^{1/4}\Big(\frac{M_{n,2}^2 s_n}{n^{1/2-1/q}}\Big)^{1/4}\Big(\frac{M_{n,1}^2 M_{n,2}^4 s_n}{n^{1-3/q}}\Big)^{1/4} \\
&= o(1).
\end{aligned}$$

This completes the proof. ■

Proof of Lemma F.3. Both results follow from Lemma B.1. We provide a proof only for the first result (the second result is simpler and follows similarly).

Recall that by Assumption 3.4(i),

$$\inf_{\|\delta\|=1} \|f_u(D', X')\delta\|_{\mathrm{P},2} \geqslant c_1.$$

Also, observe that for any $x, y \in [0, 1]$, we have

$$\Big|\sqrt{x(1-x)} - \sqrt{y(1-y)}\Big| \leqslant \sqrt{|x-y|}.$$

Therefore, since $f_u^2 = \mathrm{E}[Y_u \mid D, X](1 - \mathrm{E}[Y_u \mid D, X])$, by Assumption 3.3, for any $u, u' \in \mathcal{U}$, we have

$$|f_{u'} - f_u| \leqslant \Big(|\mathrm{E}[Y_{u'} - Y_u \mid D, X]|\Big)^{1/2} \leqslant \Big(C_1|u' - u|\Big)^{1/2}.$$

Hence, since $\ell_n \to \infty$, with probability $1-o(1)$ uniformly over $u, u' \in \mathcal{U}$ and $\delta \in \mathbb{R}^{p+\tilde{p}}$ with $\|\delta\| = 1$ and $\|\delta\|_0 \leqslant \ell_n s_n$, we have

$$\begin{aligned}
\Big| \|f_{u'}(D', X')\delta\|_{\mathbb{P}_n,2} - \|f_u(D', X')\delta\|_{\mathbb{P}_n,2} \Big| &\leqslant \|(f_{u'} - f_u)(D', X')\delta\|_{\mathbb{P}_n,2} \\
&\leqslant \|f_{u'} - f_u\|_{\mathbb{P}_n,2} \max_{1\leqslant i\leqslant n} \|(D_i', X_i')'\|_\infty \|\delta\|_1 \\
&\leqslant \|f_{u'} - f_u\|_{\mathbb{P}_n,2} n^{1/(2q)} M_{n,2} \ell_n \sqrt{s_n} \\
&\leqslant \Big(C_1 |u' - u| \Big)^{1/2} n^{1/(2q)} M_{n,2} \ell_n \sqrt{s_n}
\end{aligned}$$

by Assumption 3.4(vii). Thus, for

$$\epsilon = \epsilon_n = \frac{c_1^2}{C_1 n^{1/(2q)} M_{n,2}^2 \ell_n^4 s_n},$$

we have with probability $1 - o(1)$ that

$$\sup_{|u-u'|\leqslant\epsilon, \|\delta\|_0 \leqslant \ell_n s_n, \|\delta\|=1} \Big| \|f_{u'}(D', X')\delta\|_{\mathbb{P}_n,2} - \|f_u(D', X')\delta\|_{\mathbb{P}_n,2} \Big| \leqslant c_1/\ell_n.$$

Now, let $\mathcal{U}^\epsilon$ be an ϵ-net of $\mathcal{U}$ such that $|\mathcal{U}^\epsilon| \leqslant 3/\epsilon$. We will apply Lemma B.1 with $\mathcal{U}$ replaced by $\mathcal{U}^\epsilon$, $k = \ell_n s_n$, and X_u replaced by $f_u(D', X')'$. Since $0 \leqslant f_u \leqslant 1$, we have

$$K = \Big(\mathrm{E}_P\Big[\max_{1\leqslant i\leqslant n} \max_{u\in\mathcal{U}^\epsilon} f_{ui}^2 \|(D_i', X_i')'\|_\infty^2 \Big] \Big)^{1/2} \leqslant \Big(\mathrm{E}_P\Big[\max_{1\leqslant i\leqslant n} \|(D_i', X_i')'\|_\infty^2 \Big] \Big)^{1/2} \leqslant n^{1/(2q)} M_{n,2}$$

by Assumption 3.4(vii). Also,

$$\sup_{\|\delta\|_0\leqslant\ell_n s_n, \|\delta\|=1} \max_{u\in\mathcal{U}^\epsilon} \mathrm{E}_P[f_u^2((D', X')\delta)^2] \leqslant \sup_{\|\delta\|_0\leqslant\ell_n s_n, \|\delta\|=1} \mathrm{E}_P[\{(D', X')\delta\}^2] \leqslant \sqrt{C_1}$$

by Assumption 3.4(iv). Thus, applying Lemma B.1 gives

$$\sup_{\|\delta\|_0\leqslant\ell_n s_n, \|\delta\|=1} \max_{u\in\mathcal{U}^\epsilon} |(\mathbb{E}_n - \mathrm{E}_P)[f_u^2((D', X')\delta)^2]| \lesssim_P \tilde{\delta}_n^2 + \tilde{\delta}_n$$

where

$$\tilde{\delta}_n = n^{-1/2+1/(2q)} M_{n,2} \sqrt{\ell_n s_n} (\log^{1/2} a_n)(\log^{3/2} n).$$

Finally, by Assumption 3.4(viii),

$$\tilde{\delta}_n^2 = n^{-1+1/q} M_{n,2}^2 \ell_n s_n (\log a_n)(\log^3 n) \leqslant n^{-1/2} \ell_n \delta_n (\log^{1/2} a_n)(\log^3 n) = o(1)$$

since $\ell_n \to \infty$ slowly enough and $\log^{1/2} a_n \lesssim \delta_n n^{1/6}$ by Assumption 3.4(ii,iii). Combining presented bounds gives the asserted claim. ∎

Appendix G. Proof of Lemma B.1

Proof of Lemma B.1. For $T \subset \{1, \ldots, p\}$, let $B_T = \{\theta \in \mathbb{R}^p \colon \|\theta\| = 1, \mathrm{supp}(\theta) \subseteq T\}$. Also, for $\mathcal{T} = \cup_{|T|=k} B_T \times \mathcal{U}$, let $R := \sup_{(\theta,u)\in\mathcal{T}}(\sum_{i=1}^n (\theta' X_{ui})^2)^{1/2}$ and $M := \max_{1\leqslant i\leqslant n, u\in\mathcal{U}} \|X_{ui}\|_\infty$. By symmetrization inequality, Lemma 6.3 in [28], we have

$$n\mathrm{E}\left[\sup_{\|\theta\|_0\leqslant k, \|\theta\|=1} \max_{u\in\mathcal{U}} \left|\mathbb{E}_n\left[(\theta' X_u)^2 - \mathrm{E}[(\theta' X_u)^2]\right]\right|\right] \leqslant 2\mathrm{E}\left[\mathrm{E}\left[\sup_{(\theta,u)\in\mathcal{T}}\left|\sum_{i=1}^n \varepsilon_i(\theta' X_{ui})^2\right| \mid X\right]\right]$$

where $X = (X_{ui})_{u\in\mathcal{U}, 1\leqslant i\leqslant n}$ and $(\varepsilon_i)_{i=1}^n$ is a sequence of independent Rademacher random variables that are independent of X. A consequence of Lemma 4.5 in [28] (see equation (4.8)) gives

$$\mathrm{E}\left[\sup_{(\theta,u)\in\mathcal{T}}\left|\sum_{i=1}^n \varepsilon_i(\theta' X_{ui})^2\right| \mid X\right] \leqslant (\pi/2)^{1/2}\mathrm{E}\left[\sup_{(\theta,u)\in\mathcal{T}}\left|\sum_{i=1}^n g_i(\theta' X_{ui})^2\right| \mid X\right]$$

where $(g_i)_{i=1}^n$ is a sequence of independent standard normal random variables that are independent of X. In turn, an application of Dudley's integral gives

$$I_1 := \mathrm{E}\left[\sup_{(\theta,u)\in\mathcal{T}}\left|\sum_{i=1}^n g_i(\theta' X_{ui})^2\right| \mid X\right] \leqslant 8\int_0^{\mathrm{diam}(\mathcal{T})} \log^{1/2} N(\mathcal{T}, d, \epsilon) d\epsilon$$

where $\mathrm{diam}(\mathcal{T}) \leqslant 2\sup_{(\theta,u)\in\mathcal{T}}(\sum_{i=1}^n(\theta' X_{ui})^4)^{1/2} \leqslant 2\sqrt{k}MR$ using that $|\theta' X_{ui}| \leqslant \|X_{ui}\|_\infty \|\theta\|_1 \leqslant M\sqrt{k}$, and d is the corresponding Gaussian semi-metric. Furthermore, we have $\log N(\mathcal{T}, d, \epsilon) \leqslant \log|\mathcal{U}| + \max_{u\in\mathcal{U}} \log N(\cup_{|T|=k} B_T \times \{u\}, d, \epsilon)$, so that

$$I_1 \leqslant 16\sqrt{k}MR\log^{1/2}|\mathcal{U}| + 8\int_0^{\mathrm{diam}(\mathcal{T})} \max_{u\in\mathcal{U}} \log^{1/2} N(\cup_{|T|=k} B_T \times \{u\}, d, \epsilon) d\epsilon.$$

Now, for any (θ, u) and $(\bar{\theta}, u)$ in $D_u^k := \cup_{|T|=k} B_T \times \{u\}$, we have

$$\begin{aligned} d((\theta,u),(\bar{\theta},u)) &= \Big(\sum_{i=1}^n \{(\theta' X_{ui})^2 - (\bar{\theta}' X_{ui})^2\}^2\Big)^{1/2} \\ &\leqslant \Big(\sum_{i=1}^n \{(\theta' X_{ui}) + (\bar{\theta}' X_{ui})\}^2\Big)^{1/2} \max_{1\leqslant i\leqslant n} |(\theta - \bar{\theta})' X_{ui}| \\ &\leqslant 2\sup_{(\theta,u)\in\mathcal{T}} \Big(\sum_{i=1}^n (\theta' X_{ui})^2\Big)^{1/2} \max_{1\leqslant i\leqslant n} |(\theta - \bar{\theta})' X_{ui}| = 2R\|\theta - \bar{\theta}\|_{X_u} \end{aligned}$$

where we let $\|\delta\|_{X_u} := \max_{1\leqslant i\leqslant n} |\delta' X_{ui}|$. This implies that

$$N(D_u^k, d, \epsilon) \leqslant N\left(D_u^k/\sqrt{k}, \|\cdot\|_{X_u}, \epsilon/\{2\sqrt{k}R\}\right).$$

Therefore, since $\mathrm{diam}(\mathcal{T}) \leqslant 2\sqrt{k}MR$, we have

$$\begin{aligned} \int_0^{\mathrm{diam}(\mathcal{T})} \max_{u\in\mathcal{U}} \log^{1/2} N(D_u^k, d, \epsilon) d\epsilon &\leqslant \int_0^{2\sqrt{k}MR} \max_{u\in\mathcal{U}} \log^{1/2} N(D_u^k/\sqrt{k}, \|\cdot\|_{X_u}, \epsilon/\{2\sqrt{k}R\}) d\epsilon \\ &= 2\sqrt{k}R\int_0^M \max_{u\in\mathcal{U}} \log^{1/2} N(D_u^k/\sqrt{k}, \|\cdot\|_{X_u}, \epsilon) d\epsilon. \end{aligned}$$

Note that $B_T/\sqrt{k} \subset B_T^1$ and $D_u^k/\sqrt{k} \subset B^1 \times \{u\}$ where $B^1 := \{\theta \in \mathbb{R}^p \colon \|\theta\|_1 \leqslant 1\}$ and $B_T^1 = \{\theta \in B^1 \colon \mathrm{supp}(\theta) \subseteq T\}$. It follows from Lemma 3.9 in [41] that $N(B^1, \|\cdot\|_{X_u}, \epsilon) \leqslant (2p)^{A\epsilon^{-2}M^2 \log n}$ for all $\epsilon > 0$ and some universal constant A. Moreover, as in the discussion after Lemma 3.9 in [41], we have $N(B_T^1, \|\cdot\|_{X_u}, \epsilon) \leqslant (1 + 2M/\epsilon)^k$ for all $\epsilon > 0$ and all $T \subset \{1, \dots, p\}$ with $|T| = k$, so that $N(D_u^k/\sqrt{k}, \|\cdot\|_{X_u}, \epsilon) \leqslant \binom{p}{k}(1 + 2M/\epsilon)^k$ for all $\epsilon > 0$. Therefore,

$$
\begin{aligned}
&\int_0^M \max_{u \in \mathcal{U}} \log^{1/2} N(D_u^k/\sqrt{k}, \|\cdot\|_{X_u}, \epsilon) d\epsilon \\
&\quad \leqslant \int_0^{M/\sqrt{k}} \log^{1/2} \left(\binom{p}{k} (1 + 2M/\epsilon)^k \right) d\epsilon + \int_{M/\sqrt{k}}^M \log^{1/2} \left((2p)^{A\epsilon^{-2}M^2 \log n} \right) d\epsilon \\
&\quad \leqslant \frac{M}{\sqrt{k}} \log^{1/2} \binom{p}{k} + \sqrt{k} \int_0^{M/\sqrt{k}} \log(1 + 2M/\epsilon) d\epsilon + A^{1/2} M (\log^{1/2} n)(\log^{1/2}(2p)) \int_{M/\sqrt{k}}^M \frac{d\epsilon}{\epsilon} \\
&\quad \leqslant M \log^{1/2} p + M\Big(1 + \log(1 + 2\sqrt{k})\Big) + A^{1/2} M (\log^{1/2} n)(\log^{1/2}(2p)) \Big(\log M - \log(M/\sqrt{k}) \Big) \\
&\quad \lesssim M \Big(\log^{1/2} p + (\log k)(\log^{1/2} n)(\log^{1/2} p) \Big)
\end{aligned}
$$

up-to a universal constant where in the third inequality, we used the fact that integrating by parts gives

$$
\sqrt{k} \int_0^{M/\sqrt{k}} \log(1 + 2M/\epsilon) d\epsilon \leqslant M(1 + \log(1 + 2\sqrt{k})).
$$

Collecting the terms, we obtain

$$
I_1 \lesssim \sqrt{k} M R \Big(\log^{1/2} |\mathcal{U}| + \log^{1/2} p + (\log k)(\log^{1/2} n)(\log^{1/2} p) \Big).
$$

Therefore, since $K \geqslant (\mathrm{E}[M^2])^{1/2}$, setting

$$
\delta_n = \frac{K\sqrt{k}}{\sqrt{n}} \Big(\log^{1/2} |\mathcal{U}| + \log^{1/2} p + (\log k)(\log^{1/2} n)(\log^{1/2} p) \Big)
$$

gives

$$
\begin{aligned}
I_2 &= \mathrm{E}\left[\sup_{\|\theta\|_0 \leqslant k, \|\theta\| = 1} \max_{u \in \mathcal{U}} \left| \mathbb{E}_n \left[(\theta' X_u)^2 - \mathrm{E}[(\theta' X_u)^2] \right] \right| \right] \\
&\lesssim \frac{\delta_n \mathrm{E}[MR]}{K\sqrt{n}} \leqslant (\delta_n/K)(\mathrm{E}[M^2])^{1/2}(\mathrm{E}[R^2/n])^{1/2} \leqslant \delta_n (E[R^2/n])^{1/2} \\
&\lesssim \delta_n \Big(I_2 + \sup_{\|\theta\|_0 \leqslant k, \|\theta\| = 1, u \in \mathcal{U}} \mathbb{E}_n \mathrm{E}[(\theta' X_u)^2] \Big)^{1/2}.
\end{aligned}
$$

Thus, because $a \leqslant \delta_n (a + b)^{1/2}$ implies $a \leqslant \delta_n^2 + \delta_n b^{1/2}$, we have

$$
I_2 \lesssim \delta_n^2 + \delta_n \sup_{\|\theta\|_0 \leqslant k, \|\theta\| = 1, u \in \mathcal{U}} \sqrt{\mathbb{E}_n \mathrm{E}[(\theta' X_u)^2]}
$$

up-to an absolute constant. This completes the proof. ■

Appendix H. Double Selection method for Logistic Regression with Functional Response Data

In this section we discuss in details and provide formal results for the double selection estimator for logistic regression with functional response data.

Algorithm 5. (Based on double selection) For each $u \in \mathcal{U}$ and $j \in [\tilde{p}]$:
Step 1'. Run post-ℓ_1-penalized logistic estimator (4.2) of Y_u on D and X to compute $(\widetilde{\theta}_u, \widetilde{\beta}_u)$.
Step 2'. Define the weights $\widehat{f}_u^2 = \widehat{f}_u^2(D, X) = \Lambda'(D_i'\widetilde{\theta}_u + X_i'\widetilde{\beta}_u)$.
Step 3'. Run the lasso estimator (4.4) of $\widehat{f}_u D_j$ on $\widehat{f}_u X$ to compute $\widehat{\gamma}_u^j$.
Step 4'. Run logistic regression of Y_u on D_j and all the selected variables in Steps 1' and 2' to compute $\check{\theta}_{uj}$.

The following result establishes the Bahadur representation for the double selection estimator (analog to Theorem 3.1 for score functions).

Theorem H.1 (Uniform Bahadur representation, double selection)**.** *Suppose that Assumptions 3.1 – 3.5 hold for all $P \in \mathcal{P}_n$. Then, the estimator $(\check{\theta}_{uj})_{u\in\mathcal{U},j\in[\tilde{p}]}$, based on the double selection, obeys as $n \to \infty$*

$$\Sigma_{uj}^{-1}\sqrt{n}(\check{\theta}_{uj} - \theta_{uj}) = \mathbb{G}_n\bar{\psi}_{uj} + O_{\mathrm{P}}(\delta_n) \text{ in } \ell^\infty(\mathcal{U} \times [\tilde{p}])$$

uniformly over $u \in \mathcal{U}$, where $\Sigma_{uj}^2 := \mathrm{E}_P[f_u^2(D - X^j\gamma_u^j)^2]^{-1}$.

Proof of Theorem H.1. The analysis is reduced to the proof of Theorem 3.1. Let $\widehat{T}_{uj} = \mathrm{supp}(\widehat{\theta}_u) \cup \mathrm{supp}(\widehat{\beta}_u) \cup \mathrm{supp}(\widehat{\gamma}_u^j)$ for which by Theorems 4.1 and 4.2 satisfies $\sup_{u\in\mathcal{U},j\in[\tilde{p}]} |\widehat{T}_{uj}| \lesssim s_n$ with probability $1 - o(1)$. Therefore Step 3 is a post-selection logistic regression which yields an initial rate of convergence $|\check{\theta}_{uj} - \theta_{uj}| + \|\bar{\theta}_{u[\tilde{p}]\setminus j} - \theta_{u[\tilde{p}]\setminus j}\| + \|\bar{\beta}_u - \beta_u\| \lesssim (s_n \log a_n/n)^{1/2}$. Moreover, by the first order condition of Step 3 we have

$$\mathbb{E}_n[\{Y_{ui} - \Lambda(D_j\check{\theta}_{uj} + D'_{[\tilde{p}]\setminus j}\bar{\theta}_{u[\tilde{p}]\setminus j} + X'\bar{\beta}_u)\}(D_j, X^j_{\widehat{T}_{uj}})'] = 0 \tag{H.1}$$

so that any linear combination yields zero. By setting the parameters $(\widetilde{\theta}'_u, \widetilde{\beta}'_u) = (\bar{\theta}_{u[\tilde{p}]\setminus j}, \bar{\theta}'_{u[\tilde{p}]\setminus j}, \bar{\beta}'_u)$, and $\widetilde{z}_u^j = (D_j, X^j_{\widehat{T}_{uj}})(1, -\widetilde{\gamma}_u^j) = D_j - X^j\widetilde{\gamma}_u^j$, we recover the setting in the proof of Theorem 3.1. The rest of the proof follows similarly. ∎

The double selection procedure benefits from additional variables selected in Step 2. The (estimated) weights used in the equation ensure that selection will ensure a near orthogonality condition that is required to remove first order bias. In contrast, (naive) Post-ℓ_1-logistic regression does not select such variables which in turn translates in to first order bias in the estimation of θ_{uj}. We stress that Step 2 is tailored to the estimation of each coefficient θ_{uj} which enables the additional adaptivity.

The double selection achieves orthogonality conditions relative to all selected variables in finite samples. Although first-order equivalent to other estimator discussed here, this additional orthogonality could potentially lead to a better finite sample performance. To provide intuition why, consider the logistic regression case with $\tilde{p} = \dim(D) = 1$ and $\mathcal{U} = \{0\}$ for simplicity. In this case we have

$$E[Y|D,X] = \Lambda(D\theta_0 + X'\beta_0) \text{ and let } f = \Lambda'(D\theta_0 + X'\beta_0).$$

Letting $\widehat{\gamma}$ be the Lasso estimate of fD on fX we have that the one step estimator

$$\bar{\theta} = \widehat{\theta} - \mathbb{E}_n[(D - X'\widehat{\gamma})^2]^{-1}\mathbb{E}_n[\{Y - \Lambda(D\widehat{\theta} + X'\widehat{\beta}))\}(D - X'\widehat{\gamma})]$$

is an approximate solution for the moment condition

$$\mathbb{E}_n[\{Y - \Lambda(D\theta + X'\widehat{\beta})\}(D - X'\widehat{\gamma})] = 0. \tag{H.2}$$

Indeed, it is one Newton step from $\widehat{\theta}$. Our proposed estimator based on estimated score functions defines $\check{\theta}$ as an exact solution for (H.2), namely

$$\mathbb{E}_n[\{Y - \Lambda(D\check{\theta} + X'\widehat{\beta})\}(D - X'\widehat{\gamma})] = 0.$$

The double selection achieves that implicitly. Indeed, letting $\widehat{T} = \text{support}(\check{\beta}) \cup \text{support}(\widehat{\gamma})$, the first order condition of running a logistic regression of Y on D and $X_{\widehat{T}}$ yields

$$\mathbb{E}_n\left[\{Y - \Lambda(D\check{\theta} + X'\check{\beta})\}\begin{pmatrix} D \\ X_{\widehat{T}} \end{pmatrix}\right] = 0 \tag{H.3}$$

where $(\check{\theta}, \check{\beta})$ is the solution of the logistic regression. By multiplying the vector $(1, -\widehat{\gamma}')'$, the relation above implies

$$\mathbb{E}_n[\{Y - \Lambda(D\check{\theta} + X'\check{\beta})\}(D - X'\widehat{\gamma})] = 0.$$

since $\text{support}(\widehat{\gamma}) \subset \widehat{T}$ (the condition $\text{support}(\widehat{\beta}) \subset \widehat{T}$ ensures that $\check{\beta}$ is a good approximation of β_0). Note that (H.3) provides a more robust orthogonality condition and does not need to explicitly create the new score functions as the other two methods.

Appendix I. Generic Finite Sample Bounds for ℓ_1-Penalized M-Estimators: Nuisance Functions and Functional Data

In this section, we establish a set of results for ℓ_1-penalized M-estimators with functional data and high-dimensional parameters. These results are used in the proofs of Theorems 4.1 and 4.2 and may be of independent interest.

We start with specifying the setting. Consider a data generating process with a functional response variable $(Y_u)_{u\in\mathcal{U}}$ and observable covariates $(X_u)_{u\in\mathcal{U}}$ satisfying for each $u \in \mathcal{U} \subset \mathbb{R}^{d_u}$,

$$\theta_u \in \arg\min_{\theta\in\mathbb{R}^p} \mathrm{E}_P[M_u(Y_u, X_u, \theta, a_u)], \tag{I.1}$$

where θ_u is a p-dimensional vector of parameters, a_u is a nuisance parameter that captures potential misspecification of the model, and M_u is a known function. Here for all $u \in \mathcal{U}$, Y_u is a scalar random variable and X_u is a p_u-dimensional random vector with $p_u \leqslant p$ for some p. We assume that the solution θ_u is sparse in the sense that the process $(\theta_u)_{u\in\mathcal{U}}$ satisfies

$$\|\theta_u\|_0 \leqslant s, \quad \text{for all } u \in \mathcal{U}.$$

Because the model (I.1) allows for the nuisance parameter a_u, such sparsity assumption is very mild and formulation (I.1) encompasses many cases of interest including approximately sparse models.

Throughout this section, we assume that n i.i.d. observations, $\{(Y_{ui}, X_{ui})_{u\in\mathcal{U}}\}_{i=1}^n$, from the distribution of $(Y_u, X_u)_{u\in\mathcal{U}}$ are available to estimate $(\theta_u)_{u\in\mathcal{U}}$. In addition, we assume that an estimate $\widehat{a}_u$ of the nuisance parameter a_u is available for all $u \in \mathcal{U}$. Using the estimate $\widehat{a}_u$, we use the criterion function

$$M_u(Y_u, X_u, \theta) := M_u(Y_u, X_u, \theta, \widehat{a}_u)$$

as a proxy for $M_u(Y_u, X_u, \theta_u, a_u)$. We allow for the case where p is much larger than n.

Since p is potentially larger n, and the parameters θ_u are assumed to be sparse, we consider an ℓ_1-penalized M_u-estimator (Lasso) of θ_u:

$$\widehat{\theta}_u \in \arg\min_\theta \left(\mathbb{E}_n[M_u(Y_u, X_u, \theta)] + \frac{\lambda}{n} \|\widehat{\Psi}_u \theta\|_1 \right) \tag{I.2}$$

where λ is a penalty level and $\widehat{\Psi}_u$ a diagonal matrix of penalty loadings. Further, for each $u \in \mathcal{U}$, we also consider a post-regularized (Post-Lasso) estimator of θ_u:

$$\widetilde{\theta}_u \in \arg\min_\theta \mathbb{E}_n[M_u(Y_u, X_u, \theta)] \quad : \quad \text{supp}(\theta) \subseteq \widehat{T}_u \tag{I.3}$$

where $\widehat{T}_u = \text{supp}(\widehat{\theta}_u)$.

We assume that for each $u \in \mathcal{U}$, the matrix of penalty loadings $\widehat{\Psi}_u$ is chosen as an appropriate estimator of the following "ideal" matrix of penalty loadings: $\widehat{\Psi}_{u0} = \text{diag}(\{l_{u0k}, k = 1, \ldots, p\})$, where

$$l_{u0k} = \left(\mathbb{E}_n \left[(\partial_{\theta_k} M_u(Y_u, X_u, \theta_u, a_u))^2 \right] \right)^{1/2}, \tag{I.4}$$

and $\partial_{\theta_k} M_u(Y_u, X_u, \theta_u, a_u)$ denotes a sub-gradient of the function $\theta \mapsto M_u(Y_u, X_u, \theta, a_u)$ with respect to the kth coordinate of θ and evaluated at $\theta = \theta_u$. The properties of $\widehat{\Psi}_u$ will be specified below in lemmas. Also, we assume that the penalty level λ is chosen such that with high probability,

$$\frac{\lambda}{n} \geqslant c \sup_{u\in\mathcal{U}} \left\| \widehat{\Psi}_{u0}^{-1} \mathbb{E}_n \left[\partial_\theta M_u(Y_u, X_u, \theta_u, a_u) \right] \right\|_\infty, \tag{I.5}$$

where $c > 1$ is a fixed constant. When $\mathcal{U}$ is a singleton, the condition (I.5) is similar to that in [14], [4], and [10]. When $\mathcal{U}$ is a continuum of indices, a similar condition was previously used in [3] in the context of ℓ_1-penalized quantile regression.

For $u \in \mathcal{U}$, denote $T_u = \text{supp}(\theta_u)$. Let ℓ and L be some constants satisfying $L \geqslant \ell > 1/c$. Also, let

$$\tilde{c} = \frac{Lc + 1}{\ell c - 1} \sup_{u\in\mathcal{U}} \|\widehat{\Psi}_{u0}\|_\infty \|\widehat{\Psi}_{u0}^{-1}\|_\infty,$$

where for any diagonal matrix $A = \text{diag}(\{a_k, k = 1, \ldots, p\})$, we denote $\|A\|_\infty = \max_{1\leqslant k\leqslant p} |a_k|$. Let $(\Delta_n)_{n\geqslant 1}$ be a sequence of positive constants converging to zero, and let $(C_n)_{n\geqslant 1}$ be a sequence of random variables. Also, let $w_u = w_u(X_u)$ be some weights satisfying $0 \leqslant w_u \leqslant 1$ almost surely. Finally, let A_u be some random subset of $\mathbb{R}^p$ and $\bar{q}_{A_u}$ be a random variable possibly depending on

A_u, where both A_u and $\bar{q}_{A_u}$ are specified in the lemmas below. To state our results in this section, we need the following assumption:

Assumption I.1 (M-Estimation Conditions)**.** *The function $\theta \mapsto M_u(Y_u, X_u, \theta)$ is convex almost surely, and with probability at least $1-\Delta_n$, the following inequalities hold for all $u \in \mathcal{U}$:*

(a) $|\mathbb{E}_n[\partial_\theta M_u(Y_u, X_u, \theta_u) - \partial_\theta M_u(Y_u, X_u, \theta_u, a_u)]'\delta| \leqslant C_n \|\sqrt{w_u} X_u'\delta\|_{\mathbb{P}_n,2}$ *for all* $\delta \in \mathbb{R}^p$*;*
(b) $\ell \widehat{\Psi}_{u0} \leqslant \widehat{\Psi}_u \leqslant L\widehat{\Psi}_{u0}$*;*
(c) *for all* $\delta \in A_u$*,*

$$\mathbb{E}_n[M_u(Y_u, X_u, \theta_u + \delta)] - \mathbb{E}_n[M_u(Y_u, X_u, \theta_u)] - \mathbb{E}_n[\partial_\theta M_u(Y_u, X_u, \theta_u)]'\delta + 2C_n\|\sqrt{w_u}X_u'\delta\|_{\mathbb{P}_n,2} \geqslant \left\{\|\sqrt{w_u}X_u'\delta\|^2_{\mathbb{P}_n,2}\right\} \wedge \left\{\bar{q}_{A_u}\|\sqrt{w_u}X_u'\delta\|_{\mathbb{P}_n,2}\right\}.$$

In many applications one can take the weights to be $w_u = w_u(X_u) = 1$ but we allow for more general weights since it is useful for our results on the weighted Lasso with estimated weights. Also, in applications, we typically have $C_n \lesssim \{n^{-1}s\log(pn)\}^{1/2}$. Assumption I.1(a) bounds the impact of estimating the nuisance functions uniformly over $u \in \mathcal{U}$. The loadings $\widehat{\Psi}_u$ are assumed larger (but not too much larger) than the ideal choice $\widehat{\Psi}_{u0}$ defined in (I.4). This is formalized in Assumption I.1(b). Assumption I.1(c) is an identification condition that will be imposed for particular choices of A_u and $\bar{q}_{A_u}$. It relates to conditions in the literature derived for the case of a singleton $\mathcal{U}$ and no nuisance functions, see the restricted strong convexity[2] used in [33] and the non-linear impact coefficients used in [3] and [8].

Define the restricted eigenvalue

$$\bar{\kappa}_{2\tilde{c}} = \inf_{u\in\mathcal{U}} \inf_{\delta\in\Delta_{2\tilde{c},u}} \|\sqrt{w_u}X_u'\delta\|_{\mathbb{P}_n,2}/\|\delta_{T_u}\|$$

where $\Delta_{2\tilde{c},u} = \{\delta : \|\delta^c_{T_u}\|_1 \leqslant 2\tilde{c}\|\delta_{T_u}\|_1\}$. Also, define minimum and maximum spare eigenvalues

$$\phi_{\min}(m,u) = \min_{1\leqslant\|\delta\|_0\leqslant m} \frac{\|\sqrt{w_u}X_u'\delta\|^2_{\mathbb{P}_n,2}}{\|\delta\|^2} \quad \text{and} \quad \phi_{\max}(m,u) = \max_{1\leqslant\|\delta\|_0\leqslant m} \frac{\|X_u'\delta\|^2_{\mathbb{P}_n,2}}{\|\delta\|^2}.$$

The following results establish the rate of convergence and a sparsity bound for the ℓ_1-penalized estimator $\widehat{\theta}_u$ defined in (I.2) as well as the rate of convergence for the post-regularized estimator $\widetilde{\theta}_u$ defined in (I.3).

Lemma I.1. *Suppose that Assumption I.1 holds with* $A_u = \{\delta\colon \|\delta^c_{T_u}\|_1 \leqslant 2\tilde{c}\|\delta_{T_u}\|_1\} \cup \{\delta\colon \|\delta\|_1 \leqslant \frac{3n}{\lambda}\frac{c\|\widehat{\Psi}^{-1}_{u0}\|_\infty}{\ell c-1}C_n\|\sqrt{w_u}X_u'\delta\|_{\mathbb{P}_n,2}\}$ *and* $\bar{q}_{A_u} > (L+\frac{1}{c})\|\widehat{\Psi}_{u0}\|_\infty \frac{\lambda\sqrt{s}}{n\bar{\kappa}_{2\tilde{c}}} + 6\tilde{c}C_n$*. In addition, suppose that λ satisfies condition (I.5) with probability $1-\Delta_n$. Then, with probability at least $1-2\Delta_n$, we have*

[2]Assumption I.1 (a) and (c) could have been stated with $\{C_n/\sqrt{s}\}\|\delta\|_1$ instead of $C_n\|\sqrt{w_u}X_u'\delta\|_{\mathbb{P}_n,2}$.

for all $u \in \mathcal{U}$ that

$$\|\sqrt{w_u} X_u'(\widehat{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2} \leqslant \Big(L + \frac{1}{c}\Big)\|\widehat{\Psi}_{u0}\|_\infty \frac{\lambda\sqrt{s}}{n\bar{\kappa}_{2\tilde{c}}} + 6\tilde{c}C_n,$$

$$\|\widehat{\theta}_u - \theta_u\|_1 \leqslant \Big(\frac{(1+2\tilde{c})\sqrt{s}}{\bar{\kappa}_{2\tilde{c}}} + \frac{3n}{\lambda}\frac{c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c - 1}C_n\Big) \cdot \Big(\Big(L + \frac{1}{c}\Big)\|\widehat{\Psi}_{u0}\|_\infty \frac{\lambda\sqrt{s}}{n\bar{\kappa}_{2\tilde{c}}} + 6\tilde{c}C_n\Big).$$

Lemma I.2. *In addition to conditions of Lemma I.1, suppose that with probability $1 - \Delta_n$, we have for some random variable L_n and all $u \in \mathcal{U}$ and $\delta \in \mathbb{R}^p$ that*

$$\Big|\{\mathbb{E}_n[\partial_\theta M_u(Y_u, X_u, \widehat{\theta}_u) - \partial_\theta M_u(Y_u, X_u, \theta_u)]\}'\delta\Big| \leqslant L_n \|X_u'\delta\|_{\mathbb{P}_n,2}. \tag{I.6}$$

Further, for all $u \in \mathcal{U}$, let $\widehat{s}_u = \operatorname{supp}(\widehat{T}_u)$. Then with probability at least $1 - 3\Delta_n$, we have for all $u \in \mathcal{U}$ that

$$\widehat{s}_u \leqslant \min_{m \in \mathcal{M}_u} \phi_{\max}(m, u) L_u^2,$$

where $\mathcal{M}_u = \{m \in \mathbb{N} \colon m \geqslant 2\phi_{\max}(m,u)L_u^2\}$ and $L_u = \frac{c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{c\ell - 1}\frac{n}{\lambda}\{C_n + L_n\}$.

Lemma I.3. *Suppose that Assumption I.1 holds with $A_u = \{\delta : \|\delta\|_0 \leqslant \widehat{s}_u + s_u\}$ and*

$$\bar{q}_{A_u} > 2\max\Big\{\Big(\mathbb{E}_n[M_u(Y_u, X_u, \widetilde{\theta}_u)] - \mathbb{E}_n[M_u(Y_u, X_u, \theta_u)]\Big)_+^{1/2}, \Big(\frac{\sqrt{\widehat{s}_u + s_u}\|\mathbb{E}_n[\partial_\theta M_u(Y_u, X_u, \theta_u, a_u)]\|_\infty}{\sqrt{\phi_{\min}(\widehat{s}_u + s_u, u)}} + 3C_n\Big)\Big\}. \tag{I.7}$$

Then with probability at least $1 - \Delta_n$, we have for all $u \in \mathcal{U}$ that

$$\|\sqrt{w_u} X_u'(\widetilde{\theta}_u - \theta_u)\|_{\mathbb{P}_n,2} \leqslant \Big(\mathbb{E}_n[M_u(Y_u, X_u, \widetilde{\theta}_u)] - \mathbb{E}_n[M_u(Y_u, X_u, \theta_u)]\Big)_+^{1/2} + \frac{\sqrt{\widehat{s}_u + s_u}\|\mathbb{E}_n[\partial_\theta M_u(Y_u, X_u, \theta_u, a_u)]\|_\infty}{\sqrt{\phi_{\min}(\widehat{s}_u + s_u, u)}} + 3C_n.$$

In addition, with probability at least $1 - \Delta_n$, we have for all $u \in \mathcal{U}$ that

$$\mathbb{E}_n[M_u(Y_u, X_u, \widetilde{\theta}_u)] - \mathbb{E}_n[M_u(Y_u, X_u, \theta_u)] \leqslant \frac{\lambda L}{n}\|\widehat{\theta}_u - \theta_u\|_1 \sup_{u \in \mathcal{U}} \|\widehat{\Psi}_{u0}\|_\infty. \tag{I.8}$$

A key requirement in Lemmas I.1 and I.2 is that λ satisfies (I.5) with high probability. Therefore, below we provide a choice of λ and a set of conditions under which the proposed choice of λ satisfies this requirement. Let $d_{\mathcal{U}} \colon \mathcal{U} \times \mathcal{U} \to \mathbb{R}_+$ denote a metric on $\mathcal{U}$. Also, let

$$S_u = \partial_\theta M_u(Y_u, X_u, \theta_u, a_u), \quad u \in \mathcal{U}.$$

Moreover, let $\underline{C}$ and $\bar{C}$ be some strictly positive constants. Finally, $(\epsilon_n)_{n\geqslant 1}$, $(\varphi_n)_{n\geqslant 1}$, and $(N_n)_{n\geqslant 1}$ be some sequences of positive constants, where $\varphi_n = o(1)$.

Condition WL. *The constants ϵ_n and N_n satisfy the inequality $N_n \geqslant N(\epsilon_n, \mathcal{U}, d_{\mathcal{U}})$ and the following conditions hold:*

(i) $\sup_{u \in \mathcal{U}} \max_{k \in [p]} (\mathrm{E}_P[|S_{uk}|^3])^{1/3} \Phi^{-1}(1 - \gamma/\{2pN_n\}) \leqslant \varphi_n n^{1/6}$;

(ii) $\underline{C} \leqslant \mathrm{E}_P[|S_{uk}|^2] \leqslant \bar{C}$*, for all* $u \in \mathcal{U}$ *and* $k \in [p]$*;*
(iii) with probability at least $1 - \Delta_n$*,*

$$\sup_{d_{\mathcal{U}}(u,u') \leqslant \epsilon_n} \|\mathbb{E}_n[S_u - S_{u'}]\|_\infty \leqslant \varphi_n n^{-1/2}, \text{ and } \sup_{d_{\mathcal{U}}(u,u') \leqslant \epsilon_n} \max_{k \in [p]} \Big| |\mathrm{E}_P[S_{uk}^2 - S_{u'k}^2]| + |(\mathbb{E}_n - \mathrm{E}_P)[S_{uk}^2]| \Big| \leqslant \varphi_n.$$

Let

$$\lambda = c' \sqrt{n} \Phi^{-1}(1 - \gamma / \{2pN_n\}), \tag{I.9}$$

where $1-\gamma$ (with $\gamma = \gamma_n = o(1)$) is a confidence level associated with the probability of event (I.5), and $c' > c$ is a slack constant. The following lemma shows that this choice of λ satisfies (I.5) with high probability under Condition WL.

Lemma I.4. *Suppose that Condition WL holds. In addition, suppose that* λ *satisfies* (I.9) *for some* $c' > c$ *and* $\gamma = \gamma_n \in [1/n, 1/\log n]$*. Then*

$$\mathrm{P}_P \left(\lambda / n \geqslant c \sup_{u \in \mathcal{U}} \|\widehat{\Psi}_{u0}^{-1} \mathbb{E}_n[S_u]\|_\infty \right) \geqslant 1 - \gamma - o(\gamma) - \Delta_n.$$

Condition WL(iii) is of high level. Therefore, to conclude this section, we present a lemma that gives easy to verify conditions that imply Condition WL(iii).

Lemma I.5. *Suppose that for all* $u \in \mathcal{U}$*,* $X_u = X$ *and* $Y_u = H(Y, u)$ *where* Y *is a random variable and* $\{H(\cdot, u) \colon u \in \mathcal{U}\}$ *is a VC-subgraph class of functions bounded by one with index* C_Y *for some constant* $C_Y \geqslant 1$*. In addition, suppose that for all* $u \in \mathcal{U}$*, we have* $S_u = (Y_u - \mathrm{E}_P[Y_u \mid X]) \cdot X$*. Moreover, suppose that* $\max_{k \in [p]} \mathrm{E}_P[X_k^4] \leqslant \bar{C}$*,* $\underline{C} \leqslant \sup_{u \in \mathcal{U}, k \in [p]} \mathrm{E}_P[S_{uk}^2] \leqslant \bar{C}$*, and* $\mathrm{E}_P[\|X\|_\infty^q]^{1/q} \leqslant K_n$*, for some constants* $\underline{C}, \bar{C} > 0$ *and* $q \geqslant 4$ *and a sequence of constants* $(K_n)_{n \geqslant 1}$*. Finally, suppose that* $\mathrm{E}_P[|Y_u - Y_{u'}|^4] \leqslant C_u |u - u'|^\nu$ *for any* $u, u' \in \mathcal{U}$ *and some constants* ν *and* C*. Then we have with probability at least* $1 - (\log n)^{-1}$ *that*

$$\sup_{d_{\mathcal{U}}(u,u') \leqslant 1/n} \|\mathbb{E}_n[S_u - S_{u'}]\|_\infty \lesssim \left(\frac{\log(npK_n)}{n^{1+\nu/2}} \right)^{1/2} + \frac{K_n \log(npK_n)}{n^{1-1/q}} \tag{I.10}$$

$$\sup_{u \in \mathcal{U}} \max_{k \in p} |(\mathbb{E}_n - \mathrm{E}_P)[S_{uk}^2]| \lesssim \left(\frac{\log(npK_n)}{n} \right)^{1/2} + \frac{K_n^2 \log(npK_n)}{n^{1-2/q}} \tag{I.11}$$

$$\max_{k \in [p]} |\mathrm{E}_P[S_{uk}^2 - S_{u'k}^2]| \lesssim d_{\mathcal{U}}(u, u')^{\nu/4} \tag{I.12}$$

up-to constants that depend only on $\underline{C}, \bar{C}, C_Y, C_u$*,* q*, and* ν*.*

Appendix J. Proofs for Appendix I

Proof of Lemma I.1. For $u \in \mathcal{U}$, let $\delta_u = \widehat{\theta}_u - \theta_u$ and $S_{u,n} = \mathbb{E}_n[\partial_\theta M_u(Y_u, X_u, \theta_u, a_u)]$. Throughout the proof, we will assume that the events (a), (b), and (c) in Assumption I.1 as well as the event (I.5) hold. These events hold with probability at least $1 - 2\Delta_n$. We will show that the inequalities in the statement of Lemma I.1 hold under these events.

By definition of $\widehat{\theta}_u$, we have $\mathbb{E}_n[M_u(Y_u, X_u, \widehat{\theta}_u)] + \frac{\lambda}{n}\|\widehat{\Psi}_u\widehat{\theta}_u\|_1 \leqslant \mathbb{E}_n[M_u(Y_u, X_u, \theta_u)] + \frac{\lambda}{n}\|\widehat{\Psi}_u\theta_u\|_1$. Thus,

$$\begin{aligned}\mathbb{E}_n[M_u(Y_u, X_u, \widehat{\theta}_u)] - \mathbb{E}_n[M_u(Y_u, X_u, \theta_u)] &\leqslant \frac{\lambda}{n}\|\widehat{\Psi}_u\theta_u\|_1 - \frac{\lambda}{n}\|\widehat{\Psi}_u\widehat{\theta}_u\|_1 \\ &\leqslant \frac{\lambda}{n}\|\widehat{\Psi}_u\delta_{u,T_u}\|_1 - \frac{\lambda}{n}\|\widehat{\Psi}_u\delta_{u,T_u^c}\|_1 \\ &\leqslant \frac{\lambda L}{n}\|\widehat{\Psi}_{u0}\delta_{u,T_u}\|_1 - \frac{\lambda \ell}{n}\|\widehat{\Psi}_{u0}\delta_{u,T_u^c}\|_1. \end{aligned} \tag{J.1}$$

Moreover, by convexity of $\theta \mapsto M_u(Y_u, X_u, \theta)$, we have

$$\begin{aligned}\mathbb{E}_n[M_u(Y_u, X_u, \widehat{\theta}_u)] - \mathbb{E}_n[M_u(Y_u, X_u, \theta_u)] &\geqslant \mathbb{E}_n[\partial_\theta M_u(Y_u, X_u, \theta_u)]'\delta_u \qquad \text{(J.2)} \\ &\geqslant -\frac{\lambda}{n}\frac{1}{c}\|\widehat{\Psi}_{u0}\delta_u\|_1 - C_n\|\sqrt{w_u}X_u'\delta_u\|_{\mathbb{P}_n,2}\end{aligned}$$

where the second inequality holds by Assumption I.1(a) and $\lambda/n \geqslant c\sup_{u\in\mathcal{U}}\|\widehat{\Psi}_{u0}^{-1}S_{u,n}\|_\infty$ since

$$\begin{aligned}|\mathbb{E}_n[\partial_\theta M_u(Y_u, X_u, \theta_u)]'\delta_u| &= |S_{u,n}'\delta_u + \{\mathbb{E}_n[\partial_\theta M_u(Y_u, X_u, \theta_u)] - S_{u,n}\}'\delta_u| \\ &\leqslant |S_{u,n}'\delta_u| + |\{\mathbb{E}_n[\partial_\theta M_u(Y_u, X_u, \theta_u)] - S_{u,n}\}'\delta_u| \\ &\leqslant \|\widehat{\Psi}_{u0}^{-1}S_{u,n}\|_\infty\|\widehat{\Psi}_{u0}\delta_u\|_1 + C_n\|\sqrt{w_u}X_u'\delta_u\|_{\mathbb{P}_n,2} \\ &\leqslant \frac{\lambda}{n}\frac{1}{c}\|\widehat{\Psi}_{u0}\delta_u\|_1 + C_n\|\sqrt{w_u}X_u'\delta_u\|_{\mathbb{P}_n,2}.\end{aligned} \tag{J.3}$$

Combining (J.1) and (J.2), we have

$$\frac{\lambda}{n}\frac{c\ell-1}{c}\|\widehat{\Psi}_{u0}\delta_{u,T_u^c}\|_1 \leqslant \frac{\lambda}{n}\frac{Lc+1}{c}\|\widehat{\Psi}_{u0}\delta_{u,T_u}\|_1 + C_n\|\sqrt{w_u}X_u'\delta_u\|_{\mathbb{P}_n,2}, \tag{J.4}$$

and for $\tilde{c} = \frac{Lc+1}{\ell c-1}\sup_{u\in\mathcal{U}}\|\widehat{\Psi}_{u0}\|_\infty\|\widehat{\Psi}_{u0}^{-1}\|_\infty$, we have

$$\|\delta_{u,T_u^c}\|_1 \leqslant \tilde{c}\|\delta_{u,T_u}\|_1 + \frac{n}{\lambda}\frac{c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}C_n\|\sqrt{w_u}X_u'\delta_u\|_{\mathbb{P}_n,2}.$$

Now, suppose that $\delta_u \notin \Delta_{2\tilde{c},u}$, namely $\|\delta_{u,T_u^c}\|_1 > 2\tilde{c}\|\delta_{u,T_u}\|_1$. Then

$$2\tilde{c}\|\delta_{u,T_u}\|_1 \leqslant \tilde{c}\|\delta_{u,T_u}\|_1 + \frac{n}{\lambda}\frac{c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}C_n\|\sqrt{w_u}X_u'\delta_u\|_{\mathbb{P}_n,2},$$

and so

$$\|\delta_{u,T_u}\|_1 \leqslant \frac{n}{\lambda}\frac{c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}C_n\|\sqrt{w_u}X_u'\delta_u\|_{\mathbb{P}_n,2}$$

since $\tilde{c} \geqslant 1$. Also,

$$\|\delta_{u,T_u^c}\|_1 \leqslant \frac{1}{2}\|\delta_{u,T_u^c}\|_1 + \frac{n}{\lambda}\frac{c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}C_n\|\sqrt{w_u}X_u'\delta_u\|_{\mathbb{P}_n,2},$$

and so

$$\|\delta_{u,T_u^c}\|_1 \leqslant \frac{2n}{\lambda}\frac{c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}C_n\|\sqrt{w_u}X_u'\delta_u\|_{\mathbb{P}_n,2}.$$

Therefore,

$$\|\delta_u\|_1 \leqslant \frac{3n}{\lambda}\frac{c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}C_n\|\sqrt{w_u}X_u'\delta_u\|_{\mathbb{P}_n,2} =: I_u, \tag{J.5}$$

as long as $\delta_u \notin \Delta_{2\tilde{c},u}$. In addition, if $\delta_u \in \Delta_{2\tilde{c},u}$, then

$$\|\delta_{u,T_u}\|_1 \leqslant \sqrt{s}\|\delta_{u,T_u}\| \leqslant \frac{\sqrt{s}}{\bar{\kappa}_{2\tilde{c}}}\|\sqrt{w_u}X_u'\delta_u\|_{\mathbb{P}_n,2} =: II_u.$$

Hence, in both cases, we have

$$\|\delta_{u,T_u}\|_1 \leqslant I_u + II_u. \tag{J.6}$$

Next, for every $u \in \mathcal{U}$, since $A_u = \Delta_{2\tilde{c},u} \cup \{\delta : \|\delta\|_1 \leqslant \frac{3n}{\lambda}\frac{c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}C_n\|\sqrt{w_u}X_u'\delta\|_{\mathbb{P}_n,2}\}$, it follows that $\delta_u \in A_u$, and we have

$$\begin{aligned}
&\|\sqrt{w_u}X_u'\delta_u\|_{\mathbb{P}_n,2}^2 \wedge \left\{\bar{q}_{A_u}\|\sqrt{w_u}X_u'\delta_u\|_{\mathbb{P}_n,2}\right\} \\
&\quad\leqslant \mathbb{E}_n[M_u(Y_u,X_u,\widehat{\theta}_u)] - \mathbb{E}_n[M_u(Y_u,X_u,\theta_u)] - \mathbb{E}_n[\partial_\theta M_u(Y_u,X_u,\theta_u)]'\delta_u + 2C_n\|\sqrt{w_u}X_u'\delta_u\|_{\mathbb{P}_n,2} \\
&\quad\leqslant \Big(L+\frac{1}{c}\Big)\frac{\lambda}{n}\|\widehat{\Psi}_{u0}\delta_{u,T_u}\|_1 + 3C_n\|\sqrt{w_u}X_u'\delta_u\|_{\mathbb{P}_n,2} \\
&\quad\leqslant \Big(L+\frac{1}{c}\Big)\frac{\lambda}{n}\|\widehat{\Psi}_{u0}\|_\infty\left\{I_u + II_u\right\} + 3C_n\|\sqrt{w_u}X_u'\delta_u\|_{\mathbb{P}_n,2} \\
&\quad\leqslant \Big\{\Big(L+\frac{1}{c}\Big)\|\widehat{\Psi}_{u0}\|_\infty\frac{\lambda\sqrt{s}}{n\bar{\kappa}_{2\tilde{c}}} + 6\tilde{c}C_n\Big\}\|\sqrt{w_u}X_u'\delta_u\|_{\mathbb{P}_n,2},
\end{aligned}$$

where the second line follows from Assumption I.1(c), the third from (J.1), (J.3), and $\ell \geqslant 1/c$, the fourth from $\|\widehat{\Psi}_{u0}\delta_{u,T_u}\|_1 \leqslant \|\widehat{\Psi}_{u0}\|_\infty\|\delta_{u,T_u}\|_1$ and (J.6), and the fifth from definitions of I_u, II_u, and $\tilde{c}$. Thus, as long as

$$\bar{q}_{A_u} > \Big(L+\frac{1}{c}\Big)\|\widehat{\Psi}_{u0}\|_\infty\frac{\lambda\sqrt{s}}{n\bar{\kappa}_{2\tilde{c}}} + 6\tilde{c}C_n, \quad \text{for all } u \in \mathcal{U},$$

which is assumed, we have

$$\|\sqrt{w_u}X_u'\delta_u\|_{\mathbb{P}_n,2} \leqslant \Big(L+\frac{1}{c}\Big)\|\widehat{\Psi}_{u0}\|_\infty\frac{\lambda\sqrt{s}}{n\bar{\kappa}_{2\tilde{c}}} + 6\tilde{c}C_n =: III_u, \quad \text{for all } u \in \mathcal{U}.$$

This gives the first asserted claim. The second asserted claim follows from

$$\begin{aligned}
\|\delta_u\|_1 &\leqslant 1\{\delta_u \in \Delta_{2\tilde{c},u}\}\|\delta_u\|_1 + 1\{\delta_u \notin \Delta_{2\tilde{c},u}\}\|\delta_u\|_1 \\
&\leqslant (1+2\tilde{c})II_u + I_u \leqslant \Big(\frac{(1+2\tilde{c})\sqrt{s}}{\bar{\kappa}_{2\tilde{c}}} + \frac{3n}{\lambda}\frac{c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell c-1}C_n\Big)III_u.
\end{aligned}$$

This completes the proof. ∎

Proof of Lemma I.2. For $u \in \mathcal{U}$, let $S_{u,n} = \mathbb{E}_n[\partial_\theta M_u(Y_u,X_u,\theta_u,a_u)]$. Throughout the proof, we will assume that the events (a), (b), and (c) in Assumption I.1 as well as the events (I.5) and (I.6) hold. These events hold with probability at least $1-3\Delta_n$. We will show that the inequalities in the statement of Lemma I.2 hold under these events.

By definition of the estimator $\widehat{\theta}_u$, there is a subgradient $\partial_\theta\mathbb{E}_n[M_u(Y_u,X_u,\widehat{\theta}_u)]$ of $\mathbb{E}_n[M_u(Y_u,X_u,\widehat{\theta}_u)]$, such that for every j with $|\widehat{\theta}_{uj}| > 0$,

$$|(\widehat{\Psi}_u^{-1}\partial_\theta\mathbb{E}_n[M_u(Y_u,X_u,\widehat{\theta}_u)])_j| = \lambda/n.$$

Therefore, we have

$$\begin{aligned}\frac{\lambda}{n}\sqrt{\widehat{s}_u} &= \|(\widehat{\Psi}_u^{-1}\partial_\theta \mathbb{E}_n[M_u(Y_u,X_u,\widehat{\theta}_u)])_{\widehat{T}_u}\| \\ &\leqslant \|(\widehat{\Psi}_u^{-1}S_{u,n})_{\widehat{T}_u}\| + \|(\widehat{\Psi}_u^{-1}\{\mathbb{E}_n[\partial_\theta M_u(Y_u,X_u,\theta_u)] - S_{u,n}\})_{\widehat{T}_u}\| \\ &\quad + \|(\widehat{\Psi}_u^{-1}\{\mathbb{E}_n[\partial_\theta M_u(Y_u,X_u,\widehat{\theta}_u) - \partial_\theta M_u(Y_u,X_u,\theta_u)]\})_{\widehat{T}_u}\| \\ &\leqslant \|\widehat{\Psi}_u^{-1}\widehat{\Psi}_{u0}\|_\infty \|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[S_{u,n}]\|_\infty \sqrt{\widehat{s}_u} + \|\widehat{\Psi}_u^{-1}\|_\infty C_n \sup_{\|\delta\|=1,\|\delta\|_0\leqslant \widehat{s}_u} \|\sqrt{w_u}X_u'\delta\|_{\mathbb{P}_n,2} \\ &\quad + \|\widehat{\Psi}_u^{-1}\|_\infty \sup_{\|\delta\|=1,\|\delta\|_0\leqslant \widehat{s}_u} |\{\mathbb{E}_n[\partial_\theta M_u(Y_u,X_u,\widehat{\theta}_u) - \partial_\theta M_u(Y_u,X_u,\theta_u)]\}'\delta| \\ &\leqslant \frac{\lambda}{c\ell n}\sqrt{\widehat{s}_u} + \frac{\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{\ell}\{C_n + L_n\} \sup_{\|\delta\|=1,\|\delta\|_0\leqslant \widehat{s}_u} \|X_u'\delta\|_{\mathbb{P}_n,2}\end{aligned}$$

where the first inequality follows from the triangle inequality, the second from Assumption I.1(a), and the third from Assumption I.1(b) and inequalities (I.5) and (I.6).

Now, recall that $L_u = \frac{n}{\lambda}\frac{c\|\widehat{\Psi}_{u0}^{-1}\|_\infty}{c\ell-1}\{C_n + L_n\}$. In addition, note that $\sup_{\|\delta\|=1,\|\delta\|_0\leqslant \widehat{s}_u}\|X_u'\delta\|^2_{\mathbb{P}_n,2} = \phi_{\max}(\widehat{s}_u,u)$. Thus, we have

$$\widehat{s}_u \leqslant \phi_{\max}(\widehat{s}_u,u)L_u^2. \tag{J.7}$$

Consider any $M \in \mathcal{M}_u = \{m \in \mathbb{N} : m > 2\phi_{\max}(m,u)L_u^2\}$, and suppose that $\widehat{s}_u > M$. By the sublinearity of the maximum sparse eigenvalue (Lemma 3 in [4]), for any integer $k \geqslant 0$ and constant $\ell \geqslant 1$, we have $\phi_{\max}(\ell k,u) \leqslant \lceil \ell \rceil \phi_{\max}(k,u)$, where $\lceil \ell \rceil$ denotes the ceiling of ℓ. Therefore,

$$\widehat{s}_u \leqslant \phi_{\max}(\widehat{s}_u,u)L_u^2 = \phi_{\max}(M\widehat{s}_u/M,u)L_u^2 \leqslant \left\lceil \frac{\widehat{s}_u}{M}\right\rceil \phi_{\max}(M,u)L_u^2 \leqslant \frac{2\widehat{s}_n}{M}\phi_{\max}(M,u)L_u^2$$

since $\lceil k \rceil \leqslant 2k$ for any $k \geqslant 1$. Therefore, we have $M \leqslant 2\phi_{\max}(M,u)L_u^2$ which violates the condition that $M \in \mathcal{M}_u$. Therefore, we have $\widehat{s}_u \leqslant M$. In turn, applying (J.7) once more with $\widehat{s}_u \leqslant M$ we obtain $\widehat{s}_u \leqslant \phi_{\max}(M,u)L_u^2$. The result follows by minimizing the bound over $M \in \mathcal{M}_u$. ■

Proof of Lemma I.3. The second asserted claim, inequality (I.8), follows from the observation that with probability at least $1-\Delta_n$, for all $u \in \mathcal{U}$, we have that

$$\begin{aligned}&\mathbb{E}_n[M_u(Y_u,X_u,\widetilde{\theta}_u)] - \mathbb{E}_n[M_u(Y_u,X_u,\theta_u)] \leqslant \mathbb{E}_n[M_u(Y_u,X_u,\widehat{\theta}_u)] - \mathbb{E}_n[M_u(Y_u,X_u,\theta_u)] \\ &\quad\leqslant \frac{\lambda}{n}\|\widehat{\Psi}_u\theta_u\|_1 - \frac{\lambda}{n}\|\widehat{\Psi}_u\widehat{\theta}_u\|_1 \leqslant \frac{\lambda}{n}\|\widehat{\theta}_u - \theta_u\|_1 \cdot \sup_{u\in\mathcal{U}}\|\widehat{\Psi}_u\|_\infty \leqslant \frac{\lambda L}{n}\|\widehat{\theta}_u - \theta_u\|_1 \cdot \sup_{u\in\mathcal{U}}\|\widehat{\Psi}_{u0}\|_\infty,\end{aligned}$$

where the first inequality holds by the definition of $\widetilde{\theta}_u$, the second by the definition of $\widehat{\theta}_u$, the third by the triangle inequality, and the fourth by Assumption I.1(ii).

To prove the first asserted claim, assume that the events (a), (b), and (c) in Assumption I.1 hold. These events hold with probability at least $1-\Delta_n$. We will show that the asserted claim holds under these events.

For $u \in \mathcal{U}$, let $\tilde{\delta}_u = \widetilde{\theta}_u - \theta_u$, $S_{u,n} = \mathbb{E}_n[\partial_\theta M_u(Y_u, X_u, \theta_u, a_u)]$, and $\tilde{t}_u = \|\sqrt{w_u} X_u' \tilde{\delta}_u\|_{\mathbb{P}_n,2}$. By the inequality in Assumption I.1(c), we have

$$\begin{aligned}
\tilde{t}_u^2 \wedge \left\{\bar{q}_{A_u}\tilde{t}_u\right\} &\leqslant \mathbb{E}_n[M_u(Y_u, X_u, \widetilde{\theta}_u)] - \mathbb{E}_n[M_u(Y_u, X_u, \theta_u)] - \mathbb{E}_n[\partial_\theta M_u(Y_u, X_u, \theta_u)]'\tilde{\delta}_u + 2C_n\tilde{t}_u \\
&\leqslant \mathbb{E}_n[M_u(Y_u, X_u, \widetilde{\theta}_u)] - \mathbb{E}_n[M_u(Y_u, X_u, \theta_u)] + \|S_{u,n}\|_\infty \|\tilde{\delta}_u\|_1 + 3C_n\tilde{t}_u \\
&\leqslant \mathbb{E}_n[M_u(Y_u, X_u, \widetilde{\theta}_u)] - \mathbb{E}_n[M_u(Y_u, X_u, \theta_u)] + \tilde{t}_u\Big(\frac{\sqrt{\widehat{s}_u + s_u}\|S_{u,n}\|_\infty}{\sqrt{\phi_{\min}(\widehat{s}_u + s_u, u)}} + 3C_n\Big)
\end{aligned}$$

where the second inequality holds by calculations as in (J.3), and the third inequality follows from

$$\|\tilde{\delta}_u\|_1 \leqslant \sqrt{\widehat{s}_u + s_u}\|\tilde{\delta}_u\|_2 \leqslant \frac{\sqrt{\widehat{s}_u + s_u}}{\sqrt{\phi_{\min}(\widehat{s}_u + s_u, u)}}\|\sqrt{w_u} X_u'\tilde{\delta}_u\|_{\mathbb{P}_n,2}.$$

Next, if $\tilde{t}_u^2 > \bar{q}_{A_u}\tilde{t}_u$, then

$$\bar{q}_{A_u}\tilde{t}_u \leqslant \frac{\bar{q}_{A_u}}{2}\{\mathbb{E}_n[M_u(Y_u, X_u, \widetilde{\theta}_u)] - \mathbb{E}_n[M_u(Y_u, X_u, \theta_u)]\}_+^{1/2} + \frac{\bar{q}_{A_u}}{2}\tilde{t}_u,$$

so that $\tilde{t}_u \leqslant \{\mathbb{E}_n[M_u(Y_u, X_u, \widetilde{\theta}_u)] - \mathbb{E}_n[M_u(Y_u, X_u, \theta_u)]\}_+^{1/2}$. On the other hand, if $\tilde{t}_u^2 \leqslant \bar{q}_{A_u}\tilde{t}_u$, then

$$\tilde{t}_u^2 \leqslant \{\mathbb{E}_n[M_u(Y_u, X_u, \widetilde{\theta}_u)] - \mathbb{E}_n[M_u(Y_u, X_u, \theta_u)]\} + \tilde{t}_u\left(\frac{\sqrt{\widehat{s}_u + s_u}\|S_{u,n}\|_\infty}{\sqrt{\phi_{\min}(\widehat{s}_u + s_u)}} + 3C_n\right).$$

Since for positive numbers a, b, c, inequality $a^2 \leqslant b + ac$ implies $a \leqslant \sqrt{b} + c$, we have

$$\tilde{t}_u \leqslant \{\mathbb{E}_n[M_u(Y_u, X_u, \widetilde{\theta}_u)] - \mathbb{E}_n[M_u(Y_u, X_u, \theta_u)]\}_+^{1/2} + \left(\frac{\sqrt{\widehat{s}_u + s_u}\|S_{u,n}\|_\infty}{\sqrt{\phi_{\min}(\widehat{s}_u + s_u)}} + 3C_n\right).$$

In both cases, the inequality in the asserted claim holds. This completes the proof. ■

Proof of Lemma I.4. For brevity of notation, denote $\epsilon = \epsilon_n$. Also, let $\mathcal{A}_n$ denote the event that the inequalities in Condition WL(iii) hold. Then $\mathrm{P}_P(\mathcal{A}_n) \geqslant 1 - \Delta_n$. Further, by the triangle inequality,

$$\begin{aligned}
\sup_{u\in\mathcal{U}}\|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[S_u]\|_\infty &\leqslant \max_{u\in\mathcal{U}^\epsilon}\|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[S_u]\|_\infty \\
&\quad + \sup_{u\in\mathcal{U}^\epsilon, u'\in\mathcal{U}, d_{\mathcal{U}}(u,u')\leqslant\epsilon}\|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[S_u] - \widehat{\Psi}_{u'0}^{-1}\mathbb{E}_n[S_{u'}]\|_\infty
\end{aligned} \tag{J.8}$$

where $\mathcal{U}^\epsilon$ is a minimal ϵ-net of $\mathcal{U}$ so that $|\mathcal{U}^\epsilon| \leqslant N_n$.

For each $k = 1, \dots, p$ and $u \in \mathcal{U}^\epsilon$, we apply Lemma L.1 with $Z_i := S_{uki}$, $\mu = 1$, and $\ell_n = c''\varphi_n^{-1}$, where c'' is a small enough constant that can be chosen to depend only on $\underline{C}$ and $\bar{C}$. Then Condition WL(i,ii) implies that

$$0 \leqslant \Phi^{-1}\left(1 - \frac{\gamma}{2pN_n}\right) \leqslant \frac{n^{1/6}M_n}{\ell_n} - 1$$

where $M_n = (\mathrm{E}_P[Z_1^2])^{1/2}/(\mathrm{E}_P[Z_1^3])^{1/3}$, and so applying Lemma L.1, the union bound, and the inequality $|\mathcal{U}^\epsilon| \leqslant N_n$ gives

$$\mathrm{P}_P\left(\sup_{u\in\mathcal{U}^\epsilon}\max_{k\in[p]}\frac{|\sqrt{n}\mathbb{E}_n[S_{uk}]|}{\sqrt{\mathbb{E}_n[S_{uk}^2]}} > \Phi^{-1}\Big(1-\tfrac{\gamma}{2pN_n}\Big)\right) \leqslant 2pN_n\cdot\frac{\gamma}{2pN_n}\cdot\Big(1+O(\varphi_n^{1/3})\Big) \leqslant \gamma + o(\gamma) \quad \text{(J.9)}$$

since $\varphi_n = o(1)$. Also, observe that

$$\max_{u\in\mathcal{U}^\epsilon}\|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[S_u]\|_\infty = \sup_{u\in\mathcal{U}^\epsilon}\max_{k\in[p]}\frac{|\mathbb{E}_n[S_{uk}]|}{\sqrt{\mathbb{E}_n[S_{uk}^2]}}.$$

Therefore, (J.9) implies that with probability at least $1-\gamma-o(\gamma)$,

$$\max_{u\in\mathcal{U}^\epsilon}\|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[S_u]\|_\infty \leqslant n^{-1/2}\Phi^{-1}\left(1-\frac{\gamma}{2pN_n}\right) \quad \text{(J.10)}$$

Further, by the triangle inequality,

$$\begin{aligned}\sup_{u\in\mathcal{U}^\epsilon, u'\in\mathcal{U}, d_{\mathcal{U}}(u,u')\leqslant\epsilon}&\|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[S_u] - \widehat{\Psi}_{u'0}^{-1}\mathbb{E}_n[S_{u'}]\|_\infty\\ &\leqslant \sup_{u\in\mathcal{U}^\epsilon, u'\in\mathcal{U}, d_{\mathcal{U}}(u,u')\leqslant\epsilon}\|(\widehat{\Psi}_{u0}^{-1}-\widehat{\Psi}_{u'0}^{-1})\widehat{\Psi}_{u0}\|_\infty\|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[S_u]\|_\infty \quad \text{(J.11)}\\ &\quad + \sup_{u,u'\in\mathcal{U}, d_{\mathcal{U}}(u,u')\leqslant\epsilon}\|\mathbb{E}_n[S_u - S_{u'}]\|_\infty\|\widehat{\Psi}_{u'0}^{-1}\|_\infty. \quad \text{(J.12)}\end{aligned}$$

To control the expression in (J.11), note that by Condition WL(ii), on the event $\mathcal{A}_n$, $\widehat{\Psi}_{u0kk}$ is bounded away from zero uniformly over $u\in\mathcal{U}$ and $k\in[p]$. Thus, we have uniformly over $u\in\mathcal{U}$ and $k\in[p]$ that

$$|(\widehat{\Psi}_{u0kk}^{-1}-\widehat{\Psi}_{u'0kk}^{-1})\widehat{\Psi}_{u0kk}| = |\widehat{\Psi}_{u0kk}-\widehat{\Psi}_{u'0kk}|\widehat{\Psi}_{u'0kk}^{-1} \lesssim |\widehat{\Psi}_{u0kk}-\widehat{\Psi}_{u'0kk}| \quad \text{(J.13)}$$

on the event $\mathcal{A}_n$. Moreover, we have

$$\begin{aligned}\sup_{u,u'\in\mathcal{U}, d_{\mathcal{U}}(u,u')\leqslant\epsilon}\max_{k\in[p]}&\left|\{\mathbb{E}_n[S_{uk}^2]\}^{1/2} - \{\mathbb{E}_n[S_{u'k}^2]\}^{1/2}\right| \quad \text{(J.14)}\\ &\leqslant \sup_{u,u'\in\mathcal{U}, d_{\mathcal{U}}(u,u')\leqslant\epsilon}\max_{k\in[p]}\Big(|\mathbb{E}_n[S_{uk}^2]-\mathbb{E}_n[S_{u'k}^2]|\Big)^{1/2}\\ &\leqslant \sup_{u,u'\in\mathcal{U}, d_{\mathcal{U}}(u,u')\leqslant\epsilon}\max_{k\in[p]}\Big(2|(\mathbb{E}_n-\mathrm{E}_P)[S_{uk}^2]| + |\mathrm{E}_P[S_{uk}^2-S_{u'k}^2]|\Big)^{1/2} \lesssim \varphi_n^{1/2}\end{aligned}$$

on the event $\mathcal{A}_n$. Thus, relations (J.13) and (J.14) imply that

$$\sup_{u,u'\in\mathcal{U}, d_{\mathcal{U}}(u,u')\leqslant\epsilon}\|(\widehat{\Psi}_{u0}^{-1}-\widehat{\Psi}_{u'0}^{-1})\widehat{\Psi}_{u0}\|_\infty \lesssim \varphi_n^{1/2}$$

on the event $\mathcal{A}_n$. Also, using standard bounds for the tails of Gaussian random variables gives

$$\Phi^{-1}\left(1-\frac{\gamma}{2pN_n}\right) \lesssim \sqrt{\log(2pN_n/\gamma)}.$$

Thus, on the intersection of events $\mathcal{A}_n$ and (J.10), we have

$$\sup_{u\in\mathcal{U}^\epsilon, u'\in\mathcal{U}, d_{\mathcal{U}}(u,u')\leqslant\epsilon} \|(\widehat{\Psi}_{u0}^{-1}-\widehat{\Psi}_{u'0}^{-1})\widehat{\Psi}_{u0}\|_\infty\|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[S_u]\|_\infty \lesssim (\varphi_n/n)^{1/2}\sqrt{\log(pN_n/\gamma)}.$$

Finally, on the event $\mathcal{A}_n$, we have that the expression in (J.12) satisfies

$$\sup_{u,u'\in\mathcal{U}, d_{\mathcal{U}}(u,u')\leqslant\epsilon} \|\mathbb{E}_n[S_u-S_{u'}]\|_\infty\|\widehat{\Psi}_{u'0}^{-1}\|_\infty \leqslant \varphi_n n^{-1/2}.$$

It follows that on the intersection of events $\mathcal{A}_n$ and (J.10), for n large enough, we have

$$\sup_{u\in\mathcal{U}^\epsilon, u'\in\mathcal{U}, d_{\mathcal{U}}(u,u')\leqslant\epsilon} \|\widehat{\Psi}_{u0}^{-1}\mathbb{E}_n[S_u]-\widehat{\Psi}_{u'0}^{-1}\mathbb{E}_n[S_{u'}]\|_\infty \leqslant \frac{c'-c}{c}\cdot\Phi^{-1}\left(1-\frac{\gamma}{2pN_n}\right),$$

where we again used standard tail bounds for the tails of Gaussian random variables. The asserted claim now follows by recalling the inequality (J.8) and noting that $\mathrm{P}_P(\mathcal{A}_n)\geqslant 1-\Delta_n$ and that (J.10) holds with probability at least $1-\gamma-o(\gamma)$. ■

Proof of Lemma I.5. For $j\in[p]$, let

$$\mathcal{F}_j = \Big\{(Y,X)\mapsto Y_uX_j\colon u\in\mathcal{U}\Big\}, \qquad \mathcal{F}_j' = \Big\{(Y,X)\mapsto X_j\mathrm{E}_P[Y_u\mid X]\colon u\in\mathcal{U}\Big\},$$
$$\mathcal{G}_j = \Big\{(Y,X)\mapsto X_j^2\zeta_u^2\colon u\in\mathcal{U}\Big\}$$

where $\zeta_u = Y_u - \mathrm{E}_P[Y_u\mid X]$. Note that the function $F(Y,X)=\|X\|_\infty$ is an envelope both for $\mathcal{F}_j$ and for $\mathcal{F}_j'$ for all $j\in[\tilde p]$. By assumption, F can be chosen to satisfy $\|F\|_{P,q}\leqslant K_n$.

Because $\mathcal{F}_j$ is a product of a VC-subgraph class of functions with index bounded by C_Y and a single function, Lemma K.1(1) implies that its uniform entropy numbers obey

$$\log N(\epsilon\|F\|_{Q,2},\mathcal{F}_j,\|\cdot\|_{Q,2}) \lesssim \log(e/\epsilon),\quad 0<\epsilon\leqslant 1. \tag{J.15}$$

Also, Lemma K.2 implies that the uniform entropy numbers of $\mathcal{F}_j'$ obey

$$\log\sup_Q N(\epsilon\|F\|_{Q,2},\mathcal{F}_j',\|\cdot\|_{Q,2}) \leqslant \log\sup_Q N\Big(\frac{\epsilon}{2}\|F\|_{Q,2},\mathcal{F}_j,\|\cdot\|_{Q,2}\Big),\quad 0<\epsilon\leqslant 1. \tag{J.16}$$

Further, since $\mathcal{G}_j\subset(\mathcal{F}_j-\mathcal{F}_j')^2$, $G=4F^2$ is an envelope for $\mathcal{G}_j$, and the uniform entropy numbers of $\mathcal{G}_j$ obey for all $\epsilon\in(0,1]$,

$$\begin{aligned}\log N(\epsilon\|G\|_{Q,2},\mathcal{G}_j,\|\cdot\|_{Q,2}) &\leqslant 2\log N\Big(\frac{\epsilon}{2}\|2F\|_{Q,2},\mathcal{F}_j-\mathcal{F}_j',\|\cdot\|_{Q,2}\Big) \\ &\leqslant 2\log N\Big(\frac{\epsilon}{4}\|F\|_{Q,2},\mathcal{F}_j,\|\cdot\|_{Q,2}\Big)+2\log N\Big(\frac{\epsilon}{4}\|F\|_{Q,2},\mathcal{F}_j',\|\cdot\|_{Q,2}\Big) \\ &\leqslant 4\log\sup_Q N\Big(\frac{\epsilon}{8}\|F\|_{Q,2},\mathcal{F}_j,\|\cdot\|_{Q,2}\Big),\end{aligned} \tag{J.17}$$

where the first and the second lines follow from Lemma K.1(2), and the third from (J.16). Hence, Lemma K.1(2) implies that the uniform entropy numbers of $\mathcal{G}=\cup_{j\in[p]}\mathcal{G}_j$ obey

$$\log N(\epsilon\|G\|_{Q,2},\mathcal{G},\|\cdot\|_{Q,2}) \lesssim \log(p/\epsilon),\quad 0<\epsilon\leqslant 1,$$

where $G = 4F^2$ is its envelope. Therefore, since $|S_{uj}| \leqslant 2|X_j|$ and $\max_{j\leqslant p} \mathrm{E}_P[X_j^4] \leqslant \bar{C}$ by assumption, Lemma L.2 implies that with probability at least $1 - (\log n)^{-1}$,

$$\sup_{u\in\mathcal{U}} \max_{j\leqslant p} |(\mathbb{E}_n - \mathrm{E}_P)[S_{uj}^2]| \lesssim \sqrt{\frac{\log(npK_n)}{n}} + \frac{n^{2/q}K_n^2}{n}\log(npK_n),$$

which gives (I.11).

To verify (I.10), note that

$$\begin{aligned}\sup_{d_{\mathcal{U}}(u,u')\leqslant 1/n} \max_{j\in[p]} \mathrm{E}_P[X_j^2(\zeta_u - \zeta_{u'})^2] &\leqslant \sup_{d_{\mathcal{U}}(u,u')\leqslant 1/n} \max_{j\in[p]} \mathrm{E}_P[X_j^2(Y_u - Y_{u'})^2] \\ &\leqslant \sup_{d_{\mathcal{U}}(u,u')\leqslant 1/n} \max_{j\leqslant p} \{\mathrm{E}_P[X_j^4]\}^{1/2}\{\mathrm{E}_P[(Y_u - Y_{u'})^4]\}^{1/2} \\ &\lesssim \sup_{d_{\mathcal{U}}(u,u')\leqslant 1/n} |u - u'|^{\nu/2} \lesssim n^{-\nu/2}.\end{aligned}$$

Therefore, Lemma L.2 implies that we have with probability at least $1 - (\log n)^{-1}$,

$$\begin{aligned}\sup_{d_{\mathcal{U}}(u,u')\leqslant 1/n} \|\mathbb{E}_n[S_u - S_{u'}]\|_\infty &= \frac{1}{\sqrt{n}} \sup_{d_{\mathcal{U}}(u,u')\leqslant 1/n} \max_{j\in p} |\mathbb{G}_n(X_j(\zeta_u - \zeta_{u'}))| \\ &\lesssim \sqrt{\frac{\log(npK_n)}{n^{1+\nu/2}}} + \frac{n^{1/q}K_n\log(npK_n)}{n},\end{aligned}$$

which gives (I.10).

Finally, to verify (I.12) note that uniformly over $u, u' \in \mathcal{U}$ and $j \in [p]$, we have

$$\begin{aligned}|\mathrm{E}_P[S_{uj}^2 - S_{u'j}^2]| &= |\mathrm{E}_P[(S_{uj} - S_{u'j})(S_{uj} + S_{u'j})]| \\ &\lesssim \Big(\mathrm{E}_P[(S_{uj} - S_{u'j})^2]\Big)^{1/2} \cdot \Big(\mathrm{E}[S_{uj}^2] + \mathrm{E}_P[S_{u'j}^2]\Big)^{1/2} \\ &\lesssim \Big(\mathrm{E}_P[X_j^2(Y_u - Y_{u'})^2]\Big)^{1/2} \lesssim \Big(\mathrm{E}_P[X_j^4]\Big)^{1/4} \cdot \Big(\mathrm{E}_P[(Y_u - Y_{u'})^4]\Big)^{1/4} \lesssim d_{\mathcal{U}}(u,u')^{\nu/4}.\end{aligned}$$

This completes the proof. ■

Appendix K. Bounds on Covering entropy

Let $(W_i)_{i=1}^n$ be a sequence of independent copies of a random element W taking values in a measurable space $(\mathcal{W}, \mathcal{A}_{\mathcal{W}})$ according to a probability law P. Let $\mathcal{F}$ be a set of suitably measurable functions $f\colon \mathcal{W} \to \mathbb{R}$, equipped with a measurable envelope $F\colon \mathcal{W} \to \mathbb{R}$.

Lemma K.1 (Algebra for Covering Entropies). *Work with the setup above.*
(1) Let $\mathcal{F}$ be a VC subgraph class with a finite VC index k or any other class whose entropy is bounded above by that of such a VC subgraph class, then the uniform entropy numbers of $\mathcal{F}$ obey

$$\sup_Q \log N(\epsilon\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}) \lesssim 1 + k\log(1/\epsilon) \vee 0$$

(2) For any measurable classes of functions $\mathcal{F}$ and $\mathcal{F}'$ mapping $\mathcal{W}$ to $\mathbb{R}$,

$$\log N(\epsilon\|F+F'\|_{Q,2},\mathcal{F}+\mathcal{F}',\|\cdot\|_{Q,2}) \leqslant \log N\left(\tfrac{\epsilon}{2}\|F\|_{Q,2},\mathcal{F},\|\cdot\|_{Q,2}\right)+\log N\left(\tfrac{\epsilon}{2}\|F'\|_{Q,2},\mathcal{F}',\|\cdot\|_{Q,2}\right),$$

$$\log N(\epsilon\|F\cdot F'\|_{Q,2},\mathcal{F}\cdot\mathcal{F}',\|\cdot\|_{Q,2}) \leqslant \log N\left(\tfrac{\epsilon}{2}\|F\|_{Q,2},\mathcal{F},\|\cdot\|_{Q,2}\right)+\log N\left(\tfrac{\epsilon}{2}\|F'\|_{Q,2},\mathcal{F}',\|\cdot\|_{Q,2}\right),$$

$$N(\epsilon\|F\vee F'\|_{Q,2},\mathcal{F}\cup\mathcal{F}',\|\cdot\|_{Q,2}) \leqslant N\left(\epsilon\|F\|_{Q,2},\mathcal{F},\|\cdot\|_{Q,2}\right)+N\left(\epsilon\|F'\|_{Q,2},\mathcal{F}',\|\cdot\|_{Q,2}\right).$$

(3) For any measurable class of functions $\mathcal{F}$ and a fixed function f mapping $\mathcal{W}$ to $\mathbb{R}$,

$$\log\sup_Q N(\epsilon\||f|\cdot F\|_{Q,2}, f\cdot\mathcal{F},\|\cdot\|_{Q,2}) \leqslant \log\sup_Q N\left(\epsilon/2\|F\|_{Q,2},\mathcal{F},\|\cdot\|_{Q,2}\right)$$

(4) Given measurable classes $\mathcal{F}_j$ and envelopes F_j, $j=1,\dots,k$, mapping $\mathcal{W}$ to $\mathbb{R}$, a function $\phi\colon\mathbb{R}^k\to\mathbb{R}$ such that for $f_j,g_j\in\mathcal{F}_j$, $|\phi(f_1,\dots,f_k)-\phi(g_1,\dots,g_k)| \leqslant \sum_{j=1}^k L_j(x)|f_j(x)-g_j(x)|$, $L_j(x)\geqslant 0$, and fixed functions $\bar f_j\in\mathcal{F}_j$, the class of functions $\mathcal{L}=\{\phi(f_1,\dots,f_k)-\phi(\bar f_1,\dots,\bar f_k)\colon f_j\in\mathcal{F}_j, j=1,\dots,k\}$ satisfies

$$\log\sup_Q N\left(\epsilon\Big\|\sum_{j=1}^k L_jF_j\Big\|_{Q,2},\mathcal{L},\|\cdot\|_{Q,2}\right) \leqslant \sum_{j=1}^k \log\sup_Q N\left(\tfrac{\epsilon}{k}\|F_j\|_{Q,2},\mathcal{F}_j,\|\cdot\|_{Q,2}\right).$$

Proof. See Lemma L.1 in [11]. ∎

Lemma K.2 (Covering Entropy for Classes obtained as Conditional Expectations)**.** *Let $\mathcal{F}$ denote a class of measurable functions $f\colon\mathcal{W}\times\mathcal{Y}\to\mathbb{R}$ with a measurable envelope F. For a given $f\in\mathcal{F}$, let $\bar f\colon\mathcal{W}\to\mathbb{R}$ be the function $\bar f(w) := \int f(w,y)d\mu_w(y)$ where μ_w is a regular conditional probability distribution over $y\in\mathcal{Y}$ conditional on $w\in\mathcal{W}$. Set $\bar{\mathcal{F}}=\{\bar f\colon f\in\mathcal{F}\}$ and let $\bar F(w) := \int F(w,y)d\mu_w(y)$ be an envelope for $\bar{\mathcal{F}}$. Then, for $r,s\geqslant 1$,*

$$\log\sup_Q N(\epsilon\|\bar F\|_{Q,r},\bar{\mathcal{F}},\|\cdot\|_{Q,r}) \leqslant \log\sup_{\widetilde Q} N((\epsilon/4)^r\|F\|_{\widetilde Q,s},\mathcal{F},\|\cdot\|_{\widetilde Q,s}),$$

where Q belongs to the set of finitely-discrete probability measures over $\mathcal{W}$ such that $0<\|\bar F\|_{Q,r}<\infty$, and $\widetilde Q$ belongs to the set of finitely-discrete probability measures over $\mathcal{W}\times\mathcal{Y}$ such that $0<\|F\|_{\widetilde Q,s}<\infty$. In particular, for every $\epsilon>0$ and any $k\geqslant 1$,

$$\log\sup_Q N(\epsilon,\bar{\mathcal{F}},\|\cdot\|_{Q,k}) \leqslant \log\sup_{\widetilde Q} N(\epsilon/2,\mathcal{F},\|\cdot\|_{\widetilde Q,k}).$$

Proof. See Lemma L.2 in [11]. ∎

Lemma K.3. *Consider a mapping $\tilde u\mapsto\xi_{\tilde u}$ from $\widetilde{\mathcal{U}}=[0,1]^k$ into $\mathbb{R}^p$ and the class of functions $\mathcal{F}=\{x\mapsto\mathcal{M}(x'\xi_{\tilde u})\colon \tilde u\in\widetilde{\mathcal{U}}\}$ mapping $\mathbb{R}^p$ into $\mathbb{R}$ where $\mathcal{M}\colon\mathbb{R}\to\mathbb{R}$ is L-Lipschitz. Assume that $\|\xi_{\tilde u_2}-\xi_{\tilde u_1}\|_1\leqslant C\|\tilde u_2-\tilde u_1\|$ for all $\tilde u_1,\tilde u_2\in\widetilde{\mathcal{U}}$ for some constant $C>0$. Then, for any $M>0$ the uniform entropy numbers of $\mathcal{F}$ satisfy*

$$\sup_Q\log N(\epsilon\|F\|_{Q,2},\mathcal{F},\|\cdot\|_{Q,2}) \leqslant k\log(3LCMk/\varepsilon), \quad \text{for all } 0<\epsilon\leqslant 1,$$

where $F(x)=\sup_{\tilde u\in\widetilde{\mathcal{U}}}|\mathcal{M}(x'\xi_{\tilde u})|+M^{-1}\|x\|_\infty$, $x\in\mathbb{R}^p$, is its envelope.

Proof. Consider any $f_1, f_2 \in \mathcal{F}$. There exist $\tilde{u}_1, \tilde{u}_2 \in \widetilde{\mathcal{U}}$ such that $f_1(x) = \mathcal{M}(x'\xi_{\tilde{u}_1})$ and $f_2(x) = \mathcal{M}(x'\xi_{\tilde{u}_2})$ for all $x \in \mathbb{R}^p$. Therefore, since $\mathcal{M}$ is L-Lipschitz, we have

$$\begin{aligned}|\mathcal{M}(x'\xi_{\tilde{u}_2}) - \mathcal{M}(x'\xi_{\tilde{u}_1})| &\leqslant L\|x\|_\infty \|\xi_{u_2} - \xi_{u_1}\|_1 \leqslant L\|x\|_\infty C\|\tilde{u}_2 - \tilde{u}_1\| \\ &\leqslant LCM\|\tilde{u}_2 - \tilde{u}_1\|\{M^{-1}\|x\|_\infty + \sup_{\tilde{u}\in\widetilde{\mathcal{U}}} |\mathcal{M}(x'\xi_{\tilde{u}})|\} \\ &\leqslant LCM\|\tilde{u}_2 - \tilde{u}_1\| F(x)\end{aligned}$$

by definition of the envelope $F(x) = M^{-1}\|x\|_\infty + \sup_{\tilde{u}\in\widetilde{\mathcal{U}}} |\mathcal{M}(x'\xi_{\tilde{u}})|$. Thus, for any finitely discrete probability measure Q on $\mathbb{R}^p$,

$$\|f_2 - f_1\|_{Q,2} \leqslant LCM\|\tilde{u}_2 - \tilde{u}_1\| \cdot \|F\|_{Q,2}.$$

Recall that since $B_\infty \subset B_2\sqrt{k}$ we have $N(B_\infty, \|\cdot\|, \epsilon) \leqslant N(B_2\sqrt{k}, \|\cdot\|, \epsilon) \leqslant (1 + 2\sqrt{k}/\epsilon)^k$ where the last inequality follows from standard volume arguments. Furthermore, for any $\epsilon \leqslant \sqrt{k}$ we have $1 + 2\sqrt{k}/\epsilon \leqslant 3k/\epsilon$. Therefore $\log N(\epsilon\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}) \leqslant k \log(3LCMk/\epsilon)$. ■

Lemma K.4. *Let $\mathcal{F}$ be a class of functions with an envelope F. Also, let $\mathcal{M}\colon \mathbb{R} \to \mathbb{R}$ be an L-Lipschitz function bounded in absolute value by a constant M. Assume that for some positive constants C_1 and C_2, the uniform entropy numbers of $\mathcal{F}$ obey*

$$\sup_Q \log N(\epsilon\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}) \leqslant C_1 \log(C_2/\epsilon), \quad \textit{for all } 0 < \epsilon \leqslant 1.$$

Then for any constant $K > 0$, the uniform entropy numbers of the class of functions $\mathcal{M}(\mathcal{F}) = \{\mathcal{M}(f)\colon f \in \mathcal{F}\}$ obey

$$\sup_Q \log N(\epsilon\|F_{\mathcal{M}(\mathcal{F})}\|_{Q,2}, \mathcal{M}(\mathcal{F}), \|\cdot\|_{Q,2}) \leqslant C_1 \log(C_2K/\epsilon), \quad \textit{for all } 0 < \epsilon \leqslant 1,$$

where $F_{\mathcal{M}(\mathcal{F})} = M + LF/K$ is its envelope.

Proof. The result follows from the observation that for any $f', f'' \in \mathcal{F}$ and any finitely-discrete probability measure Q,

$$\|\mathcal{M}(f') - \mathcal{M}(f'')\|_{Q,2} \leqslant L\|f_1 - f_2\|_{Q,2},$$

so that if $\mathcal{F}$ can be covered by k balls of radius $\epsilon\|F\|_{Q,2}$ (in the $\|\cdot\|_{Q,2}$ norm), then $\mathcal{M}(\mathcal{F})$ can be covered by k balls of radius $\epsilon L\|F\|_{Q,2}$ (in the same norm). ■

Appendix L. Some Probabilistic Inequalities

Lemma L.1 (Moderate deviations for self-normalized sums, [26])**.** *Let $Z_1, \ldots, Z_n$ be independent, zero-mean random variables and $\mu \in (0, 1]$. Let $S_{n,n} = \sum_{i=1}^n Z_i$, $V_{n,n}^2 = \sum_{i=1}^n Z_i^2$,*

$$M_n = \left\{\frac{1}{n}\sum_{i=1}^n \mathrm{E}[Z_i^2]\right\}^{1/2} \Big/ \left\{\frac{1}{n}\sum_{i=1}^n \mathrm{E}[|Z_i|^{2+\mu}]\right\}^{1/\{2+\mu\}} > 0$$

and $0 < \ell_n \leqslant n^{\frac{\mu}{2(2+\mu)}} M_n$. *Then for some absolute constant* A,

$$\left| \frac{\mathrm{P}(|S_{n,n}/V_{n,n}| \geqslant x)}{2(1-\Phi(x))} - 1 \right| \leqslant \frac{A}{\ell_n^{2+\mu}}, \quad 0 \leqslant x \leqslant n^{\frac{\mu}{2(2+\mu)}} \frac{M_n}{\ell_n} - 1.$$

Let $(W_i)_{i=1}^n$ be a sequence of independent copies of a random element W taking values in a measurable space $(\mathcal{W}, \mathcal{A}_{\mathcal{W}})$ according to a probability law P. Let $\mathcal{F}$ be a set of suitably measurable functions $f\colon \mathcal{W} \to \mathbb{R}$, equipped with a measurable envelope $F\colon \mathcal{W} \to \mathbb{R}$.

Lemma L.2 (Maximal Inequality I, [18])**.** *Work with the setup above. Suppose that* $F \geqslant \sup_{f\in\mathcal{F}} |f|$ *is a measurable envelope for* $\mathcal{F}$ *with* $\|F\|_{P,q} < \infty$ *for some* $q \geqslant 2$. *Let* $M = \max_{i\leqslant n} F(W_i)$ *and* $\sigma^2 > 0$ *be any positive constant such that* $\sup_{f\in\mathcal{F}} \|f\|^2_{P,2} \leqslant \sigma^2 \leqslant \|F\|^2_{P,2}$. *Suppose that there exist constants* $a \geqslant e$ *and* $v \geqslant 1$ *such that*

$$\log \sup_Q N(\epsilon\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}) \leqslant v \log(a/\epsilon),\ 0 < \epsilon \leqslant 1.$$

Then

$$\mathrm{E}_P[\|\mathbb{G}_n\|_{\mathcal{F}}] \leqslant K \left(\sqrt{v\sigma^2 \log\left(\frac{a\|F\|_{P,2}}{\sigma}\right)} + \frac{v\|M\|_{P,2}}{\sqrt{n}} \log\left(\frac{a\|F\|_{P,2}}{\sigma}\right) \right),$$

where K *is an absolute constant. Moreover, for every* $t \geqslant 1$, *with probability* $> 1 - t^{-q/2}$,

$$\|\mathbb{G}_n\|_{\mathcal{F}} \leqslant (1+\alpha)\mathrm{E}_P[\|\mathbb{G}_n\|_{\mathcal{F}}] + K(q)\Big[(\sigma + n^{-1/2}\|M\|_{P,q})\sqrt{t} + \alpha^{-1} n^{-1/2}\|M\|_{P,2} t\Big],\ \forall \alpha > 0,$$

where $K(q) > 0$ *is a constant depending only on* q. *In particular, setting* $a \geqslant n$ *and* $t = \log n$, *with probability* $> 1 - c(\log n)^{-1}$,

$$\|\mathbb{G}_n\|_{\mathcal{F}} \leqslant K(q,c) \left(\sigma \sqrt{v \log\left(\frac{a\|F\|_{P,2}}{\sigma}\right)} + \frac{v\|M\|_{P,q}}{\sqrt{n}} \log\left(\frac{a\|F\|_{P,2}}{\sigma}\right) \right), \tag{L.1}$$

where $\|M\|_{P,q} \leqslant n^{1/q}\|F\|_{P,q}$ *and* $K(q,c) > 0$ *is a constant depending only on* q *and* c.

Lemma L.3 (Maximal Inequality II, [18])**.** *Work with the setup above. Suppose that the conditions of Lemma L.2 are satisfied. Then*

$$\mathrm{E}_P\Big[\|\mathbb{E}_n[f^2(W)]\|_{\mathcal{F}}\Big] - \sup_{f\in\mathcal{F}} \mathrm{E}_P[f^2(W)] \leqslant \frac{K\|M\|_{P,2}}{\sqrt{n}} \left(\sigma \sqrt{v \log\left(\frac{a\|F\|_{P,2}}{\sigma}\right)} + \frac{v\|M\|_{P,2} \log\left(\frac{a\|F\|_{P,2}}{\sigma}\right)}{\sqrt{n}} \right)$$

where K *is an absolute constant.*

Proof. The proof of the asserted claim coincides one-by-one with that given for the corresponding inequality in Lemma 2.2 of [18], with the constant 3 replaced everywhere by the constant 2. At the end of the proof, the entropy integral

$$J(\delta) = \int_0^\delta \sup_Q \sqrt{1 + \log N(\epsilon\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2})} d\epsilon$$

is bounded by $\delta(v\log(a/\delta))^{1/2}$ under our condition on the uniform entropy numbers of $\mathcal{F}$. ∎