MANY AVERAGE PARTIAL EFFECTS: WITH AN APPLICATION TO TEXT REGRESSION

HAROLD D. CHIANG

Abstract. We study estimation, pointwise and simultaneous inference, and confidence intervals for many average partial effects of lasso Logit. Focusing on high-dimensional cluster-sampling environments, we propose a new average partial effect estimator and explore its asymptotic properties. Practical penalty choices compatible with our asymptotic theory are also provided. The proposed estimator allow for valid inference without requiring oracle property. We provide easy-to-implement algorithms for cluster-robust high-dimensional hypothesis testing and construction of simultaneously valid confidence intervals using a multiplier cluster bootstrap. We apply the proposed algorithms to the text regression model of Wu (2018) to examine the presence of gendered language on the internet.

1. Introduction

Binary response models are some of the most commonly used nonlinear econometric models. When studying such models, the average partial effect, henceforth APE, is a popular target parameter of interest. Under big data environments, as often happens in text analysis, dimension reduction via lasso, or other type of machine learning algorithms, is often unavoidable. Failure to account for the model selection step often leads to severely biased estimates, which invalidate the usual inference procedures (see Figure 1 for an illustration). Few results are available for valid post-selection inference for a single nonlinear functional of high-dimensional nuisance parameters, such as APE, let alone simultaneous inference for potentially many of such parameters. To fill this void, this paper considers simultaneous

Key words and phrases. average partial effect, post-selection inference, machine learning, text analysis, text regression, cluster-robust inference, lasso Logit.

Job market paper. I am indebted to Alex Belloni, Otilia Boldea, Irene Botosaru, Mehmet Caner, Atsushi Inoue, Hugo Jales, Kengo Kato, Frank Kleibergen, Tong Li, Pedro Sant'Anna, Yuya Sasaki, Takuya Ura, Matt Webb, and the participants at AMES 2019, Bristol Econometrics Study Group 2019, Canadian Econometric Study Group 2019, ESEM 2019, International PhD Conference at Erasmus University Rotterdam, NY Camp Econometrics XIV, TER 2019 for their insightful comments and discussions. I also benefited from the programming discussions with Terry Kim and Bin Yang Tan.

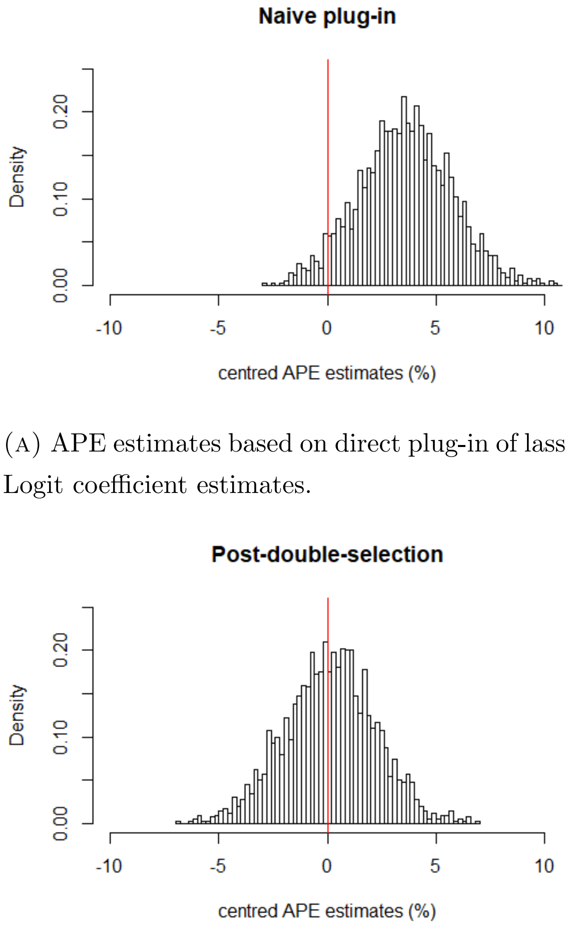
inference and confidence intervals for lasso Logit APEs. All results stay valid for cluster-sampled data.

To our knowledge, this is the first paper handling multiple testing and simultaneous confidence interval problems for more than a single APE under high-dimensional or big-data environments. In addition, cluster sampling with heterogeneous cluster sizes is allowed. Using the Neyman orthogonalization technique, we propose a new lasso-based post-double selection APE estimator. To accompany the main theoretical results, we propose valid nuisance parameter estimators as well as their practical tuning parameter selection algorithms that are compatible with our theory. To address the multiple-testing problem, we develop a new, simple-to-implement, multiplier cluster bootstrap. We provide simple algorithms for testing high-dimensional hypotheses and constructing simultaneously valid confidence intervals. Simulation studies suggest the proposed methods have favorable finite-sample performance. We illustrate the applicability of our theoretical results through examining a claim of Wu (2018) on the presence of genderally biased use of language following Wu's text regression model using internet forum textual data from Economics Job Market Rumors (EJMR) forum - see the following section.

2. Motivation: Text Analysis and Gendered Language on the Internet

Text analysis using machine learning algorithms has become a useful alternative to the more traditional data analysis used in economics and other social sciences. Popular categories of text analysis models include text regression models, generative models, dictionary-based methods and word embeddings. The first two categories link attributes and word counts through conditional probabilities[1] and, therefore, naturally relate to common econometric models. Notable examples of applications using text regression include stock prices prediction (e.g. Jegadeesh and Wu (2013)) and the Google Flu Trends, which is summarized in Ginsberg, Mohebbi, Patel, Brammer, Smolinski and Brilliant (2009), among others. Gentzkow, Shapiro and Taddy (2019) is a representative recent example for generative models applied to economics. For more details and applications, see Gentzkow, Kelly and Taddy (2019) for an up-to-date review.

[1] Roughly speaking, given attributes v_i and word counts c_i, a text regression model considers $\mathrm{P}(v_i|c_i)$ and a generative model considers $\mathrm{P}(c_i|v_i)$.



(A) APE estimates based on direct plug-in of lasso Logit coefficient estimates.

(B) APE estimates based on the proposed post-double-selection estimator.

FIGURE 1. Simulations for low-dimensional lasso Logit APE estimation based on $2,000$ iterations. Each iteration has sample size $n = 200$. The dimensionality of covariates is set to be $p = 10$. We set true parameter vector as $\beta^0 = [.1, -1, 1, 0, ..., 0]$. Covariates X are generated as i.i.d. zero-mean multivariate normal random vectors with Toeplitz covariance matrix Σ with $\Sigma_{ij} = 0.5^{|i-j|}$. Outcome variables are generated following $Y = \mathbb{1}\{X'\beta^0 + U\}$ with i.i.d. U following standard logistic distribution. The lasso estimations are implemented using R package **glmnet** with penalty selection algorithms discussed in Section 6.

Table 1. Top 10 most predictive words for female/male from Wu (2018)

Female		*Male*	
Word	APE	Word	APE
Pregnancy	0.292	Knocking	−0.329
Hotter	0.289	Testosterone	−0.204
Pregnant	0.258	Blog	−0.183
Hp	0.238	Hateukbro	−0.176
Vagina	0.228	Adviser	−0.175
Breast	0.220	Hero	−0.174
Plow	0.219	Cuny	−0.173
Shopping	0.207	Handsome	−0.166
Marry	0.207	Mod	−0.166
Gorgeous	0.201	Homo	−0.160

(pronoun sample; a replication of Table 2 in Wu (2018))

Using a text regression model, Wu (2018) examines how women and men are discussed and depicted in the anonymous Economics Job Market Rumors forum. The author first extracted a list of female/male classifier vocabularies. According to Wu, a post is considered to be female if it contains any female classifier and male if it contains any male classifier[2]. Let $Female_i$ be an indicator of whether post i is female. X_i denotes a vector of counts for each of the top 10,000 most common words[3] (excluding all gender classifiers) that are present in gendered post i. Wu considers the text regression model with the logistic[4] specification,

$$\mathrm{P}(Female_i|X_i) = \Lambda(X_i'\beta)$$

where Λ is the logistic function, using a lasso Logit procedure. The $Male$ counterpart is estimated analogously. For interpretability, Wu computes estimates for the APE for each of the $9,540$ words, where the APE for the word count of the k-th word is defined as

$$\mathrm{APE}_k = \mathrm{E}[\beta_k \Lambda'(X_i'\beta)].$$

[2] Wu makes use of a classification procedure to decide the posts that contains both female and male classifiers. See Section II A of Wu (2018) for more details

[3] It is also possible to use frequency and n-grams in place of word count and words, respectively, as suggested in Gentzkow, Kelly and Taddy (2019).

[4] For text regression models with binary attributes, a penalized logistic model is recommended by Gentzkow, Kelly and Taddy (2019); see their Section 3.1.1 for more details.

Based on these estimates, Wu concludes the words that predict a post about a woman are typically about physical appearance or personal information, whereas those most predictive of a post about a man tend to focus on academic or professional characteristics.

Wu (2018) focuses on estimation. To further investigate magnitude and statistical significance of these estimates, the researcher may be interested in conducting hypotheses testing or constructing confidence intervals. To do so, several issues need to be carefully accounted for. First, as posts in EJMR data of Wu (2018) are sampled from different threads of various discussion topics, it is likely that posts coming from the same thread are highly correlated. Therefore, statistical testing should be conducted using a cluster robust inference method. Secondly, Wu (2018) highlights that females are often described with words about appearance or personal information. To formally examine such statements, one may want to conduct multiple testing for APEs of a (potentially large) set of vocabularies related to appearance or personal information. Furthermore, in many cases, words with the same or close meaning are double-counted in this data set, e.g. "attractive" and "attractiveness" or "homo", "homosexual," and "gay."[5] Thus, the researcher may want to consider a joint test that controls family-wise error rates for APEs of these words. This results in a multiple testing problem. Therefore, the testing procedure needs to be able to control the family-wise error rate while testing potentially many variables. To our best knowledge, no method in the literature is capable of addressing all these issues simultaneously. This paper attempts to provide a useful and easy-to-implement method that can be applied to such problems.

3. Background and Literature Review

3.1. **Contributions.** Our main contribution is to provide a theory for high-dimensional multiple-testing and simultaneous confidence intervals for APEs of binomial and fractional response regression models under clustered data. To our best knowledge, no results were previously available for this purpose. As a by-product, this paper also complements existing papers by proposing a practical method for studying low-dimensional APEs of interest under high-dimensional settings. Furthermore, cluster sizes are allowed to be heterogeneous - this is essential to our application as number of posts varies from thread to thread. Inference and construction of confidence intervals for such models are practically challenging; despite that methods are proposed in the literature, no simulation evidence for inference of even a single APE under lasso-regularization with these methods is available. In addition, we present

[5] If the researcher is only concerned about joint testing, an easy alternative is to combine these words. However, this is not desirable when one wants to obtain separate estimates.

practical and theoretically justified penalty choices for all the lasso estimators. Furthermore, easy-to-implement bootstrap procedures are also provided for inference/confidence intervals that hold valid, regardless of whether the researcher is interested in one or multiple APEs.

3.2. **Relations to the Literature.** The past decade has seen an explosive development in the literature of post-selection inference for lasso-based high-dimensional methods. This includes Belloni, Chernozhukov, Chen and Hansen (2012) for instrumental variable models, Belloni, Chernozhukov and Hansen (2014), Javanmard and Montanari (2014), Zhang and Zhang (2014), Farrell (2015), Caner and Kock (2018) and Athey, Imbens and Wager (2018) for linear regression/treatment effects models. Post-selection inference for generalized linear models such as Logit has been studied by van de Geer, Bühlmann, Ritov and Dezeure (2014), Belloni, Chernozhukov and Kato (2015), Belloni, Chernozhukov and Wei (2016), Belloni, Chernozhukov, Fernández-Val and Hansen (2017) and Belloni, Chernozhukov, Chetverikov and Wei (2018), to list a few. This line of research predominately focuses on regression coefficients of the generalized linear models rather than nonlinear functionals such as an APE. Recently, Chernozhukov, Newey and Singh (2018) study L_2-continuous functionals using lasso and Dantzig selector. While focusing on affine-functionals, they provide an extension of their method to nonlinear functionals. Their method makes use of a linear Riesz representer to approximate the linearization of a nonlinear functional, which differs from our approach. In addition, all of the aforementioned papers are based on i.i.d. or independent sampling assumptions. On the other hand, there are some results available for high-dimensional linear panel data. This includes Belloni, Chernozhukov, Hansen and Kozbur (2016), Kock (2016) and Kock and Tang (2018).

Cluster-robust inference under various fixed-dimensional parametric settings has been well-studied and widely applied in the literature. See Wooldridge (2010) and Cameron and Miller (2015) for textbook treatment and comprehensive reviews. There has been recent literary focus on cluster-robust bootstrap inference. This includes Kline and Santos (2012), Hagemann (2017), MacKinnon and Webb (2017) and Djogbenou, MacKinnon and Nielsen (2018), among others. These results cannot be generalized in a straightforward manner to high-dimensional settings as the delta-method does not, in general, hold in an asymptotic framework with increasing dimensionality, see Caner (2017) for more details.

APE for binomial/fractional regression models has been discussed extensively in the literature (cf Chamberlain (1984), Wooldridge (2005) and Wooldridge (2018), etc). Inference for APEs of lasso-based binomial regression models are first studied by Wooldridge and Zhu (2017) under a short (balanced) panel data setting. They make use of a single-selection step with a lasso Probit estimator and propose a de-biased estimator for a single APE and obtain asymptotic normality. More recently, Hirshberg and Wager (2018) highlight the estimator of Wooldridge and Zhu (2017) for its requirement of a “soft” beta-min assumption that rules out regularization bias asymptotically[6]. For i.i.d. data, Hirshberg and Wager (2018) provide an alternative augmented minimax estimator based on the novel framework for linear functionals developed in Hirshberg and Wager (2017). However, no variance estimator for this approach is proposed. Also, the aforementioned results are available only for a single APE; multiple testing and simultaneous confidence intervals for more than one APE remain unavailable. In addition, implementing inference for even a single APE under such settings presents practical challenges; to our best knowledge, there has been no simulation evidence presented for the proposed estimators in the aforementioned papers.

This paper aims to address all the aforementioned issues simultaneously. To do so, we extend the general framework for i.i.d. data developed in the important works of Belloni, Chernozhukov and Kato (2015) and Belloni, Chernozhukov, Chetverikov and Wei (2018) to allow for cluster sampling and adapt it to the studies of APEs. The pointwise/simultaneous inference and confidence intervals are based on a multiplier cluster bootstrap which is built upon the high-dimensional central limit theorem of Chernozhukov, Chetverikov and Kato (2013).

3.3. **Notations.** Denote $(\Omega, \mathcal{A})$ the underlying measurable space and for each $G \in \mathbb{N}$, $\mathcal{P}_G$ is a set of probability measures $\mathrm{P} \in \mathcal{P}_G$ defined on $\mathcal{A}$. Consider triangular array data $\{W_g^G : g = 1, ..., G, G = 1, 2, 3, ...\}$ defined on probability space $(\Omega, \mathcal{A}, \mathrm{P})$, where P depends on G through $\mathcal{P}_G$. Each $W_g^G = \{W_{ig}^G : 1 \leq i \leq n_g\}$, is a random vector that is independent across g, but not necessarily identically distributed. All parameters that characterize the distribution of $\{W_g^G; g = 1, ..., G\}$ are implicitly indexed by P_G and thus by G. This dependence is henceforth omitted for simplicity. $W_{ig} = (Y_{ig}, X_{ig}')'$ takes values in $\mathbb{R}^{p+1}$. For each $g \leq G$, $G \in \mathbb{N}$, the deterministic size of cluster n_g satisfies $1 \leq n_g \leq \bar{n}$ for a constant $\bar{n}$ independent of G. Therefore, for i such that $n_g < i \leq \bar{n}$, we can set $W_{ig} = 0$ and thus each W_g can be represented as a $\bar{n}(p+1)$-dimensional random vector. Let E_{P} be the expectation with respect to law P.

For a vector β, the k-th component is denoted as β_k. For vectors, denote the ℓ_1-norm as $\|\cdot\|_1$, l_2-norm as $\|\cdot\|$, ℓ_∞-norm as $\|\cdot\|_\infty$, and the “ℓ_0-norm” as $\|\cdot\|_0$ to denote the number

[6] Such post model selection inference issues are widely discussed in the literature (see e.g. Pötscher and Leeb (2009) and the reference within).

of non-zero components. For a matrix A, let A' be the transpose of A. For $1 \leq q < \infty$, $\|A\|_q$ denotes the induced l_q-norm and $\|A\|_\infty = \max_{1\leq j,k\leq p} |A_{j,k}|$. For a vector $\delta \in \mathbb{R}^p$ and given data, $\|X'_{ig}\delta\|_G = \sqrt{\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(X'_{ig}\delta)^2}$ denotes the prediction norm of δ. Let e_j be the j-th vector of the standard basis for $\mathbb{R}^p$. Given a vector $\delta \in \mathbb{R}^p$, and a set of indices $T \subseteq \{1, \ldots, p\}$, denote $\delta_T \in \mathbb{R}^p$ the vector such that $(\delta_T)_j = \delta_j$ if $j \in T$ and $(\delta_T)_j = 0$ if $j \notin T$. The support of δ is defined as $\text{support}(\delta) = \{j \in \{1, ..., p\} : \delta_j \neq 0\}$. We denote $a \vee b = \max\{a, b\}$, and $a \wedge b = \min\{a, b\}$. The notaion $[k] = \{1, ..., k\}$ is used for $k \in \mathbb{N}$. We use c, C to denote strictly positive constants that is independent of G and $\mathrm{P} \in \mathcal{P}_G$. Their values may change at each presence. The notation $a_G \lesssim b_G$ denotes $a_G \leq Cb_G$ for all G and some $C > 0$ that does not depend on G. $a_G = o(1)$ means that there exists a sequence $(b_G)_{G\geq 1}$ of positive numbers that do not depend on $\mathrm{P} \in \mathcal{P}_G$ for all G such that $|a_G| \leq b_G$ for all G and $b_G = o(1)$ as G converges to zero. $a_G \lesssim_P b_G$ means that for any $\epsilon > 0$, there exists C such that $\mathrm{P}_\mathrm{P}(a_G > Cb_G) \leq \epsilon$ for all G. Throughout the paper we assume $G \geq 3$. Let (T, d) be a pseudomaetric space. For any $\varepsilon > 0$, denote $N(T, d, \varepsilon)$ for the ε-covering number of T.

3.4. **Outline.** The rest of the paper is structured as follows. In Section 4, an overview of the method and algorithms are given. Section 5 contains the main asymptotic results. Section 6 covers algorithms for penalty choices and the auxiliary results for theoretical performance of nuisance parameters. Results of simulation studies are demonstrated in Section 7. In Section 8, we apply the proposed method to conduct simultaneous testing to verify a statement about gendered language in Wu (2018). We concludes in Section 9. All the mathematical proofs and additional details are delegated to the appendix.

4. An Overview

Recall $W_{ig} = (Y_{ig}, X'_{ig})'$. Suppose that the researcher observes data sampled from G clusters, $\{W_{ig} : i = 1, ..., n_g,\ g = 1, ..., G\}$. Each cluster size n_g is considered non-random, and $1 \leq n_g \leq \bar{n} < \infty$ for a constant $\bar{n}$ that does not depend on G. Denote $n = \sum_{g=1}^{G} n_g$. Throughout the paper, we assume that the conditional expectation of Y given X follows the following single-index structure

$$\mathrm{E}_\mathrm{P}(Y_{ig}|X_{ig}) = \Lambda(X'_{ig}\beta^0).$$

for each cluster g. Any $W_{i_1 g}, W_{i_2 g}$ can be arbitrarily correlated while any $W_{i_1 g_1}, W_{i_2 g_2}$ are independent if $g_1 \neq g_2$. The dimensionality of β^0 is allowed to increase with G. This is the population-averaged approach as β^0 represents an averaged parameter after integrating out heterogeneity. The target parameter is the APE with respect to the k-th continuous covariate of interest,

$$\mathrm{APE}_k = \mathrm{E}_{\mathrm{P}}\Big[\frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\beta_k^0\Lambda'(X_{ig}'\beta^0)\Big]$$

where Λ' stands for the derivative of Λ. As $n_g \leq \bar{n}$, it suffices to consider α_k, the rescaled APE[7] defined as

$$\alpha_k = \mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\beta_k^0\Lambda'(X_{ig}'\beta^0)\Big].$$

4.1. **Estimation and Inference Procedures.** We now summarize the estimation, inference and construction of simultaneous confidence intervals procedures based on the theoretical results to be presented in Section 5 and 6 ahead. First, we describe the procedures for computing the proposed APE estimators. Set α_k as the parameter of interest. The post-double-selection estimator for α_k is defined as

$$\widetilde{\alpha}_k = \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\check{\beta}_k^k\Lambda'(X_{ig}'\check{\beta}^k) \tag{4.1}$$

where $\check{\beta}^k$ is the pooled Logit estimate with its support restricted to the set of covariates

$$\widetilde{T}_k = \{k\} \cup \mathrm{support}(\widehat{\beta}) \cup \mathrm{support}(\widehat{\zeta}^k) \cup \mathrm{support}(\widehat{\gamma}^k), \tag{4.2}$$

and $\widehat{\beta}$, $\widehat{\zeta}^k$ and $\widehat{\gamma}^k$ are nuisance parameter estimators to be defined below. Therefore, once $\widetilde{T}_k$ is obtained, estimation of $\widetilde{\alpha}_k$ becomes a standard pooled Logit problem.

Suppose that we have some generic penalty tuning parameters λ, λ_k^γ and λ_k^ζ and, in addition, $\widehat{\Psi}$, $\widehat{\Psi}_k^\gamma$, $\widehat{\Psi}_k^\zeta$, diagonal normalization matrices of dimensions p, $p-1$ and p, respectively. Formal and theoretically justified choices of these objects are delayed to Section 6.

First, $\widehat{\beta}$ and its two post-lasso counterparts are defined as

$$\widehat{\beta} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\{-Y_{ig}X_{ig}'\beta + \log(1+\exp(X_{ig}'\beta))\} + \frac{\lambda}{G}\|\widehat{\Psi}\beta\|_1, \tag{4.3}$$

$$\widetilde{\beta} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p:\mathrm{support}(\beta)\subset\mathrm{support}(\widehat{\beta})} \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\{-Y_{ig}X_{ig}'\beta + \log(1+\exp(X_{ig}'\beta))\}, \tag{4.4}$$

$$\widetilde{\beta}^k \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p:\mathrm{support}(\beta)\subset\mathrm{support}(\widehat{\beta}_{-k})} \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\{-Y_{ig}X_{ig}'\beta + \log(1+\exp(X_{ig}'\beta))\}. \tag{4.5}$$

[7] The original APE can be simply recovered by $\mathrm{APE}_k = (G/n)\cdot\alpha_k$.

Using the above post-lasso estimates, we compute $\widehat{f}_{ig}^2 = \Lambda'(X_{ig}'\widetilde{\beta})$ and $\widehat{S}_{ig}^k = \widetilde{\beta}_k^k \cdot \{1 - 2\Lambda(X_{ig}'\widetilde{\beta})\}$. Throughout the rest of this paper, denote $D_{ig}^j = X_{ig,j}$, the j-th component of X_{ig}, and $X_{ig}^j = X_{ig,-j}'$, the remaining $p-1$ variables. Using these quantities, the remaining two nuisance parameter estimates can be obtained as

$$\widehat{\gamma}^k = \operatorname*{argmin}_{\gamma \in \mathbb{R}^{p-1}} \frac{1}{G} \sum_{g=1}^{G} \sum_{i=1}^{n_g} \widehat{f}_{ig}^2 (D_{ig}^k - X_{ig}^{k\prime}\gamma)^2 + 2\frac{\lambda_k^\gamma}{G} \|\widehat{\Psi}_k^\gamma \gamma\|_1, \tag{4.6}$$

$$\widehat{\zeta}^k = \operatorname*{argmin}_{\zeta \in \mathbb{R}^{p}} \frac{1}{G} \sum_{g=1}^{G} \sum_{i=1}^{n_g} \widehat{f}_{ig}^2 (\widehat{S}_{ig}^k - X_{ig}'\zeta)^2 + 2\frac{\lambda_k^\zeta}{G} \|\widehat{\Psi}_k^\zeta \zeta\|_1. \tag{4.7}$$

Now $\widetilde{T}_k$ can be calculated following (4.2) and thus

$$\check{\beta}^k = \operatorname*{argmin}_{\beta \in \mathbb{R}^p : \beta_j = 0 \text{ for all } j \in \widetilde{T}_k^c} \frac{1}{G} \sum_{g=1}^{G} \sum_{i=1}^{n_g} \{-Y_{ig}X_{ig}'\beta + \log(1 + \exp(X_{ig}'\beta))\}, \tag{4.8}$$

and $\widetilde{\alpha}_k$ can be obtained following equation (4.1).

Suppose that the researcher is interested in α_k for a set of continuous covariates with $k \in A$ for an index set $A \subset [p]$[8]. We present a concrete estimation procedure as the following algorithm.

Algorithm 4.1 (Post-Double-Selection Estimator)**.** *For each $k \in A$,*

(1) Run lasso and post-lasso Logit to compute $\widetilde{\beta}$ following (4.3) and (4.4).
(2) Define generated weights $\widehat{f}_{ig}^2 = \Lambda'(X_{ig}'\widetilde{\beta})$.
(3) Run lasso to compute $\widehat{\gamma}^k$ following (4.6).
(4) Run lasso to compute $\widehat{\zeta}^k$ following (4.7).
(5) Let $\widetilde{T}_k = \{k\} \cup \text{support}(\widehat{\beta}) \cup \text{support}(\widehat{\zeta}^k) \cup \text{support}(\widehat{\gamma}^k)$ and compute $\check{\beta}^k$ following (4.8).
(6) Compute plug-in estimator $\widetilde{\alpha}_k$ following (4.1).

Remark 4.1. The post-double-selection estimator is theoretically related to the post-double-selection estimators for linear models in Belloni, Chernozhukov and Hansen (2014) and for Logit regression coefficients in Belloni, Chernozhukov and Wei (2016) and Belloni, Chernozhukov, Chetverikov and Wei (2018). However, because our target parameters of interest are APEs, the nonlinear transformations of high-dimensional nuisance parameters, rather than regression coefficients themselves, the structure of our nuisance parameters are

[8] There is no restriction on the cardinality of A. $A = [p]$ is also allowed.

fundamentally different. Estimation of these nuisance parameters requires different strategies and therefore presents extra challenges. We discuss the theory of nuisance parameters estimation in Section 6.

For inference, let us define the post-lasso counterparts of $\widehat{\gamma}^k$ and $\widehat{\zeta}^k$

$$\widetilde{\gamma}^k = \underset{\text{support}(\gamma)\subset\text{support}(\widehat{\gamma}^k)}{\text{argmin}} \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} \widehat{f}_{ig}^2(D_{ig}^k - X_{ig}^k\gamma)^2, \tag{4.9}$$

$$\widetilde{\zeta}^k = \underset{\text{support}(\zeta)\subset\text{support}(\widehat{\zeta}^k)}{\text{argmin}} \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} \widehat{f}_{ig}^2(\widehat{S}_{ig}^k - X_{ig}'\zeta)^2, \tag{4.10}$$

and the nuisance parameter estimate

$$\widetilde{\theta}^k = [-\widetilde{\gamma}_1^k, ..., -\widetilde{\gamma}_{k-1}^k, 1, -\widetilde{\gamma}_k^k, ..., -\widetilde{\gamma}_{p-1}^k]' \cdot \Big\{\frac{1}{G\widehat{\tau}_k^2}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{f}_{ig}^2\Big\}, \tag{4.11}$$

where each $\widehat{\tau}_k^2$ is calculated using

$$\widehat{\tau}_k^2 := \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{f}_{ig}^2(D_{ig}^k - X_{ig}^k\widetilde{\gamma}^k)^2. \tag{4.12}$$

Define the additional nuisance parameter estimate

$$\widetilde{\mu}^k = \widetilde{\zeta}^k + \widetilde{\theta}^k. \tag{4.13}$$

Finally, define the variance estimate as

$$\widetilde{\sigma}_k^2 = \frac{1}{G}\sum_{g=1}^{G}\Big\{\sum_{i=1}^{n_g}\Big(\widetilde{\alpha}_k\Big(\frac{G}{n}\Big) - \widetilde{\beta}_k\Lambda'(X_{ig}'\widetilde{\beta}) + \widetilde{\mu}^{k\prime}X_{ig}\{Y_{ig} - \Lambda(X_{ig}'\widetilde{\beta})\}\Big)\Big\}^2. \tag{4.14}$$

We are now ready to introduce a procedure for simultaneous inference. Suppose that the null hypothesis of interest is

$$\mathrm{H}_0 : \alpha_k = \alpha_k^0 \text{ for all } k \in A$$

for some values $(\alpha_k^0)_{k\in A}$. We present a concrete simultaneous inference procedure as the following algorithm.

Algorithm 4.2 (Simultaneous Inference via Multiplier Cluster Bootstrap)**.** *For each $k \in A$,*

(1) Compute $\widetilde{\sigma}_k$ for $k \in A$ following (4.14).
(2) Compute the test statistic $T = \max_{k\in[p]} \sqrt{G}\widetilde{\sigma}_k^{-1}|\widetilde{\alpha}_k - \alpha_k^0|$.
(3) For each $k \in A$, compute $\widetilde{\mu}^k$ following (4.13).

(4) Set the number of bootstrap iterations to B. For each $b \in [B]$, generate i.i.d. standard normal random variables $\{\xi_g^b\}_{g=1}^G$ independently from data.

(5) For each $k \in A$ and $b \in [B]$, compute

$$W^b = \max_{k\in A} \Big| \frac{1}{\sqrt{G}\widetilde{\sigma}_k} \sum_{g=1}^{G} \xi_g^b \sum_{i=1}^{n_g} \Big(\widetilde{\alpha}_k \Big(\frac{G}{n}\Big) - \widetilde{\beta}_k^k \Lambda'(X_{ig}'\widetilde{\beta}) + \widetilde{\mu}^{k\prime} X_{ig}\{Y_{ig} - \Lambda(X_{ig}'\widetilde{\beta})\} \Big) \Big| \tag{4.15}$$

and c_a, the $(1-a)$-th quantile of $\{W^b\}_{b=1}^B$.

(6) If $T > c_a$, reject H_0. Otherwise do not reject H_0.

Finally, we illustrate the procedure for constructing simultaneously valid confidence intervals with $(1-a)$ coverage probability for α_k, $k \in A$.

Algorithm 4.3 (Simultaneous Confidence Intervals via Multiplier Cluster Bootstrap)**.** *For each $k \in A$,*

(1) Compute $\widetilde{\sigma}_k^2$ for $k \in A$ following (4.14).

(2) Set the number of bootstrap iterations to B. For each $b \in [B]$, generate i.i.d. standard normal random variables $\{\xi_g^b\}_{g=1}^G$ independently from data.

(3) For each $k \in A$ and $b \in [B]$, compute W^b following (4.15) and c_a, the $(1-a)$-th quantile of $\{W^b\}_{b=1}^B$.

(4) Compute simultaneous confidence intervals $I = \times_{k\in A} I_k$, where $I_k = \widetilde{\alpha}_k \pm \widetilde{\sigma}_k \cdot c_a/\sqrt{G}$.

Remark 4.2. Note that it is also possible to conduct multiple testing and simultaneous confidence intervals without normalization (studentization). To do so, one simply follows every step in Algorithms 4.2 and 4.3 with 1 in place of $\widehat{\sigma}_k$ for all k.

5. Main Theoretical Results

In this section, we present our main theoretical results for simultaneous inference and construction of confidence intervals. These results justify the validity of the algorithms proposed in Section 4. First, we introduce some notations. Recall

$$\mathrm{E_P}(Y_{ig}|X_g) = \mathrm{E_P}(Y_{ig}|X_{ig}) = \Lambda(X_{ig}'\beta^0).$$

Define the Neyman orthogonal score for α_k by

$$\begin{aligned} \bar{\psi}_k(W_{ig}, \alpha, \eta) =& \alpha \cdot \frac{G}{n} - \beta_k \Lambda'(X_{ig}'\beta) + \mu' X_{ig}\{Y_{ig} - \Lambda(X_{ig}'\beta)\} \qquad (5.16) \\ =& \alpha \cdot \frac{G}{n} - \psi_k(W_{ig}, \eta), \end{aligned}$$

where $\psi_k(W_{ig},\eta) = \beta_k\Lambda'(X'_{ig}\beta) - \mu' X_{ig}\{Y_{ig} - \Lambda(X'_{ig}\beta)\}$. In addition, let the "ideal" population nuisance parameters[9] for α_k be $\eta^k = (\beta^{0\prime}, \mu^{k\prime})' \in \mathbb{R}^{2p}$ with

$$\mu^k = \zeta^k + \theta^k, \tag{5.17}$$

$$\zeta^k = \Big\{\mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} f_{ig}^2 X_{ig}X'_{ig}\Big]\Big\}^{-1}\mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} f_{ig}^2 X_{ig}S_{ig}^k\Big], \tag{5.18}$$

$$\theta^k = \Big\{\mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} f_{ig}^2 X_{ig}X'_{ig}\Big]\Big\}^{-1}\mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} f_{ig}^2 e_k\Big], \tag{5.19}$$

where $S_{ig}^k = \beta_k^0(1 - 2\Lambda(X'_{ig}\beta^0))$ is an auxiliary regressor and $f_{ig}^2 = \Lambda'(X'_{ig}\beta^0)$ is a regression weight. Also denote the population nodewise regression coefficients for the j-th covariate as γ^j. We can also rewrite the population nuisance parameter regression coefficients ζ^j as a weighted projection of S_{ig}^j on X_{ig}. Thus, we have the following

$$\gamma^j = \underset{\gamma\in\mathbb{R}^{p-1}}{\mathrm{argmin}}\mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} f_{ig}^2(D_{ig}^j - X_{ig}^j\gamma)^2\Big], \tag{5.20}$$

$$\zeta^j = \underset{\gamma\in\mathbb{R}^{p}}{\mathrm{argmin}}\mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} f_{ig}^2(S_{ig}^j - X_{ig}\zeta)^2\Big]. \tag{5.21}$$

Denote the projection errors by $Z_{ig}^j = D_{ig}^j - X_{ig}^j\gamma^j$ and $\varepsilon_{ig}^j = S_{ig}^j - X_{ig}\zeta^j$. Let $q > 4$ be a constant independent of G. Let c_1 and C_1 be some strictly positive constants independent of G. Furthermore, let $a_G = p \vee G$ and $\check{\delta}_G$ be a sequence of positive constants that converge to zero. $M_{G,1} \geq 1$ and $M_{G,2} \geq 1$ be some sequence of positive constants possibly diverging to infinity. $s = s_G$ is a non-decreasing sequence of constants. We make the follow assumptions.

Assumption 1 (Parameters)**.**

$$\|\beta^0\|_2 + \max_{j\in[p]}\|\gamma^j\|_2 + \max_{k\in[p]}\|\zeta^k\|_2 \leq C_1.$$

Also, for all $k \in [p]$, H_k contains a ball of radius $(s\log a_G)/G^{1/2}$ centered at η^k.

Assumption 2 (Sparsity)**.** *There exist vectors $\bar{\gamma}^j \in \mathbb{R}^{p-1}$ and $\bar{\zeta}^k \in \mathbb{R}^p$ for all $j,k \in [p]$ such that*

$$\|\beta^0\|_0 + \max_{j\in[p]}\|\bar{\gamma}^j\|_0 + \max_{k\in[p]}\|\bar{\zeta}^k\|_0 \leq s$$

[9] See Section A in the Appendix for derivation of this moment condition.

and

$$\max_{j,k\in[p]}(\|\bar{\gamma}^j-\gamma^j\|_2\vee\|\bar{\theta}^k-\theta^k\|_2+s^{-1/2}\|\bar{\gamma}^j-\gamma^j\|_1\vee\|\bar{\zeta}^k-\zeta^k\|_1)\le C_1(s\log a_G/G)^{1/2}.$$

Remark 5.1. Assumption 1 requires bounded ℓ_2 norm of nuisance parameters, which is mild and standard in the lasso literature. The ℓ_1 norm of the nuisance parameters are allowed to be growing with G. Note that we do not require exact sparsity of γ^j and ζ^k in Assumption 2 since the exact sparsity of nodewise lasso coefficients could be more difficult to justify in many applications. Also, note that for each $j\in[p]$, we can without loss of generality assume $\bar{\gamma}^j=\gamma^j_T$, where $T=\text{support}(\gamma^j)$. The same applies to $\bar{\zeta}^k$ and ζ^k.

For the following assumption, define $U_{gk}=\bar{n}\cdot\max_{i\in[n_g]}|X_{ig,k}|$, $U_g=[U_{gk}]_{k\in[p]}$ and $V_g^j=\max_{i\in[n_g]}(|Z_{ig}^j|\vee|\varepsilon_{ig}^j|)$.

Assumption 3 (Covariates)**.** *At least one coordinate of X_{ig} is continuously distributed. There exist finite positive constants c_1, C_1 such that the following moment conditions hold for all G,*

(1) $\inf_{\|\xi\|_2=1}\mathrm{E}_{\mathrm{P}}[\frac{1}{G}\sum_{g=1}^G\sum_{i=1}^{n_g}(f_{ig}X_{ig}'\xi)^2]\wedge\inf_{\|\xi\|_2=1}\mathrm{E}_{\mathrm{P}}[\frac{1}{G}\sum_{g=1}^G(\sum_{i=1}^{n_g}\{Y_{ig}-\Lambda(X_{ig}'\beta^0)\}X_{ig}'\xi)^2]\ge c_1$.

(2) $\min_{j,k}\mathrm{E}_{\mathrm{P}}[\frac{1}{G}\sum_{g=1}^G(\sum_{i=1}^{n_g}f_{ig}^2Z_{ig}^jX_{ig,k})^2]\wedge\min_{j,k}\mathrm{E}_{\mathrm{P}}[\frac{1}{G}\sum_{g=1}^G(\sum_{i=1}^{n_g}f_{ig}^2X_{ig,j}X_{ig,k})^2]\ge c_1$.

(3) $\max_{j,k}\{\mathrm{E}_{\mathrm{P}}[\frac{1}{G}\sum_{g=1}^G|V_g^jU_{gk}|^3]\}^{1/3}\log^{1/2}a_G\le\check{\delta}_GG^{1/6}$.

(4) $\sup_{\|\xi\|_2=1}\mathrm{E}_{\mathrm{P}}[\frac{1}{G}\sum_{g=1}^G(U_g'\xi)^4]+\max_{j\in[p]}\mathrm{E}_{\mathrm{P}}[\frac{1}{G}\sum_{g=1}^G(V_g^j)^4]\le C_1$.

(5) $M_{G,1}\ge\{\mathrm{E}_{\mathrm{P}}[\frac{1}{G}\sum_{g=1}^G\max_{j\in[p]}|V_g^j|^{2q}]\}^{1/2q}$.

(6) $M_{G,1}^2s\log a_G\le\check{\delta}_GG^{1/2-1/q}$.

(7) $M_{G,2}\ge\{\mathrm{E}_{\mathrm{P}}[\frac{1}{G}\sum_{g=1}^G\|U_g\|_\infty^{2q}]\}^{1/2q}$.

(8) $M_{G,2}^2s\log a_G\le\check{\delta}_GG^{1/2-1/q}$.

(9) $(M_{G,1}^2\vee s\log^2a_G)M_{G,2}^4s\le\check{\delta}_GG^{1-3/q}$.

Assumption 4 (Sparse Eigenvalues)**.** *Let $\Delta(m)=\{\delta\in\mathbb{R}^p:\|\delta\|_0\le m,\|\delta\|_2=1\}$. With probability at least $1-C(\log G)^{-1}$, we have*

$$1\lesssim\min_{j\in[p]}\min_{\delta\in\Delta(Cs)}\frac{1}{G}\sum_{g=1}^G\sum_{i=1}^{n_g}(Z_{ig}^jX_{ig}'\delta)^2\le\max_{j\in[p]}\max_{\delta\in\Delta(Cs)}\frac{1}{G}\sum_{g=1}^G\sum_{i=1}^{n_g}(Z_{ig}^jX_{ig}'\delta)^2\lesssim1,$$

$$1\lesssim\min_{\delta\in\Delta(Cs)}\frac{1}{G}\sum_{g=1}^G\sum_{i=1}^{n_g}(X_{ig}'\delta)^2\le\max_{\delta\in\Delta(Cs)}\frac{1}{G}\sum_{g=1}^G\sum_{i=1}^{n_g}(X_{ig}'\delta)^2\lesssim1.$$

Remark 5.2. Assumptions 1, 2, 3 are the cluster sampling counterpart of the Assumptions 3.1, 3.2, 3.4 and 3.5 of Belloni, Chernozhukov, Chetverikov and Wei (2018). To deal with APEs, however, we do need extra conditions on the growth of some moments that are listed below in the statement of Theorem 1. These growth conditions are satisfied when, for example, the covariates are sub-gaussian and/or uniformly bounded. When regressors are uniformly bounded, which is assumed in both Wooldridge and Zhu (2017) and Hirshberg and Wager (2018), the rate requirement would be $s \log p/G^{1/2} = o(1)$ ($s^{3/2} \log p/G^{1/2} = o(1)$ is required by Wooldridge and Zhu (2017)). Assumption 4 is analogous to condition SE in Belloni, Chernozhukov, Hansen and Kozbur (2016) for the linear panel data model.

Theorem 1 (Main Result)**.** *Suppose that Assumptions 1, 2, 3, 4 hold and* $(M_{G,1} \vee M_{G,2})^4 (\log a_G)^7 \lesssim G^{1-2/q-c_1}$ *for some* $c_1 \in (0, 1-2/q)$*,*

(1) The following uniform Bahadur representation holds with probability at least $1 - C(\log G)^{-1}$

$$\sup_{\mathrm{P} \in \mathcal{P}_G} \max_{1 \le k \le p} \left| \sqrt{G} \sigma_k^{-1} (\widehat{\alpha}_k - \alpha_k) - \frac{1}{\sqrt{G}} \sum_{g=1}^{G} \sum_{i=1}^{n_g} \varphi_k(W_{ig}, \alpha_k, \eta^k) \right| \lesssim \delta_G,$$

where $\varphi_k(W_{ig}, \alpha, \eta) = -\bar{\psi}_k(W_{ig}, \alpha, \eta)/\sigma_k$ *and* $\eta^k = (\beta^{0\prime}, \mu^{k\prime})'$*.*

(2) Let $c_W(a)$ *be the a-th quantile of W, we have, with probability at least* $1 - C(\log G)^{-1}$*,*

$$\sup_{\mathrm{P} \in \mathcal{P}_G} \sup_{\alpha \in (0,1)} \left| \mathrm{P}_{\mathrm{P}} \Big(\max_{1 \le k \le p} |\sqrt{G} \sigma_k^{-1} (\widehat{\alpha}_k - \alpha_k)| \le c_W(a) \Big) - a \right| = o(1).$$

That is to say, the algorithms in Section 4 provide valid simultaneous inference and confidence intervals asymptotically.

A proof can be found in Section D.1 in the Appendix. Now, it remains to find a valid variance estimator. Recall the variance estimator $\widetilde{\sigma}_k^2$ defined in (4.14). Denote $\widetilde{\sigma}_k = \{\widetilde{\sigma}_k^2\}^{1/2}$.

Lemma 1 (Variance Estimator)**.** *Suppose that the conditions for Theorem 1 hold. Then*

$$\max_{k \in [p]} |\widetilde{\sigma}_k - \sigma_k| \lesssim (\log a_G)^{-1}$$

with probability at least $1 - C(\log G)^{-1}$*.*

A proof can be found in Section D.2 in the Appendix.

6. Nuisance Parameters

Recall that the "ideal" nuisance parameter vector $\eta^k = (\beta^{0\prime}, \mu^{k\prime})'$, where

$$\mu^k = \Big\{ \mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} f_{ig}^2 X_{ig}X_{ig}'\Big]\Big\}^{-1} \cdot \mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\Big(\beta_k^0 \Lambda''(X_{ig}'\beta^0)X_{ig} + \Lambda'(X_{ig}'\beta^0)e_k\Big)\Big] = \zeta^k + \theta^k.$$

In this section, we propose estimators for these nuisance parameters as well as some theoretically justified choices of penalty tuning parameters. The choices here are based on the moderate deviation theory of self-normalized sums, which is first adapted for penalty selection of lasso by Belloni, Chernozhukov, Chen and Hansen (2012). Throughout this section, we fix a positive integer $\bar{m} \geq 1$ as the number of iterations used in the algorithms for choosing penalty tuning parameters.

6.1. **Post-Lasso Logit and Estimation of β^0.**

We now establish an asymptotic theory for estimation of β^0, which plays a central role in estimation of APE. The identification of β^0 follows from quasi-maximum likelihood and the assumption of population-averaged approach $\mathrm{E}[Y_{ig}|X_{ig}] = \Lambda(X_{ig}'\beta^0)$. Define the negative partial log-likelihood function by

$$M(Y_{ig}, X_{ig}, \beta) = -\{Y_{ig}X_{ig}'\beta - \log(1 + \exp(X_{ig}'\beta))\}. \tag{6.22}$$

Then, one has

$$\beta^0 = \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} M(Y_{ig}, X_{ig}, \beta)\Big].$$

We propose the following algorithm for the choice of $\widehat{\Psi}$.

Algorithm 6.1 (Penalty Choice: Clustered Lasso Logit β^0)**.** *Define $\lambda = c\sqrt{G}\Phi^{-1}(1-\gamma/2p)$ and set $c = 1.1$ and $\gamma = 0.1/\log G$. For $m = 0$, let*

$$\widehat{l}_{j,0} = \frac{1}{2}\Big\{\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g} n_g X_{ig,j}^2\Big)\Big\}^{1/2}$$

and for $1 \leq m \leq \bar{m}$,

$$\widehat{l}_{j,m} = \Big\{\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\{Y_{ig} - \Lambda(X_{ig}\widetilde{\beta})\}X_{ig,j}\Big)^2\Big\}^{1/2}$$

with $\widetilde{\beta}$ coming from iteration $m-1$. Let $\widehat{\Psi} = \mathrm{diag}\{\widehat{l}_{j,m} : j \in [p]\}$.

The following result provides convergence rates of $\widetilde{\beta}$ and $\widetilde{\beta}^k$.

Theorem 2. *Suppose that the Assumption 1, 2, 3 and 4 are satisfied. If* $\check{\delta}_G^2 \log a_G = o(1)$*, then with penalty chosen according to Algorithm (6.1), with probability* $1-\gamma$, $\gamma = O(\frac{1}{\log G})$,

$$\|\widetilde{\beta} - \beta^0\|_1 \vee \max_{k\in[p]} \|\widetilde{\beta}^k - \beta^0\|_1 \lesssim s\sqrt{\frac{\log a_G}{G}} \text{ and } \|\widetilde{\beta} - \beta^0\|_2 \vee \max_{k\in[p]} \|\widetilde{\beta}^k - \beta^0\|_2 \lesssim \sqrt{\frac{s\log a_G}{G}}.$$

A proof can be found in Section E.1 in the Appendix.

6.2. **Weighted Post-Lasso with Estimated Weights.** We now establish asymptotic theory for weighted post-lasso with estimated weights that will be essential for Sections 6.3 and 6.4. We propose the following algorithm for the choices of penalty tuning parameters.

Algorithm 6.2 (Penalty Choice: Weighted Clustered Lasso γ^j)**.** *Define* $\lambda^\gamma = c\sqrt{G}\Phi^{-1}(1-\gamma/2p(p-1))$ *and set* $c = 1.1$ *and* $\gamma = 0.1/\log G$*. For each* $j \in [p]$*, for* $m = 0$*, set*

$$\widehat{l}_{jk,0} = 2\max_{g\in[G]}\max_{i\in[n_g]} |\widehat{f}_{ig}X_{ig,k}| \Big\{\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\widehat{f}_{ig}D_{ig}^j\Big)^2\Big\}^{1/2}$$

and $1 \le m \le \bar{m}$,

$$\widehat{l}_{jk,m} = 2\Big\{\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\widehat{f}_{ig}^2(D_{ig}^j - X_{ig}^j\widetilde{\gamma}^j)X_{ig,k}^j\Big)^2\Big\}^{1/2}$$

and $\widehat{\Psi}_j^\gamma = \operatorname{diag}\{\widehat{l}_{jk,m} : k \in [p-1]\}$.

The following result provides convergence rates of $\widetilde{\gamma}^j$, which plays an important role in Section 6.3.

Theorem 3. *Suppose that Assumption 1, 2, 3, 4 are satisfied and if* $\check{\delta}_G^2 \log a_G = o(1)$*, then with penalty chosen according to Algorithm (6.2), with probability* $1-\gamma$, $\gamma = O(\frac{1}{\log G})$

$$\max_{j\in[p]} \|\widetilde{\gamma}^j - \gamma^j\|_1 \lesssim s\sqrt{\frac{\log a_G}{G}} \text{ and } \max_{j\in[p]} \|\widetilde{\gamma}^j - \gamma^j\|_2 \lesssim \sqrt{\frac{s\log a_G}{G}}.$$

A proof can be found in Section E.2 in the Appendix.

6.3. **Nodewise Post-Lasso and Estimation of θ^k.** Now we provide estimators for θ^k that are built upon the method of cluster nodewise post-lasso estimator for approximately inverting a singular matrix. The theory developed here is based on applying the weighted post-lasso with estimated weights from Belloni, Chernozhukov, Chetverikov and Wei (2018) to the panel nodewise regressions of Kock (2016). Recall that each nuisance parameter vector θ^k contains the matrix

$$\Theta := \Big\{ \mathrm{E}_{\mathrm{P}} \Big[\frac{1}{G} \sum_{g=1}^{G} \sum_{i=1}^{n_g} f_{ig}^2 X_{ig} X_{ig}' \Big] \Big\}^{-1}.$$

Its sample counterpart is not invertible if $p > n$ and could be very unstable if p is only moderately larger than n. Here, we take advantage of Assumption 2 to construct a high quality approximate inverse estimate. Denote Θ_j for the j-th row written as a column vector. If we can find some reasonable estimator $\widehat{\Theta}_k$ for Θ_k, then intuitively an estimator for θ^k can be defined as

$$\widetilde{\theta}^k = \widehat{\Theta}_k \cdot \frac{1}{G} \sum_{g=1}^{G} \sum_{i=1}^{n_g} \widehat{f}_{ig}^2.$$

We propose a cluster nodewise post-lasso procedure to estimate Θ. Recall that the error $Z_{ig}^j = D_{ig}^j - X_{ig}^j \gamma^j$ which satisfies $\mathrm{E}_{\mathrm{P}}[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} f_{ig}^2 X_{ig}^j Z_{ig}^j] = 0$. Define the error variance $\tau_j^2 = \mathrm{E}_{\mathrm{P}}[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} f_{ig}^2 (Z_{ig}^j)^2]$. Some properties of τ_j^2 can be found in Section F.1. Note that γ^j has a sparse approximation $\bar{\gamma}^j$ under Assumption 1. Then, we can use post-lasso estimate $\widetilde{\gamma}^j$ for γ^j from Section 6.2 and construct a $p \times p$ matrix $\widehat{C}$ by

$$\widehat{C} = \begin{bmatrix} 1 & -\widetilde{\gamma}_1^1 & \dots & -\widetilde{\gamma}_{p-1}^1 \\ -\widetilde{\gamma}_1^2 & 1 & \dots & -\widetilde{\gamma}_{p-1}^2 \\ \vdots & \vdots & \ddots & \vdots \\ -\widetilde{\gamma}_1^p & -\widetilde{\gamma}_2^p & \dots & 1 \end{bmatrix}.$$

That is, the off-diagonal spots of the j-th row of $\widehat{C}$ consist of components of $-\widetilde{\gamma}^j$ and the diagonal entries are set to 1. Also, denote

$$\widehat{T}^2 = \mathrm{diag}\{\widehat{\tau}_1^2, ..., \widehat{\tau}_p^2\},$$

where $\widehat{\tau}_j^2$ is defined in (4.12). Now, the cluster nodewise post-lasso estimator for Θ is defined as

$$\widehat{\Theta} = \widehat{T}^{-2} \widehat{C},$$

which in turn gives the expression of (4.11). The following results provide validity of $\widehat{\Theta}$ and $\widetilde{\theta}^k$.

Lemma 2. *Suppose that the Assumption 1, 2, 3, 4 are satisfied. If* $\check{\delta}_G^2 \log a_G = o(1)$*, then with penalty chosen according to Algorithm 6.2, with probability* $1-\gamma$, $\gamma = O(\frac{1}{\log G})$,

$$\max_{j\in[p]} \|\widehat{\Theta}_j - \Theta_j\|_1 \lesssim s\sqrt{\frac{\log a_G}{G}} \text{ and } \max_{j\in[p]} \|\widehat{\Theta}_j - \Theta_j\|_2 \lesssim \sqrt{\frac{s\log a_G}{G}}.$$

Theorem 4. *Suppose that all assumptions required by Lemma 2 are satisfied. Then, with probability* $1-\gamma$, $\gamma = O(\frac{1}{\log G})$*, we have*

$$\max_{k\in[p]} \|\widetilde{\theta}^k - \theta^k\|_1 \lesssim s\sqrt{\frac{\log a_G}{G}} \text{ and } \max_{k\in[p]} \|\widetilde{\theta}^k - \theta^k\|_2 \lesssim \sqrt{\frac{s\log a_G}{G}}.$$

Proofs for the above two results can be found in Sections E.4 and E.5 in the Appendix.

6.4. **Weighted Post-Lasso and Estimation of ζ^k.** Recall that the nuisance parameters ζ^k is identified by

$$\zeta^k = \operatorname*{argmin}_{\zeta\in\mathbb{R}^p} \mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} f_{ig}^2 (S_{ig}^k - X_{ig}'\zeta)^2\Big].$$

We propose the following algorithm for choice of the penalty tuning parameters.

Algorithm 6.3 (Penalty Choice: Weighted Clustered Lasso ζ^k)**.** *Define* $\lambda_j^\zeta = c\sqrt{G}\Phi^{-1}(1-\gamma/2p^2)$ *and set* $c = 1.1$ *and* $\gamma = 0.1/\log G$*. For each* $k \in [p]$*, for* $m = 0$*, set*

$$\widehat{l}_{kj,0} = 2\max_{g\in[G]}\max_{i\in[n_g]} |\widehat{f}_{ig}X_{ig,j}|\Big\{\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\widehat{f}_{ig}\widehat{S}_{ig}^k\Big)^2\Big\}^{1/2}$$

and $1 \le m \le \bar{m}$,

$$\widehat{l}_{kj,m} = 2\Big\{\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\widehat{f}_{ig}^2(\widehat{S}_{ig}^k - X_{ig}'\widetilde{\zeta}^k)X_{ig,j}\Big)^2\Big\}^{1/2}$$

and $\widehat{\Psi}_k^\zeta = \mathrm{diag}\{\widehat{l}_{kj,m} : j \in [p]\}$.

The following result provides convergence rates of $\widetilde{\zeta}^k$.

Corollary 1. *Suppose that Assumptions 1, 2, 3, 4 hold. If* $\check{\delta}_G^2 \log a_G = o(1)$*, then with penalty chosen according to Algorithm 6.3, with probability* $1-\gamma$, $\gamma = O(\frac{1}{\log G})$,

$$\max_{k\in[p]} \|\widetilde{\zeta}^k - \zeta^k\|_1 \lesssim s\sqrt{\frac{\log a_G}{G}} \text{ and } \max_{k\in[p]} \|\widetilde{\zeta}^k - \zeta^k\|_2 \lesssim \sqrt{\frac{s\log a_G}{G}}.$$

A proof can be found in Section E.3 in the Appendix.

7. Simulation Studies

In this section, we conduct simulation studies to examine the finite-sample performance of the proposed procedures. We set the number of total observations to n, and each observation is then randomly assigned into G_0 clusters. The empty clusters, if they exist, are then discarded and thus $G \leq G_0$. For DGP1, let the number of covariates for each observation be $p = 1.5 \cdot G_0$ and

$$\beta^0 = [1, \beta_2, 1/3, 1/4, 1/5, .., 1/19, 1/20, 0, ..., 0]' \in \mathbb{R}^p.$$

The first component of each covariate vector is set to 1 and the rest of the subvector, $X_{ig,-1}$, can be decomposed into an idiosyncratic part X^1_{ig} and a cluster-wise component X^2_g as

$$X_{ig,-1} = X^1_{ig} + X^2_g$$

and both X^1_{ig} and X^2_g are i.i.d. following a multivariate normal distribution with mean 0 and a Toeplitz covariance matrix:

$$\Sigma_{ij}(\rho) := \rho^{|i-j|},\ \rho = 0.1,\ 0.3,\ 0.5,\ 0.7,\ 0.9,\ i, j, \in [p-1].$$

So the larger ρ is, the more correlated the covariates are. The outcome variable is generated by

$$Y_{ig} = \mathbb{1}\Big\{X'_{ig}\beta^0 + U_{ig} > 0\Big\},$$

where the error term can also be decomposed into an idiosyncratic term and a cluster-wise term as

$$U_{ig} = \Lambda\Big(\Phi^{-1}(U^1_{ig} + U^2_g)\Big),$$

where both U^1_{ig} and U^2_g are i.i.d. following the normal $N(0, 1/2)$ distribution. Thus, U_{ig} is a standard logistic distribution. Thus both covariates and errors are correlated within each cluster. To consider "outliers" and substantial skew and kurtosis in marginal distribution of independent variables, we also consider alternative DGPs inspired by Kline and Santos (2012) by setting X^1_{ig} and X^2_g to follow a mixture between two distributions, $N(0, \Sigma(\rho))$ with probability 0.9 and a $N(0, \Sigma(\rho)) - 1.5 \times N(1, \Sigma(\rho))$ with probability 0.1.

Table 2. List of DGPs in Simulation Studies.

Model DGP	Descriptions
M1	X^1_{ig}, $X^2_g \sim N(0, \Sigma(\rho))$ with $\rho = 0.1$
M2	Same as M1 except $\rho = 0.3$
M3	Same as M1 except $\rho = 0.5$
M4	Same as M1 except $\rho = 0.7$
M5	Same as M1 except $\rho = 0.9$
M6	X^1_{ig}, $X^2_g \sim \Big(N(0, \Sigma(\rho)) - 1.5 * B(1, 0.1) * N(1, \Sigma(\rho))\Big)$ with $\rho = 0.1$
M7	Same as M6 except $\rho = 0.3$
M8	Same as M6 except $\rho = 0.5$
M9	Same as M6 except $\rho = 0.7$
M10	Same as M6 except $\rho = 0.9$

Note that for the DGPs with high ρ, such as M4, M5, M9 and M10, the approximate sparsity conditions in Assumption 1 are violated. We conduct three sets of simulations. First we examine one-dimensional confidence interval coverage for α_2 with true underlying $\beta_2 \in \{0, 0.25, 0.5, 0.75, 1\}$. Our second goal is to construct simultaneous confidence intervals that control the family-wise error rate for α_k for $k \in A$, A is set to be

$$A_1 = \{2\},\ A_2 = \{2,3\},\ A_3 = \{2,3,4\},\ A_5 = \{2,3,...,6\},\ A_{10} = \{2,3,...,10\},\ A_{20} = \{2,3,...,20\},$$
$$A_{30} = \{2,3,...,31\},\ A_{40} = \{2,3,...,41\},\ A_{50} = \{2,3,...,51\},\ A_{100} = \{2,3,...,101\},$$

where the APE with respect to the intercept is always omitted. In this group of simulations, we set $\beta_2 = 0.5$. Finally, we examine the asymptotic behaviors of coverage probabilities of simultaneous intervals for A_{10}.

The estimation of all lasso and lasso Logit are conducted using R package **glmnet** and the penalty choices follow Algorithms 6.1, 6.2 and 6.3 in Section 6 with $\bar{m} = 1$. For each iteration of the simulation, we set the number of bootstrap iterations to $B = 600$. We then simulate $1,000$ times for each DGP. The simultaneous confidence intervals are constructed following Algorithm 4.3 and without normalization by $\widetilde{\sigma}_k$ for simplicity. The true α_k are computed using $3,000,000$ additional observations generated independently from data following the same marginal distribution as X_{ig}. The nominal coverage probability is set to be 0.95. The results for one-dimensional confidence intervals are presented in the tables below.

TABLE 3. Coverage probability for one-dimensional 95% confidence intervals for α_2 under each DGP with $G_0 = 200$, $n = 500$ and $p = 300$:

Model DGP	$\beta_2 = 0$	$\beta_2 = .25$	$\beta_2 = .5$	$\beta_2 = .75$	$\beta_2 = 1$
M1	0.953	0.935	0.943	0.959	0.972
M2	0.949	0.944	0.937	0.956	0.969
M3	0.939	0.938	0.941	0.951	0.968
M4	0.938	0.944	0.937	0.944	0.954
M5	0.928	0.931	0.928	0.940	0.923
M6	0.928	0.919	0.920	0.946	0.970
M7	0.921	0.920	0.912	0.926	0.957
M8	0.934	0.925	0.929	0.954	0.956
M9	0.926	0.929	0.933	0.941	0.958
M10	0.938	0.935	0.944	0.938	0.947

We now present the coverage probabilities for simultaneous confidence intervals for different sets of covariates. For this part, we focus on models M1 to M5 with $\beta_2 = 0.5$.

TABLE 4. Coverage probability for 95% simultaneous confidence intervals for α_k, $k \in A$ under each DGP with $G_0 = 200$, $n = 500$ and $p = 300$:

Model DGP	A_1	A_2	A_3	A_5	A_{10}	A_{20}	A_{30}	A_{40}	A_{50}	A_{100}
M1	0.943	0.941	0.933	0.926	0.920	0.940	0.951	0.959	0.957	0.943
M2	0.937	0.945	0.932	0.930	0.907	0.927	0.917	0.920	0.940	0.950
M3	0.941	0.943	0.956	0.914	0.882	0.900	0.901	0.914	0.920	0.930
M4	0.937	0.928	0.925	0.876	0.865	0.863	0.865	0.891	0.897	0.925
M5	0.928	0.926	0.930	0.891	0.865	0.861	0.874	0.898	0.894	0.904

Finally, we investigate the asymptotic behaviors of the case with A_{10}, one of the worst-performing cases in the above simulations for simultaneous confidence intervals, to examine whether the performance improves as sample size increases. In this set of simulations, set $\beta_2 = 0.5$ for number of nominal clusters $G_0 = 200$, 400, 600 and 800, $p = 1.5 \cdot G_0$, and $n = 2.5 \cdot G_0$.

Table 5. Asymptotic behaviors of coverage probability for 95% simultaneous confidence intervals for α_k, $k \in A_{10}$ under each DGP:

Model DGP	$G_0 = 200$	$G_0 = 400$	$G_0 = 600$	$G_0 = 800$
M1	0.920	0.914	0.930	0.945
M2	0.907	0.917	0.920	0.921
M3	0.882	0.894	0.898	0.903
M4	0.865	0.856	0.864	0.876
M5	0.865	0.859	0.847	0.857

In all of the three sets of simulations, the coverage probabilities are mostly fairly close to the nominal coverage rate when ρ is not very high. When ρ is high, the approximate sparsity of nuisance parameters in Assumption 1 is violated. Thus some of the coverage probabilities are not close to the nominal rate. In addition, the coverage probabilities improve as sample size increases. In summary, the outcomes of the simulations are consistent with our theoretical results.

8. Application: Testing Gendered Language on the Internet

In this section, we apply our method of simultaneous inference for APEs in the text regression model of Wu (2018) introduced in Section 2. We make use of the pronoun sample (gendered posts including either female or male pronouns) from Wu (2018). Following Wu, using the EJMR dataset[10], we exclude the same list of words from the 10,000, including all gender classifiers, plus names of non-economist celebrities. We conduct our analysis based on the subset of non-duplicate posts that are used as the test sample for selecting optimal probability threshold in the original paper (the posts with index labelled as test0) for classification of posts that contains both female and male classifiers. We consider only pronoun sample. This leaves 46,502 posts sampled from 31,739 threads and 9541 covariates[11] that consists of an intercept and the word counts of 9,540 non-excluded vocabularies.

[10] The dataset is publicly available at url:https://www.aeaweb.org/articles?id=10.1257/pandp.20181101

[11] Since the number of observations is larger than dimensionality of parameters, regular Logit and even OLS can be applied here. We have attempted to implement Logit using **glm** package in R. However, it did not finish after 70 minutes. OLS on the other hand takes 55 minutes to complete. In contrast, the proposed estimation and inference algorithms, when applied to the testing problem in this section, takes about two minutes to complete.

Wu (2018) highlights that posts about males include more academically and professionally oriented vocabularies, such as “adviser,” “supervisor,” and “Nobel.” To see the joint significance of these words' APE in terms of predicting female, we test

$$\mathrm{H}_0 : \alpha_{\text{adviser}} = \alpha_{\text{supervisor}} = \alpha_{\text{nobel}} = 0.$$

Following the penalty choices of Algorithms 6.1, 6.2 and 6.3, the estimates of APEs of these words calculated using Algorithm 4.1 are listed as follows:

Table 6. APE estimates for “adviser,” “supervisor,” and “Nobel.”

	adviser	supervisor	Nobel
APE estimate	-0.1414	-0.1214	-0.1214

These estimates are qualitatively similar to the corresponding estimates in Wu (2018). Using multiplier cluster bootstrap with $10,000$ bootstrap iterations, we obtain the follow test results

Table 7. Multiple Testing Results under $1-\alpha\%$ Confidence level.

α	MCB critical value	test statistic
10%	16.0889	25.1867
5%	18.2870	25.1867
1%	22.2930	25.1867

Thus, under all three confidence levels, we reject the null hypothesis and the statistical evidence supports Wu's statement [12].

9. Conclusion

In this paper, we study logistic average partial effects with lasso regularization when data is sampled under clustering. We proposed two valid estimators along with their theoretically justified lasso penalty choices. Based on these estimators, we provide easy-to-implement algorithms for simultaneous inference and confidence intervals and establish their asymptotic validity. Simulation studies demonstrate that the proposed procedures work as predicted by the theory in finite sample. We then apply the proposed method to conduct analysis of textual data to examine the presence of gendered language on the EJMR forum following

[12] One may be concerned about the high-correlation between ”supervisor” and ”adviser.” However, removing either one of them does not change the significance of the tests at 99% confidence level.

the text regression model of Wu (2018). Our analysis provides further statistical evidence to support Wu's finding.

References

Athey, Susan, Guido W. Imbens, and Stefan Wager. "Approximate residual balancing: debiased inference of average treatment effects in high dimensions." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80, no. 4 (2018): 597-623.

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen. "Sparse models and methods for optimal instruments with an application to eminent domain." *Econometrica* 80, no. 6 (2012): 2369-2429.

Belloni, Alexandre, Victor Chernozhukov, and Kengo Kato. "Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems." *Biometrika* 102, no. 1 (2015): 77-94.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. "Inference on treatment effects after selection among high-dimensional controls." *The Review of Economic Studies* 81, no. 2 (2014): 608-650.

Belloni, Alexandre, Victor Chernozhukov, Denis Chetverikov, and Ying Wei. "Uniformly valid post-regularization confidence regions for many functional parameters in Z-estimation framework." *The Annals of Statistics* 46, no. 6B (2018): 3643-3675.

Belloni, Alexandre, Victor Chernozhukov, Ivan Fernández-Val, and Christian Hansen. "Program evaluation and causal inference with high-dimensional data."*Econometrica* 85, no. 1 (2017): 233-298.

Belloni, Alexandre, Victor Chernozhukov, Christian Hansen, and Damian Kozbur. "Inference in high-dimensional panel models with an application to gun control." *Journal of Business and Economic Statistics* 34.4 (2016): 590-605.

Belloni, Alexandre, Victor Chernozhukov, and Ying Wei. "Post-selection inference for generalized linear models with many controls." *Journal of Business and Economic Statistics* 34, no. 4 (2016): 606-619.

Bickel, Peter J., Ya'acov Ritov, and Alexandre B. Tsybakov. "Simultaneous analysis of Lasso and Dantzig selector." *The Annals of Statistics* 37, no. 4 (2009): 1705-1732.

Bühlmann, Peter, and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science and Business Media, 2011.

Cameron, A. Colin, and Douglas L. Miller. "A practitioner's guide to cluster-robust inference." *Journal of Human Resources* 50.2 (2015): 317-372.

Caner, Mehmet. "Delta Theorem in the Age of High Dimensions." arXiv preprint arXiv:1701.05911 (2017).

Caner, Mehmet, and Anders Bredahl Kock. "Asymptotically honest confidence regions for high dimensional parameters by the desparsified conservative lasso." *Journal of Econometrics* 203, no. 1 (2018): 143-168.

Chamberlain, Gary. "Panel data." *Handbook of econometrics* 2 (1984): 1247-1318.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. "Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors." *The Annals of Statistics* 41, no. 6 (2013): 2786-2819.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. "Gaussian approximation of suprema of empirical processes." *The Annals of Statistics* 42, no. 4 (2014): 1564-1597.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal* 21, no. 1 (2018): C1-C68.

Chernozhukov, Victor, Chris Hansen, and Martin Spindler. "hdm: High-dimensional metrics." arXiv preprint arXiv:1608.00354 (2016).

Chernozhukov, Victor, Whitney K. Newey, and Rahul Singh. "Learning L2 Continuous Regression Functionals via Regularized Riesz Representers." arXiv preprint arXiv:1809.05224 (2018).

Djogbenou, Antoine A., James G. MacKinnon, and Morten Orregard Nielsen. "Asymptotic theory and wild bootstrap inference with clustered errors." working paper. (2018).

Farrell, Max H. "Robust inference on average treatment effects with possibly more covariates than observations," *Journal of Econometrics*, 189, no. 1 (2015): 1-23.

Gentzkow, Matthew, Bryan T. Kelly, and Matt Taddy. "Text as data." forthcoming at *Journal of Economic Literature.* (2019).

Gentzkow, Matthew, Jesse Shapiro, and Matt Taddy. "Measuring polarization in high-dimensional data: Method and application to congressional speech." *Econometrica,* forthcoming, (2019).

Giné, Evarist, and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models.* Cambridge University Press, (2016).

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. "Detecting influenza epidemics using search engine query data." *Nature* 457, no. 7232 (2009): 1012.

Hagemann, Andreas. "Cluster-robust bootstrap inference in quantile regression models." *Journal of the American Statistical Association* 112, no. 517 (2017): 446-456.

Hirshberg, David A and Stefan Wager. "Balancing out regression error: efficient treatment effect estimation without smooth propensities." arXiv preprint arXiv:1712.00038 (2017)

Hirshberg, David A and Stefan Wager. "Debiased inference of average partial effects in single-index models." arXiv preprint arXiv:1811.02547 (2018)

Javanmard, Adel, and Andrea Montanari. "Confidence intervals and hypothesis testing for high-dimensional regression." *The Journal of Machine Learning Research* 15, no. 1 (2014): 2869-2909.

Jegadeesh, Narasimhan, and Di Wu. "Word power: A new approach for content analysis." *Journal of Financial Economics* 110, no. 3 (2013): 712-729.

Kato, Kengo. "Lecture notes in empirical process theory" technical report (2017).

Kline, Patrick, and Andres Santos. "A score based approach to wild bootstrap inference." *Journal of Econometric Methods* 1.1 (2012): 23-41.

Kock, Anders Bredahl. "Oracle inequalities, variable selection and uniform inference in high-dimensional correlated random effects panel data models." *Journal of Econometrics* 195.1 (2016): 71-85.

Kock, Anders Bredahl, and Haihan Tang. "Uniform inference in high-dimensional dynamic panel data models with approximately sparse fixed effects." *Econometric Theory* (2018): 1-65.

MacKinnon, James G., and Matthew D. Webb. "Wild bootstrap inference for wildly different cluster sizes." Journal of Applied Econometrics 32, no. 2 (2017): 233-254.

Pötscher, Benedikt M., and Hannes Leeb. "On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding." *Journal of Multivariate Analysis* 100, no. 9 (2009): 2065-2082.

Wooldridge, Jeffrey M. "Unobserved heterogeneity and estimation of average partial effects." Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg (2005): 27-55.

Wooldridge, Jeffrey M. *Econometric analysis of cross section and panel data.* MIT press, (2010).

Wooldridge, Jeffrey M. "Correlated random effects models with unbalanced panels." *Journal of Econometrics* (2018) forthcoming.

Wooldridge, Jeff, and Ying Zhu. "Inference in Approximately Sparse Correlated Random Effects Probit Models." *Journal of Economic and Business Statistics* (2017) forthcoming.

Van de Geer, Sara, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. "On asymptotically optimal confidence regions and tests for high-dimensional models." *The Annals of Statistics* 42, no. 3 (2014): 1166-1202.

van der Vaart, Aad W., and Jon A. Wellner. *Weak convergence and empirical processes.* Springer, New York, NY, (1996).

Wu, Alice H. ”Gendered Language on the Economics Job Market Rumors Forum.” In AEA Papers and Proceedings, vol. 108, pp. 175-79. 2018.

Zhang, Cun-Hui, and Stephanie S. Zhang. ”Confidence intervals for low dimensional parameters in high dimensional linear models.”*Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76, no. 1 (2014): 217-242.

Appendix A. Orthogonalization of the Score

In this Section, we derive the Neyman orthogonal score for α_k, as defined in (5.16), following the methodology in Section 2.2 of Belloni, Chernozhukov, Chetverikov and Wei (2018) (see also Section 2 of Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey and Robins (2018)). The first order condition of the population quasi-maximal likelihood and definition of the k-th APE give $\mathrm{E}_{\mathrm{P}}[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} m(W_{ig},\alpha_k,\beta^0)] = 0$, where

$$m_k(W_{ig},\alpha,\beta) = \begin{bmatrix} \partial_\alpha \widetilde{\ell}(W_{ig},\alpha,\beta) \\ \partial_\beta \widetilde{\ell}(W_{ig},\alpha,\beta) \end{bmatrix} := \begin{bmatrix} \alpha \cdot \frac{G}{n} - \beta_k \Lambda'(X_{ig}'\beta) \\ \ell'(Y_{ig}, X_{ig}'\beta) X_{ig} \end{bmatrix},$$

where $\ell(a,b) = a\log\Lambda(b) + (1-a)\log(1-\Lambda(b))$, $\ell'(a,b) = \frac{\partial}{\partial b}\ell(a,b)$, $\ell''(a,b) = \frac{\partial^2}{\partial b^2}\ell(a,b)$. Note that the order of integral and derivative are interchangeable in this case. Let us define

$$\begin{aligned} J =& \partial_{(\alpha,\beta')'} \mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} m_k(W_{ig},\alpha,\beta)\Big]\big|_{\alpha=\alpha_k,\beta=\beta^0} \\ =& \begin{bmatrix} 1 & -\mathrm{E}_{\mathrm{P}}[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(\beta_k\Lambda''(X_{ig}'\beta)X_{ig}' + \Lambda'(X_{ig}'\beta)e_k')] \\ 0 & \mathrm{E}_{\mathrm{P}}[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\ell''(Y_{ig},X_{ig}'\beta)X_{ig}X_{ig}'] \end{bmatrix}_{\alpha=\alpha_k,\beta=\beta^0} \\ =& \begin{bmatrix} J_{\alpha\alpha} & J_{\alpha\beta} \\ J_{\beta\alpha} & J_{\beta\beta} \end{bmatrix}. \end{aligned}$$

Now define population nuisance parameter

$$\begin{aligned} \mu^k =& -J_{\beta\beta}^{-1}J_{\alpha\beta}' \\ =& \Big\{\mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\ell''(Y_{ig},X_{ig}'\beta^0)X_{ig}X_{ig}'\Big]\Big\}^{-1}\mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(\beta_k^0\Lambda''(X_{ig}'\beta^0)X_{ig} + \Lambda'(X_{ig}'\beta^0)e_k)\Big] \\ =& \Big\{\mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}f_{ig}^2X_{ig}X_{ig}'\Big]\Big\}^{-1}\mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}f_{ig}^2X_{ig}S_{ig}^k\Big] + \Big\{\mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}f_{ig}^2X_{ig}X_{ig}'\Big]\Big\}^{-1}\mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}f_{ig}^2e_k\Big] \\ =& \zeta^k + \theta^k, \end{aligned}$$

where $S_{ig}^k = \beta_k(1-2\Lambda(X_{ig}'\beta^0))$. Here we have used the property of the logistic function $\Lambda''(X_{ig}'\beta^0) = \Lambda(X_{ig}'\beta^0)(1-\Lambda(X_{ig}'\beta^0))(1-2\Lambda(X_{ig}'\beta^0))$ and thus

$$\beta_k^0\Lambda''(X_{ig}'\beta^0) = f_{ig}^2\beta_k^0(1-2\Lambda(X_{ig}'\beta^0)) = f_{ig}^2S_{ig}^k.$$

We now define Neyman orthogonal score for α_k as

$$\begin{aligned}\bar{\psi}_k(W_{ig},\alpha,\eta) =&\partial_\alpha \widetilde{\ell}(W_{ig},\alpha,\beta) - \mu'\partial_\beta \widetilde{\ell}(W_{ig},\alpha,\beta)\\ =&\alpha\cdot\frac{G}{n} - \beta_k\Lambda'(X_{ig}'\beta) + \mu' X_{ig}\{Y_{ig} - \Lambda(X_{ig}'\beta)\}\\ =&\alpha\cdot\frac{G}{n} - \psi_k(W_{ig},\eta),\end{aligned}$$

where β_k is the k-th coordinate of β and $\eta = (\beta', \mu') \in \mathbb{R}^{2p}$. It is straightforward to verify the followings,

$$\mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\bar{\psi}_k(W_{ig},\alpha_k,\eta^k)\Big] = 0, \qquad \text{(existance condition)}$$

$$\partial_\eta \mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\bar{\psi}_k(W_{ig},\alpha_k,\eta^k)\Big] = 0, \qquad \text{(Neyman orthogonality condition)}$$

$$\partial_\alpha \mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\bar{\psi}_k(W_{ig},\alpha_k,\eta^k)\Big] = 1 \neq 0. \ \text{(uniqueness condition)}$$

Appendix B. Main Results under High-level Assumptions

In this section we introduce a version of our asymptotic results under high-level conditions. They serve as building blocks for results in Section 5. Suppose that we have some generic nuisance parameter estimators $\widehat{\eta}^k$ such that $\eta^k \in \mathcal{H}_k$ for G large enough. Denote A_k, a bounded interval of α_k shrinking with G, and $\mathcal{H}_k \subset H_k$, a sparse neighborhood of η^k shrinking with G, where $H_k \subset \mathbb{R}^p$ a compact and convex set that contains η^k. Write B_{1G} and B_{2G} as some positive sequences of constants that can possibly diverge to infinity. Let a_G, v_G, K_G be some positive sequences of constants that can possibly grow to infinity, where $a_G \geq G \vee K_G$ and $v_G \geq 1$ for all $G \geq 1$. Let $q \geq 2$ be some constant. Further, let τ_G, δ_G and Δ_G be some positive sequences of constants that converge to zero and $\Delta_G < 1$.

Assumption 5. *For each $G \in \mathbb{N}$, $G \geq G_0$, $\mathrm{P} \in \mathcal{P}_G$ and $1 \leq k \leq p$, define sequences of positive constants $\delta_G = o(1)$, and $\tau_G = o(1)$. The true parameters $(\alpha_k, \eta^k) \in A_k \times \mathcal{H}_k$ for some A_k and $\mathcal{H}_k$ and the following are satisfied:*

(i) $\eta \mapsto \mathrm{E_P}[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\psi_k(W_{ig},\eta)]$ *is twice continuously differentiable.*

(ii) It holds that

(a) $\sup_{\eta\in\mathcal{H}_k}\mathrm{E_P}[\frac{1}{G}\sum_{g=1}^{G}(\sum_{i=1}^{n_g}\{\psi_k(W_{ig},\eta) - \psi_k(W_{ig},\eta^k)\ \})^2] \leq C_0\|\eta - \widehat{\eta}^k\|_2^2$,

(b) $\sup_{\eta\in\mathcal{H}_k}\|\partial_{\eta'}\partial_\eta \mathrm{E_P}[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\psi_k(W_{ig},\eta)]\|_2 \leq B_{1G}$.

Assumption 6. *At least one coordinate of X_{ig} is continuously distributed. Furthermore, there exists sequences of positive constants $\Delta_G = o(1)$, $\tau_G = o(1)$ such that for the same $\mathcal{H}_k$ from Assumption 5, the following holds uniformly in k.*

(i) $\widehat{\eta}^k \in \mathcal{H}_k$ *with probability at least* $1 - \Delta_G$*and* $\sup_{\eta\in\mathcal{H}_k} \|\eta - \eta^k\|_2 \le \tau_G$.

(ii) The collection of functions

$$\mathcal{F}_0 = \{\bar{\psi}_k(\cdot, \alpha_k, \eta) : k \in [p],\ \eta \in \mathcal{H}_k\} \cup \{0\}$$

is pointwise measurable and satisfies that for all $0 < \varepsilon \le 1$,

$$\sup_Q \log N(\mathcal{F}_0, L_2(Q), \varepsilon\|F_0\|_{Q,2}) \le v_G \log(a_G/\varepsilon)$$

where the supremum is taken over the set of all finite measures and F_0 *is a measurable envelope of* $\mathcal{F}_0$ *such that* $\{\mathrm{E_P}[\frac{1}{G}\sum_{g=1}^G |\sum_{i=1}^{n_g} F_0(W_{ig})|^q]\}^{1/q} \le K_G$.

(iii) For all $f \in \mathcal{F}_0$, *we have* $c_0 \le \{\mathrm{E_P}[\frac{1}{G}\sum_{g=1}^G(\sum_{i=1}^{n_g} f(W_{ig}))^2]\}^{1/2} \le C_0$.

(iv) $\tau_G\sqrt{v_G \log a_G} \vee G^{-1/2+1/q}K_G v_G \log a_G \lesssim \delta_G$ *and* $\sqrt{G}B_{1G}\tau_G^2 \lesssim \delta_G$.

Remark B.1. While been adapted to our cluster sampling setting, Assumptions 5, 6 are similar to Condition 2, 3 of Belloni, Chernozhukov and Kato (2015) and Assumption 2.1, 2.2 of Belloni, Chernozhukov, Chetverikov and Wei (2018). However, due to the additive separability of $\widehat{\alpha}_k$, we do not need to assume Assumption 2.1(b) of Belloni, Chernozhukov, Chetverikov and Wei (2018). Also, differentiability of the orthogonal score comes directly from smoothness of logistic function.

The following result builds upon the ideas of the main results in Belloni, Chernozhukov and Kato (2015) and Belloni, Chernozhukov, Chetverikov and Wei (2018) while allowing for cluster sampling. Given some generic nuisance parameters estimate $\widehat{\eta}^k$, we define the generic APE estimator for the k-th continuous covariate as

$$\widehat{\alpha}_k = \frac{1}{G}\sum_{g=1}^G \sum_{i=1}^{n_g} \psi_k(W_{ig}, \widehat{\eta}^k). \tag{B.23}$$

It is easy to verify the fact that the post-double-section estimator $\widetilde{\alpha}_k$, as defined in (4.1), satisfies (B.23) for $\widehat{\eta}^k = (\widetilde{\beta}^{k\prime}, \widetilde{\mu}^{k\prime})'$ following the first order condition of (4.8), the definition of $\widetilde{T}_k$ and the definition of ψ_k.

Theorem 5 (Uniform Bahadur Representation)**.**
Suppose that we have nuisance parameter estimates $(\widehat{\eta}^k)_{k\in[p]}$ *such that Assumptions 5 and*

6 are satisfied. For the generic $(\widehat{\alpha}_k)_{k\in[p]}$ defined based on $(\widehat{\eta}^k)_{k\in[p]}$ following (B.23), with probability at least $1-\Delta_G-(\log G)^{-1}$,

$$\sup_{\mathrm{P}\in\mathcal{P}_G}\max_{1\le k\le p}\left|\sqrt{G}\sigma_k^{-1}(\widehat{\alpha}_k-\alpha_k)-\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\varphi_k(W_{ig},\alpha_k,\eta^k)\right|\lesssim\delta_G,$$

where $\varphi_k(W_{ig},\alpha,\eta)=-\bar{\psi}_k(W_{ig},\alpha,\eta)/\sigma_k$ and $\eta^k=(\beta^{0\prime},\mu^{k\prime})'$.

A proof can be found in Appendix C.1.

Let $\{\xi_g\}_{g=1}^G$ be independent standard normal random variables generated independently from data. Define

$$W=\max_{1\le k\le p}\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\xi_g\sum_{i=1}^{n_g}\widehat{\varphi}_k(W_{ig},\widehat{\alpha}_k,\widehat{\eta}^k)\text{ and }W_0=\max_{1\le k\le p}\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\xi_g\sum_{i=1}^{n_g}\varphi_k(W_{ig},\alpha_k,\eta^k).$$

We also denote $\bar{A}_G\ge G$ and $\bar{\rho}_G\ge\log G$ be sequences of positive constants that grow to infinity.

Assumption 7. *For all $G\ge G_0$ and $\mathrm{P}\in\mathcal{P}_G$, the following holds.*

(i) There exists $B_G\ge 1$ such that $B_G^4(\log(p\cdot G))^7/G\le C_1G^{-c_1}$ for positive constants c_1, C_1 and for all $1\le g\le G$ and $k\in[p]$

$$\max_{b=1,2}\mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^{G}\Big|\sum_{i=1}^{n_g}\varphi_k(W_{ig},\alpha_k,\eta^k)\Big|^{2+b}/B_G^b\Big]+\mathrm{E}_{\mathrm{P}}\Big[\Big(\max_{1\le k\le p}\Big|\sum_{i=1}^{n_g}\varphi_k(W_{ig},\alpha_k,\eta^k)\Big|/B_G\Big)^4\Big]\le 4.$$

(ii) The collection

$$\widehat{\mathcal{F}}_0=\Big\{W_{ig}\mapsto\sum_{i=1}^{\bar{n}}(W_{ig},\alpha_k,\eta^k)-\widehat{\varphi}_k(W_{ig},\widehat{\alpha}_k,\widehat{\eta}^k))\cdot\mathbb{1}\{\|||W_{ig}|\|_\infty>0\}\ :k\in\{1,...,p\}\Big\}$$

satisfies with probability at least $1-\Delta_G$,

$$\log N(\widehat{\mathcal{F}}_0,L_2(\mathbb{P}_G),\epsilon)\le\bar{\rho}_G\log(\bar{A}_G/\epsilon),\ \textit{for all}\ 0<\epsilon\le 1,$$

and $\{\mathrm{E}_{\mathrm{P}}[\frac{1}{G}\sum_{g=1}^{G}f^2]\}^{1/2}\le\bar{\delta}_G$ for all $f\in\widehat{\mathcal{F}}_0$.

(iii) $\bar{\delta}_G^2\bar{\rho}_G\log\bar{A}_G\log p=o(1)$ and $\delta_G^2\log p=o(1)$.

Remark B.2. Assumption 7 (i) is required by the high-dimensional central limit theorem of Chernozhukov, Chetverikov and Kato (2013) (see their Corollary 2.1). Assumption 7 (ii) is discussed in the next remark. Assumption 7 (iii) is a technical assumption that turns out to be mild, as shown in the sufficient conditions in Section 5.

Remark B.3 (Double/debiased Machine Learning)**.** One could potentially employ sample splitting to eliminate the dependence between the orthogonal score and nuisance parameters. This procedure is known as "double/debiased machine learning" (cf Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey and Robins (2018)). This would allow us to relax Assumption 7 (ii). We did not make use of sample splitting due to the following considerations. First, we do not assume each cluster is identically distributed since it is not suitable for the sampling method used in the motivating example in Section 2. Second, even if identical distribution is assumed, when we have binary outcome variable Y_{ig}, sample-splitting may results in subsamples with high percentages of outcomes equal to 1 or 0. In such case, the estimate for $\widehat{\eta}^k$ could be very unreliable. Finally, relaxing Assumption 7 (ii) does not appear to allow us to relax any sufficient conditions presented in Secion 5. Therefore, we do not consider sample splitting in this paper.

Corollary 2 (Multiplier Cluster Bootstrap of Maxima)**.** *Suppose that Assumptions 5, 6 and 7 are satisfied, then let $c_W(a)$ be the a-th quantile of W, we have*

$$\sup_{\mathrm{P}\in\mathcal{P}_G}\sup_{\alpha\in(0,1)}\left|\mathrm{P}_{\mathrm{P}}\Big(\max_{1\leq k\leq p}|\sqrt{G}\sigma_k^{-1}(\widehat{\alpha}_k-\alpha_k)|\leq c_W(a)\Big)-a\right|=o(1).$$

A proof can be found in Section C.2 in the Appendix.

Remark B.4 (Uniform in DGP)**.** Note that all the above results are valid uniformly over $\mathcal{P}_G$, the set of DGP's such that Assumptions 5, 6, 7 are satisfied. This is due to the fact that Lemma A.1 of Belloni, Chernozhukov, Fernández-Val and Hansen (2017) implies that it suffices to show that these results hold for any sequence $\mathrm{P}_G\in\mathcal{P}_G$, which is satisfied since all the bounds in this paper are established independently of DGP.

Appendix C. Proofs for Results in Section B

C.1. **Proof for Theorem 5.**

Proof. By (B.23), it holds that $\widehat{\alpha}_k = \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\psi_k(W_{ig},\widehat{\eta}^k)$ for an $\widehat{\eta}^k$ from Assumption 6. The fact that $\alpha_k = \mathrm{E_P}[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\psi_k(W_{ig},\eta^k)]$ implies

$$\begin{aligned}\widehat{\alpha}_k - \alpha_k =& \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\psi_k(W_{ig},\widehat{\eta}^k) - \alpha_k\\ =& \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\psi_k(W_{ig},\eta^k) - \alpha_k + \underbrace{\Big(\mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\psi_k(W_{ig},\widehat{\eta}^k)\Big] - \alpha_k\Big)}_{I_k}\\ &+ \underbrace{\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\Big\{\psi_k(W_{ig},\widehat{\eta}^k) - \psi_k(W_{ig},\eta^k) - \mathrm{E_P}[\psi_k(W_{ig},\widehat{\eta}^k) - \psi_k(W_{ig},\eta^k)]\Big\}}_{II_k}.\end{aligned}$$

It suffices to show that $|I_k|$ and $|II_k|$ are of order $o_{\mathrm{P}}(1/\sqrt{G})$ uniformly over $k \in [p]$ and uniformly in $\mathcal{P}_G$.

Step 1: Bound for $|I_k|$

To upperbound $|I_k|$, by applying the mean-value expansion and under Assumption 5 (i), there exists a vector $\ddot{\eta}^k$ with each of its coordinates lies between those of η^k and $\widehat{\eta}^k$ such that

$$\begin{aligned}I_k =& \mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\psi_k(W_{ig},\widehat{\eta}^k)\Big] - \alpha_k\\ =& \Big(\mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\psi_k(W_{ig},\eta^k)\Big] - \alpha_k\Big) + \partial_\eta \mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\psi_k(W_{ig},\eta^k)\Big](\widehat{\eta}^k - \eta^k)\\ &+ (\widehat{\eta}^k - \eta^k)'\Big\{\partial_{\eta'}\partial_\eta \mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\psi_k(W_{ig},\eta)\Big]\Big|_{\eta=\ddot{\eta}^k}\Big\}(\widehat{\eta}^k - \eta^k)\\ =& 0 + 0 + (\widehat{\eta}^k - \eta^k)'\Big\{\partial_{\eta'}\partial_\eta \mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\psi_k(W_{ig},\eta)\Big]\Big|_{\eta=\ddot{\eta}^k}\Big\}(\widehat{\eta}^k - \eta^k),\end{aligned}$$

where the last equality follows from existence condition and Neyman orthogonality condition defined in the end of Section A of this Appendix. Hence by the definition of induced matrix ℓ_2-norm and Assumptions 5 (ii)(b) and 6 (i), one has

$$|I_k| \le \Big\|\partial_{\eta'}\partial_\eta \mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\psi_k(W_{ig},\eta)\Big]\Big|_{\eta=\ddot{\eta}^k}\Big\|_2 \tau_G^2 \le B_{1G}\tau_G^2$$

uniformly in k with probability at least $1-\Delta_G$. This, together with Assumption 6 (iv), implies $\sqrt{G}I_k \leq \sqrt{G}B_{1G}\tau_G^2 \lesssim \delta_G$ with the same probability.

Step 2: Bound for $|II_k|$

Next, we upper bound $|II_k|$. Note that $1 \leq n_g \leq \bar{n}$ and n_g is predetermined. For each $g \in \{1, ..., G\}$, recall that W_g can be written as

$$W_g = \begin{cases} (W'_{1g}, \quad \dots \quad, W'_{\bar{n}g})' & \text{if } n_g = \bar{n}, \\ (W'_{1g}, ..., W'_{n_g g}, 0, ..., 0)' & \text{if } n_g < \bar{n}. \end{cases}$$

We can assume without loss of generality that some coordinate of X_{ig} is a random variable that is positive a.s. (otherwise replace $\mathbb{1}\{\|W_{ig}\|_\infty > 0\}$ by $\mathbb{1}\{\||W_{ig}|\|_\infty > 0\}$). Our goal is to find a uniform entropy bound for the class

$$\mathcal{F} = \Big\{W_g \mapsto \sum_{i=1}^{\bar{n}} \psi_k(W_{ig}, \eta) \cdot \mathbb{1}\{\|W_{ig}\|_\infty > 0\} \ : k = 1, ..., p, \ \eta \in \mathcal{H}_k\Big\}.$$

It allows us to apply the maximal inequality of Corollary 3 to obtain the desired bound. First, let us define

$$\mathcal{G}_j = \{W_g \mapsto \mathbb{1}\{\|W_{jg}\|_\infty > 0\}\}$$

For each j, such a class contains only one function. Thus each of them is a VC-subgraph class with VC index equals unity and themselves as their envelopes. Thus for any $0 < \epsilon \leq 1$, it holds that

$$\sup_Q \log N(\epsilon \|G_j\|_{Q,2}, \mathcal{G}_j, \|\cdot\|_{Q,2}) \lesssim 1 + \log(1/\epsilon).$$

Now we define

$$\mathcal{F}_j = \{W_g \mapsto \psi_k(W_{jg}, \eta) : k = 1, ..., p, \ \eta \in \mathcal{H}_k\}.$$

Apply Lemma K1(2) of Belloni, Chernozhukov, Fernández-Val and Hansen (2017) under Assumption 6 (ii), we have for each $1 \leq j \leq \bar{n}$, for all $0 < \epsilon \leq 1$,

$$\begin{aligned} &\sup_Q \log N(\epsilon \|F_j \cdot G_j\|_{Q,2}, \mathcal{F}_j \cdot \mathcal{G}_j, \|\cdot\|_{Q,2}) \\ \leq &\sup_Q \log N(\epsilon/2 \|F_j\|_{Q,2}, \mathcal{F}_j, \|\cdot\|_{Q,2}) + \sup_Q \log N(\epsilon/2 \|G_j\|_{Q,2}, \mathcal{G}_j, \|\cdot\|_{Q,2}) \\ \lesssim &v_G \log(a_G/\epsilon) + \log(1/\epsilon) + C. \end{aligned}$$

Also, we have $\mathcal{F} \subset \bar{\mathcal{F}} := (\mathcal{F}_1 \cdot \mathcal{G}_1 + ... + \mathcal{F}_{\bar{n}} \cdot \mathcal{G}_{\bar{n}})$.

Define the transformation $\phi(f_1, ..., f_{\bar{n}}) = \sum_{j=1}^{\bar{n}} f_j$. By the triangle inequality, one has $|\phi(f_1, ..., f_{\bar{n}}) - \phi(g_1, ..., g_{\bar{n}})| \leq \sum_{j=1}^{\bar{n}} 1 \cdot |f_j - g_j|$. Applying Lemma K1(4) of Belloni, Chernozhukov, Fernández-Val and Hansen (2017), for the envelope $F(W_g) = \sum_{j=1}^{\bar{n}} F_j(W_g) \cdot G_j(W_g) = \sum_{j=1}^{\bar{n}} F_0(W_{jg}) \mathbb{1}\{\|W_{jg}\|_\infty > 0\}$ for $\mathcal{F}$, we have

$$\begin{aligned}
&\sup_Q \log N(\epsilon \|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}) \\
\lesssim &\sum_{j=1}^{\bar{n}} \sup_Q \log N\Big(\frac{\epsilon}{\bar{n}} \|F_j \cdot G_j\|_{Q,2}, \mathcal{F}_j \cdot \mathcal{G}_j, \|\cdot\|_{Q,2}\Big) \\
\lesssim &\bar{n} v_G \log(a_G/\epsilon) + \bar{n} \log(1/\epsilon) + C.
\end{aligned}$$

Under Assumption 6 (ii), we have $\sqrt{\mathrm{E_P}\Big[\max_{g\in[G]} F^2(W_g)\Big]} \leq G^{1/q} K_G$. Now let

$$\bar{\mathcal{F}} = \Big\{ W_g \mapsto \sum_{j=1}^{\bar{n}} \Big(\psi_k(W_{jg}, \eta) - \psi_k(W_{jg}, \eta^k) \Big) \cdot \mathbb{1}\{\|W_{jg}\|_\infty > 0\} \ : k = 1, ..., p, \ \eta \in \mathcal{H}_k \Big\}.$$

Observe that since $\bar{\mathcal{F}} \subset \mathcal{F} - \mathcal{F}$, it holds that $\sup_{f\in\bar{\mathcal{F}}} |f| \leq 2 \sup_{f\in\mathcal{F}} |f| \leq 2F$. Assumption 6 (i),(ii) now implies

$$\begin{aligned}
\sup_{f\in\bar{\mathcal{F}}} \mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G} f^2(W_g)\Big] \lesssim &\sup_{\eta\in\mathcal{H}_k} \mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\{\psi_k(W_{ig}, \eta) - \psi_k(W_{ig}, \eta^k)\ \}\ \Big)^2\Big] \\
\leq &C_0 \|\eta - \widehat{\eta}^k\|_2^2 \lesssim \tau_G^2.
\end{aligned}$$

Under Assumptions 5 (ii)(a), 6 (ii), apply Lemma 8 (2) and Corollary 3, we have

$$\begin{aligned}
\sqrt{G}|II_k| \leq &\sup_{f\in\bar{\mathcal{F}}} \Big| \frac{1}{\sqrt{G}} \sum_{g=1}^{G}\sum_{i=1}^{n_g} [\, f(W_{ig}) - \mathrm{E} f(W_{ig})\,] \Big| \\
\lesssim &\tau_G \sqrt{v_G \log a_G} + \frac{K_G v_G \log a_G}{G^{-1/2+1/q}} \lesssim \delta_G,
\end{aligned}$$

uniformly in k with probability at least $1 - C(\log G)^{-1}$ as long as $G \geq 3$, where the last inequality follows from Assumption 6 (iv). Finally, the conclusion follows that $0 < \sigma_k < \infty$ uniformly from Assumptions 6 (i)(iii). ∎

C.2. **Proof for Corollary 2.**

Proof. Throughout the proof, let us use the notations of $\widehat{\varphi}_{gk} := \sum_{i=1}^{n_g} \widehat{\varphi}_j(W_{ig}, \widehat{\alpha}_k, \widehat{\eta}^k)$ and $\varphi_{gk} := \sum_{i=1}^{n_g} \varphi_j(W_{ig}, \alpha_k, \eta^k)$. Define

$$T = \max_{1\le k\le p} |\sqrt{G}\sigma_k^{-1}(\widehat{\alpha}_k - \alpha_k)|, \text{ and } T_0 = \max_{1\le k\le p} \frac{1}{\sqrt{G}} \sum_{g=1}^{G}\sum_{i=1}^{n_g} \varphi_k(W_{ig}, \alpha_k, \eta^k).$$

Step 1: In this step, we bound

$$\rho := \sup_{t\in\mathbb{R}} |\mathrm{P}_{\mathrm{P}}(T_0 \le t) - \mathrm{P}_{\mathrm{P}}(Z_0 \le t)|.$$

First invoke Corollary 2.1 of Chernozhukov, Chetverikov and Kato (2013) under Assumption 7 (i) and obtain

$$\rho = \sup_{t\in\mathbb{R}} |\mathrm{P}_{\mathrm{P}}(T_0 \le t) - \mathrm{P}_{\mathrm{P}}(Z_0 \le t)| \le CG^{-c}$$

for $Z_0 = \max_{1\le k\le p} G^{-1/2}\sum_{g=1}^{n} Y_{gk}$, where $\{Y_g\}_{g=1}^G$ are independently distributed p-dimensional centered Gaussian random vector such that $G^{-1/2}\sum_{g=1}^G Y_{gk}$ has the same covariance matrix as $G^{-1/2}\sum_{g=1}^G \varphi_{gk}$.

Step 2: In this step, we show

$$\mathrm{P}_{\mathrm{P}}(|T - T_0| > \theta_1) < \theta_2 \tag{C.24}$$

$$\mathrm{P}_{\mathrm{P}}(\mathrm{P}_{\xi}(|W - W_0| > \theta_1) > \theta_2) < \theta_2. \tag{C.25}$$

for some appropriate θ_1, $\theta_2 = o(1)$. Set $\theta_1 = \delta_G \vee C\bar{\delta}_G\sqrt{\bar{\rho}_G \log \bar{A}_G} \ge \delta_G$ and $\theta_2 = \Delta_G + (\log G)^{-1} \ge \Delta_G + (\log G)^{-1}$. By Theorem 5 (recall $q \ge 2$ and $G \ge 3$), (C.24) holds. We now claim (C.25). We first show that

$$|W - W_0| \le \max_{1\le k\le p} \Big|\frac{1}{\sqrt{G}}\sum_{g=1}^G \xi_g(\widehat{\varphi}_{gk} - \varphi_{gk})\Big| = O_{\mathrm{P}}\Big(\bar{\delta}_G\sqrt{\bar{\rho}_G \log \bar{A}_G}\Big) \tag{C.26}$$

Call the event in Assumption 7 (ii) Ω_1. Note that $\mathrm{P}_{\mathrm{P}}(\Omega_1) \ge 1 - \Delta_G$. Conditional on Ω_1, $G^{-1/2}\sum_{g=1}^G(\widehat{\varphi}_{gk} - \varphi_{gk})\xi_g$ is zero-mean Gaussian with variance $\mathrm{E}_{\mathrm{P}}\frac{1}{G}\sum_{g=1}^G[\widehat{\varphi}_{gk} - \varphi_{gk}]^2 \le \bar{\delta}_G^2$ uniformly in k with probability at least $1 - \Delta_G$ following Assumption 7 (iii). Therefore, applying Corollary 2.2.8 of van der Vaart and Wellner (1996), conditional on Ω_1, one has

$$\mathrm{E}_{\xi}\Big[\max_{1\le k\le p}\Big|\frac{1}{\sqrt{G}}\sum_{g=1}^G \xi_g(\widehat{\varphi}_{gk} - \varphi_{gk})\Big|\Big] \lesssim \bar{\delta}_G\sqrt{\bar{\rho}_G \log \bar{A}_G}.$$

where the expectation is taken with respect to the law of ξ_g's (recall that ξ_g's are generated independently from the data). By Corollary 3, conditional on Ω_1, the left hand side of equation (C.26) is less than

$$2C\bar{\delta}_G\sqrt{\bar{\rho}_G \log \bar{A}_G} + K\bar{\delta}_G \log G \lesssim \bar{\delta}_G\sqrt{\bar{\rho}_G \log \bar{A}_G}$$

with probability at least $1-(\log G)^{-1}$. Thus for some constant C large enough, conditional on Ω_1, one has

$$\mathrm{P}_\xi\Big(\max_{1\le k\le p}\Big|\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\xi_g(\widehat{\varphi}_{gk}-\varphi_{gk})\Big| > C\bar{\delta}_G\sqrt{\bar{\rho}_G \log \bar{A}_G}\Big) \lesssim (\log G)^{-1} < (\log G)^{-1} + \Delta_G,$$

which means the left hand side is greater than $(\log G)^{-1} + \Delta_G$ only if Ω_1^c is true. Recall that $\theta_1 \ge C\bar{\delta}_G\sqrt{\bar{\rho}_G \log \bar{A}_G}$ and $\theta_2 = (\log G)^{-1} + \Delta_G$, therefore

$$\mathrm{P}_{\mathrm{P}}\Big(\mathrm{P}_\xi\Big(\max_{1\le k\le p}\Big|\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\xi_g(\widehat{\varphi}_{gk}-\varphi_{gk})\Big| > \theta_1\Big) > \theta_2\Big) \le \mathrm{P}_{\mathrm{P}}(\Omega_1^c) \le \Delta_G < \theta_2.$$

This verifies condition (C.25).

Step 3: Here we establish bootstrap validity based on the results from preceding steps. Under Assumptions 5 (ii), 6 (ii), one can apply Theorem 3.2 of Chernozhukov, Chetverikov and Kato (2013) and obtains that for every $\vartheta > 0$,

$$\sup_{\alpha\in(0,1)}|\mathrm{P}_{\mathrm{P}}(T \le c_W(\alpha)) - \alpha| \le \rho_{\ominus} + \rho,$$

where

$$\rho_{\ominus} \le 2(\rho + \pi(\vartheta) + \mathrm{P}_{\mathrm{P}}(\Delta > \vartheta)) + C_3\theta_1\sqrt{1 \vee \log(p/\theta_1)} + 5\theta_2,$$

$\pi(\vartheta) := C_2\vartheta^{1/3}(1 \vee \log(p/\vartheta))^{2/3}$ and $\Delta := \max_{1\le j,l\le p}|\frac{1}{G}\sum_{g=1}^{G}([\varphi_{gj}\varphi_{gl}] - \mathrm{E}_{\mathrm{P}}[\varphi_{gj}\varphi_{gl}])|$. It then suffices to show that each component on the right hand side goes to zero.

First, set $\vartheta = B_G^2(\log p)^{3/2}/\sqrt{G}$ and note that Assumption 7 (i) implies $B_G^4(\log p)^7/G = o(1)$ and thus $\vartheta^{1/2}\log p = o(1)$. By l'Hôspital's rule, $\vartheta = o(1)$ implies $\vartheta^{1/2}\log\vartheta = o(1)$. So $\pi(\vartheta) \lesssim \vartheta^{1/3}(\log p - \log\vartheta)^{2/3} = o(1)$. Similarly, $\theta_1^2\log\theta_1 = o(1)$ as long as $\theta_1 = o(1)$. By Assumption 7 (iii), set $\theta_1 = \delta_G \vee \bar{\delta}_G\sqrt{\bar{\rho}_G \log \bar{A}_G}$, then $\theta_1\sqrt{\log p} = o(1)$ and we conclude that $\theta_1\sqrt{1 \vee \log(p/\theta_1)} = o(1)$. Secondly, we verify $\mathrm{P}_{\mathrm{P}}(\Delta > \vartheta) = o(1)$. Under Assumption 7 (i), Lemma C.1. of Chernozhukov, Chetverikov and Kato (2013) implies

$$\mathrm{E}_{\mathrm{P}}[\Delta] \lesssim (B_G^2(\log p)/G)^{1/2} \vee B_G^2(\log p)/\sqrt{G}.$$

Apply the Markov's inequality, as long as $B_G^4(\log p)^7/G = o(1)$, we have $\mathrm{P}_{\mathrm{P}}(\Delta > \vartheta) = o(1)$ and $\vartheta(\log p)^2 = o(1)$. This concludes the proof. ■

Appendix D. Proofs for Results in Section 5

D.1. **Proof for Theorem 1.**

Proof. For each $k \in [p]$, let us define $\mathcal{H}_k = \mathcal{H}_k^G$, the bounded and convex sparse subset in $\mathbb{R}^p$ shrinking with G, as follows

$$\begin{aligned}\mathcal{H}_k = \{\eta^k\} \cup \Big\{ (\eta^{(1)}, \eta^{(0)}) \in \mathbb{R}^{2p} : \eta^{(0)} = \eta^{(2)} + \eta^{(3)},\ &\|\eta^{(1)}\|_0 \vee \|\eta^{(2)}\|_0 \vee \|\eta^{(3)}\|_0 \leq Cs, \\ &\|\eta^{(1)} - \beta^0\|_1 \vee \|\eta^{(2)} - \zeta^k\|_1 \vee \|\eta^{(3)} - \theta^k\|_1 \leq C\sqrt{s}\tau_G, \\ &\|\eta^{(1)} - \beta^0\|_2 \vee \|\eta^{(2)} - \zeta^k\|_2 \vee \|\eta^{(3)} - \theta^k\|_2 \leq C\tau_G \qquad \Big\},\end{aligned}$$

where $\tau_G = C'(s \log a_G / G)^{1/2}$ for some sufficiently large positive constants C and C'. The rest of this proof is divided into three steps corresponding to the verification of the three high-level assumptions.

Step 1. In this step, we examine Assumption 5. Assumption 5 (i) is clear since Λ is infinitely continuously differentiable. To verify Assumption 5 (ii)(a), since for all $\eta^k \in \mathcal{H}_k$, $\|\eta^k - \eta\|_2 \lesssim 1$, using a mean value expansion and the definition of the induced matrix ℓ_2 norm,

$$\begin{aligned}&\mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\psi_k(W_{ig},\eta) - \psi_k(W_{ig},\eta^k)\Big)^2\Big] \\ =&\mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\partial_\eta\psi_k(W_{ig},\widetilde{\eta})'(\eta - \eta^k)\Big)^2\Big] \\ =&(\eta - \eta^k)'\mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\partial_\eta\psi_k(W_{ig},\widetilde{\eta}))\Big)\Big(\sum_{i=1}^{n_g}\partial_\eta\psi_k(W_{ig},\widetilde{\eta})\Big)'\Big](\eta - \eta^k) \\ \leq&\|\eta - \eta^k\|_2^2\Big\|\mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\partial_\eta\psi_k(W_{ig},\widetilde{\eta}))\Big)\Big(\sum_{i=1}^{n_g}\partial_\eta\psi_k(W_{ig},\widetilde{\eta})\Big)'\Big]\Big\|_2.\end{aligned}$$

where each coordinate of $\widetilde{\eta}$ lies between the corresponding coordinate of η and η^k. By Assumption 1 and the definition of τ_G, we know $\|\widetilde{\eta}\|_2 \lesssim 1$. Notice

$$\begin{aligned}\sum_{i=1}^{n_g}\partial_\eta\psi_k(W_{ig},\widetilde{\eta}) =& \sum_{i=1}^{n_g}\begin{bmatrix}\partial_\beta\psi_k(W_{ig},\widetilde{\eta}) \\ \partial_\mu\psi_k(W_{ig},\widetilde{\eta})\end{bmatrix} = \sum_{i=1}^{n_g}\begin{bmatrix}\widetilde{\beta}_k\Lambda''(X_{ig}'\widetilde{\beta})X_{ig} + \Lambda'(X_{ig}'\widetilde{\beta})e_k + \widetilde{\mu}'X_{ig}\Lambda'(X_{ig}'\widetilde{\beta})X_{ig} \\ -\{Y_{ig} - \Lambda(X_{ig}'\widetilde{\beta})\}X_{ig}\end{bmatrix} \\ =& \begin{bmatrix}A_g + B_g + C_g \\ D_g\end{bmatrix}.\end{aligned}$$

Thus, the sum of cross products can be denoted by

$$\Big\|\mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\partial_\eta\psi_k(W_{ig},\widetilde{\eta})\Big)\Big(\sum_{i=1}^{n_g}\partial_\eta\psi_k(W_{ig},\widetilde{\eta})\Big)'\Big]\Big\|_2$$

$$=\left\|\mathrm{E_P}\frac{1}{G}\sum_{g=1}^{G}\begin{bmatrix}(A_g+B_g+C_g)(A_g+B_g+C_g)' & (A_g+B_g+C_g)D_g' \\ (A_g+B_g+C_g)'D_g & D_gD_g'\end{bmatrix}\right\|_2.$$

To further bound the right-hand side, it suffices to bound the matrix ℓ_2 norm for each of the product terms. Under Assumption 3 (4) and $\|\widetilde{\mu}\|_2 \lesssim 1$, we have

$$\Big\|\mathrm{E_P}\frac{1}{G}\sum_{g=1}^{G}C_gC_g'\Big\|_2 \le \Big\|\mathrm{E_P}\frac{1}{G}\sum_{g=1}^{G}(\widetilde{\mu}'U_g)^2U_gU_g'\Big\|_2 \lesssim \sup_{\|\xi\|_2=1}\mathrm{E_P}\frac{1}{G}\sum_{g=1}^{G}[(\widetilde{\mu}'U_g)^2\xi'U_gU_g'\xi]$$

$$\le\Big\{\mathrm{E_P}\frac{1}{G}\sum_{g=1}^{G}(\widetilde{\mu}'U_g)^4\Big\}^{1/2}\max_{\|\xi\|_2=1}\Big\{\mathrm{E_P}\frac{1}{G}\sum_{g=1}^{G}(\xi'U_g)^4\Big\}^{1/2} \le C_1, \text{ and}$$

$$\Big\|\mathrm{E_P}\frac{1}{G}\sum_{g=1}^{G}A_gA_g'\Big\|_2 \vee \Big\|\mathrm{E_P}\frac{1}{G}\sum_{g=1}^{G}B_gB_g'\Big\|_2 \vee \Big\|\mathrm{E_P}\frac{1}{G}\sum_{g=1}^{G}D_gD_g'\Big\|_2 \lesssim \Big\|\sup_{\|\xi\|_2=1}\mathrm{E_P}\frac{1}{G}\sum_{g=1}^{G}(U_g'\xi)^2\Big\|_2 \le C_1.$$

This shows Assumption 5 (ii)(a).

To verify Assumption 5 (ii)(b), note that we can write the matrix

$$\sum_{i=1}^{n_g}\partial_{\mu'}\partial_\mu\psi_k(W_{ig},\widetilde{\eta}) = \begin{bmatrix}A_g & B_g \\ B_g & 0\end{bmatrix}$$

$$=\sum_{i=1}^{n_g}\begin{bmatrix}\widetilde{\beta}_k\Lambda'''(X_{ig}'\widetilde{\beta})X_{ig}X_{ig}' + \Lambda''(X_{ig}'\widetilde{\beta})e_kX_{ig}' + \Lambda''(X_{ig}'\widetilde{\beta})X_{ig}e_k' + \widetilde{\mu}'X_{ig}\Lambda''(X_{ig}'\widetilde{\beta})X_{ig}X_{ig}' & \Lambda'(X_{ig}'\widetilde{\beta})X_{ig}X_{ig}' \\ \Lambda'(X_{ig}'\widetilde{\beta})X_{ig}X_{ig}' & 0\end{bmatrix}.$$

So for $\eta = [\beta', \mu']'$, we have

$$\Big\|\partial_{\mu'}\mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\partial_\mu\psi_k(W_{ig},\widetilde{\eta})\Big]\Big\|_2 = \max_{0<\|\xi\|_2\le 1}\xi'\mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\partial_{\mu'}\partial_\mu\psi_k(W_{ig},\widetilde{\eta})\Big]\xi$$

$$\le \max_{0<\|\xi\|_2\le 1}\Big(\beta'\mathrm{E_P}\frac{1}{G}\sum_{g=1}^{G}A\beta + 2\beta'\mathrm{E_P}\frac{1}{G}\sum_{g=1}^{G}B\eta\Big).$$

By Cauchy-Schwarz and the definition of induced matrix ℓ_2 norm, $\beta'\mathrm{E_P}[\frac{1}{G}\sum_{g=1}^{G}B_g]\eta \le \|\beta'\mathrm{E_P}[\frac{1}{G}\sum_{g=1}^{G}B_g]\|_2\|\eta\|_2 \le \|\beta\|_2\|\mathrm{E_P}[\frac{1}{G}\sum_{g=1}^{G}B_g]\|_2\|\eta\|_2$. Thus a bound can be obtained similarly to 5 (ii)(a).

Step 2. In this step, we check Assumption 6. To verify Assumption 6 (i), note that set $\Delta_G = C(\log G)^{-1}$ and $\tau_G = (s\log a_G/G)^{1/2}$, it follows from the convergence rate results of Theorems 2, 4 and Corollary 1.

To verify Assumption 6 (ii), recall that

$$\bar{\psi}_k(W_{ig},\alpha_k,\eta) = \alpha_k - \beta_k\Lambda'(X_{ig}'\beta) + \mu'\{Y_{ig} - \Lambda(X_{ig}'\beta)\}X_{ig}$$

Pointwise measurability follows from its continuity. Let

$$\begin{aligned}
\mathcal{G}_{1iT} =& \Big\{W_g \mapsto Y_{ig} - \Lambda(X_{ig}'\beta) : \beta \in \mathbb{R}^p, T = \text{support}(\beta),\ \|\beta-\beta^0\|_2 \le C\tau_G\Big\},\\
\mathcal{G}_{2iT} =& \Big\{W_g \mapsto \Lambda'(X_{ig}'\beta) : \beta \in \mathbb{R}^p, T = \text{support}(\beta),\ \|\beta-\beta^0\|_2 \le C\tau_G\Big\},\\
\mathcal{G}_{3ikT} =& \Big\{W_g \mapsto \mu' X_{ig} : \mu \in \mathbb{R}^p, T = \text{support}(\mu),\ \|\mu-\mu^k\|_2 \le C\tau_G\Big\},\\
\mathcal{G}_4 =& \Big\{W_g \mapsto b : |b| \le C\Big\},\\
\mathcal{G}_5 =& \Big\{W_g \mapsto \alpha : |\alpha| \le C\Big\},\\
\mathcal{G}_{6i} =& \Big\{W_g \mapsto \Lambda'(X_{ig}'\beta^0)\Big\},\\
\mathcal{G}_{7ik} =& \Big\{W_g \mapsto \mu^{k\prime} X_{ig}\Big\},\\
\mathcal{G}_{8i} =& \Big\{W_g \mapsto Y_{ig} - \Lambda(X_{ig}'\beta^0)\Big\}.
\end{aligned}$$

Then the following holds

$$\begin{aligned}
&\mathcal{F}_0 \subset \mathcal{F}_0' \cup \mathcal{F}_0'' \cup \{0\},\\
&\mathcal{F}_0' = \mathcal{G}_5 - (\mathcal{G}_4)\cdot \sum_{i=1}^{\bar{n}} (\cup_{T\subset[p],|T|\le Cs}\mathcal{G}_{2iT}) + \sum_{i=1}^{\bar{n}} (\cup_{k\in[p]} \cup_{T\subset[p],|T|\le Cs} \mathcal{G}_{3ikT})\cdot(\cup_{T\subset[p],|T|\le Cs}\mathcal{G}_{1iT}),\\
&\mathcal{F}_0'' = \mathcal{G}_5 - (\mathcal{G}_4)\cdot \sum_{i=1}^{\bar{n}} \mathcal{G}_{6i} + \sum_{i=1}^{\bar{n}} (\cup_{k\in[p]}\mathcal{G}_{7ik})\cdot(\mathcal{G}_{8i}).
\end{aligned}$$

Note that all these classes are uniformly bounded with the exceptions of $\mathcal{G}_{3ikT}$ and $\mathcal{G}_{7ik}$. To obtain envelopes for them, note that all classes are uniformly bounded except for $\mathcal{G}_{3ikT}$ and $\mathcal{G}_{7ik}$. To obtain an envelope for $\mathcal{G}_{3ikT}$, notice for any ikT, $\|\mu^k\|_1 \le \sqrt{s}C_1$ since $\|\mu^k\|_2 \le C_1$ following Assumption 1. Therefore, $\|\mu\|_1 \le \|\mu^k\|_1 + \|\mu-\mu^k\|_1 \le \sqrt{s}C_1 + \sqrt{s}C_1\tau_G \lesssim \sqrt{s}$. Set envelope G to be such that

$$G(W_g) = \max_{k\in[p]} \max_{i\in[\bar{n}]} \sup_{,\mu\in\mathbb{R}^p:\|\mu-\mu^k\|_1\le C\sqrt{s}\tau_G} |\mu' X_{ig}|,$$

then for any μ in the index set, one has

$$|\mu' X_{ig}| \le |(\mu-\mu^k)' X_{ig}| + |\mu^{k\prime} X_{ig}| \lesssim C\sqrt{s}\tau_G\|U_g\|_\infty + |\mu^{k\prime} X_{ig}|.$$

Since $\mu^k = \zeta^k + \theta^k$ and $\theta^k = [-\gamma^k_1, ..., -\gamma^k_{k-1}, 1, -\gamma^k_{k+1}, ..., -\gamma^k_{p-1}]/\tau^2_k$ and $\tau_k^{-2} = O(1)$ following from Assumption 3 (1), we have

$$\begin{aligned}|\mu^{k\prime} X_{ig}| \leq & |\zeta^{k\prime} X_{ig}| + |\theta^{k\prime} X_{ig}| \\ \lesssim & |S^k_{ig}| + |\varepsilon^k_{ig}| + |D^k_{ig}| + |\gamma^{k\prime} X^k_{ig}| \\ \lesssim & 1 + \|V_g\|_\infty + \|U_g\|_\infty,\end{aligned}$$

where the last inequality is due to $|\gamma^{k\prime} X^k_{ig}| \leq |D^k_{ig}| + |Z^k_{ig}|$ and the definition of V_g, U_g. Now, Assumption 3 (6) implies $\sqrt{s}\tau_G = o(1)$, thus the above implies

$$|\mu' X_{ig}| \lesssim \|V_g\|_\infty + \|U_g\|_\infty \leq 2(\|V_g\|_\infty \vee \|U_g\|_\infty).$$

Under Assumption 3 (5)(7),

$$\begin{aligned}\Big\{\mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^G G^q(W_g)\Big]\Big\}^{1/q} \lesssim & \Big\{\mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^G (\|V_g\|_\infty \vee \|U_g\|_\infty)^q\Big]\Big\}^{1/q} \leq \Big\{\mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^G (\|V_g\|_\infty \vee \|U_g\|_\infty)^{2q}\Big]\Big\}^{1/2q} \\ \lesssim & M_{G,1} \vee M_{G,2}. \qquad \text{(D.27)}\end{aligned}$$

Similar argument holds for $\mathcal{G}_{7ik}$ as well.

To obtain a bound for the uniform covering entropy number, first let us consider $\mathcal{G}_{3ikT}$ for some fixed $i, k \in [p]$ and $|T| \leq Cs$. Applying Lemma 21 of Kato (2017), we have that each $\mathcal{G}_{3ikT}$ is a VC-subgraph class of functions with VC-index $Cs + 2 = O(s)$. Thus the union of these $p \cdot \binom{p}{Cs}$ class of functions is a VC-type class and has uniform covering number satisfying that for any $0 < \varepsilon \leq 1$,

$$\sup_Q N(\varepsilon \|\widetilde{G}\|_{Q,2}, \cup_{k\in[p]} \cup_{T\subset[p],|T|\leq Cs} \mathcal{G}_{3ikT}, \|\cdot\|_{Q,2}) \lesssim a_G^{Cs+2}\Big(\frac{A}{\varepsilon}\Big)^{Cs}.$$

Thus we have

$$\log \sup_Q N(\varepsilon \|\widetilde{G}\|_{Q,2}, \cup_{k\in[p]} \cup_{T\subset[p],|T|\leq Cs} \mathcal{G}_{3ikT}, \|\cdot\|_{Q,2}) \lesssim s\log(a_G/\varepsilon).$$

Similar entropy calculations hold for $\mathcal{G}_{1iT}$ and $\mathcal{G}_{2iT}$ as well since Λ is monotone and $\Lambda' = \Lambda \cdot (1 - \Lambda)$ and thus Lemma 22 of Kato (2017) and Lemma 8 can be applied.

To verify Assumption 6 (iii), note that the lower bound is implied by Assumption 3 (1). It then suffices to bound $\mathrm{E_P}\frac{1}{G}\sum_{g=1}^G(\mu' X_{ig})^2 \leq \{\mathrm{E_P}\frac{1}{G}\sum_{g=1}^G(\mu' X_{ig})^4\}^{1/2} \lesssim 1$ for all $\|\mu\|_2 \leq C$. This follows directly from Assumption 3 (4).

To verify Assumption 6 (iv), note that set $v_G = s$, $\tau_G = (s \log a_G/G)^{1/2}$, $K_G = M_{G,1} \vee M_{G,2}$, then by Assumption 3 (6)(8) we have

$$\frac{s \log a_G}{G^{1/2}} \vee \frac{s(M_{G,1} \vee M_{G,2}) \log a_G}{G^{1/2-1/q}} \lesssim \frac{s(M_{G,1} \vee M_{G,2}) \log a_G}{G^{1/2-1/q}} \lesssim \delta_G$$

for $\delta_G = (\log a_G)^{-2}$.

Step 3. In this step, we examine Assumption 7. To verify Assumption 7 (i), we need to find B_G such that $B_G^4(\log(pG))^7/G \leq C_1 G^{-c_1}$ for positive constants c_1, C_1 and for all $1 \leq g \leq G$ and $k \in [p]$

$$\max_{b=1,2} \mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\Big|\sum_{i=1}^{n_g}(\alpha_k - \beta_k^0\Lambda(X_{ig}'\beta^0) + \mu^{k\prime}X_{ig}\{Y_{ig} - \Lambda(X_{ig}'\beta^0)\})\Big|^{2+b}/\sigma_k^{2+b}B_G^b\Big]$$

$$+\mathrm{E_P}\Big(\max_{1\leq k\leq p}\Big|\sum_{i=1}^{n_g}(\alpha_k - \beta_k^0\Lambda(X_{ig}'\beta^0) + \mu^{k\prime}X_{ig}\{Y_{ig} - \Lambda(X_{ig}'\beta^0)\})\Big|/\sigma_k B_G\Big)^4 \leq 4.$$

Assumption 6 (iii) implies that $1 \lesssim \sigma_k \lesssim 1$. For the first term, note that it suffices to bound $\sup_{\|\mu\|_2=1} \mathrm{E_P}\frac{1}{G}\sum_{g=1}^{G}(U_g'\mu)^{2+b}$, which is bounded under Assumption 3 (4). So the entire first term is bounded by a constant. Now, for the second term, note that using (D.27),

$$\begin{aligned}\max_{g\in[G]}\mathrm{E_P}\Big(\max_{1\leq k\leq p}\Big|\sum_{i=1}^{n_g}(\alpha_k - \beta_k^0\Lambda(X_{ig}'\beta^0) + \mu^{k\prime}X_{ig}\{Y_{ig} - \Lambda(X_{ig}'\beta^0)\})\Big|\Big)^4 \lesssim& \mathrm{E_P}\Big[\sum_{g=1}^{G}G^4(W_g)\Big]\\ \lesssim& G^{2/q}(M_{G,1} \vee M_{G,2})^4.\end{aligned}$$

So take $B_G = CG^{1/2q}(M_{G,1} \vee M_{G,2})$ for some C large enough, we have

$$\frac{B_G^4(\log(pG))^7}{G} \lesssim \frac{(M_{G,1} \vee M_{G,2})^4(\log a_G)^7}{G^{1-2/q}} \lesssim G^{-c_1}$$

under the rate condition in the statement of Theorem 1.

To verify Assumption 7 (ii), note that both

$$\bar{\mathcal{F}} = \Big\{W_g \mapsto \sum_{i=1}^{\bar{n}}\varphi_k(W_{ig}, \alpha_k, \eta^k)\mathbb{1}\{\||W_{ig}|\|_\infty > 0\} : k \in [p]\Big\},$$

$$\widehat{\mathcal{F}} = \Big\{W_g \mapsto \sum_{i=1}^{\bar{n}}\varphi_k(W_{ig}, \widehat{\alpha}_k, \widehat{\eta}^k)\mathbb{1}\{\||W_{ig}|\|_\infty > 0\} : k \in [p]\Big\},$$

contains only at most p functions. Thus, for any $0 < \varepsilon \leq 1$

$$\log N(\widehat{\mathcal{F}}_0, \|\cdot\|_{\mathbb{P}_G}, \varepsilon) \lesssim \log(p/\varepsilon) \leq \bar{\rho}\log(\bar{A}_G/\varepsilon)$$

for $\bar{\rho} = \log G$ and $\bar{A}_G = a_G$. Also, by Assumption 6 (i), with probability $1 - C(\log G)^{-1}$, we have $\widehat{\eta}^k \in \mathcal{H}_k$ and thus by Assumption 5 (ii)(a), for any $f \in \widehat{\mathcal{F}}_0$, $\mathrm{E_P}\frac{1}{G}\sum_{g=1}^{G} f^2 \lesssim \tau_G^2$ so Assumption 7 (ii) holds by setting $\bar{\delta}_G = \tau_G = (s\log a_G/G)^{1/2}$.

Finally, Assumption 7 (iii) is satisfied by setting $\bar{\delta}_G = \tau_G = (s\log a_G/G)^{1/2}$, $v_G = \bar{\rho}_G = s$, $\bar{A}_G = a_G$ and $\delta_G = (\log a_G)^{-2}$ under Assumption 3 (8). ■

D.2. **Proof for Lemma 1.**

Proof. Notice that Assumption 3 (4) implies that $\sigma_k \lesssim 1$. By the continuous mapping theorem, it suffices to bound

$$\begin{aligned}
|\widetilde{\sigma}_k^2 - \sigma_k^2| \leq & \Big|\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\bar{\psi}_k(W_{ig},\widetilde{\alpha}_k,\widetilde{\eta}^k)\Big)^2 - \mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\bar{\psi}_k(W_{ig},\alpha_k,\eta^k)\Big)^2\Big]\Big| \\
\lesssim & |\widetilde{\alpha}_k^2 - \alpha_k^2| + \sup_{\eta\in\mathcal{H}_k}|\widetilde{\alpha}_k - \alpha_k|\Big|\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\psi_k(W_{ig},\eta)\Big| \\
& + \sup_{\eta\in\mathcal{H}_k}\Big|\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\psi_k(W_{ig},\eta)\Big)^2 - \mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\psi_k(W_{ig},\eta)\Big)^2\Big]\Big| \\
& + \sup_{\eta\in\mathcal{H}_k}\Big|\mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\Big\{\Big(\sum_{i=1}^{n_g}\psi_k(W_{ig},\eta)\Big)^2 - \Big(\sum_{i=1}^{n_g}\psi_k(W_{ig},\eta^k)\Big)^2\Big\}\Big]\Big| \\
= & (I) + (II) + (III) + (IV)
\end{aligned}$$

uniformly over $k \in [p]$. First of all, $(IV) = o_{\mathrm{P}}(1)$ by the Lipschitzness of $\mathrm{E_P}[\frac{1}{G}\sum_{g=1}^{G}(\sum_{i=1}^{n_g}\psi_k(W_{ig},\cdot))^2]$ and Assumption 5, which is verified in Theorem 1. To bound (III), let the collection of functions

$$\mathcal{F} = \Big\{W_g \mapsto \sum_{i=1}^{\bar{n}}\sum_{j=1}^{\bar{n}}\psi_k(W_{ig},\eta)\psi_k(W_{jg},\eta)\mathbb{1}\{|||W_{ig}|||_\infty \wedge |||W_{jg}|||_\infty > 0\} : k \in [p], \eta \in \mathcal{H}_k\Big\}.$$

Under Assumption 3 (1)(5)(6)(7), using a similar argument as in Theorem 1, we obtain

$$\sup_Q N(\varepsilon\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}) \lesssim \Big(\frac{A}{\varepsilon}\Big)^{Cs}$$

with an envelope F defined by

$$F(W_g) = \max_{k\in[p]}\max_{i\in[\bar{n}]}\sup_{\mu\in\mathbb{R}^p:\|\mu-\mu^k\|_1\leq C\sqrt{s}\tau_G}\bar{n}^2|\mu' X_{ig}|^2 + C_3$$

for some constant C_3. Furthermore, under Assumption 3 (5)-(8), it holds that

$$\begin{aligned}
\mathrm{E_P}[\max_{g\in[G]} F^2(W_g)] \lesssim & \mathrm{E_P}[\max_{g\in[G]}\max_{k\in[p]} \sup_{\mu\in\mathbb{R}^p:\|\mu-\mu^k\|_1\leq C\sqrt{s}\tau_G} |(\mu-\mu^k)'U_g|^4] + \mathrm{E_P}[\max_{g\in[G]}\max_{i\in[\bar{n}]}\max_{k\in[p]}(\mu^{k\prime}X_{ig})^4] \\
\lesssim & G^{2/q}\Big(\mathrm{E_P}\frac{1}{G}\sum_{g=1}^{G} s^2\tau_G^4\|U_g\|_\infty^{2q} + (M_{G,1}\vee M_{G,2})^{2q}\Big)^{2/q} \\
\lesssim & G^{2/q}\{s^2\tau_G^4 M_{G,2}^4 + (M_{G,1}\vee M_{G,2})^4\} \lesssim G^{2/q}(M_{G,1}\vee M_{G,2})^4.
\end{aligned}$$

Applying Corollary 3 under Assumption 3 (4)(5)(6)(7)(8), with probability at least $1-C(\log G)^{-1}$, we have

$$\begin{aligned}
\sup_{\eta\in\mathcal{H}_k}\Big|\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\psi_k(W_{ig},\eta)\Big)^2 - \mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\psi_k(W_{ig},\eta^k)\Big)^2\Big]\Big| \lesssim & \sqrt{\frac{s\log a_G}{G}} + \frac{(M_{G,1}\vee M_{G,2})^2 s\log a_G}{G^{1-1/q}} \\
\lesssim & o(\log^{-1} a_G).
\end{aligned}$$

To bound (I) and (II), note that Theorem 5 suggests

$$\widetilde{\alpha}_k - \alpha_k = \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\psi_k(W_{ig},\eta^k) + o_{\mathrm{P}}(G^{-1/2})$$

uniformly in $k\in[p]$. We may apply Corollary 3 to

$$\mathcal{F} = (\mathcal{G}_4)\cdot\sum_{i=1}^{\bar{n}}\mathcal{G}_{6i} + \sum_{i=1}^{\bar{n}}(\cup_{k\in[p]}\mathcal{G}_{7ik})\cdot(\mathcal{G}_{8i})$$

as the components $\mathcal{G}$'s are defined in the proof of Theorem 1. An envelope can be

$$F(W_g) = \max_{k\in[p]}\max_{i\in[\bar{n}]}\sup_{\mu\in\mathbb{R}^p:\|\mu-\mu^k\|_1\leq C\sqrt{s}\tau_G} C(1+\mu' X_{ig})$$

for some C that does not depend on G. It is then implied by (D.27) that $\{\mathrm{E_P}\frac{1}{G}\sum_{g=1}^{G}F^q\}^{1/q}\lesssim M_{G,1}\vee M_{G,2}$ and thus $\sqrt{\mathrm{E_P}[\max_{g\in[G]}F^2(W_g)]}\lesssim G^{1/2q}(M_{G,1}\vee M_{G,2})$. We also have $\sup_{f\in\mathcal{F}}\mathrm{E_P}\frac{1}{G}\sum_{g=1}^{G}f^2\lesssim C+\sup_{\xi\in\mathbb{R}^p:\|\xi\|_2=1}\mathrm{E_P}\frac{1}{G}\sum_{g=1}^{G}(U_g'\xi)^2\lesssim 1$.

Applying Corollary 3 under Assumption 3 (4)(5)(6)(7)(8) leads to

$$\max_{k\in[p]}|\widetilde{\alpha}_k-\alpha_k| \lesssim \sqrt{\frac{\log a_G}{G}} + \frac{s(M_{G,1}\vee M_{G,2})\log a_G}{G^{1-1/2q}} = o(\log^{-1} a_G)$$

with probability at least $1-C(\log G)^{-1}$. This implies $(I)+(II)=o_{\mathrm{P}}(1)$. ■

Appendix E. Proofs for Results in Section 6

E.1. **Proof for Theorem 2.** The steps of this proof are analogous to the one of Theorem 4.1 in Belloni, Chernozhukov, Chetverikov and Wei (2018) with modifications to account for cluster sampling. The major difference lies in the verification of Assumption 8 (2).

Proof. We will apply Lemma 3, 4 and 5 after verifying the required assumptions. Let $w_{ig} = f_{ig}^2$ and

$$M(Y_{ig}, X_{ig}, \beta) = \widehat{M}(Y_{ig}, X_{ig}, \beta) = -\{Y_{ig} X_{ig}' \beta - \log(1 + \exp(X_{ig}' \beta))\}.$$

In order to apply Lemma 6, we verify Assumption 9. Since

$$\begin{aligned} S_g =& \partial_\beta M(Y_{ig}, X_{ig}, \beta)|_{\beta=\beta^0} \\ =& -\sum_{i=1}^{n_g} \{Y_{ig} - \Lambda(X_{ig}'\beta^0)\} X_{ig}, \end{aligned}$$

we have $|S_{gj}| \le \bar{n} \max_{i \in [n_g]} |X_{ig,j}| = U_{gj}$. In addition, since $\gamma \ge 1/G$, using the fact that $1 - \Phi(t) \le \frac{1}{\sqrt{2\pi}} \frac{1}{t} e^{t^2/2}$, we have

$$\Phi^{-1}(1 - \gamma/2p) \lesssim \sqrt{\log(pG)} \lesssim \sqrt{\log a_G}.$$

Using Assumption 3 (2),(3),(4), it follows that $\log^{1/2} a_G \lesssim \check{\delta}_G G^{1/6}$ and $\{\mathrm{E_P} \frac{1}{G} \sum_{g=1}^G |S_{gj}|^3\}^{1/3} \lesssim 1$ uniformly over $j \in [p]$. Thus Assumption 9 (1) is satisfied. Under Assumption 3 (1),(4), it holds uniformly over $j \in [p]$ that

$$\begin{aligned} &\mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^G S_{gj}^2\Big] \ge \inf_{\|\xi\|_2=1} \mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^G \Big(\sum_{i=1}^{n_g} \{Y_{ig} - \Lambda(X_{ig}'\beta^0)\} X_{ig}'\xi\Big)^2\Big] \gtrsim 1 \text{ and} \\ &\mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^G S_{gj}^2\Big] \le \mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^G U_{gj}^2\Big] \lesssim 1. \end{aligned}$$

So Assumption 9 (2) is verified.

To apply Lemma 3, we verify Assumption 8. The convexity is trivial. To show Assumption 8 (2) holds, note that $S_{gj}^2 \le U_{gj}^2$ and Assumption 3 (4)(7) implies that if we let

$$\begin{aligned} \mathcal{F} =& \Big\{ W_g \mapsto \Big(-\sum_{i=1}^{\bar{n}} \{Y_{ig} - \Lambda(X_{ig}'\beta^0)\} X_{ig,j} \Big)^2 : j \in [p] \Big\}, \\ \mathcal{F}_j =& \Big\{ W_g \mapsto \Big(-\sum_{i=1}^{\bar{n}} \{Y_{ig} - \Lambda(X_{ig}'\beta^0)\} X_{ig,j} \Big)^2 \Big\}, \end{aligned}$$

then each $\mathcal{F}_j$ is of VC-subgraph class since it consists of a single function, and $\mathcal{F} \subset \cup_{j\in[p]}\mathcal{F}_j$, and $\mathcal{F}$ has an envelope F such that $F(W_g) = \bar{n}^2 \max_{j\in[p]} \max_{i\in[n_g]} |X_{ig,j}|^2$. Note that under Assumption 3 (4)(7)(8), one has $\max_{j\in[p]} \mathrm{E}_{\mathrm{P}} \frac{1}{G}\sum_{g=1}^G U_{gj}^4 \lesssim 1$, $(\mathrm{E}_{\mathrm{P}} \frac{1}{G}\sum_{g=1}^G F^2(W_g))^{1/2} = (\mathrm{E}_{\mathrm{P}} \frac{1}{G}\sum_{g=1}^G \|U_g\|_\infty^4)^{1/2} \leq M_{G,2}^2$, $M_{G,2} \leq (\mathrm{E}_{\mathrm{P}} \frac{1}{G}\sum_{g=1}^G \|U_g\|_\infty^{2q})^{1/2q} \leq (\check{\delta}_G G^{1/2-1/q})^{1/2}$, and $\sqrt{\mathrm{E}_{\mathrm{P}}[(\max_{1\leq g\leq G} F(W_g))^2]} \leq G^{1/q} M_{G,2}^2$. Applying Corollary 2.1 of Chernozhukov, Chetverikov and Kato (2013), with probability at least $1 - c(\log G)^{-1}$, it holds that

$$\Big|\frac{1}{G}\sum_{g=1}^G (S_{gj}^2 - \mathrm{E}_{\mathrm{P}}[S_{gj}^2])\Big| \lesssim \sqrt{\frac{\log(pa_G M_{G,2}^2/M_{G,2}^2)}{G}} + \frac{M_{G,2}^2}{G^{1-1/q}}\log(pa_G M_{G,2}^2/M_{G,2}^2)$$
$$\lesssim \sqrt{\frac{\log(a_G)}{G}} + \frac{M_{G,2}^2}{G^{1-1/q}}\log(a_G) = o(1),$$

where the last equality follows from Assumption 3 (9). This implies $\frac{1}{G}\sum_{g=1}^G S_{gj}^2 = (1 - o(1))\mathrm{E}_{\mathrm{P}}[\frac{1}{G}\sum_{g=1}^G S_{gj}^2]$ uniformly in $j \in [p]$. Similar arguments can be used to establish the statement that $\frac{1}{G}\sum_{g=1}^G (n_g \sum_{i=1}^{n_g} X_{ig,j}^2) = (1 - o(1))\mathrm{E}_{\mathrm{P}}[\frac{1}{G}\sum_{g=1}^G (n_g \sum_{i=1}^{n_g} X_{ig,j}^2)]$ with probability at least $1 - C(\log G)^{-1}$. Now it suffices to show that

$$(1 - o(1))\mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^G S_{gj}^2\Big] \leq \widehat{\Psi}_j^2 \lesssim 1 \tag{E.28}$$

with probability $1 - c(\log G)^{-1}$ uniformly over $j \in [p]$. The case of $\bar{m} = 0$ follows from the calculations that

$$\widehat{l}_{j,0}^2 = \frac{1}{4}\frac{1}{G}\sum_{g=1}^G \Big(n_g \sum_{i=1}^{n_g} X_{ig,j}^2\Big) \lesssim \frac{1 - o(1)}{4}\mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^G\sum_{i=1}^{n_g} X_{ig,j}^2\Big] \leq \frac{1 - o(1)}{4}\mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^G U_{gj}^2\Big] \lesssim 1$$

with probability $1 - c(\log G)^{-1}$ under Assumption 3 (4) and

$$\frac{1}{4}\mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^G\sum_{i=1}^{n_g} X_{ig,j}^2\Big] \geq \mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^G\sum_{i=1}^{n_g} f_{ig}^2 X_{ig,j}^2\Big] = \mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^G\sum_{i=1}^{n_g} \{Y_{ig} - \Lambda(X_{ig}'\beta^0)\}^2 X_{ig,j}^2\Big]$$
$$\gtrsim \mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^G \Big(\sum_{i=1}^{n_g} \{Y_{ig} - \Lambda(X_{ig}'\beta^0)\} X_{ig,j}\Big)^2\Big] = \mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^G S_{gj}^2\Big],$$

where the Cauchy-Schwarz inequalty, the law of iterated expectations and the fact that $f_{ig}^2 \leq \|\Lambda(1 - \Lambda)\|_\infty \leq 1/4$ are used.

To show (E.28) with $m \geq 1$, suppose that (E.28) holds for $\bar{m} - 1$, we can complete the proof and has $\|f_{ig} X_{ig}(\widehat{\beta} - \beta^0)\|_G \lesssim (s \log a_G/G)^{1/2}$ with probability $1 - C(\log G)^{-1}$. For $m = \bar{m}$, denote $\Lambda_{ig} = \Lambda(X_{ig}'\beta^0)$ and $\widetilde{\Lambda}_{ig} = \Lambda(X_{ig}'\widetilde{\beta})$, use the fact that for positive a, b,

$|\sqrt{a}-\sqrt{b}|\leq\sqrt{|a-b|}$, we have

$$|\widehat{l}_j - l_j| \leq \Big(\Big| \frac{1}{G}\sum_{g=1}^{G} \widehat{S}_{gj}^2 - \frac{1}{G}\sum_{g=1}^{G} S_{gj}^2 \Big| \Big)^{1/2}.$$

In addition, it holds uniformly over $j\in[p]$ that with probability at least $1-C(\log G)^{-1}$,

$$\begin{aligned}
&\Big|\frac{1}{G}\sum_{g=1}^{G}\widehat{S}_{gj}^2-\frac{1}{G}\sum_{g=1}^{G}S_{gj}^2\Big| \\
\leq&\Big|\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\{\Lambda_{ig}-\widetilde{\Lambda}_{ig}\}X_{ig,j}\Big)^2\Big|+2\Big|\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\{Y_{ig}-\Lambda_{ig}\}X_{ig,j}\Big)\Big(\sum_{i=1}^{n_g}\{\Lambda_{ig}-\widetilde{\Lambda}_{ig}\}X_{ig,j}\Big)\Big| \\
\leq&\Big|\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}X_{ig,j}^2\Big)\Big(\sum_{i=1}^{n_g}\{\Lambda_{ig}-\widetilde{\Lambda}_{ig}\}^2\Big)\Big|+2\Big|\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\{Y_{ig}-\Lambda_{ig}\}X_{ig,j}\Big)\Big(\sum_{i=1}^{n_g}\{\Lambda_{ig}-\widetilde{\Lambda}_{ig}\}X_{ig,j}\Big)\Big| \\
\lesssim&\max_{1\leq g\leq G}\|U_g\|_\infty^2\Big|\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\{(f_{ig}X_{ig}'(\beta^0-\widetilde{\beta})\}^2\Big|+\max_{1\leq g\leq G}\|U_g\|_\infty\Big|\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\{(f_{ig}X_{ig}'(\beta^0-\widetilde{\beta})\}^2\Big|^{1/2} \\
\lesssim&\frac{M_{G,2}s\log a_G}{G^{1-1/q}}+\frac{M_{G,2}(s\log a_G)^{1/2}}{G^{1/2-1/2q}}=o(1),
\end{aligned}$$

where the second inequality follows Cauchy-Schwarz inequality and the third follows Assumption 3 (7) and the last holds following $|\Lambda(t+\Delta t)-\Lambda(t)|\lesssim\Lambda'(t)\Delta t$ for $|\Delta t|\leq 1$ as in inequality (I6) in Belloni, Chernozhukov, Chetverikov and Wei (2018), the rates from $m=1$, Assumption 3 (8), and the fact Λ' is Lipschitz. This verifies Assumption 8 (2).

We now apply Lemma 6 and obtain that with some $c'>c$ and $\gamma=\gamma_G\in[1/G,1/\log G]$, one has

$$\mathrm{P}_{\mathrm{P}}\Big(\frac{\lambda}{G}\geq c\Big\|\widehat{\Psi}^{-1}\frac{1}{G}\sum_{g=1}^{G}S_g\Big\|_\infty\Big)\geq 1-\gamma-o(\gamma).$$

Assumption 8 (1) is trivial since we have $\widehat{M}(y,x,\beta)=M(y,x,\beta)$ in this case. Assumption 8 (3) holds for any A and $C_G\lesssim(s\log a_G/G)^{1/2}$ following Lemma 9.

Now, let us define

$$\bar{q}_A=\inf_{\delta\in A}\frac{(\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}w_{ig}|X_{ig}'\delta|^2)^{3/2}}{\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}w_{ig}|X_{ig}'\delta|^3}.$$

To apply Lemma 3, we need to verify the condition

$$\bar{q}_A=\bar{q}_{A_1}\wedge\bar{q}_{A_2}\geq(L+\frac{1}{c})\|\widehat{\Psi}_0\|_\infty\frac{\lambda\sqrt{s}}{G\bar{\kappa}_{2\widetilde{c}}}+6\widetilde{c}C_G$$

for $A=\Delta_{2\widetilde{c}}\cup\{\delta\in\mathbb{R}^p:\|\delta\|_1\leq\frac{3G}{\lambda}\frac{c\|\widehat{\Psi}_0^{-1}\|_\infty}{\ell c-1}C_G\|\sqrt{w_{ig}}X_{ig}'\delta\|_G\}=A_1\cup A_2$.

Note that under Assumptions 3 (6)(7)(8) and 4, we have

$$\begin{aligned}\bar{q}_{A_1} \geq & \inf_{\delta \in A_1} \frac{(\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} w_{ig}|X_{ig}'\delta|^2)^{1/2}}{\max_{1\leq g\leq G}\|U_g\|_\infty\|\delta\|_1} \gtrsim_{\mathrm{P}} \inf_{\delta \in A_1} \frac{(\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} w_{ig}|X_{ig}'\delta|^2)^{1/2}}{G^{1/2q}M_{G,2}\|\delta\|_1} \\ \geq & \inf_{\delta \in A_1} \frac{(\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} w_{ig}|X_{ig}'\delta|^2)^{1/2}}{G^{1/2q}M_{G,2}(1+2\widetilde{c})\sqrt{s}\|\delta_T\|_2} \gtrsim \frac{\bar{\kappa}_{2\widetilde{c}}}{G^{1/2q}M_{G,2}(1+2\widetilde{c})\sqrt{s}} \\ \gtrsim & \frac{1}{\check{\delta}^{1/2}G^{1/4}} \gtrsim \sqrt{\frac{s\log a_G}{\check{\delta}G}}.\end{aligned}$$

Next, using Assumptions 3 (7)(8), since $\lambda \lesssim \sqrt{G\log a_G}$ and $C_G \lesssim (s\log a_G/G)^{1/2}$, some calculations yield

$$\begin{aligned}\bar{q}_{A_2} \geq & \inf_{\delta \in A_2} \frac{(\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} w_{ig}|X_{ig}'\delta|^2)^{1/2}}{\max_{1\leq g\leq G}\|U_g\|_\infty\|\delta\|_1} \gtrsim_{\mathrm{P}} \inf_{\delta \in A_2} \frac{(\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} w_{ig}|X_{ig}'\delta|^2)^{1/2}}{G^{1/2q}M_{G,2}\|\delta\|_1} \\ \geq & \frac{\lambda}{3GC_G}\frac{\ell c-1}{c}\frac{\|\widehat{\Psi}_0^{-1}\|_\infty^{-1}}{G^{1/2q}M_{G,2}} \gtrsim_{\mathrm{P}} \frac{\lambda}{C_G G^{1+1/2q}M_{G,2}} \\ \gtrsim_{\mathrm{P}} & \frac{1}{G^{1/2q}M_{G,2}\sqrt{s}} \geq \frac{1}{\check{\delta}^{1/2}G^{1/4}} \gtrsim \sqrt{\frac{s\log a_G}{\check{\delta}G}}.\end{aligned}$$

Furthermore, we have

$$(L+\frac{1}{c})\|\widehat{\Psi}_0\|_\infty \frac{\lambda\sqrt{s}}{G\bar{\kappa}_{2\widetilde{c}}} + 6\widetilde{c}C_G \lesssim \sqrt{\frac{s\log a_G}{G}}$$

since $\|\widehat{\Psi}_0\|_\infty \lesssim 1$ with probability $1-C(\log G)^{-1}$. So all conditions required by Lemma 3 are satisfied. An application of the Lemma leads to

$$\|\sqrt{w_{ig}}X_{ig}'(\widehat{\beta}-\beta^0)\|_G \lesssim \sqrt{\frac{s\log a_G}{G}} \text{ and } \|\widehat{\beta}-\beta^0\|_1 \lesssim \sqrt{\frac{s^2\log a_G}{G}}.$$

Now, to apply Lemma 4, we need to verify condition (F.43). First, using Assumption 3 (7)(8), we have

$$\max_{1\leq g\leq G}\max_{i\in[n_g]}|X_{ig}'(\widehat{\beta}-\beta^0)| \lesssim_{\mathrm{P}} G^{1/2q}M_{G,2}\|\widehat{\beta}-\beta^0\|_1 \lesssim \sqrt{\frac{M_{G,2}^2 s^2\log a_G}{G^{1-1/q}}} \lesssim \check{\delta}_G = o(1).$$

Also, following equation (I.6) of Belloni, Chernozhukov, Chetverikov and Wei (2018), one has $|\Lambda(t+\Delta t)-\Lambda(t)| \lesssim \Lambda'(t)|\Delta t|$ uniformly over t and Δt with $|\Delta t| \leq 1$. It holds uniformly over ig that

$$\begin{aligned}[\partial_\beta\widehat{M}(Y_{ig},X_{ig},\widehat{\beta}) - \partial_\beta\widehat{M}(Y_{ig},X_{ig},\beta^0)]\}'\delta \lesssim & |\Lambda(X_{ig}'\widehat{\beta})-\Lambda(X_{ig}'\beta^0)|\cdot|X_{ig}'\delta| \\ \lesssim & \Lambda'(X_{ig}'\beta^0)\cdot|X_{ig}'(\widehat{\beta}-\beta^0)|\cdot|X_{ig}'\delta|.\end{aligned}$$

Since $|\Lambda'(X_{ig}'\beta^0)| \lesssim w_{ig} \leq \sqrt{w_{ig}}$, with probability at least $1 - C(\log G)^{-1}$, we have

$$\begin{aligned} &\Big|\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}[\partial_\beta \widehat{M}(Y_{ig}, X_{ig}, \widehat{\beta}) - \partial_\beta \widehat{M}(Y_{ig}, X_{ig}, \beta^0)]\}'\delta\Big| \\ \leq & C\|\sqrt{w_{ig}}X_{ig}'(\widehat{\beta} - \beta^0)\|_G \|X_{ig}'\delta\|_G \leq L_G\|X_{ig}'\delta\|_G \end{aligned}$$

for some $L_G \lesssim (s\log a_G/G)^{1/2}$. Thus condition (F.43) is satisfied. In addition, Lemma 4 implies $\|\widehat{\beta}\|_0 \lesssim s$.

Finally, to establish the convergence rates for $\widetilde{\beta}$, we apply Lemma 5. We verify condition (F.44) on $\bar{q}_A$ for $A = \{\delta \in \mathbb{R}^p : \|\delta\|_0 \leq Cs\}$ for a constant $\widehat{s} + s \leq Cs$ with probability $1 - o(1)$. Note it holds that

$$\begin{aligned} \bar{q}_A = & \inf_{\delta \in A} \frac{(\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} w_{ig}|X_{ig}'\delta|^2)^{1/2}}{\max_{1\leq g\leq G}\|U_g\|_\infty \|\delta\|_1} \\ \geq & \inf_{\|\delta\|_0 \leq Cs} \frac{(\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} w_{ig}|U_g'\delta|^2)^{1/2}}{\max_{1\leq g\leq G}\|U_g\|_\infty\sqrt{Cs}\|\delta\|_2} \gtrsim_{\mathrm{P}} \inf_{\|\delta\|_0 \leq Cs} \frac{\sqrt{\phi_{\max}(Cs)}}{\sqrt{s}G^{1/2q}M_{G,2}} \gtrsim \frac{\log^{1/4} a_G}{\check{\delta}_G G^{1/4}} \end{aligned}$$

under Assumptions 3 (7)(8) and 4. On the other hand, it follows from (F.45) that with probability $1 - C(\log G)^{-1}$,

$$\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\partial_\beta \widehat{M}(Y_{ig}, X_{ig}, \widetilde{\beta}) - \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\partial_\beta \widehat{M}(Y_{ig}, X_{ig}, \beta^0) \lesssim \frac{s\log a_G}{G}$$

since $\lambda/G \lesssim (\log a_G/G)^{1/2}$, $\|\widehat{\beta} - \beta^0\|_1 \lesssim (s\log a_G/G)^{1/2}$ and $\|\widehat{\Psi}_0\|_\infty \lesssim 1$ with probability $1 - C(\log G)^{-1}$. Also $C_G \lesssim (s^2\log a_G/G)^{1/2}$,

$$\Big\|\frac{1}{G}\sum_{g=1}^{G} S_g\Big\|_\infty \leq \Big\|\widehat{\Psi}_0\Big\|_\infty \Big\|\widehat{\Psi}_0^{-1}\frac{1}{G}\sum_{g=1}^{G} S_g\Big\|_\infty \lesssim \frac{\lambda}{G}$$

with probability $1 - C(\log G)^{-1}$. So right-hand side of (F.44) is bounded by $(s\log a_G/G)^{1/2}$. So by Lemma 5, we have the desired results.

Finally, since $s \geq 1$, we can without loss of generality assume the k-th coordinate is always in the support of $\widehat{\beta}$ and this does not affect the rate of convergence in post-lasso (see Comment D.1. of Belloni, Chernozhukov and Kato (2015)). Also, since all $k \in [p]$ share the same regularized event, the convergence rate holds uniformly for all $k \in [p]$. ■

E.2. **Proof for Theorem 3.** The proof follows analogously of the Proof of Theorem 4.2 in Belloni, Chernozhukov, Chetverikov and Wei (2018) with modifications to account for cluster sampling. The major difference lies in the verification of our Assumption 8 (2).

Proof. Let $\bar{r}^j_{ig} = X^j_{ig}(\gamma^j - \bar{\gamma}^j)$, $w_{ig} = \widehat{f}^2_{ig}$ and

$$\begin{aligned} M(D^j_{ig}, X^j_{ig}, \gamma) =& f^2_{ig}(D^j_{ig} - X^j_{ig}\gamma - \bar{r}^j_{ig})^2, \\ \widehat{M}(D^j_{ig}, X^j_{ig}, \gamma) =& \widehat{f}^2_{ig}(D^j_{ig} - X^j_{ig}\gamma)^2. \end{aligned} \tag{E.29}$$

Then, the sparse approximation $\bar{\gamma}^j$ can be identified by

$$\bar{\gamma}^j = \operatorname*{argmin}_{\gamma \in \mathbb{R}^{p-1}} \mathrm{E}_{\mathrm{P}} \frac{1}{G} \sum_{g=1}^{G} \sum_{i=1}^{n_g} M(D^j_{ig}, X^j_{ig}, \gamma).$$

We will first show that the regularized events (F.42) holds uniformly over $j \in [p]$. Subsequently, we apply Lemmas 3, 4 and 5 to bound different norms of $(\widetilde{\gamma}^j - \bar{\gamma}^j)$. Then bounds for $(\widehat{\gamma}^j - \gamma^j)$ follow from Assumption 2.

First, we verify Assumption 9. For Assumption 9 (1), note that

$$S^j_g = \sum_{i=1}^{n_g} \partial_\gamma M(D^j_{ig}, X^j_{ig}, \bar{\gamma}^j) = 2\sum_{i=1}^{n_g} f^2_{ig}(D^j_{ig} - X^j_{ig}\bar{\gamma}^j - \bar{r}^j_{ig})(X^j_{ig})' = 2\sum_{i=1}^{n_g} f^2_{ig} Z^j_{ig}(X^j_{ig})'.$$

where $a^j = \beta^0$. Since $\Phi^{-1}(1-\gamma/2p) \leq \sqrt{\log(1/t)}$ for all $t \in (0, 1/2)$, along with Assumption 3 (3), we have

$$\Big\{\mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^{G}|S^j_{gk}|^3\Big]\Big\}^{1/3}\Phi^{-1}(1-\gamma/2p) \lesssim \Big\{\mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^{G}|V^j_g U_{gk}|^3\Big]\Big\}^{1/3}\log^{1/2} a_G \leq \check{\delta}_G G^{1/6}$$

uniformly in $j \in [p]$ and $k \in [p] \setminus \{j\}$. This shows Assumption 9 (1).

To show Assumption 9 (2), notice that Assumption 3 (2) implies $\mathrm{E}_{\mathrm{P}}\frac{1}{G}\sum_{g=1}^{G}|S^j_{gk}|^2 \gtrsim 1$ and

$$\mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^{G}(S^j_{gk})^2\Big] \leq \mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^{G}(V^j_g U_{gk})^2\Big] \leq \mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^{G}(|V^j_g|^4 + |U_{gk}|^4)\Big] \lesssim 1$$

uniformly over $j \in [p]$ and $k \in [p] \setminus \{j\}$ by Assumption 3 (4).

The convexity requirement is trivially satisfied. To show Assumptions 8 (1), we first claim that with probability $1 - C(\log G)^{-1}$,

$$\max_{j\in[p]} \|(\widehat{f}^2_{ig} - f^2_{ig})Z^j_{ig}/\widehat{f}_{ig}\|_G \lesssim (s \log a_G/G)^{1/2}. \tag{E.30}$$

Now, since by Theorem 2 and Assumption 3 (7)(8), one has

$$\max_{i,g}|X'_{ig}(\widehat{\beta} - \beta^0)| \leq \max_{i,g}\|X_{ig}\|_\infty \|\widehat{\beta} - \beta^0\|_1 \lesssim_{\mathrm{P}} G^{1/2q} M_{G,2}(s^2 \log a_G/G)^{1/2} \leq \check{\delta}_G = o(1)$$

with probability $1 - C(\log G)^{-1}$, we then have with probability $1 - C(\log G)^{-1}$

$$|\widehat{f}_{ig}^2 - f_{ig}^2| \leq |\Lambda(X_{ig}'\widetilde{\beta}) - \Lambda(X_{ig}'\beta^0)| \lesssim \Lambda'(X_{ig}'\beta^0)|X_{ig}'(\widehat{\beta} - \beta^0)| \leq f_{ig}^2/2 \leq 1 \tag{E.31}$$

uniformly over all i, g. Note we have used the fact that for $|\widetilde{t} - t| \leq 1$, $|\Lambda(t) - \Lambda(\widetilde{t})| \lesssim \Lambda'(t)|t - \widetilde{t}|$. Also, some calculations give that for G large enough, let $\widetilde{t}_{ig} = X_{ig}\widetilde{\beta}$, $t_{ig} = X_{ig}\beta^0$, then it holds that

$$\begin{aligned}
|\widehat{f}_{ig}^2 - f_{ig}^2| =& |\Lambda(\widetilde{t}_{ig}) - \Lambda^2(\widetilde{t}_{ig}) - \Lambda(t_{ig}) + \Lambda^2(t_{ig})| \\
\leq& |\Lambda(\widetilde{t}_{ig}) - \Lambda(t_{ig})| + |\Lambda^2(\widetilde{t}_{ig}) - \Lambda^2(t_{ig})| \\
\leq& |\Lambda(\widetilde{t}_{ig}) - \Lambda(t_{ig})| + \Lambda(\widetilde{t}_{ig})|\Lambda(\widetilde{t}_{ig}) - \Lambda(t_{ig})| + \Lambda(t_{ig})|\Lambda(\widetilde{t}_{ig}) - \Lambda(t_{ig})| \\
\lesssim& |\Lambda(\widetilde{t}_{ig}) - \Lambda(t_{ig})| \lesssim \Lambda'(\widetilde{t}_{ig})|\widetilde{t}_{ig} - t_{ig}| = f_{ig}^2|\widetilde{t}_{ig} - t_{ig}|.
\end{aligned}$$

Thus, with probability at least $1 - C(\log G)^{-1}$, one has

$$\begin{aligned}
\max_{j\in[p]} \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(\widehat{f}_{ig}^2 - f_{ig}^2)^2(Z_{ig}^j/\widehat{f}_{ig})^2 \lesssim& \max_{j\in[p]} \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(\Lambda'(X_{ig}'\widetilde{\beta}))^2|X_{ig}'(\beta^0 - \widetilde{\beta})|^2(Z_{ig}^j/\widehat{f}_{ig})^2 \\
\lesssim& \max_{j\in[p]} \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}|X_{ig}'(\beta^0 - \widetilde{\beta})|^2(Z_{ig}^j)^2 \\
\leq& \max_{j\in[p]} \sup_{\|\delta\|_0\leq Cs, \|\delta\|_2=1} \frac{s\log a_G}{G}\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(X_{ig}'\delta)^2(Z_{ig}^j)^2 \\
\leq& \frac{s\log a_G}{G}O(1),
\end{aligned}$$

where the last inequality follows from Assumption 4. Therefore, with probability $1 - C(\log G)^{-1}$, we have

$$\max_{j\in[p]} \|(\widehat{f}_{ig}^2 - f_{ig}^2)Z_{ig}^j/\widehat{f}_{ig}\|_G \lesssim \max_{j\in[p]} \|(\widehat{f}_{ig}^2 - f_{ig}^2)Z_{ig}^j/f_{ig}\|_G.$$

Recall that $\bar{r}^j_{ig} = X^j_{ig}(\gamma^j - \bar{\gamma}^j)$. Assumption 8 (1) can be examined by noting that it holds uniformly in $j \in [p]$ that

$$\begin{aligned}
&\Big|\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\Big(\partial_\gamma\widehat{M}(Y_{ig}, X_{ig}, \bar{\gamma}^j) - \partial_\gamma M(Y_{ig}, X_{ig}, \bar{\gamma}^j)\Big)\Big]'\delta\Big| \\
\leq&\Big|2\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(\widehat{f}^2_{ig}\bar{r}^j_{ig} + (\widehat{f}^2_{ig} - f^2_{ig})Z^j_{ig})X^j_{ig}\delta\Big| \\
\leq&2\Big(\|\widehat{f}_{ig}\bar{r}^j_{ig}\|_G + \|(\widehat{f}^2_{ig} - f^2_{ig})Z^j_{ig}/\widehat{f}_{ig}\|_G\Big)\|\sqrt{w_{ig}}X^j_{ig}\delta\|_G \\
\leq&C_G\|\sqrt{w_{ig}}X'_{ig}\delta\|_G.
\end{aligned}$$

We now verify the condition $C_G \lesssim (s\log a_G/G)^{1/2}$ in Assumption 8 (1). Notice that one has

$$\|\widehat{f}_{ig}\bar{r}^j_{ig}\|_G \leq \|\bar{r}^j_{ig}\|_G \lesssim (s\log a_G/G)^{1/2} \tag{E.32}$$

with probability at least $1 - C(\log G)^{-1}$ following the same arguments as in Lemma J1 of Belloni, Chernozhukov, Chetverikov and Wei (2018) under Assumption 1, 2, 3, 4. To see this, let

$$\begin{aligned}
\mathcal{G} &= \{W_g \mapsto \sum_{i=1}^{\bar{n}} X^j_{ig}(\gamma^j - \bar{\gamma}^j) : j \in [p]\}, \\
\mathcal{G}_{ijT} &= \{W_g \mapsto X^j_{ig}(\gamma^j - \gamma^j_T) : j \in [p]\}.
\end{aligned}$$

Note $\bar{\gamma}^j = \gamma^j_T$ for some T by Assumption 1. Thus one has $\mathcal{G} \subset \cup_{j\in[p], T\leq s}\sum_{i=1}^{\bar{n}}\mathcal{G}_{ijT}$. So for $\mathcal{G}^2$, we have an envelope $G(\mathbf{w}) = \|u(\mathbf{w})\|_\infty \max_{j\in[p]}\|\bar{\gamma}^j - \gamma^j\|_1^2$ with

$$\Big\{\mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^{G}\max_{1\leq g\leq G}G^2(W_g)\Big]\Big\}^{1/2} \lesssim \frac{s^2 M_{G,2}\log a_G}{G^{1-1/q}}.$$

In addition, for all finite discrete measures Q and $0 < \epsilon \leq 1$, it holds that

$$\sup_Q \log N(\epsilon\|G\|_{Q,2}, \mathcal{G}^2, \|\cdot\|_{Q,2}) \lesssim s\log(a_G/\epsilon).$$

Thus by applying Corollary 3, one has with probability at least $1 - C(\log G)^{-1}$,

$$\max_{j\in[p]}\Big|\Big(\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(\bar{r}^2_{ig} - \mathrm{E}_{\mathrm{P}}\bar{r}^2_{ig}\Big)\Big| \lesssim \frac{s\log a_G}{G}.$$

Finally, $\mathrm{E}_{\mathrm{P}}\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\bar{r}^2_{ig} \lesssim \sup_{\|\xi\|_2=1}\mathrm{E}_{\mathrm{P}}\frac{1}{G}\sum_{g=1}^{G}(U'_g\xi)^2\|\gamma^j - \bar{\gamma}^j\|_2^2 \lesssim s\log a_G/G$ by Assumption 1. This shows (E.32) and thus Assumption 8 (1).

Note that Assumption 8 (3) holds with $\check{\Delta}_G = 0$ and $\bar{q}_A = \infty$ for any A since

$$\begin{aligned}\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}_j(D_{ig}^j, X_{ig}^j, \bar{\gamma}^j+\delta) - \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}(Y_{ig}, X_{ig}, \bar{\gamma}^j)\\ -2\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{f}_{ig}^2(D_{ig}^j - X^j\bar{\gamma}^j)X_{ig}^j\delta = \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(\widehat{f}_{ig}X_{ig}^j\delta)^2\end{aligned}$$

and $\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(\widehat{f}_{ig}X_{ig}^j\delta)^2 = \|\sqrt{w_{ig}}X_{ig}^j\delta\|_G^2$.

To check Assumption 8 (2), note that under Assumption 3 (5)(6)(7)(8), one has

$$\begin{aligned}&\mathrm{E_P}\Big[\max_{1\le g\le G}\max_{j\in[p]}\Big\|\sum_{i=1}^{n_g} f_{ig}^2 Z_{ig}^j X_{ig}^j\Big\|_\infty^2\Big]\\ \le&\mathrm{E_P}\Big[\max_{1\le g\le G}\max_{j\in[p]}|V_g^j|^2\|U_g\|_\infty^2\Big]\\ \lesssim& G^{2/q}(M_{G,1}+M_{G,2}).\end{aligned}$$

Thus, an application of Lemma 7 gives

$$\begin{aligned}\max_{k\in[p]}\max_{j\in[p]\setminus\{k\}}\Big|\frac{1}{G}\sum_{g=1}^{G}\Big[\Big(\sum_{i=1}^{n_g}S_{gk}^j\Big)^2 - \mathrm{E_P}\Big(\sum_{i=1}^{n_g}S_{gk}^j\Big)^2\Big]\Big| \lesssim_{\mathrm{P}}& G^{-(1/2-1/q)}(M_{G,1}^2+M_{G,2}^2)\log a_G\\ \le&\check{\delta}_G\log a_G = o(1).\end{aligned}$$

where the last equality follows the rate assumption in statement of the Theorem. Therefore, since $l_{j0k} = \{\frac{1}{G}\sum_{g=1}^{G}(\sum_{i=1}^{n_g}S_{gk}^j)^2\}^{1/2}$, we have $1 \lesssim \widehat{\Psi}_{j0k}^\gamma \lesssim 1$ with probability at least $1-C(\log G)^{-1}$ uniformly over $j\in[p]$ and $k\in[p]\setminus\{j\}$.

For $m=0$, with probability $1-C(\log G)^{-1}$, we have

$$\widehat{l}_{jk,0} \gtrsim 2\Big\{\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\widehat{f}_{ig}^2 Z_{ig}^j X_{ig,k}^j\Big)^2\Big\}^{1/2} \gtrsim 2\Big\{\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}f_{ig}^2 Z_{ig}^j X_{ig,k}^j\Big)^2\Big\}^{1/2} \gtrsim 1$$

uniformly over $j\in[p]$ and $k\in[p]\setminus\{j\}$. This follow from the fact that $|\widehat{f}_{ig}^2 - f_{ig}^2| \le f_{ig}^2$ with probability $1-C(\log G)^{-1}$. To obtain an upperbound, note under Assumption 3 (4)(7) and the fact that $\widehat{f}_{ig}\le 1$, one has

$$\widehat{l}_{jk,0} \lesssim 2\max_{g\in[G]}\max_{i\in[n_g]}|\widehat{f}_{ig}X_{ig,k}|\Big\{\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\widehat{f}_{ig}D_{ig}^j\Big)^2\Big\}^{1/2} \lesssim_{\mathrm{P}} G^{1/2q}M_{G,2}.$$

Thus for $m=0$, Assumption 8 (2) holds with $L \lesssim G^{1/2q}M_{G,2}\log^{1/2}a_G$ and $\ell \gtrsim 1$. For $m\ge 1$, suppose that the statement holds for $m=\bar{m}-1$, we can complete the proof and obtain the bound

$$\max_{j\in[p]}\|\widehat{f}_{ig}(\widetilde{\gamma}^j-\bar{\gamma}^j)\|_G\lesssim(L+1)(s^2\log a_G/G)^{1/2}$$

for $L\lesssim G^{1/2q}M_{G,2}\log^{1/2}a_G$. In addition, under Assumption 2 and 3 (7), it holds uniformly over $j\in[p]$ that

$$\begin{aligned}\|\widehat{f}_{ig}X_{ig}^j(\bar{\gamma}^j-\gamma^j)\|_G\le&\max_{1\le g\le G}\|U_g\|_\infty\cdot\|\bar{\gamma}^j-\gamma^j\|_1\\ \lesssim_{\mathrm{P}}&G^{1/2q}M_{G,2}(s^2\log a_G/G)^{1/2}.\end{aligned}$$

Thus by the triangle inequality, we have

$$\max_{j\in[p]}\|\widehat{f}_{ig}X_{ig}^j(\widetilde{\gamma}^j-\gamma^j)\|_G\lesssim(L+1)(s^2\log a_G/G)^{1/2}$$

for $L\lesssim G^{1/2q}M_{G,2}\log^{1/2}a_G$. Using the fact that for positive a,b, $|\sqrt{a}-\sqrt{b}|\le\sqrt{|a-b|}$, we have, for $m=\bar{m}$, it holds uniformly over $j\in[p]$, $k\in[p-1]$ that

$$\begin{aligned}|\widehat{l}_{jk,m}-l_{j0k}|=&2\Big|\Big\{\frac{1}{G}\sum_{g=1}^G\Big(\sum_{i=1}^{n_g}\widehat{f}_{ig}^2(D_{ig}^j-X_{ig}^j\widetilde{\gamma}^j)X_{ig,k}^j\Big)^2\Big\}^{1/2}-\Big\{\frac{1}{G}\sum_{g=1}^G\Big(\sum_{i=1}^{n_g}f_{ig}^2(D_{ig}^j-X_{ig}^j\gamma^j)X_{ig,k}^j\Big)^2\Big\}^{1/2}\Big|\\ \le&2\Big\{\Big|\frac{1}{G}\sum_{g=1}^G\Big(\sum_{i=1}^{n_g}\widehat{f}_{ig}^2(D_{ig}^j-X_{ig}^j\widetilde{\gamma}^j)X_{ig,k}^j\Big)^2-\frac{1}{G}\sum_{g=1}^G\Big(\sum_{i=1}^{n_g}f_{ig}^2(D_{ig}^j-X_{ig}^j\gamma^j)X_{ig,k}^j\Big)^2\Big|\Big\}^{1/2}\\ =&2\Big\{\Big|\frac{1}{G}\sum_{g=1}^G\Big(\sum_{i=1}^{n_g}\widehat{f}_{ig}^2X_{ig,k}^j\widehat{Z}_{ig}^j\Big)^2-\frac{1}{G}\sum_{g=1}^G\Big(\sum_{i=1}^{n_g}f_{ig}^2X_{ig,k}^jZ_{ig}^j\Big)^2\Big|\Big\}^{1/2},\end{aligned}$$

where $\widehat{Z}_{ig}^j=D_{ig}^j-X_{ig}^j\widetilde{\gamma}^j$. To bound the right-hand side, note by adding and subtracting terms and the triangle inequality,

$$\begin{aligned}&\Big|\frac{1}{G}\sum_{g=1}^G\Big(\sum_{i=1}^{n_g}\widehat{f}_{ig}^2X_{ig,k}^j\widehat{Z}_{ig}^j\Big)^2-\frac{1}{G}\sum_{g=1}^G\Big(\sum_{i=1}^{n_g}f_{ig}^2X_{ig,k}^jZ_{ig}^j\Big)^2\Big|\\ \le&\Big|\frac{1}{G}\sum_{g=1}^G\Big(\sum_{i=1}^{n_g}\widehat{f}_{ig}^2X_{ig,k}^jX_{ig}^j(\gamma^j-\widetilde{\gamma}^j)\Big)^2\Big|+2\Big|\frac{1}{G}\sum_{g=1}^G\Big(\sum_{i=1}^{n_g}\widehat{f}_{ig}^2X_{ig,k}^jX_{ig}^j(\gamma^j-\widetilde{\gamma}^j)\Big)\Big(\sum_{i=1}^{n_g}\widehat{f}_{ig}^2X_{ig,k}^jZ_{ig}^j\Big)\Big|\\ &+\Big|\frac{1}{G}\sum_{g=1}^G\Big(\sum_{i=1}^{n_g}(\widehat{f}_{ig}^2-f_{ig}^2)X_{ig,k}^jZ_{ig}^j\Big)^2\Big|+2\Big|\frac{1}{G}\sum_{g=1}^G\Big(\sum_{i=1}^{n_g}(\widehat{f}_{ig}^2-f_{ig}^2)X_{ig,k}^jZ_{ig}^j\Big)\Big(\sum_{i=1}^{n_g}f_{ig}^2X_{ig,k}^jZ_{ig}^j\Big)\Big|=o(1)\end{aligned}$$

uniformly over $j\in[p]$, $k\in[p-1]$ with probability at least $1-C(\log G)^{-1}$. The inequality holds following the Cauchy-Schwarz inequality. Then under Assumption 3 (7)(8), with probability at least $1 - C(\log G)^{-1}$, one has

$$\begin{aligned}\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}(\widehat{f}_{ig}^2 - f_{ig}^2)X_{ig,k}^j Z_{ig}^j\Big)^2 \leq & \frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}(\widehat{f}_{ig}^2 - f_{ig}^2)^2(Z_{ig}^j)^2/\widehat{f}_{ig}^2\Big)\Big(\sum_{i=1}^{n_g}\widehat{f}_{ig}^2(X_{ig,k}^j)^2\Big)\\ \leq & \max_{1\leq g\leq G}\|U_{gk}\|_\infty^2\|(\widehat{f}_{ig}^2 - f_{ig}^2)Z_{ig}^j/\widehat{f}_{ig}\|_G^2 \leq \frac{M_{G,2}^2 s\log a_G}{G^{1-1/q}} = o(1)\end{aligned}$$

uniformly over $j \in [p]$, $k \in [p-1]$. Here, we have used $\|(\widehat{f}_{ig}^2 - f_{ig}^2)Z_{ig}^j/\widehat{f}_{ig}\|_G \lesssim (s\log a_G/G)^{1/2}$ with probability at least $1 - C(\log G)^{-1}$ by equation (E.30). Similar arguments show that by Assumption 3 (8), with probability at least $1 - C(\log G)^{-1}$, we have

$$\begin{aligned}\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\widehat{f}_{ig}^2 X_{ig,k}^j X_{ig}^j(\gamma^j - \widetilde{\gamma}^j)\Big)^2 \leq & \frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}(\gamma^j - \widetilde{\gamma}^j)'\widehat{f}_{ig}^2(X_{ig}^j)'X_{ig}^j(\gamma^j - \widetilde{\gamma}^j)\Big)\Big(\sum_{i=1}^{n_g}\widehat{f}_{ig}^2(X_{ig,k}^j)^2\Big)\\ \leq & \max_{1\leq g\leq G}\|U_{gk}\|_\infty^2\|\widehat{f}_{ig}X_{ig}^j(\widetilde{\gamma}^j - \gamma^j)\|_G^2\\ \leq & L\frac{M_{G,2}s\log a_G}{G^{1-1/q}} \leq \frac{M_{G,2}^3 s\log^{3/2} a_G}{G^{1-3/2q}} = o(1)\end{aligned}$$

uniformly over $j \in [p]$, $k \in [p-1]$. Furthermore, by the Cauchy-Schwarz inequality, we have

$$\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}(\widehat{f}_{ig}^2 - f_{ig}^2)X_{ig,k}^j Z_{ig}^j\Big)\Big(\sum_{i=1}^{n_g}f_{ig}^2 X_{ig,k}^j Z_{ig}^j\Big) \leq \Big\{\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}(\widehat{f}_{ig}^2 - f_{ig}^2)X_{ig,k}^j Z_{ig}^j\Big)^2\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}f_{ig}^2 X_{ig,k}^j Z_{ig}^j\Big)^2\Big\}^{1/2}$$

and

$$\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\widehat{f}_{ig}^2 X_{ig,k}^j X_{ig}^j(\gamma^j - \widetilde{\gamma}^j)\Big)\Big(\sum_{i=1}^{n_g}\widehat{f}_{ig}^2 X_{ig,k}^j Z_{ig}^j\Big) \leq \Big\{\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\widehat{f}_{ig}^2 X_{ig,k}^j X_{ig}^j(\gamma^j - \widetilde{\gamma}^j)\Big)^2\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\widehat{f}_{ig}^2 X_{ig,k}^j Z_{ig}^j\Big)^2\Big\}^{1/2}$$

From the preceding results, it suffices to show the claim uniformly over $j \in [p]$ and $k \in [p-1]$,

$$\frac{M_{G,2}^3 s\log^{3/2} a_G}{G^{1-3/2q}}\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\widehat{f}_{ig}^2 X_{ig,k}^j Z_{ig}^j\Big)^2 = o_p(1).$$

Under Assumption 3 (4)(7), since $\widehat{f}_{ig}^2 \leq 1$, the Cauchy-Schwarz inequality gives

$$\max_{j\in[p]}\max_{k\in[p]\backslash\{j\}}\frac{1}{G}\sum_{g=1}^{G}\Big(\sum_{i=1}^{n_g}\widehat{f}_{ig}^2 X_{ig,k}^j Z_{ig}^j\Big)^2 \lesssim \max_{k\in[p]}\frac{1}{G}\sum_{g=1}^{G}U_{gk}^4 + \max_{k\in[p]}\frac{1}{G}\sum_{g=1}^{G}(V_g^j)^4 \lesssim_{\mathrm{P}} 1$$

by Assumption 3 (4). The same bound holds even if $\widehat{f}_{ig}$ is used in place of f_{ig}. The claim then follows from Assumption 3 (9). Thus for $m = \bar{m}$, the result holds for some $L, \ell, \check{\Delta}_G$ with $L \lesssim 1$, $\ell \gtrsim 1$ and $\check{\Delta}_G = o(1)$. This verifies Assumption 8 (2).

Note that $\|\widehat{\Psi}_0\|_\infty \lesssim 1$ and $\|\widehat{\Psi}_0^{-1}\|_\infty \lesssim 1$ with probability $1 - C(\log G)^{-1}$ following the preceding arguments. By Lemma 6, (F.42) holds with probability $1 - C(\log G)^{-1}$. Furthermore, following Assumption 4 and the fact $|\widehat{f}_{ig}^2 - f_{ig}^2| \le f_{ig}^2/2$ with probability $1 - C(\log G)^{-1}$, we have, for some $\ell_G \to \infty$, it holds that

$$1 \lesssim \min_{\|\delta\|_0 \le \ell_G s} \frac{\|f_{ig} X_{ig}'\delta\|_G^2}{\|\delta\|_2^2} \le \max_{\|\delta\|_0 \le \ell_G s} \frac{\|X_{ig}'\delta\|_G^2}{\|\delta\|_2^2} \lesssim 1.$$

Thus, by Lemma 3, one has

$$\|\widehat{f}_{ig} X_{ig}^j(\widehat{\gamma}^j - \bar{\gamma}^j)\|_G \lesssim (s \log a_G/G)^{1/2} \text{ and } \|\widehat{\gamma}^j - \bar{\gamma}^j\|_1 \lesssim (s^2 \log a_G/G)^{1/2}$$

with probability $1 - C(\log G)^{-1}$ uniformly over $j \in [p]$.

By the Cauchy-Schwarz inequality and the fact that $\widehat{f}_{ig} \le 1$, we have

$$\Big|\Big\{\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}[\partial_\gamma \widehat{M}(Y_{ig}, X_{ig}, \widehat{\gamma}^j) - \partial_\gamma \widehat{M}(Y_{ig}, X_{ig}, \bar{\gamma}^j)]\Big\}'\delta\Big| \le \|\widehat{f}_{ig} X_{ig}^j(\widehat{\gamma}^j - \bar{\gamma}^j)\|_G \|\widehat{f}_{ig} X_{ig}^j \delta\|_G \le L_G \|X_{ig}'\delta\|_G$$

with probability $1 - C(\log G)^{-1}$ uniformly over $j \in [p]$ for some $L_G \lesssim (s \log a_G/G)^{1/2}$. Since Assumption 4 implies that there is a $\ell_G \to \infty$ such that $\phi_{\max}(\ell_G s) \lesssim 1$ with probability $1 - C(\log G)^{-1}$, it follows Lemma 4 that $\|\widehat{\gamma}^j\|_0 \lesssim s$ with probability $1 - C(\log G)^{-1}$ uniformly over $j \in [p]$.

Note that condition (F.44) holds with $\bar{q}_A = \infty$. Also, with probability $1 - C(\log G)^{-1}$, it holds that

$$\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} \widehat{M}(Y_{ig}, X_{ig}, \widehat{\gamma}^j) - \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} \widehat{M}(Y_{ig}, X_{ig}, \bar{\gamma}^j) \lesssim s \log a_G/G$$

since $\lambda/G \lesssim (s \log a_G/G)^{1/2}$, $\max_{j\in[p]} \|\widehat{\gamma}^j - \bar{\gamma}^j\|_1 \lesssim (s^2 \log a_G)^{1/2}$ and $\max_{j \in [p]} \|\widehat{\Psi}_{j0}\|_\infty \lesssim 1$ with probability $1 - C(\log G)^{-1}$. Finally, one has $C_G \lesssim (s^2 \log a_G/G)^{1/2}$,

$$\Big\|\frac{1}{G}\sum_{g=1}^{G} S_g^j\Big\|_\infty \le \|\widehat{\Psi}_0\|_\infty \Big\|\widehat{\Psi}_0^{-1}\frac{1}{G}\sum_{g=1}^{G} S_g^j\Big\|_\infty \lesssim \frac{\lambda}{G}$$

with probability $1 - C(\log G)^{-1}$. This concludes the proof. ∎

E.3. **Proof for Corollary 1.**

Proof. Define $\ddot{r}_{ig}^k = X_{ig}'(\bar{\zeta}^k - \zeta^k)$ and the lost functions be

$$\begin{aligned} M(S_{ig}^k, X_{ig}, \zeta) =& f_{ig}^2(S_{ig}^k - X_{ig}'\zeta - \ddot{r}_{ig}^k)^2, \\ \widehat{M}(S_{ig}^k, X_{ig}', \zeta) =& \widehat{f}_{ig}^2(\widehat{S}_{ig}^k - X_{ig}\zeta)^2. \end{aligned}$$

The sparse approximation $\bar{\zeta}^k$ is identified by

$$\bar{\zeta}^k = \operatorname*{argmin}_{\zeta\in\mathbb{R}^p}\mathrm{E}_{\mathrm{P}}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} M(S^k_{ig}, X_{ig}, \zeta)\Big].$$

Then the proof follows the same steps in the proof for Theorem 3 as long as one can verify that Assumption F.40 (1) is still satisfied with $\widehat{S}^k_{ig}$ in place of S^k_{ig}. Thus, it suffices to show

$$\begin{aligned}\Big|\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\{\widehat{f}^2_{ig}(\widehat{S}^k_{ig} - S^k_{ig})X_{ig}\}\Big]'\delta\Big| =&\Big|\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\{\widehat{f}^2_{ig}(\widehat{S}^k_{ig} - X'_{ig}\zeta^k)X_{ig} - \widehat{f}^2_{ig}(S^k_{ig} - X'_{ig}\zeta^k)X_{ig}\}\Big]'\delta\Big| \\ \lesssim&\|\widetilde{\beta} - \beta^0\|_2\Big\{\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(\widehat{f}_{ig}X'_{ig}\delta)^2\Big\}^{1/2}.\end{aligned}$$

Observe that the left-hand side equals

$$\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\{\widehat{f}^2_{ig}(\widehat{S}^k_{ig} - \widetilde{S}^k_{ig})X_{ig} + \widehat{f}^2_{ig}(\widetilde{S}^k_{ig} - S^k_{ig})X_{ig}\}\Big]'\delta = (i) + (ii).$$

Notice that

$$|(ii)| \le 2\|\Lambda\|_\infty|\widetilde{\beta}^k_k - \beta^0|\Big|\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{f}^2_{ig}X_{ig}\Big]'\delta\Big| \lesssim |\widetilde{\beta}^k_k - \beta^0_k|\max_{i,g}|f_{ig}|\Big\{\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(\widehat{f}_{ig}X'_{ig}\delta)^2\Big\}^{1/2}.$$

A mean value expansion and an application of Hölder's inequality give that with probability at least $1 - C(\log G)^{-1}$,

$$\begin{aligned}|(i)| \le&2|\widetilde{\beta}^k_k|\|\Lambda'\|_\infty\frac{1}{G}\sum_{g=1}^{G}\Big[\sum_{i=1}^{n_g}\{\widehat{f}^2_{ig}X'_{ig}(\widetilde{\beta} - \beta^0)X_{ig}\Big]'\delta \\ \lesssim&\Big\{\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\Big(\widehat{f}_{ig}X'_{ig}(\widetilde{\beta} - \beta^0)\Big)^2\Big\}^{1/2}\Big\{\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\Big(\widehat{f}_{ig}X'_{ig}\delta\Big)^2\Big\}^{1/2} \\ \le&(s\log a_G/G)^{1/2}\|\sqrt{w_{ig}}X'_{ig}\delta\|_G \\ \le&C_G\|\sqrt{w_{ig}}X'_{ig}\delta\|_G.\end{aligned}$$

This concludes the proof. ■

E.4. **Proof for Lemma 2.**

Proof. Throughout the proof, we denote $\|v\|^2_G = v'v/G$ and $(u, v)_G = u'v/G$ for $u, v \in \mathbb{R}^n$. For each $j = 1, ..., p$, denote $D^j = \{D^j_{ig} : 1 \le i \le n_g, 1 \le g \le G\}$, an $n \times 1$ vector, $\mathbf{X}^j = \{X^j_{ig} : 1 \le i \le n_g, 1 \le g \le G\}$, a $n \times (p-1)$ matrix. We also make use of the notations $\widehat{F} = \mathrm{diag}\{\widehat{f}_{ig} : i \in [n_g], g \in [G]\}$ and $F = \mathrm{diag}\{f_{ig} : i \in [n_g], g \in [G]\}$.

Step 1. We first derive the identity

$$\widehat{\tau}_j^2 = D^{j\prime}\widehat{F}^2(D^j - \mathbf{X}^j\widetilde{\gamma}^j)/G. \tag{E.33}$$

The first order condition of nodewise post-lasso gives

$$-\mathbf{X}^{j\prime}_{\widehat{T}^j}\widehat{F}^2(D^j - \mathbf{X}^j\widetilde{\gamma}^j)/G = 0 \tag{E.34}$$

where $\widehat{T}^j = \text{support}(\widetilde{\gamma}^j)$.

Multiplying both sides by $\widetilde{\gamma}_j'$, we have

$$-\widetilde{\gamma}^{j\prime}\mathbf{X}^{j\prime}\widehat{F}^2D^j/G + \widetilde{\gamma}^{j\prime}\mathbf{X}^{j\prime}\widehat{F}^2\mathbf{X}^j\widetilde{\gamma}^j/G = 0. \tag{E.35}$$

Using its definition, some calculations yield that

$$\widehat{\tau}_j^2 = D^{j\prime}\widehat{F}^2D^j/G - 2\widehat{\gamma}^{j\prime}\mathbf{X}^{j\prime}\widehat{F}^2D^j/G + \widehat{\gamma}^{j\prime}\mathbf{X}^{j\prime}\widehat{F}^2\mathbf{X}^j\widehat{\gamma}^j/G.$$

Subtracting (E.35) from this gives (E.34).

Step 2. Applying Theorem 3, we have the convergence rates

$$\|\widetilde{\gamma}^j - \gamma^j\|_1 \lesssim s\sqrt{\frac{\log a_G}{G}} \qquad \text{and} \qquad \|\widehat{f}_{ig}X_{ig}^{j\prime}(\widetilde{\gamma}^j - \gamma^j)\|_G \vee \|\widetilde{\gamma}^j - \gamma^j\|_2 \lesssim \sqrt{\frac{s\log a_G}{G}}$$

uniformly in j with probability $1 - C(\log G)^{-1}$.

Step 3. Since $FD^j = F\mathbf{X}^j\gamma^j + FZ^j$, by Step 1, one has

$$\begin{aligned}\widehat{\tau}_j^2 =& D^{j\prime}\widehat{F}^2(D^j - \mathbf{X}^j\widetilde{\gamma}^j)/G\\ =& D^{j\prime}(\widehat{F}^2 - F^2)(D^j - \mathbf{X}^j\widetilde{\gamma}^j)/G + D^{j\prime}F^2(D^j - \mathbf{X}^j\widetilde{\gamma}^j)/G.\end{aligned}$$

Note we only need to consider bounding $D^{j\prime}F^2(D^j - \mathbf{X}^j\widetilde{\gamma}^j)/G$ term since the first term is of smaller order following the fact that $|\widehat{f}_{ig} - f_{ig}| \lesssim f_{ig}$ holds with probability at least $1 - C(\log G)^{-1}$ by (E.31) in the proof of Theorem 3. Now, decompose it into

$$\begin{aligned}D^{j\prime}F^2(D^j - \mathbf{X}^j\widetilde{\gamma}^j)/G =& D^{j\prime}F^2\mathbf{X}^j(\gamma^j - \widetilde{\gamma}^j)/G + \gamma^j\mathbf{X}^jF^2Z^j/G + Z^{j\prime}F^2Z^j/G\\ =&(I)_j + (II)_j + (III)_j.\end{aligned}$$

First, we bound $(I)_j$. Under Assumption 2, 3 (4) and Cauchy-Schwarz inequality, it holds uniformly that

$$\begin{aligned}\max_{j\in[p]}|(I)_j| \leq& \max_{j\in[p]}(D^{j\prime}F, F\mathbf{X}^j(\widetilde{\gamma}^j-\gamma^j))_G\\ \leq& \max_{k\in[p]}\Big\{\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}X_{ig,k}^2\Big\}^{1/2}\max_{j\in[p]}\|f_{ig}X_{ig}^{j\prime}(\widetilde{\gamma}^j-\gamma^j)\|_G\\ \lesssim_{\mathrm{P}}& \, O_{\mathrm{P}}(1)\cdot\sqrt{\frac{s\log a_G}{G}}\end{aligned}$$

with probability $1-C(\log G)^{-1}$.

We now bound $(II)_j$. The property of projection implies $\mathrm{E}_{\mathrm{P}}\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}f_{ig}^2X_{ig}^jZ_{ig}^j=0$,

$$\begin{aligned}\max_{j\in[p]}|(II)_j| \leq& \max_{j\in[p]}\|\gamma^j\|_1\|\mathbf{X}^jF^2Z^j/G\|_\infty\\ \leq& \max_{j\in[p]}\|\gamma^j\|_1\max_{j,k}\Big|\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}f_{ig}^2X_{ig,k}^jZ_{ig}^j\Big|\\ \leq& C_1\sqrt{s}\max_{j,k}\Big|\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\Big(f_{ig}^2X_{ig,k}^jZ_{ig}^j-\mathrm{E}_{\mathrm{P}}f_{ig}^2X_{ig,k}^jZ_{ig}^j\Big)\Big|.\end{aligned}$$

For each $j,k\in[p]$, denote the classes of functions

$$\begin{aligned}\mathcal{G}=&\Big\{W_g\mapsto\sum_{i=1}^{\bar{n}}\Lambda'(X_{ig}'\beta^0)X_{ig,k}^jZ_{ig}^j\mathbb{1}\{\||W_{ig}|\|_\infty>0\}:j,k\in[p]\Big\},\\ \mathcal{G}_{j,k}=&\Big\{W_g\mapsto\sum_{i=1}^{\bar{n}}\Lambda'(X_{ig}'\beta^0)X_{ig,k}^jZ_{ig}^j\mathbb{1}\{\||W_{ig}|\|_\infty>0\}\Big\}.\end{aligned}$$

Then each $\mathcal{G}_{j,k}$ contains only one function and thus is a VC-subgraph class with VC index equals unity with itself as an envelope. Also $\mathcal{G}\subset\cup_{j,k\in[p]}\mathcal{G}_{j,k}$. Since $|f_{ig}|\leq 1$, a measurable envelope for $\mathcal{G}$ is $H(W_g)=\max_{j,k}|U_{gk}V_g^j|$.

Some calculations and Assumption 3 (5)(6)(7)(8) give

$$\mathrm{E}_{\mathrm{P}}[\max_g|H(W_g)|^2]\lesssim\mathrm{E}_{\mathrm{P}}[\max_g\|U_g\|_\infty^4]+\mathrm{E}_{\mathrm{P}}[\max_g\max_{j\in[p]}|V_g^j|^4]\lesssim G^{2/q}(M_{G,1}^4+M_{G,2}^4).$$

The fact that $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$ for a, $b>0$ suggests $\{\mathrm{E}_{\mathrm{P}}\max_g|H(W_g)|^2\}^{1/2}\lesssim G^{1/q}(M_{G,1}^2+M_{G,2}^2)$. Similarly, under Assumption 3 (4), we have $\sup_{g\in\mathcal{G}}\mathrm{E}_{\mathrm{P}}\frac{1}{G}\sum_{g=1}^{G}[G^2(W_g)]\lesssim 1$. Applying Lemma 8 (1) and (2), we have for any $0<\epsilon\leq 1$,

$$N\left(\epsilon\|H\|_{Q,2},\mathcal{G},\|\cdot\|_{Q,2}\right)\lesssim p^2\max_{j,k}N(\epsilon\|G_{j,k}\|_{Q,2},\mathcal{G}_{j,k},\|\cdot\|_{Q,2})\lesssim p^2\Big(\frac{1}{\epsilon}\Big).$$

Thus one has $\sup_Q \log N\left(\epsilon\|H\|_{Q,2}, \mathcal{G}, \|\cdot\|_{Q,2}\right) \lesssim \log p \lesssim \log a_G$. Applying Corollary 3, we have with probability at least $1 - C(\log G)^{-1}$,

$$\max_{j,k\in[p]}\Big|\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(f_{ig}^2 X_{ig,k}^j Z_{ig}^j - \mathrm{E_P} f_{ig}^2 X_{ig,k}^j Z_{ig}^j\right)\Big| \lesssim \sqrt{\frac{\log a_G}{G}} + \frac{(M_{G,1}^2 \vee M_{G,2}^2)\log a_G}{G^{1-1/q}}.$$

Therefore, under Assumption 3 (6)(8),

$$\max_{j\in[p]}|(II)_j| \lesssim \sqrt{\frac{s\log a_G}{G}} + \frac{\sqrt{s}\log a_G(M_{G,1}^2 \vee M_{G,2}^2)}{G^{1-1/q}} \lesssim (s\log a_G/G)^{1/2}.$$

Now, we show $|(III)_j - \tau_j^2| = o_{\mathrm{P}}(1)$. Under Assumption 3 (4)(5)(6), using Lemma 8 (1) and (2), a similar argument leads to that with probability at least $1 - C(\log G)^{-1}$,

$$\max_{j\in[p]}|Z^{j\prime}F^2 Z^j/G - \tau_j^2| \lesssim \sqrt{\frac{\log a_G}{G}} + \frac{M_{G,1}^2\log a_G}{G^{1-1/q}} \lesssim \sqrt{\frac{s\log a_G}{G}}.$$

Therefore, we conclude that with probability at least $1 - C(\log G)^{-1}$, one has

$$\max_{j\in[p]}|\widehat{\tau}_j^2 - \tau_j^2| \lesssim \sqrt{\frac{s\log a_G}{G}}.$$

Step 4. By invoking Assumption 3 (1), we have for any G, one has $\tau_j^2 = 1/\Theta_{j,j} \geq 1/\Lambda_{\max}(\Theta) = \Lambda_{\min}(\Sigma) = \min_{\|\xi\|_2=1}\mathrm{E_P}[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(f_{ig}X_{ig}'\xi)^2] = c_1 > 0$. This implies that with probability at least $1 - C(\log G)^{-1}$, one has

$$\max_{j\in[p]}|1/\widehat{\tau}_j^2 - 1/\tau_j^2| \lesssim \sqrt{\frac{s\log a_G}{G}}.$$

Step 5. We now conclude the proof by deriving a bound for $\max_{j\in[p]}\|\widehat{\Theta}_j - \Theta_j\|_2$. By (F.38), Assumption 2 and use preceding steps, we have

$$\begin{aligned}\max_{j\in[p]}\|\widehat{\Theta}_j - \Theta_j\|_2 =& \max_{j\in[p]}\|\widehat{C}_j/\widehat{\tau}_j^2 - C_j/\tau_j^2\|_2\\ \leq& \max_{j\in[p]}\|\widetilde{\gamma}^j - \gamma^j\|_2/\widehat{\tau}_j^2 + \max_{j\in[p]}(\|\bar{\gamma}^j\|_2 + \|\gamma^j - \bar{\gamma}^j\|_2)|1/\widehat{\tau}_j^2 - 1/\tau_j^2|\\ \lesssim& \sqrt{\frac{s\log a_G}{G}}\cdot O_{\mathrm{P}}(1) + O_{\mathrm{P}}(1)\cdot\sqrt{\frac{s\log a_G}{G}} \lesssim \sqrt{\frac{s\log a_G}{G}}\end{aligned}$$

with probability at least $1 - C(\log G)^{-1}$. Similar arguments give $\max_{j\in[p]}\|\widehat{\Theta}_j - \Theta_j\|_1 \lesssim s\sqrt{\frac{\log a_G}{G}}$ with probability at least $1 - C(\log G)^{-1}$. ∎

E.5. **Proof for Theorem 4.**

Proof. Since $\|\Lambda''\|_\infty \lesssim 1$, one has

$$\max_{k\in[p]} \|\widehat{\theta}^k - \theta^k\|_2 \leq \max_{k\in[p]} \|\widehat{\Theta}_k - \Theta_k\|_2 \Big| \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} \Lambda'(X_{ig}'\widetilde{\beta})\Big| + \max_{k\in[p]} \|\Theta_k\|_2 \Big|\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\Big(\Lambda'(X_{ig}'\widetilde{\beta}) - \Lambda'(X_{ig}'\beta^0)\Big)\Big|.$$

Assumptions 1 and 3 (2) imply $\max_{k\in[p]} \|\Theta_k\|_2 \leq C_1$. Furthermore, using equation (I.6) of Belloni, Chernozhukov, Chetverikov and Wei (2018) and the fact $\Lambda' = \Lambda \cdot (1-\Lambda)$, suppose that $|X_{ig}'(\widetilde{\beta} - \beta^0)| \leq 1$, it holds that

$$\begin{aligned}\Big|\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\Big(\Lambda'(X_{ig}'\widetilde{\beta}) - \Lambda'(X_{ig}'\beta^0)\Big)\Big| =& \Big|\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\Big(\widetilde{\Lambda}_{ig}(1-\widetilde{\Lambda}_{ig}) - \Lambda_{ig}(1-\widetilde{\Lambda}_{ig}) + \Lambda_{ig}(1-\widetilde{\Lambda}_{ig}) - \Lambda_{ig}(1-\Lambda_{ig})\Big)\Big| \\ \lesssim& (\|\Lambda\|_\infty + \|1-\Lambda\|_\infty)\max_{i,g}|f_{ig}|\Big|\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} f_{ig}X_{ig}'(\widetilde{\beta}-\beta^0)\Big| \\ \lesssim& O(1)\cdot \|f_{ig}X_{ig}'(\widetilde{\beta}-\beta^0)\|_G \\ \lesssim& \sqrt{\frac{s\log a_G}{G}}\end{aligned}$$

with probability $1 - C(\log G)^{-1/2}$, where $\Lambda_{ig} = \Lambda'(X_{ig}'\beta)$, $\widetilde{\Lambda}_{ig} = \Lambda'(X_{ig}'\widetilde{\beta})$ and the second inequality is due to an application of Cauchy-Schwarz inequality. The condition $|X_{ig}'(\widetilde{\beta} - \beta^0)| \leq 1$ holds asymptotically with probability $1 - o(1)$ since

$$\max_{i,g}\|X_{ig}\|_\infty \|\widetilde{\beta} - \beta^0\|_1 \lesssim_{\mathrm{P}} \frac{M_{G,2}s(\log a_G)^{1/2}}{G^{1/2-1/2q}} = o(1)$$

with probability $1 - C(\log G)^{-1}$ under Assumption 3 (7)(8) and Theorem 2. Furthermore,

$$\max_{k\in[p]}\|\widehat{\Theta}_k - \Theta_k\|_2 \lesssim_{\mathrm{P}} \sqrt{\frac{s\log a_G}{G}}$$

following Theorem 3 and $\widehat{\tau}^{-2} = O(\tau^{-2}) = O(1)$. So

$$\max_{k\in[p]}\|\widetilde{\theta}^k - \theta^k\|_2 \leq \sqrt{\frac{s\log a_G}{G}}.$$

The bound $\max_{k\in[p]}\|\widetilde{\theta}^k - \theta^k\|_1 \leq s\sqrt{\frac{\log a_G}{G}}$ with probability at least $1 - C(\log G)^{-1}$ can be established following similar arguments and the fact that $\max_{k\in[p]}\|\Theta_k\|_1 \leq \sqrt{s}C_1$. ■

Supplementary Appendix to "Many Average Partial Effects: with an Application to Text Regressions"

Harold D. Chiang

Vanderbilt University

This supplementary appendix includes sections that contain additional theoretical results used in proving the main results in the previous sections as well as their proofs. Most of these results follow closely from existing results in the literature under some minor modifications. We include them for the sake of completeness.

Appendix F. Additional Theoretical Results

F.1. Properties of τ_j^2.

In this Section, we derives some important properties of τ_j^2, which is based on the work of Kock (2016), a panel data generalization of the nodewise lasso in van de Geer, Bühlmann, Ritov and Dezeure (2014). Denote $\Sigma = \mathrm{E_P}\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} f_{ig}^2 X_{ig}X_{ig}'$. Let $\Sigma_{-j,-j}$ be the $(p-1)\times(p-1)$ submatrix of Σ with the j-th column and row removed. $\Sigma_{j,-j}$ represents the j-th row of Σ with its j-th element removed and $\Sigma_{-j,j}$ is defined analogously. From the inverse formula of a partitioned matrix, we have

$$\begin{aligned}\Theta_{j,j} &= (\Sigma_{j,j} - \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j})^{-1}\\ \Theta_{j,-j} &= (\Sigma_{j,j} - \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{j,-j})\Sigma_{j,-j}\Sigma_{-j,-j}^{-1} = -\Theta_{j,j}\Sigma_{j,-j}\Sigma_{-j,-j}^{-1}.\end{aligned}$$

Now, by solving (5.20), we have

$$\begin{aligned}\gamma^j =&\Big\{\mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} f_{ig}^2 X_{ig}^{j\prime}X_{ig}^{j}\Big]\Big\}^{-1}\cdot \mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} f_{ig}^2 X_{ig}^{j\prime}D_{ig}^{j}\Big]\\ =&\Sigma_{-j,-j}^{-1}\Sigma_{j,-j}'\end{aligned}$$

Combining with above, we have

$$\Theta_{j,-j} = -\Theta_{j,j}\gamma^{j\prime}. \tag{F.36}$$

Furthermore, using $D^j = \mathbf{X}^j\gamma^j + Z^j$ and $\mathrm{E_P}[Z^{j\prime}F^2X^j] = 0$, we have

$$\begin{aligned}\Sigma_{j,j} =& \mathrm{E_P}[D^{j\prime}F^2D^j]\\ =& \gamma^{j\prime}\mathrm{E_P}[\mathbf{X}^{j\prime}F^2\mathbf{X}^j]\gamma^j + \mathrm{E_P}[Z^{j\prime}F^2Z^j] + 2\mathrm{E_P}[Z^{j\prime}F^2\mathbf{X}^j]\gamma^j\\ =& \Sigma_{j,-j}\Sigma^{-1}_{-j,-j}\Sigma'_{-j,j} + \tau_j^2 + 0.\end{aligned}$$

Therefore we have

$$\tau_j^2 = \Sigma_{j,j} - \Sigma_{j,-j}\Sigma^{-1}_{-j,-j}\Sigma'_{-j,j} = 1/\Theta_{j,j}. \tag{F.37}$$

Now define

$$C = \begin{bmatrix} 1 & -\gamma^1_1 & \dots & -\gamma^1_{p-1} \\ -\gamma^2_1 & 1 & \dots & -\gamma^2_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ -\gamma^p_1 & -\gamma^p_2 & \dots & 1 \end{bmatrix}$$

and $T^2 = \mathrm{diag}\{\tau_1^2, ..., \tau_p^2\}$, using (F.36) and (F.37), we have

$$\Theta = T^{-2}C. \tag{F.38}$$

F.2. **Results for Nuisance Parameters Estimation.**

The following results generalizes lemmas in Appendix L of Belloni, Chernozhukov, Chetverikov and Wei (2018) to cluster sampling. Their proofs follow closely those of Lemma L1-L4 of Belloni, Chernozhukov, Chetverikov and Wei (2018) but we only consider an increasing finite index set for simplicity.

F.2.1. *ℓ-1 Penalized M-Estimation with Clustered Data.* Consider a data generating process with an outcome variable Y^k_{ig} and p-dimensional covariates X^k_{ig}, both indexed by $k \in \mathcal{U}_G$ for some $\mathcal{U}_G \subset [p]$. We maintain the cluster sampling setting as before. The parameter of interest

$$\mu^k \in \underset{\mu\in\mathbb{R}^p}{\mathrm{argmin}}\mathrm{E_P}\Big[\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} M_k(Y^k_{ig}, X^k_{ig}, \mu)\Big]. \tag{F.39}$$

Define the lasso and post-lasso estimators

$$\widehat{\mu}^k \in \underset{\mu\in\mathbb{R}^p}{\mathrm{argmin}}\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}(Y^k_{ig}, X^k_{ig}, \mu) + \frac{\lambda}{G}\|\widehat{\Psi}_k\mu\|_1, \tag{F.40}$$

$$\widetilde{\mu}^k \in \underset{\mu\in\mathrm{support}(\widehat{\mu}_k)}{\mathrm{argmin}}\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}_k(Y^k_{ig}, X^k_{ig}, \mu). \tag{F.41}$$

For each $k \in \mathcal{U}_G$, denote the *ideal* penalty loadings $\widehat{\Psi}_{k0} = \operatorname{diag}(\{l_{k0j} : j \in [p]\})$, where

$$l_{k0j} = \Big\{ \frac{1}{G} \sum_{g=1}^{G} \Big(\sum_{i=1}^{n_g} \partial_{\mu_j} M_k(Y_{ig}^k, X_{ig}^k, \mu^k) \Big)^2 \Big\}^{1/2} = \Big\{ \frac{1}{G} \sum_{g=1}^{G} (S_{gj}^k)^2 \Big\}^{1/2},$$

where $S_{gj}^k = \sum_{i=1}^{n_g} \partial_{\mu_j} M_k(Y_{ig}^k, X_{ig}^k, \mu^k)$. We also denote the feasible penalty loadings by $\widehat{\Psi}_k = \operatorname{diag}(\{l_{kj} : j \in [p]\})$ for some l_{kj}

$$l_{kj} = \Big\{ \frac{1}{G} \sum_{g=1}^{G} \Big(\sum_{i=1}^{n_g} \partial_{\mu_j} \widehat{M}_k(Y_{ig}^k, X_{ig}^k, \widehat{\mu}^k) \Big)^2 \Big\}^{1/2} = \Big\{ \frac{1}{G} \sum_{g=1}^{G} (\widehat{S}_{gj}^k)^2 \Big\}^{1/2},$$

where $\widehat{S}_{gj}^k = \sum_{i=1}^{n_g} \partial_{\mu_j} \widehat{M}_k(Y_{ig}^k, X_{ig}^k, \widehat{\mu}^k)$. Also write $S_g^k = (\{S_{gj}^k : j \in [p]\})$ and $\widehat{S}_g^k = (\{\widehat{S}_{gj}^k : j \in [p]\})$. Denote $T_k = \operatorname{support}(\mu^k)$ and $\widehat{T}_k = \operatorname{support}(\widehat{\mu}^k)$. We assume λ is chosen such that with high probability,

$$\frac{\lambda}{G} \geq c \max_{k \in \mathcal{U}_G} \|\widehat{\Psi}_0^{-1} \sum_{i=1}^{n_g} \partial_\mu M(Y_{ig}^k, X_{ig}^k, \mu^k)\|_\infty, \tag{F.42}$$

for a fixed constant $c > 1$. This will be shown to happen under some sufficient conditions in Section F.2.2. Let $L \geq \ell > 1/c$ be some fixed constants and let

$$\widetilde{c} = \frac{Lc + 1}{\ell c - 1} \max_{k \in \mathcal{U}_G} \|\widehat{\Psi}_{k0}\|_\infty \|\widehat{\Psi}_{k0}^{-1}\|_\infty.$$

Denote $s_k = \|\mu^k\|_0$ and let $\widetilde{\Delta}_G$ be a sequence of positive constants converging to zero, let $\widetilde{C}_G$ be a sequence of random variables and $w_{ig} = w(X_{ig})$ be some weights such that $0 \leq w_{ig} \leq 1$ almost surely. Finally, let A_k be a random subset of $\mathbb{R}^p$ and $\bar{q}_{A_k}$ a random variable depends possibly on A_k.

Assumption 8. *Suppose that* $\max_{k \in \mathcal{U}_G} \|\mu^k\|_0 = s$ *and for each* $k \in [p]$ $\mu \mapsto \widehat{M}_k(y, x, \mu)$ *is convex almost surely and with probability at least* $1 - \widetilde{\Delta}_G$ *for all* $\delta \in \mathbb{R}^p$, *it holds that for all* $k \in \mathcal{U}_G$,

- *(1)* $\Big|\Big\{ \frac{1}{G} \sum_{g=1}^{G} \sum_{i=1}^{n_g} [\partial_\mu \widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k) - \partial_\mu M_k(Y_{ig}^k, X_{ig}^k, \mu^k)] \Big\}' \delta\Big| \leq C_G \|\sqrt{w_{ig}} X_{ig}^{k\prime} \delta\|_G$ *for all* $\delta \in \mathbb{R}^p$;
- *(2)* $\ell \widehat{\Psi}_{k0} \leq \widehat{\Psi}_k \leq L \widehat{\Psi}_{k0}$;
- *(3) for all* $\delta \in A_k$,

$$\frac{1}{G} \sum_{g=1}^{G} \sum_{i=1}^{n_g} \widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k + \delta) - \frac{1}{G} \sum_{g=1}^{G} \sum_{i=1}^{n_g} \widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k) - \frac{1}{G} \sum_{g=1}^{G} \sum_{i=1}^{n_g} [\partial_\mu \widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k + \delta)]' \delta$$

$$+ 2C_G \|\sqrt{w_{ig}} X_{ig}^{k\prime} \delta\|_G \geq \{\|\sqrt{w_{ig}} X_{ig}^{k\prime} \delta\|_G^2\} \wedge \{\bar{q}_{A_k} \|\sqrt{w_{ig}} X_{ig}^{k\prime} \delta\|_G\}.$$

Define the restricted eigenvalue

$$\bar{\kappa}_{2\widetilde{c}} = \min_{k \in \mathcal{U}_G} \inf_{\delta \in \Delta_{2\widetilde{c},k}} \frac{\|\sqrt{w_{ig}} X_{ig}' \delta\|_G}{\|\delta_{T_k}\|_2},$$

where $\Delta_{2\widetilde{c},k} = \{\delta \in \mathbb{R}^p : \|\delta_{T_k^c}\|_1 \leq 2\widetilde{c}\|\delta_{T_k}\|_1\}$. In addition, define minimum and maximum sparse eigenvalues

$$\phi_{\min}(m,k) = \min_{1 \leq \|\delta\|_0 \leq m} \frac{\|\sqrt{w_{ig}} X_{ig}^{k\prime} \delta\|_G^2}{\|\delta\|_2^2} \text{ and } \phi_{\max}(m,k) = \max_{1 \leq \|\delta\|_0 \leq m} \frac{\|\sqrt{w_{ig}} X_{ig}^{k\prime} \delta\|_G^2}{\|\delta\|_2^2}.$$

Boundedness of minimum and maximum sparse eigenvalues with probability goes to 1 implies that restricted eigenvalue is bounded away from 0 with probability goes to 1. For its proof, see Lemma 4.1 of Bickel, Ritov and Tsybakov (2009).

Lemma 3. *Suppose that Assumption 8 holds with*

$$A_k = \Delta_{2\widetilde{c},k} \cup \{\delta \in \mathbb{R}^p : \|\delta\|_1 \leq \frac{3G}{\lambda} \frac{c\|\widehat{\Psi}_{k0}^{-1}\|_\infty}{\ell c - 1} C_G \|\sqrt{w_{ig}} X_{ig}^{k\prime} \delta\|_G\},$$

and $\bar{q}_{A_k} \geq (L + \frac{1}{c})\|\widehat{\Psi}_{k0}\|_\infty \frac{\lambda\sqrt{s}}{G\bar{\kappa}_{2\widetilde{c}}} + 6\widetilde{c}C_G$. *In addition, suppose that* λ *satisfies condition F.42 with probability at least* $1 - \widetilde{\Delta}_G$. *Then, with probability at least* $1 - 2\widetilde{\Delta}_G$, *we have*

$$\begin{aligned}
\|\sqrt{w_{ig}} X_{ig}^{k\prime}(\widehat{\mu}^k - \mu^k)\|_G \leq& \Big(L + \frac{1}{c}\Big)\|\widehat{\Psi}_{k0}\|_\infty \frac{\lambda\sqrt{s}}{G\bar{\kappa}_{2\widetilde{c}}} + 6\widetilde{c}C_G, \\
\|\widehat{\mu}^k - \mu^k\|_1 \leq& \Big(\frac{(1+2\widetilde{c})\sqrt{s}}{\bar{\kappa}_{2\widetilde{c}}} + \frac{3G}{\lambda}\frac{c\|\widehat{\Psi}_{k0}^{-1}\|_\infty}{\ell c - 1} C_G\Big)\Big(\Big(L + \frac{1}{c}\Big)\|\widehat{\Psi}_{k0}^{-1}\|_\infty \frac{\lambda\sqrt{s}}{G\bar{\kappa}_{2\widetilde{c}}} + 6\widetilde{c}C_G\Big)
\end{aligned}$$

uniform for $k \in \mathcal{U}_G$.

Lemma 4. *In addition to conditions of Lemma 3, suppose that with probability* $1 - \widetilde{\Delta}_G$, *for some random variable* L_G *such that for all* $\delta \in \mathbb{R}^p$, *it holds that*

$$\Big|\Big\{\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}[\partial_\mu \widehat{M}_k(Y_{ig}^k, X_{ig}^k, \widehat{\mu}^k) - \partial_\mu \widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k)]\Big\}'\delta\Big| \leq L_G\|X_{ig}^{k\prime}\delta\|_G. \tag{F.43}$$

Then with probability $1 - 3\widetilde{\Delta}_G$, *we have for all* $k \in \mathcal{U}_G$,

$$\widehat{s}_k \leq \min_{m \in \mathcal{M}_k} \phi_{\max}(m,k) L_k^2,$$

where $\mathcal{M}_k = \{m \in \mathbb{N} : m \geq 2\phi_{\max}(m,k)L_k^2\}$ *and* $L_k = \frac{c\|\widehat{\Psi}_{k0}^{-1}\|_\infty}{c\ell - 1}\frac{G}{\lambda}\{C_G + L_G\}$.

Lemma 5. *Suppose that Assumption 8 holds with* $A_k = \{\delta \in \mathbb{R}^p : \|\delta\|_0 \leq \widehat{s}_k + s_k\}$ *and*

$$\bar{q}_{A_k} > 2\max\Big\{\Big(\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}[\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \widetilde{\mu}^k) - \widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k)]\Big)_+^{1/2}, \\ \Big(\frac{\sqrt{\widehat{s}_k + s_k}\|\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\partial_\mu M_k(Y_{ig}^k, X_{ig}^k, \mu^k)\|_\infty}{\sqrt{\phi_{\min}(\widehat{s}_k + s_k)}} + 3C_G\Big)\Big\}. \tag{F.44}$$

Then with probability at least $1 - \widetilde{\Delta}_G$*,*

$$\|\sqrt{w_{ig}}X_{ig}^{k\prime}(\widetilde{\mu}^k - \mu^k)\|_G \leq \Big\{\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}[\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \widetilde{\mu}^k) - \widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k)]\Big\}_+^{1/2} \\ + \frac{\sqrt{\widehat{s}_k + s_k}\|\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\partial_\mu M_k(Y_{ig}^k, X_{ig}^k, \mu^k)\|_\infty}{\sqrt{\phi_{\min}(\widehat{s}_k + s_k)}} + 3C_G$$

uniform for $k \in \mathcal{U}_G$*. In addition, with probability at least* $1 - \widetilde{\Delta}_G$*, one has*

$$\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\partial_\mu\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \widetilde{\mu}^k) - \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\partial_\mu\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k) \leq L\frac{\lambda}{G}\|\widehat{\mu}^k - \mu^k\|_1\|\widehat{\Psi}_{k0}\|_\infty. \tag{F.45}$$

Therefore, with probability at least $1 - \widetilde{\Delta}_G$*, we have*

$$\|\widetilde{\mu}^k - \mu^k\|_1 \leq \frac{\sqrt{\widehat{s}_k + s_k}}{\sqrt{\phi_{\max}(\widehat{s}_k + s_k)}\min_{i,g} w_{ig}^2}\Big(L\frac{\lambda}{G}\|\widehat{\mu}^k - \mu^k\|_1\|\widehat{\Psi}_{k0}\|_\infty + \frac{\lambda\sqrt{\widehat{s}_k + s_k}}{cG\sqrt{\phi_{\min}(\widehat{s}_k + s_k)}} + 3C_G\Big)$$

uniform for $k \in \mathcal{U}_G$*.*

F.2.2. *Concentration for Regularized Events.* We now provide sufficient conditions for F.42. Denote $|\mathcal{U}_G| = \widetilde{p}$.

Assumption 9. *Suppose that the following holds for each* G*,*

(1) $\max_{k\in\mathcal{U}_G}\max_{j\in[p]}(\mathrm{E}_{\mathrm{P}}\frac{1}{G}\sum_{g=1}^{G}|S_{gj}^k|^3)^{1/3}\Phi^{-1}(1-\gamma/2p) \leq \widetilde{\varphi}_G G^{1/6}$ *for* $j \in [\widetilde{p}]$*.*
(2) $\underline{C} \leq (\mathrm{E}_{\mathrm{P}}\frac{1}{G}\sum_{g=1}^{G}|S_{gj}^k|^2)^{1/2} \leq \overline{C}$ *for all* $k \in \mathcal{U}_G$ *for* $j \in [\widetilde{p}]$*.*

Let

$$\lambda = c'\sqrt{G}\Phi^{-1}(1 - \gamma/2p\widetilde{p}), \tag{F.46}$$

where $\gamma = \gamma_G = o(1)$.

Lemma 6. *Suppose that 9 holds and λ satisfies (F.46) with some $c' > c$ and $\gamma = \gamma_G \in [1/G, 1/\log G]$. Then*

$$\mathrm{P_P}\left(\frac{\lambda}{G} \geq c \max_{k \in \mathcal{U}_G} \left\| \widehat{\Psi}_k^{-1} \frac{1}{G} \sum_{g=1}^{G} S_g^k \right\|_\infty \right) \geq 1 - \gamma - o(\gamma).$$

Appendix G. Proof for Additional Results

G.1. Proof for Lemma 3.

Proof. Denote $\delta^k = \widehat{\mu}^k - \mu^k$. Assume the events of Assumption 8 and (F.42) holds. This happens with probability at least $1 - 2\widetilde{\Delta}_G$. By definition of $\widehat{\mu}$,

$$\begin{aligned}\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \widehat{\mu}^k) - \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k) \leq & \frac{\lambda}{G}\|\widehat{\Psi}\mu^k\|_1 - \frac{\lambda}{G}\|\widehat{\Psi}\widehat{\mu}^k\|_1 \\ \leq & L\frac{\lambda}{G}\|\widehat{\Psi}_{k0}\delta_{k,T_k}\|_1 - \ell\frac{\lambda}{G}\|\widehat{\Psi}_{k0}\delta_{k,T_k^c}\|_1.\end{aligned} \tag{G.47}$$

Furthermore, Assumption 8 (a) and the convexity of M in μ as well as condition (F.42) suggest

$$\begin{aligned}&\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \widehat{\mu}^k) - \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k) \\ \geq & \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}[\partial_\mu \widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k)]'\delta_k \geq -\frac{\lambda}{G}\frac{1}{c} - C_G\|\sqrt{w_{ig}}X_{ig}^{k\prime}\delta_k\|_G.\end{aligned} \tag{G.48}$$

Combining (G.47) and (G.48) gives

$$\frac{\lambda}{G}\frac{\ell c - 1}{c}\|\widehat{\Psi}_{k0}\delta_{k,T_k^c}\|_1 \leq \frac{\lambda}{G}\frac{Lc+1}{c}\|\widehat{\Psi}_{k0}\delta_{k,T_k}\|_1 + C_G\|\sqrt{w_{ig}}X_{ig}^{k\prime}\delta_k\|_G. \tag{G.49}$$

Thus

$$\|\delta_{k,T_k^c}\|_1 \leq \widetilde{c}\|\delta_{k,T_k}\|_1 + \frac{G}{\lambda}\frac{c\|\widehat{\Psi}_{k0}^{-1}\|_\infty}{\ell c - 1}C_G\|\sqrt{w_{ig}}X_{ig}^{k\prime}\delta_k\|_G.$$

Consider the case that $\delta \notin \Delta_{2\widetilde{c},k}$, then since $\widetilde{c} \geq 1$,

$$\|\delta_{k,T_k}\|_1 \leq \frac{G}{\lambda}\frac{c\|\widehat{\Psi}_{k0}^{-1}\|_\infty}{\ell c - 1}C_G\|\sqrt{w_{ig}}X_{ig}^{k\prime}\delta_k\|_G.$$

Also from above,

$$\|\delta_{k,T_k^c}\|_1 \leq \frac{1}{2}\|\delta_{k,T_k^c}\|_1 + \frac{G}{\lambda}\frac{c\|\widehat{\Psi}_{k0}^{-1}\|_\infty}{\ell c - 1}C_G\|\sqrt{w_{ig}}X_{ig}^{k\prime}\delta_k\|_G,$$

and thus

$$\|\delta_{k,T_k^c}\|_1 \leq \frac{2G}{\lambda}\frac{c\|\widehat{\Psi}_{k0}^{-1}\|_\infty}{\ell c-1}C_G\|\sqrt{w_{ig}}X_{ig}^{k\prime}\delta_k\|_G.$$

Adding them up, one has

$$\|\delta_k\|_1 \leq \frac{3G}{\lambda}\frac{c\|\widehat{\Psi}_{k0}^{-1}\|_\infty}{\ell c-1}C_G\|\sqrt{w_{ig}}X_{ig}^{k\prime}\delta_k\|_G := I_k.$$

Now suppose that $\delta \in \Delta_{2\widetilde{c},k}$, the definition of $\bar{\kappa}_{2\widetilde{c}}$ gives

$$\|\delta_{k,T_k}\|_1 \leq \sqrt{s}\|\delta_{k,T_k}\|_2 \leq \frac{\sqrt{s}}{\bar{\kappa}_{2\widetilde{c}}}\|\sqrt{w_{ig}}X_{ig}^{k\prime}\delta_k\|_G = II_k.$$

So by combining two cases, we have

$$\|\delta_{k,T_k}\|_1 \leq I_k + II_k. \tag{G.50}$$

Recall that

$$A_k = \left\{\delta \in \mathbb{R}^p : \|\delta\|_1 \leq \frac{3G}{\lambda}\frac{c\|\widehat{\Psi}_{k0}^{-1}\|_\infty}{\ell c-1}C_G\|\sqrt{w_{ig}}X_{ig}^{k\prime}\delta\|_G\right\}.$$

By invoking Assumption 8 (3), we have

$$\begin{aligned}
&\{\|\sqrt{w_{ig}}X_{ig}^{k\prime}\delta_k\|_G^2\} \wedge \{\bar{q}_{A_k}\|\sqrt{w_{ig}}X_{ig}^{k\prime}\delta_k\|_G\}\}\\
\leq&\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k+\delta_k) - \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k) - \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}[\partial_\mu\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k+\delta_k)]'\delta_k\\
&+ 2C_G\|\sqrt{w_{ig}}X_{ig}^{k\prime}\delta_k\|_G\\
\leq&\Big(L+\frac{1}{c}\Big)\frac{\lambda}{G}\|\widehat{\Psi}_{k0}\delta_{k,T_k}\|_1 + 3C_G\|\sqrt{w_{ig}}X_{ig}^{k\prime}\delta_k\|_G\\
\leq&\Big(L+\frac{1}{c}\Big)\frac{\lambda}{G}\|\widehat{\Psi}_{k0}\|_\infty(I_k+II_k) + 3C_G\|\sqrt{w_{ig}}X_{ig}^{k\prime}\delta_k\|_G\\
\leq&\Big\{\Big(L+\frac{1}{c}\Big)\|\widehat{\Psi}_{k0}\|_\infty\frac{\lambda\sqrt{s}}{G\bar{\kappa}_{2\widetilde{c}}} + 6\widetilde{c}C_G\Big\}\|\sqrt{w_{ig}}X_{ig}^{k\prime}\delta_k\|_G.
\end{aligned}$$

The definition of A implies that the minimum on the left-hand side must be achieved by the quadratic term and thus

$$\|\sqrt{w_{ig}}X_{ig}^{k\prime}\delta_k\|_G \leq \Big\{\Big(L+\frac{1}{c}\Big)\|\widehat{\Psi}_{k0}\|_\infty\frac{\lambda\sqrt{s}}{G\bar{\kappa}_{2\widetilde{c}}} + 6\widetilde{c}C_G\Big\}.$$

Finally,

$$\|\delta_k\|_1 \leq (1+2\widetilde{c})II_k + I_k \leq \Big(\frac{(1+2\widetilde{c})\sqrt{s}}{\bar{\kappa}_{2\widetilde{c}}} + \frac{3G}{\lambda}\frac{c\|\widehat{\Psi}_{k0}^{-1}\|_\infty}{\ell c-1}C_G\Big)$$

uniform for $k \in \mathcal{U}_G$. ■

G.2. **Proof for Lemma 4.**

Proof. Let $S_G^k = \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} M_k(Y_{ig}^k, X_{ig}^k, \mu^k)$. Assume the events of Assumption 8, conditions (F.42) and (F.43) holds. This happens with probability at least $1-3\widetilde{\Delta}_G$.

By definition of $\widehat{\mu}^k$, for all $j \in \widehat{T}_k$,

$$\Big|(\widehat{\Psi}_k^{-1}\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\partial_{\mu_j}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \widehat{\mu}^k)\Big| = \frac{\lambda}{G}.$$

Therefore, using Assumption 8 (1),(2), and inequalities (F.42),(F.43),

$$\begin{aligned}
&\frac{\lambda}{G}\sqrt{s_k} = \|(\widehat{\Psi}_k^{-1}\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\partial_{\mu}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \widehat{\mu}^k))_{\widehat{T}_k}\|_2 \\
\leq&\|(\widehat{\Psi}_k^{-1}S_G^k)_{\widehat{T}_k}\|_2 + \|(\widehat{\Psi}_k^{-1}\{\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\partial_{\mu}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k) - S_G^k\})_{\widehat{T}_k}\|_2 \\
&+ \|(\widehat{\Psi}_k^{-1}\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\{\partial_{\mu}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \widehat{\mu}^k) - \partial_{\mu}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k)\})_{\widehat{T}_k}\|_2 \\
\leq&\sqrt{s}\|\widehat{\Psi}_k^{-1}\widehat{\Psi}_{k0}\|_\infty\|S_G^k\|_\infty + \|\widehat{\Psi}_k^{-1}\|_\infty C_G \sup_{\|\delta\|_2=1,\|\delta\|_0\leq\widehat{s}_k}\|\sqrt{w_{ig}}X_{ig}^{k\prime}\delta\|_G \\
&+ \|\widehat{\Psi}_k^{-1}\|_\infty \sup_{\|\delta\|_2=1,\|\delta\|_0\leq\widehat{s}_k}\Big|\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}[\partial_{\mu}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \widehat{\mu}^k) - \partial_{\mu}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k)]'\delta\Big| \\
\leq&\frac{\lambda}{c\ell G}\sqrt{s_k} + \frac{\|\widehat{\Psi}_{k0}^{-1}\|_\infty}{\ell}\{C_G + L_G\}\sup_{\|\delta\|_2=1,\|\delta\|_0\leq\widehat{s}_k}\|X_{ig}^{k\prime}\delta\|_G.
\end{aligned}$$

Note that $\sup_{\|\delta\|_2=1,\|\delta\|_0\leq\widehat{s}_k}\|X_{ig}^{k\prime}\delta\|_G = \phi_{\max}(\widehat{s}_k, k)$,

$$\widehat{s}_k \leq \phi_{\max}(\widehat{s}_k)L_k^2.$$

The rest follows from the sublinearity of maximum sparse eigenvalue and minimizing over $M \in \mathcal{M}_k$. ■

G.3. **Proof for Lemma 5.**

Proof. First, note that by definition of $\widetilde{\mu}^k$ and $\widehat{\mu}^k$

$$
\begin{aligned}
&\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\partial_\mu \widehat{M}_k(Y_{ig}^k, X_{ig}^k, \widetilde{\mu}^k - \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\partial_\mu \widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k)\\
\leq&\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\partial_\mu \widehat{M}_k(Y_{ig}^k, X_{ig}^k, \widehat{\mu}^k) - \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\partial_\mu \widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k)\\
\leq&L\frac{\lambda}{G}\|\widehat{\mu}^k - \mu^k\|_1\|\widehat{\Psi}_{k0}\|_\infty
\end{aligned}
$$

with probability at least $1 - \widetilde{\Delta}_G$.

To show the first claim, let us suppose the events of Assumption 8 holds with probability $1 - \widetilde{\Delta}_G$. Denote $\delta_k = \widetilde{\mu}^k - \mu^k$ and $S_G^k = \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} M_k(Y_{ig}^k, X_{ig}^k, \mu^k)$ and $t_k = \|\sqrt{w_{ig}}X_{ig}^{k\prime}\delta_k\|_G$. Assumption 8 (3) gives

$$
\begin{aligned}
t_k^2 \wedge \{\bar{q}_{A_k} t_k\} \leq&\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \widetilde{\mu}^k) - \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k)\\
&-\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}[\partial_\mu \widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k)]'\delta_k + 2C_G t_k\\
\leq&\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \widetilde{\mu}^k) - \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k)\\
&+\|S_G^k\|_\infty\|\delta_k\|_1 + 3C_G t_k\\
\leq&\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \widetilde{\mu}^k) - \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k)\\
&+\Big(\frac{\sqrt{\widehat{s}_k + s_k}\|S_G^k\|_\infty}{\sqrt{\phi_{\min}(\widehat{s}_k + s_k, k)}} + 3C_G\Big)t_k.
\end{aligned}
$$

where the last inequality follows from

$$
\|\delta_k\|_1 \leq \sqrt{\widehat{s}_k + s_k}\|\delta_k\|_2 \leq \frac{\sqrt{\widehat{s}_k + s_k}}{\sqrt{\phi_{\min}(\widehat{s}_k + s_k, k)}}\|\sqrt{w_{ig}}X_{ig}^{k\prime}\delta_k\|_G.
$$

We then consider two cases. First, suppose $t_k^2 > \bar{q}_{A_k} t_k$, by definition of $\bar{q}_{A_k}$

$$
\bar{q}_{A_k} t_k \leq \frac{\bar{q}_{A_k}}{2}\Big\{\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \widetilde{\mu}^k) - \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k)\Big\}_+^{1/2} + \frac{\bar{q}_{A_k}}{2} t_k,
$$

and thus $t_k \le \{\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \widetilde{\mu}^k) - \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k)\}_+^{1/2}$. Now suppose $t_k^2 \le \bar{q}_{A_k} t_k$, then

$$t_k^2 \le \Big\{\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \widetilde{\mu}^k) - \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k)\Big\} + \Big(\frac{\sqrt{\widehat{s}_k + s_k}\|S_G^k\|_\infty}{\sqrt{\phi_{\min}(\widehat{s}_k + s_k, k)}} + 3C_G\Big)t_k.$$

Since for any positive numbers a, b, c, $a^2 \le b + ac$ implies $a \le \sqrt{b} + c$, one has

$$t_k \le \Big\{\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \widetilde{\mu}^k) - \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{M}_k(Y_{ig}^k, X_{ig}^k, \mu^k)\Big\}_+^{1/2} + \Big(\frac{\sqrt{\widehat{s}_k + s_k}\|S_G^k\|_\infty}{\sqrt{\phi_{\min}(\widehat{s}_k + s_k, k)}} + 3C_G\Big).$$

■

G.4. **Proof for Lemma 9.**

Proof. By Assumption 9, we have for $\ell_G = c''/\widetilde{\varphi}_G$, c'' a constant depends only on $\underline{C}$, $\overline{C}$,

$$0 \le \Phi^{-1}(1-\gamma/2p) \le \frac{G^{1/6}(\mathrm{E}_{\mathrm{P}}\frac{1}{G}\sum_{g=1}^{G}|S_g^k|^2)^{1/2}/(\mathrm{E}_{\mathrm{P}}\frac{1}{G}\sum_{g=1}^{G}|S_g^k|^3)^{1/3}}{\ell_G} - 1.$$

for all $k \in \mathcal{U}_G$. Applying inequalty for self-normalized sums (Lemma 5 in Belloni, Chernozhukov, Chen and Hansen (2012)), we have

$$\begin{aligned}
&\mathrm{P}_{\mathrm{P}}\Big(\frac{\lambda}{G} \ge c\max_{k\in\mathcal{U}_G}\|\widehat{\Psi}_k^{-1}\frac{1}{G}\sum_{g=1}^{G}S_g^k\|_\infty\Big)\\
\ge&\mathrm{P}_{\mathrm{P}}\left(\Phi^{-1}(1-\gamma/2p\widetilde{p}) \ge \max_{k\in\mathcal{U}_G}\max_{j\in[p]}\frac{|\sqrt{G}\frac{1}{G}\sum_{g=1}^{G}S_{gj}^k|}{\sqrt{\frac{1}{G}\sum_{g=1}^{G}(S_{gj}^k)^2}}\right)\\
\ge&1-\gamma-o(\gamma).
\end{aligned}$$

■

Corollary 3. *Given the Assumptions of Lemma 7. Denote* $M = \mathrm{E}_{\mathrm{P}}\frac{1}{G}\sum_{g=1}^{G}[F]^{1/2}$. *Suppose there exist constants* $a \ge n$ *and* $v \ge 1$ *such that*

$$\log\sup_Q N(\epsilon\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}) \le v\log(a/\epsilon),\ 0 < \epsilon \le 1.$$

Then with probability $> 1 - C(\log n)^{-1}$, *one has*

$$\Big\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(X_i) - \mathrm{E}f)\Big\|_{\mathcal{F}} \lesssim \sigma\sqrt{v\log\Big(\frac{aM}{\sigma}\Big)} + \frac{vB}{\sqrt{n}}\log\Big(\frac{aM}{\sigma}\Big).$$

Proof. It follows immediately from Lemma 7 and the Proof of Lemma 2.2 of Chernozhukov, Chetverikov and Kato (2014). ■

Appendix H. Technical Lemmas

For completeness, we collect some of the technical results used in our proofs in this Section. They are either direct restated from other papers or their straightforward modifications.

H.1. A Maximal Inequality.

In this section we present a slight modification of Theorem 5.2 in Chernozhukov, Chetverikov and Kato (2014). The main difference is that we assume independence instead of i.i.d. of data. Let $\mathcal{F}$ be a pointwise measurable class of measurable functions $\mathcal{S} \mapsto \mathbb{R}$ with measurable envelope F. For all $0 < \delta < \infty$, define the integrated Koltchinskii-Pollard entropy of $\mathcal{F}$ as

$$J(\mathcal{F}, F, \delta) := \int_0^\delta \sup_Q \sqrt{\log 2N(\mathcal{F}, L_2(Q), \varepsilon\|F\|_{L_2(Q)})}d\varepsilon$$

where the supremum is taken over all discrete probabilities with a finite number of atoms and rational weights.

Lemma 7.
Given $X_1, ..., X_n$ independent $\mathcal{S}$-valued random variables. Suppose $0 < \mathrm{E_P}\frac{1}{G}\sum_{g=1}^G F^2 < \infty$, and let $\sigma^2 > 0$ be any positive constant such that $\sup_{f\in\mathcal{F}} \mathrm{E_P}\frac{1}{G}\sum_{g=1}^G f^2 \leq \sigma^2 \leq \mathrm{E_P}\frac{1}{G}\sum_{g=1}^G F^2$. Let $\delta = \sigma/(\mathrm{E_P}\frac{1}{G}\sum_{g=1}^G F^2)^{1/2}$. Define $B = \sqrt{\mathrm{E}[\max_{1\leq i\leq n} F^2(X_i)]}$. Then

$$\mathrm{E}\Big[\Big\|\frac{1}{\sqrt{n}}\sum_{i=1}^n (f(X_i) - \mathrm{E}f)\Big\|_{\mathcal{F}}\Big] \leq C\Big\{J(\delta, \mathcal{F}, F)(\mathrm{E_P}\frac{1}{G}\sum_{g=1}^G F^2)^{1/2} + \frac{BJ^2(\delta, \mathcal{F}, F)}{\delta^2\sqrt{n}}\Big\}$$

where $C > 0$ is a universal constant.

Proof. The proof follows almost exactly the same steps as in Chernozhukov, Chetverikov and Kato (2014). We provide the proof for completeness.

In this proof, denote C as a universal constant that the value may change from place to place. We assume F is positive everywhere without loss of generality and abbreviate $J(\mathcal{F}, F, \delta)$ as $J(\delta)$. Let $\sigma_n^2 = \sup_{f\in\mathcal{F}} \mathbb{E}_n f^2$. Given any i.i.d. Rademacher random variables $\varepsilon_1, ..., \varepsilon_n$ independent of $X_1, ..., X_n$, the symmetrization inequality (Theorem 3.1.21 in Giné and Nickl (2016)) implies

$$\mathrm{E}\Big[\Big\|\frac{1}{\sqrt{n}}\sum_{i=1}^n (f(X_i) - \mathrm{E}f)\Big\|_{\mathcal{F}}\Big] \leq \mathrm{E}\Big[\Big\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(X_i)\Big\|_{\mathcal{F}}\Big].$$

Using Remark 3.5.2 in Giné and Nickl (2016),

$$\begin{aligned}\mathrm{E}_\varepsilon\Big[\Big\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(X_i)\Big\|_{\mathcal{F}}\Big] &\le C\int_0^{\sigma_n}\sqrt{1+\log N(\mathcal{F},\|\cdot\|_{P_n,2},\varepsilon)}d\varepsilon \\ &\le C\|F\|_{P_n,2}\int_0^{\sigma_n/\|F\|_{P_n,2}}\sqrt{1+\log N(\mathcal{F},\|\cdot\|_{P_n,2},\varepsilon\|F\|_{P_n,2})}d\varepsilon \\ &\le C\|F\|_{P_n,2}J(\sigma_n/\|F\|_{P_n,2}).\end{aligned}$$

Hence by Lemma 3.5.3 part (c) of Giné and Nickl (2016) and applying Jensen's inequality,

$$Z := \mathrm{E}\Big[\Big\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(X_i)\Big\|_{\mathcal{F}}\Big] \le C(\mathrm{E}_{\mathrm{P}}\frac{1}{G}\sum_{g=1}^G F^2)^{1/2}J(\mathcal{F},F,\{\mathrm{E}[\sigma_n^2]/\mathrm{E}_{\mathrm{P}}\frac{1}{G}\sum_{g=1}^G F^2\}^{1/2}).$$

Now we bound $\mathrm{E}[\sigma_n^2]$ by the contraction principle (Corollary 3.2.2 of Giné and Nickl (2016)) and the Cauchy-Schwarz inequality,

$$\begin{aligned}\mathrm{E}[\sigma_n^2] &\le \sigma^2 + 8\mathrm{E}\Big[\max_{1\le i\le n}F(X_i)\Big\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(X_i)\Big\|_{\mathcal{F}}\Big] \\ &\le \sigma^2 + 8\sqrt{\mathrm{E}\Big[\max_{1\le i\le n}F^2(X_i)\Big]}\sqrt{\mathrm{E}\Big[\Big\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(X_i)\Big\|_{\mathcal{F}}^2\Big]}.\end{aligned}$$

Further by Hoffmann-Jørgensen inequality (Theorem A.1 in Chernozhukov, Chetverikov and Kato (2014)),

$$\sqrt{\mathrm{E}\Big[\Big\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(X_i)\Big\|_{\mathcal{F}}^2\Big]} \le C\Big\{\frac{1}{\sqrt{n}}Z + \frac{1}{n}B\Big\}.$$

Hence we obtain

$$\sqrt{\mathrm{E}[\sigma_n^2]} \le C(\mathrm{E}_{\mathrm{P}}\frac{1}{G}\sum_{g=1}^G F^2)^{1/2}J(\Delta\vee\sqrt{DZ}),$$

where $\Delta^2 := \max\{\sigma^2, B^2/n\}/\mathrm{E}_{\mathrm{P}}\frac{1}{G}\sum_{g=1}^G F^2 \ge \delta^2$ and $D := B/(\sqrt{n}\mathrm{E}_{\mathrm{P}}\frac{1}{G}\sum_{g=1}^G F^2)$. Therefore, applying Lemma A.2 (ii) of Chernozhukov, Chetverikov and Kato (2014), we have

$$Z \le C(\mathrm{E}_{\mathrm{P}}\frac{1}{G}\sum_{g=1}^G F^2)^{1/2}J(\Delta\vee\sqrt{DZ}).$$

The rest follows exactly the same analysis of two cases as in Chernozhukov, Chetverikov and Kato (2014)) with only difference being $(\mathrm{E}_{\mathrm{P}}\frac{1}{G}\sum_{g=1}^G F^2)^{1/2}$ in place of their $\|F\|_{P,2}$. ■

H.2. **Additional Technical Lemmas.**

The following is a restate of Lemma K.1 in Belloni, Chernozhukov, Fernández-Val and Hansen (2017).

Lemma 8.

Let $\mathcal{F}$ denote a class of measurable functions $f : \mathcal{W} \to \mathbb{R}$ with a measurable envelope F.

(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 $\mathbb{R}$,

$$\begin{aligned}
&\log N(\epsilon \|F + F'\|_{Q,2}, \mathcal{F} + \mathcal{F}', \|\cdot\|_{Q,2}) \\
&\quad \le \log N\left(\tfrac{\epsilon}{2}\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}\right) + \log N\left(\tfrac{\epsilon}{2}\|F'\|_{Q,2}, \mathcal{F}', \|\cdot\|_{Q,2}\right), \\
&\log N(\epsilon \|F \cdot F'\|_{Q,2}, \mathcal{F} \cdot \mathcal{F}', \|\cdot\|_{Q,2}) \\
&\quad \le \log N\left(\tfrac{\epsilon}{2}\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}\right) + \log N\left(\tfrac{\epsilon}{2}\|F'\|_{Q,2}, \mathcal{F}', \|\cdot\|_{Q,2}\right), \\
&N(\epsilon \|F \vee F'\|_{Q,2}, \mathcal{F} \cup \mathcal{F}', \|\cdot\|_{Q,2}) \\
&\quad \le N\left(\epsilon\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}\right) + N\left(\epsilon\|F'\|_{Q,2}, \mathcal{F}', \|\cdot\|_{Q,2}\right).
\end{aligned}$$

(3) For any measurable class of functions $\mathcal{F}$ and a fixed function f mapping $\mathcal{W}$ to $\mathbb{R}$,

$$\log \sup_Q N(\epsilon \| |f| \cdot F\|_{Q,2}, f \cdot \mathcal{F}, \|\cdot\|_{Q,2}) \le \log \sup_Q N\left(\epsilon/2 \|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}\right)$$

(4) Given measurable classes $\mathcal{F}_j$ and envelopes F_j, $j = 1, \dots, k$, mapping $\mathcal{W}$ to $\mathbb{R}$, a mapping $\phi\colon \mathbb{R}^k \to \mathbb{R}$ such that for $f_j, g_j \in \mathcal{F}_j$, the following Lipschitz condition holds: $|\phi(f_1, \dots, f_k) - \phi(g_1, \dots, g_k)| \le \sum_{j=1}^k L_j(x)|f_j(x) - g_j(x)|$ for $L_j(x) \ge 0$, and some 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

$$\begin{aligned}
&\log \sup_Q N\left(\epsilon \Big\| \sum_{j=1}^k L_j F_j \Big\|_{Q,2}, \mathcal{L}, \|\cdot\|_{Q,2}\right) \\
&\quad \le \sum_{j=1}^k \log \sup_Q N\left(\tfrac{\epsilon}{k}\|F_j\|_{Q,2}, \mathcal{F}_j, \|\cdot\|_{Q,2}\right).
\end{aligned}$$

The following generalizes Lemma 9 of Belloni, Chernozhukov and Wei (2016) to allow for cluster sampling. The proof follows closely to the orginal. Denote $M = \frac{1}{G}\sum_{g=1}^G \sum_{i=1}^{n_g} f_{ig}^2 X_{ig} X_{ig}'$.

Lemma 9 (Minoration Lemma)**.**
Suppose that for each G*,* $L(\beta) = -\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\{Y_{ig}X_{ig}'\beta - \log(1+\exp(X_{ig}'\beta))\}$*. For any* $\delta \in A \subset \mathbb{R}^p$*,*

$$L(\beta^0+\delta) - L(\beta^0) - \nabla L(\beta^0)'\delta \geq \frac{1}{3G}\delta' M\delta \wedge \frac{1}{3G}\bar{q}_A\sqrt{\delta' M\delta}$$

Proof. The proof is divided into two steps.

Step 1. (Minoration) Write $F(\delta) = L(\beta^0+\delta) - L(\beta^0) - \nabla L(\beta^0)'\delta$. Define

$$r_A =: \sup\Big\{r \in \mathbb{R} : F(\delta) \geq \frac{1}{3G}\delta' M\delta \text{ for all } \delta \in A, \sqrt{\delta' M\delta} \leq r\Big\}$$

So for any $\delta \in A$, if $\sqrt{\delta' M\delta} \leq r_A$, then by construction of r_A,

$$F(\delta) \geq \frac{1}{3G}\delta' M\delta.$$

Otherwise if $\sqrt{\delta' M\delta} > r_A$, by convexity of $t \mapsto F(t\delta)$ and the fact that $\frac{r_A}{\sqrt{\delta' M\delta}} < 1$,

$$F(\delta) \geq \frac{\sqrt{\delta' M\delta}}{r_A}F\Big(\frac{r_A}{\sqrt{\delta' M\delta}}\delta\Big)$$

Now, let $\bar{\delta} = \frac{r_A}{\sqrt{\delta' M\delta}}\delta$, then $\sqrt{\bar{\delta}' M\bar{\delta}} \leq r_A$ and thus

$$F(\delta) \geq \frac{\sqrt{\delta' M\delta}}{r_A}F(\bar{\delta}) \geq \frac{\sqrt{\delta' M\delta}}{r_A}\frac{1}{3G}r_A^2 \geq \frac{1}{3G}\bar{q}_A\sqrt{\delta' M\delta}.$$

where the last inequality follows from $r_A \geq \bar{q}_A$ that is shown in the next step. Combining these two cases, we have

$$F(\delta) \geq \frac{1}{3G}\delta' M\delta \wedge \frac{1}{3G}\bar{q}_A\sqrt{\delta' M\delta}.$$

Step 2. We now prove $r_A \geq \bar{q}_A$. Define $f_{ig}(t) = \log\{1+\exp(X_{ig}'\beta^0)\}$, then

$$F(\delta) = \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}[f_{ig}(1) - f_{ig}(0) - 1 \cdot f_{ig}'(0)].$$

By Lemma 7 and 8 of Belloni, Chernozhukov and Wei (2016), we have

$$f_{ig}(1) - f_{ig}(0) - 1 \cdot f_{ig}'(0) \geq f_{ig}^2\Big\{\frac{|X_{ig}'\delta|^2}{2} - \frac{|X_{ig}'\delta|^3}{6}\Big\}.$$

Summing over i, we have

$$F(\delta) \geq \frac{1}{2}\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}f_{ig}^2|X_{ig}'\delta|^2 - \frac{1}{6}\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g}f_{ig}^2|X_{ig}'\delta|^3.$$

Now, for any $\delta \in A$ such that $\sqrt{\delta' M \delta} \leq \bar{q}_A$, the definition of $\bar{q}_A$ gives

$$\sqrt{\delta' M \delta} \leq \bar{q}_A \leq \frac{(\delta' M \delta)^{3/2}}{\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} f_{ig}^2 |X_{ig}' \delta|^3}$$

This implies $\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} f_{ig}^2 |X_{ig}'\delta|^3 \leq \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} f_{ig}^2 |X_{ig}'\delta|^2$ and thus

$$F(\delta) \geq \frac{1}{2}\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} f_{ig}^2 |X_{ig}'\delta|^2 - \frac{1}{6}\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} f_{ig}^2 |X_{ig}'\delta|^3 \geq \frac{1}{3}\frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n_g} f_{ig}^2 |X_{ig}'\delta|^2 = \frac{1}{3G}\delta' M \delta.$$

The definition of r_A then suggests $r_A \geq \bar{q}_A$. ■

(Harold D. Chiang) Department of Economics, Vanderbilt University, United States