# Supplementary Materials of "Transfer Learning with High-dimensional Generalized Linear Models"

## Contents

# S.1 More details

## S.1.1 A schematic to illustrate $\mathcal{A}$-Trans-GLM

To better illustrate Algorithm 1, we draw a schematic in Figure 5. The blue point represents the target coefficient $\boldsymbol{\beta} = \boldsymbol{w}^{(0)}$ and the surrounding blue circle represents the estimation error $\mathcal{O}_p\left(\sqrt{\frac{s\log p}{n_0}}\right)$. The purple point denotes the estimator $\hat{\boldsymbol{\beta}}_{\text{naïve-Lasso}}$ from the classical Lasso with only target data. The orange point represents $\boldsymbol{w}^{\mathcal{A}_h}$, which is the population version of the rough estimator we obtain from the transferring step by pooling target and source data in $\mathcal{A}$ (see Section 2.3), and the surrounding orange circle denotes its estimation error. It can be seen that $\boldsymbol{w}^{\mathcal{A}_h}$ is a pooled version of $\{\boldsymbol{w}^{(k)}\}_{k\in\{0\}\cup\mathcal{A}_h}$, which is close to $\boldsymbol{\beta}$ when $h$ is small. Starting from an initial estimate of $\boldsymbol{\beta}$, the transferring step of $\mathcal{A}$-Trans-GLM algorithm updates the estimate to $\hat{\boldsymbol{w}}^{\mathcal{A}_h}$ (an estimate of $\boldsymbol{w}^{\mathcal{A}_h}$ based on source data in $\mathcal{A}$ and the target data), then the debiasing step yields the final estimator $\hat{\boldsymbol{\beta}}_{\mathcal{A}\text{-Trans-GLM}}$.

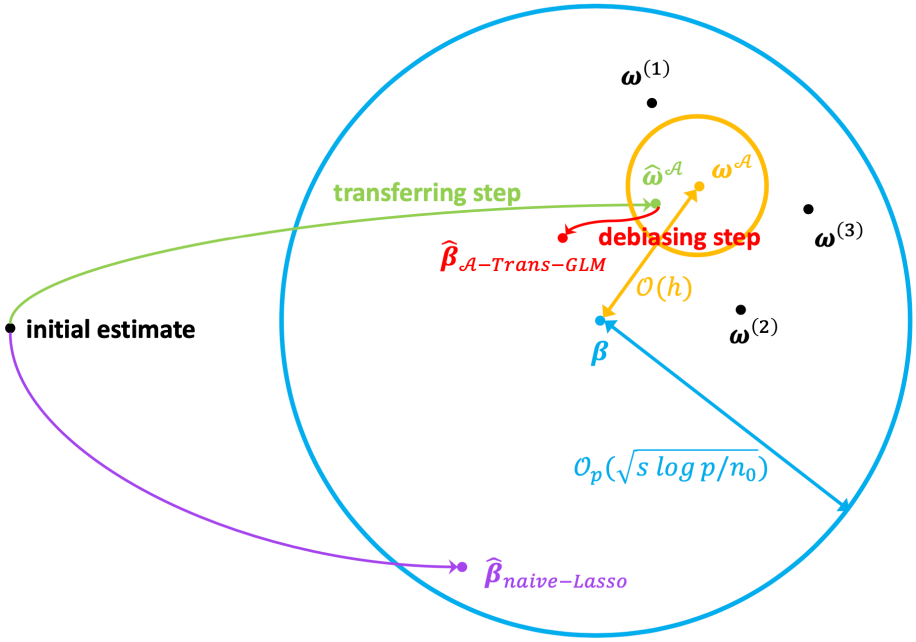


Figure 5: A schematic of $\mathcal{A}$-Trans-GLM (Algorithm 1). $\mathcal{A} = \{1, 2, 3\}$.

## S.1.2 Theory

### S.1.2.1 Explicit forms of convergence rates in Assumption 5

**Proposition 1** (Explicit forms of $\Upsilon_1^{(k)}$, $\Gamma_1^{(k)}$, $\Gamma_2^{(k)}$, $g_1^{(k)}$ and $g_2^{(k)}$ for certain families)**.** *Denote*

$$\Omega_k = \begin{cases} \sqrt{\frac{s\log p}{n_0}}, & k = 0, \\ \sqrt{\frac{s\log p}{n_k+n_0}} + \left(\frac{\log p}{n_k+n_0}\right)^{1/4}\sqrt{h} + \sqrt{s}h, & k \in \mathcal{A}_h \\ h'\sqrt{\frac{\log p}{n_k+n_0}} + \sqrt{\frac{s'\log p}{n_k+n_0}}\cdot W_k + \left(\frac{\log p}{n_k+n_0}\right)^{1/4}\sqrt{h'W_k}, & k \in \mathcal{A}_h^c, \end{cases}$$

*where* $W_k = 1 \vee \|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}\|_2 \vee \|\boldsymbol{\beta}^{(k)} - \boldsymbol{w}^{(k)}\|_2$. *Assume Assumptions 1 and 2 hold. For Poisson model, it is further required that* $h \le U^{-1}\bar{U}$ *and* $U \sup_{k\in\mathcal{A}^c}\{\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}\|_1 \vee \|\boldsymbol{\beta}^{(k)} - \boldsymbol{w}^{(k)}\|_1\} \le \bar{U}$. *With* $\lambda^{(k)[r]} = C\left(\sqrt{\frac{\log p}{n_k+n_0}} + h\right)$ *when* $k \in \mathcal{A}_h$, $\lambda^{(k)[r]} = C\sqrt{\frac{\log p}{n_k+n_0}}\cdot W_k$ *when* $k \in \mathcal{A}_h^c$ *and*

$\lambda^{(0)} = C\sqrt{\frac{\log p}{n_0}}$ *for some sufficiently large constant* $C > 0$*, we have the following explicit forms of* $\Gamma_1^{(k)}$, $\Gamma_2^{(k)}$, $\Upsilon_1^{(k)}$, $g_1^{(k)}$ *and* $g_2^{(k)}$ *for logistic, linear and Poisson regression models.*

*(i) For the logistic regression model:*

$$\Gamma_1^{(0)} = \sqrt{\frac{s\log p}{n_0}}, \quad \Gamma_2^{(0)} = \|\boldsymbol{\beta}\|_2/\sqrt{n_0},$$
$$\Upsilon_1^{(k)} = \Omega_k, \quad \Gamma_1^{(k)} = \sqrt{\frac{1}{n_0}}\Omega_k, \quad \Gamma_2^{(k)} = \sqrt{\frac{1}{n_0}} \cdot \left[\|\boldsymbol{w}^{(k)}\|_2 \mathbb{1}(k \in \mathcal{A}_h) + \|\boldsymbol{\beta}^{(k)}\|_2 \mathbb{1}(k \in \mathcal{A}_h^c)\right],$$

$$g_1^{(k)}(\zeta) = g_2^{(k)}(\zeta) = \exp(-\zeta^2).$$

*(ii) For the linear model:*

$$\Gamma_1^{(0)} = \sqrt{\frac{s\log p}{n_0}} \cdot \|\boldsymbol{\beta}\|_2, \quad \Gamma_2^{(0)} = (\|\boldsymbol{\beta}\|_2^2 \vee \|\boldsymbol{\beta}\|_2)/\sqrt{n_0},$$
$$\Upsilon_1^{(k)} = \Omega_k \cdot \left[\|\boldsymbol{w}^{(k)}\|_2 \mathbb{1}(k \in \mathcal{A}_h) + \|\boldsymbol{\beta}^{(k)}\|_2 \mathbb{1}(k \in \mathcal{A}_h^c)\right],$$
$$\Gamma_1^{(k)} = \sqrt{\frac{1}{n_0}}\Omega_k \cdot \left[\|\boldsymbol{w}^{(k)}\|_2 \mathbb{1}(k \in \mathcal{A}_h) + \|\boldsymbol{\beta}^{(k)}\|_2 \mathbb{1}(k \in \mathcal{A}_h^c)\right],$$
$$\Gamma_2^{(k)} = \sqrt{\frac{1}{n_0}} \left[(\|\boldsymbol{w}^{(k)}\|_2^2 \vee \|\boldsymbol{w}^{(k)}\|_2)\mathbb{1}(k \in \mathcal{A}_h) + (\|\boldsymbol{\beta}^{(k)}\|_2^2 \vee \|\boldsymbol{\beta}^{(k)}\|_2)\mathbb{1}(k \in \mathcal{A}_h^c)\right],$$
$$g_1^{(k)}(\zeta) = g_2^{(k)}(\zeta) = \exp(-\zeta^2), k \neq 0; \quad g_1^{(0)}(\zeta) = \exp(-\zeta^2), g_2^{(0)}(\zeta) = \exp\{-n_0\} + \exp(-\zeta^2).$$

*(iii) For the Poisson regression model with bounded predictors (*$\sup_k \|\boldsymbol{x}^{(k)}\|_\infty \leq U < \infty$*):*

$$\Gamma_1^{(0)} = \sqrt{\frac{s\log p}{n_0}} \cdot \exp\left(U\|\boldsymbol{\beta}\|_1\right), \quad \Gamma_2^{(0)} = \exp\left(U\|\boldsymbol{\beta}\|_1\right) \cdot \frac{1 + \|\boldsymbol{\beta}\|_2 + U\|\boldsymbol{\beta}\|_1}{\sqrt{n_0}},$$
$$\Upsilon_1^{(k)} = \Omega_k \cdot \exp\left\{U\|\boldsymbol{w}^{(k)}\|_1 \cdot \mathbb{1}(k \in \mathcal{A}_h) + U\|\boldsymbol{\beta}^{(k)}\|_1 \cdot \mathbb{1}(k \in \mathcal{A}_h^c)\right\},$$
$$\Gamma_1^{(k)} = \sqrt{\frac{1}{n_0}}\Omega_k \cdot \exp\left\{U\|\boldsymbol{w}^{(k)}\|_1 \cdot \mathbb{1}(k \in \mathcal{A}_h) + U\|\boldsymbol{\beta}^{(k)}\|_1 \cdot \mathbb{1}(k \in \mathcal{A}_h^c)\right\},$$

$$\Gamma_2^{(k)} = \sqrt{\frac{1}{n_0}} \left[ \exp\left(U\|\boldsymbol{w}^{(k)}\|_1\right)\left(1+\|\boldsymbol{w}^{(k)}\|_2 + U\|\boldsymbol{w}^{(k)}\|_1\right)\cdot \mathbb{1}(k\in\mathcal{A}_h) \right.$$
$$\left. + \exp\left(U\|\boldsymbol{\beta}^{(k)}\|_1\right)\left(1+\|\boldsymbol{\beta}^{(k)}\|_2 + U\|\boldsymbol{\beta}^{(k)}\|_1\right)\cdot \mathbb{1}(k\in\mathcal{A}_h^c)\right],$$
$$g_1^{(k)}(\zeta) = g_2^{(k)}(\zeta) = \exp(-\zeta^2), k\neq 0; \quad g_1^{(0)}(\zeta) = \exp(-\zeta^2), g_2^{(0)}(\zeta) = \zeta^{-2}.$$

#### S.1.2.2 Bound of a prediction error measure

Denote $L_{n_0}^{(0)}(\boldsymbol{w}) = -\frac{1}{n_0}\sum_{i=1}^{n_0}\log\rho(\boldsymbol{x}_i^{(0)}) - \frac{1}{n_0}(\boldsymbol{y}^{(0)})^T\boldsymbol{X}^{(0)}\boldsymbol{w} + \frac{1}{n_0}\sum_{i=1}^{n_0}\psi(\boldsymbol{w}^T\boldsymbol{x}_i^{(0)})$. Suggested by Loh and Wainwright (2015), we consider a special measure of the prediction error, which is defined by $\langle\nabla L_{n_0}^{(0)}(\hat{\boldsymbol{\beta}}) - \nabla L_{n_0}^{(0)}(\boldsymbol{\beta}), \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rangle$, where $\nabla L_{n_0}^{(0)}(\boldsymbol{w}) = -\frac{1}{n_0}(\boldsymbol{X}^{(0)})^T\boldsymbol{y}^{(0)} + \frac{1}{n_0}\sum_{i=1}^{n_0}\boldsymbol{x}_i^{(0)}\psi(\boldsymbol{w}^T\boldsymbol{x}_i^{(0)}) \in \mathbb{R}^p$. Note that the loss function $L_{n_0}^{(0)}$ is convex, therefore this quantity is non-negative. As argued in their paper, $\langle\nabla L_{n_0}^{(0)}(\hat{\boldsymbol{\beta}}) - \nabla L_{n_0}^{(0)}(\boldsymbol{\beta}), \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rangle$ can be easily interpreted in GLMs. For example, in the case of linear models where $\psi(u) = u^2/2$, this measure equals to the in-sample square loss $\|\boldsymbol{X}^{(0)}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|_2^2$. For general GLMs, it is equivalent to the symmetrized Bregman divergence defined by $\psi$.

Next we would like to present the bounds of $\langle\nabla L_{n_0}^{(0)}(\hat{\boldsymbol{\beta}}) - \nabla L_{n_0}^{(0)}(\boldsymbol{\beta}), \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rangle$ with and without Assumption 4 for $\mathcal{A}_h$-Trans-GLM.

**Theorem 6** (Bound of a prediction error measure for $\mathcal{A}_h$-Trans-GLM with Assumption 4)**.** *By imposing the same conditions in Theorem 1, we have*

$$\sup_{\boldsymbol{\xi}\in\Xi(s,h)}\langle\nabla L_{n_0}^{(0)}(\hat{\boldsymbol{\beta}}) - \nabla L_{n_0}^{(0)}(\boldsymbol{\beta}), \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rangle \lesssim s\left(\frac{\log p}{n_0}\right)^{1/2}\left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/2} + h\left(\frac{\log p}{n_0}\right)^{1/2}$$
$$+ \left(\frac{s\log p}{n_0}\right)^{1/2}\left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/4} h^{1/2},$$

*with probability at least* $1-n_0^{-1}$.

**Remark 11.** *When* $h \ll s\sqrt{\frac{\log p}{n_0}}$ *and* $n_{\mathcal{A}_h} \gg n_0$*, the upper bounds in (i) are better than the classical Lasso bound* $\mathcal{O}_p\left(\frac{s\log p}{n_0}\right)$ *with target data (Loh and Wainwright, 2015).*

**Theorem 7** (Bound of a prediction error measure for $\mathcal{A}_h$-Trans-GLM without Assumption 4)**.** *By imposing the same conditions in Theorem 3, we have*

$$\sup_{\boldsymbol{\xi}\in\Xi(s,h)} \langle \nabla L_{n_0}^{(0)}(\hat{\boldsymbol{\beta}}) - \nabla L_{n_0}^{(0)}(\boldsymbol{\beta}), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle \lesssim s\left(\frac{\log p}{n_0}\right)^{1/2}\left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/2} + sh\left(\frac{\log p}{n_0}\right)^{1/2} + \left(\frac{s\log p}{n_0}\right)^{1/2}\left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/4} h^{1/2}$$

*with probability at least* $1 - n_0^{-1}$.

**Remark 12.** *When* $h \ll \sqrt{\frac{\log p}{n_0}}$ *and* $n_{\mathcal{A}_h} \gg n_0$*, the upper bounds in (i) are better than the classical Lasso bound* $\mathcal{O}_p\left(\frac{s\log p}{n_0}\right)$ *with target data (Loh and Wainwright, 2015).*

## S.1.3 Additional numerical results

### S.1.3.1 More details about the implementation of numerical experiments

All experiments in this paper are conducted in R. The GLM Lasso is implemented via R package `glmnet` (Friedman et al., 2010). We summarize R codes for GLM transfer learning algorithms in a new R package `glmtrans`, which is available on CRAN. We use 10-fold cross-validation to choose the penalty parameter for naïve-Lasso and our GLM transfer learning algorithms. The largest $\lambda$ which achieves one standard error within the minimum cross-validation error will be chosen for the transferring step, which is sometimes called *"lambda.1se"* (Friedman et al., 2010). To effectively debias the transferring step, we choose the lambda achieving minimal cross-validation error, which is often called *"lambda.min"*. Since in transferable source detection, the first step is the same as the transferring step of $\{k\}$-Trans-GLM, therefore we keep the same setting as the transferring step in Algorithm 1, i.e. take *"lambda.1se"*. And in Algorithm 2, we set the constant $C_0 = 2$. In the two-step

transfer learning procedure of Algorithm 3, we use "lambda.min" in both transferring and debiasing steps.

In real-data studies, SVM with RBF kernel is implemented by package `e1071`, and decision trees are implemented through package `rpart`. We fit the random forest via package `randomForest`, and implement boosting trees through package `fastAdaboost`. The number of weak classifiers in boosting trees is set to be 50. Since the sample size of each state is small and some states have very imbalanced responses, we change the cross-validation folds from default 10 to 5 for all Lasso-based methods. All the other parameters are kept the same as the default settings.

#### S.1.3.2 Transfer learning on $\mathcal{A}_h$

In this section, we supplement more numerical results about the performance of Algorithm 1 under different $h$ and $(\{n_k\}_{k=0}^K, p, s)$ settings. In addition to the previous $(\{n_k\}_{k=0}^K, p, s)$ setting studied in Section 4.1.1, the following two settings of $(\{n_k\}_{k=0}^K, p, s)$ are considered:

(i) $n_k = 150$ for all $k = 0, \ldots, K$, $p = 1000$, $s = 15$;

(ii) $n_k = 200$ for all $k = 0, \ldots, K$, $p = 2000$, $s = 20$.

Given each $(\{n_k\}_{k=0}^K, p, s)$ setting, consider the same setting we use in Section 4.1.1. All the experiments are replicated 200 times and the average $\ell_2$-estimation errors of $\mathcal{A}_h$-Trans-GLM and naïve-Lasso under linear, logistic and Poisson regression models are shown in Figure 6 and 7.

The trends in Figures 6 and 7 are similar to that in Figure 1. As $K$ increases, the estimation error of $K_{\mathcal{A}_h}$-Trans-GLM continues declining and is lower that of naïve-Lasso on target data only.

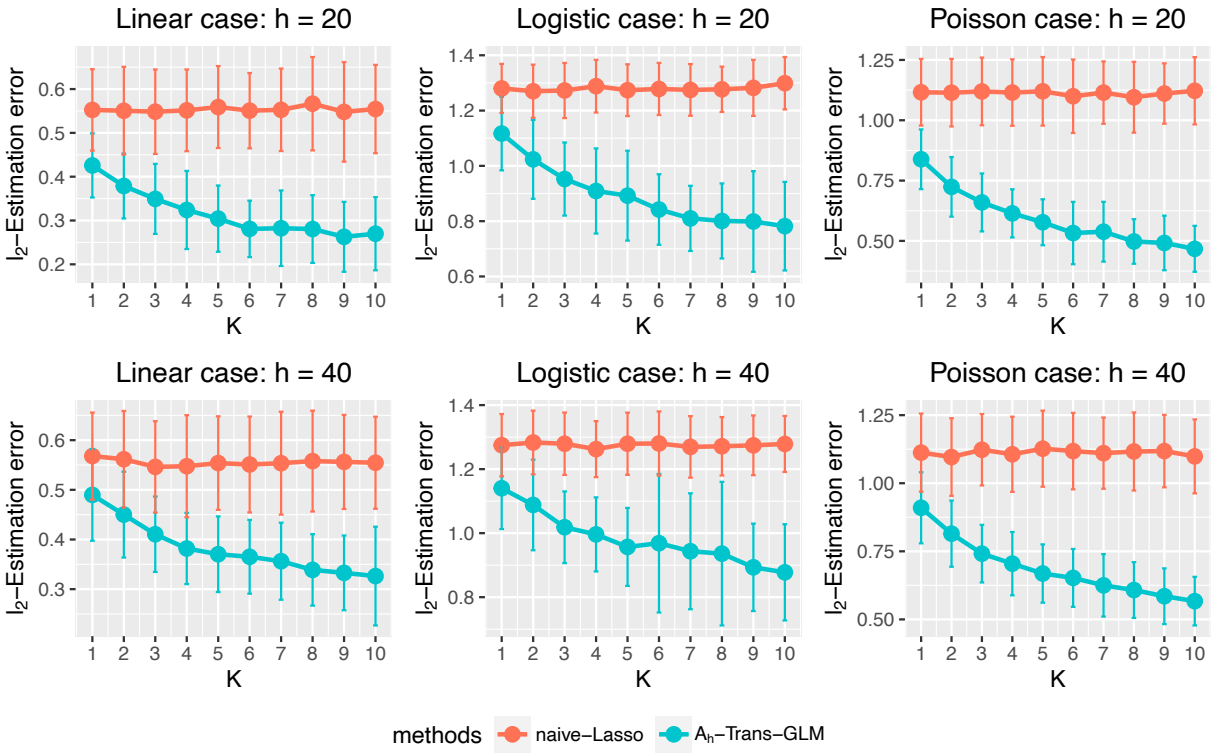


Figure 6: The average $\ell_2$-estimation of $\mathcal{A}_h$-Trans-GLM and naïve-Lasso under linear, logistic and Poisson regression models with different settings of $h$ and $K$. $n_k = 150$ for all $k = 0, \ldots, K$, $p = 1000$, $s = 15$. Error bars denote the standard deviations.

### S.1.3.3 Transferable source detection

In this section we supplement more experimental results in the case that some sources are not in the level-$h$ transferring set $\mathcal{A}_h$. The model settings are the same as those in Section 4.1.2. In addition to the setting used in Section 4.1.2, two more ones are considered:

(i) $n_k = 100$ for all $k = 0, \ldots, K$, $p = 500$, $s = 10$;

(ii) $n_k = 150$ for all $k = 0, \ldots, K$, $p = 1000$, $s = 15$.

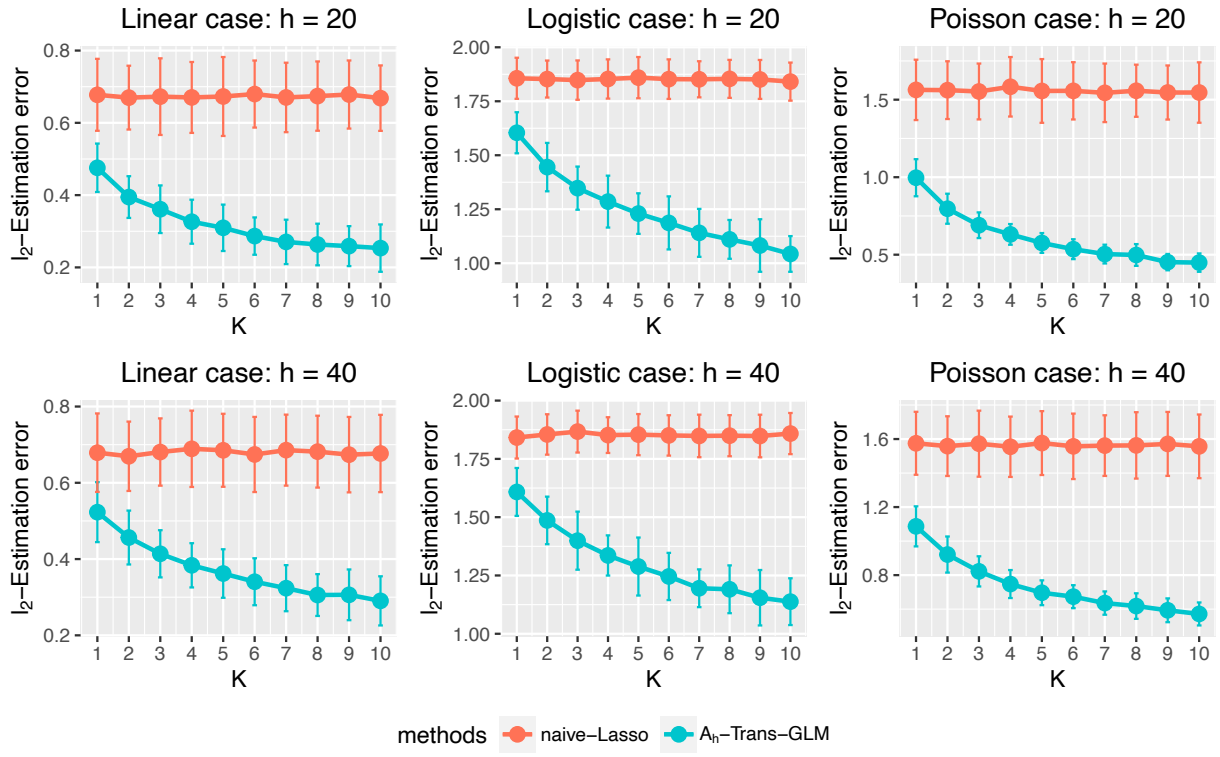


Figure 7: The average $\ell_2$-estimation of $\mathcal{A}_h$-Trans-GLM and naïve-Lasso under linear, logistic, and Poisson regression models with different settings of $h$ and $K$. $n_k = 200$ for all $k = 0, \ldots, K$, $p = 2000$, $s = 20$. Error bars denote the standard deviations.

We vary the values of $|K_{\mathcal{A}_h}|$ and $h$, and repeat each setting for 200 times. The average $\ell_2$-estimation errors are summarized in Figures 8 and 9.

Similar to Figure 3, it can be seen that $\mathcal{A}_h$-Trans-GLM always achieves the best performance. Trans-GLM mimics the behavior of $\mathcal{A}_h$-Trans-GLM very well in most cases, implying that the detection algorithm can accurately identify $\mathcal{A}$. We also observe that for linear models and logistic regression models, when $h = 40$, there is a gap between the estimation error of $\mathcal{A}_h$-Trans-GLM and Trans-GLM, meaning that when $h$ increases,

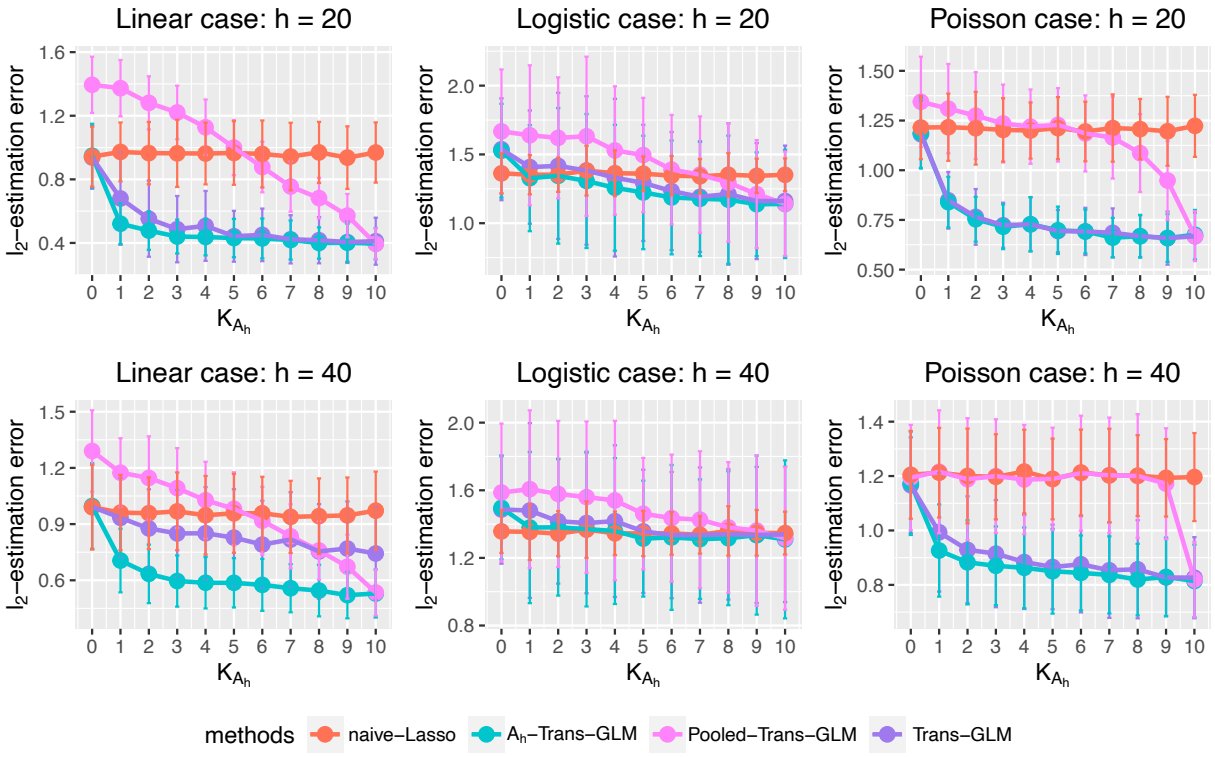


Figure 8: The average $\ell_2$-estimation error of various models with different settings of $h$ and $K_{\mathcal{A}_h}$ when $K = 10$. $n_k = 100$ for all $k = 0, \ldots, K$, $p = 500$, $s = 10$. Error bars denote the standard deviations.

Trans-GLM might begin missing sources in $\mathcal{A}_h$ or wrongly including sources in $\mathcal{A}_h$.

Furthermore, we pick the setting when $p = 2000$ and try different numbers of folds in the cross-validation procedure of Algorithm 2 (steps 2-5). The results are displayed in Figure 10. The findings suggest that more cross-validation folds may lead to better performance of Trans-GLM. When the cross-validation folds are large, the detection is likely to be more accurate. In the meantime, this may cause more computational burdens. Therefore, we may choose a moderate fold number like 3 or 5 in practice to achieve a good trade-off

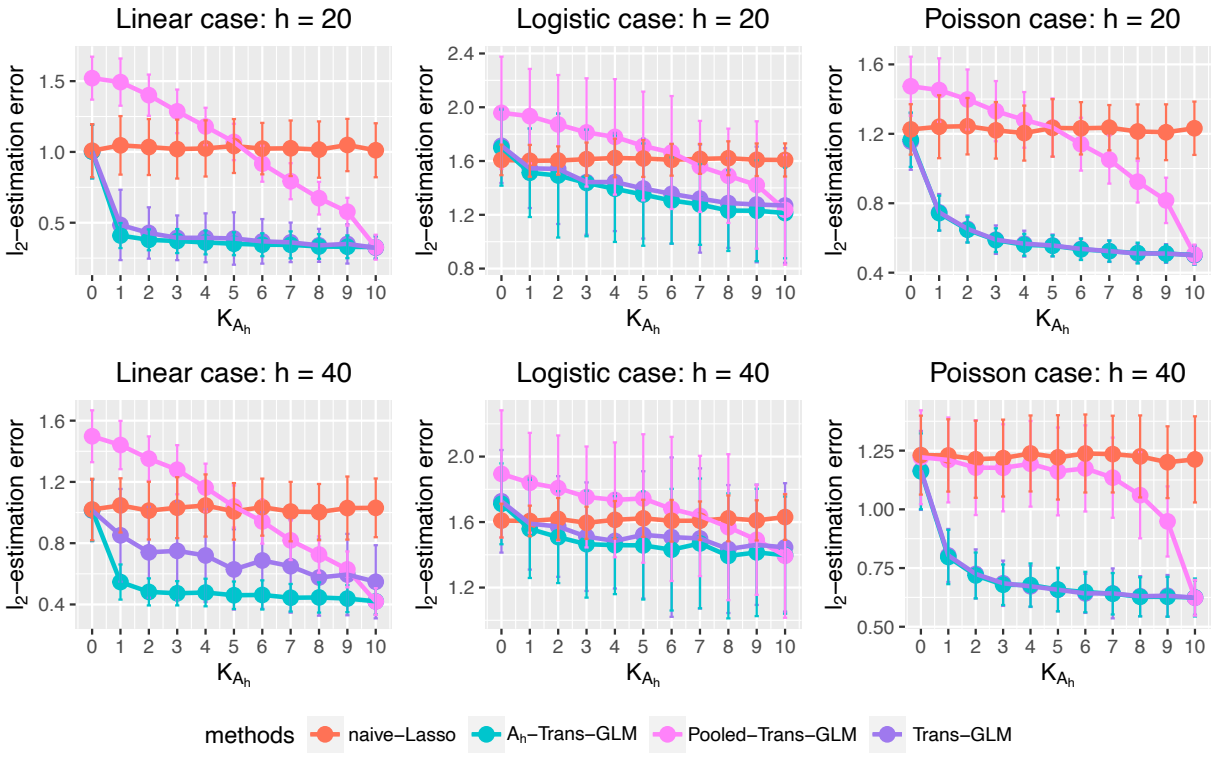


Figure 9: The average $\ell_2$-estimation error of various models with different settings of $h$ and $K_{\mathcal{A}_h}$ when $K = 10$. $n_k = 150$ for all $k = 0, \ldots, K$, $p = 1000$, $s = 15$. Error bars denote the standard deviations.

between the accuracy and computational costs.

#### S.1.3.4 Additional results of real data analysis

In this section, we aim to identify variables with significant effects for different targets in the real-data study (Section 4.2), by applying Algorithm 3. Taking the randomness caused by the cross-validation procedure in algorithms, we repeat the experiment 500 times. In

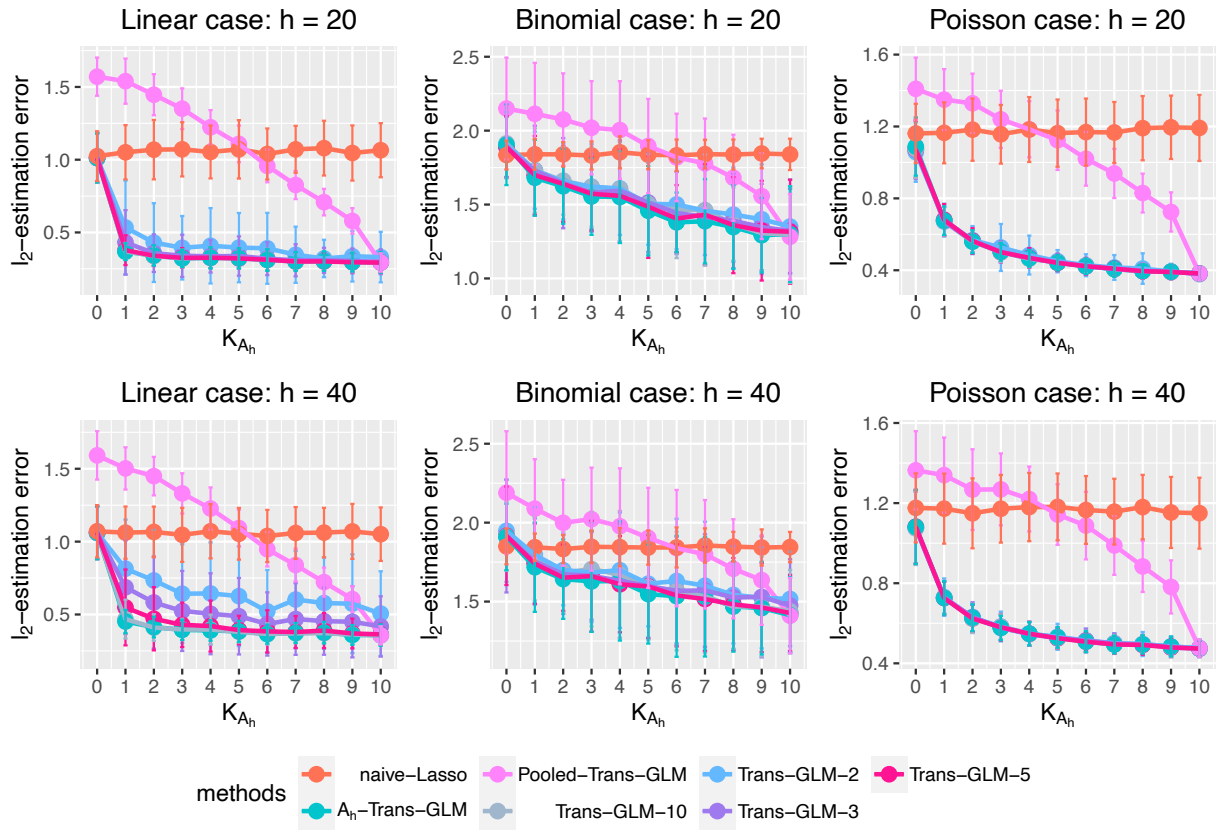


Figure 10: The average $\ell_2$-estimation error of various models with different settings of $h$ and $K_{\mathcal{A}_h}$ when $K = 10$. $n_k = 200$ for all $k = 0, \ldots, K$, $p = 2000$, $s = 20$. Error bars denote the standard deviations. The numbers after Trans-GLM indicate the number of folds used in cross-validation procedure of Algorithm 2 (steps 2-5)

each replication, for each target state, we first run Algorithm 2 to get a point estimate $\hat{\boldsymbol{\beta}}$ of the coefficient and the estimated informative source index set $\widehat{\mathcal{A}}$. Then, we run Algorithm 3 with $\hat{\boldsymbol{\beta}}$ and $\widehat{\mathcal{A}}$ to get the significant variables under 90% confidence level. Equivalently speaking, we identify coefficient components whose 90% confidence interval (CI) does not

cover zero and divided them into two parts based on the sign of CI center. Recall our recoding rule of the response: 0 denotes Republicans and 1 denotes Democrats. Therefore, given all other variables fixed, increasing the variable with a positive CI center gives rise to the chance of a county to vote Democrats. In contrast, increasing the variable with a negative CI center gives rise to the chance of a county to vote Republicans. The results are summarized in Figure 11, where we list all variables which are significant under 90% level in at least 5% of 500 replications. We use different colors and shapes to distinguish the effect of these variables (the sign of their CI centers). The $x$-axis and $y$-axis display the target state [1] and the variable names. The description of some main effects is presented in Table 2. It reveals that RHI825214 is positively significant in 6 of 8 target states, which means that when other variables are fixed, a larger White percentage in a county leads to a higher chance to vote Republicans. In opposition to this, EDU685213 is negatively significant in 4 of 8 target states, showing that when other variables are fixed, a larger Bachelor or higher degree holder percentage benefits Democrats under county-level. More interesting findings can be obtained from Figure 11, which are expected to provide some insights to better understand the election results.

## S.2 Proofs

Define $\hat{\boldsymbol{u}}^{\mathcal{A}_h} = \hat{\boldsymbol{w}}^{\mathcal{A}_h} - \boldsymbol{w}^{\mathcal{A}_h}$ and $\mathcal{D} = \{(\boldsymbol{X}^{(k)}, \boldsymbol{y}^{(k)})\}_{k\in\{0\}\cup\mathcal{A}_h}$. In the following, we will use bolded $\boldsymbol{\psi}'$ to represent the vector whose each component comes from the scalar function

[1] Target state AR is removed because none of the variables are significant in 5% of 500 replications

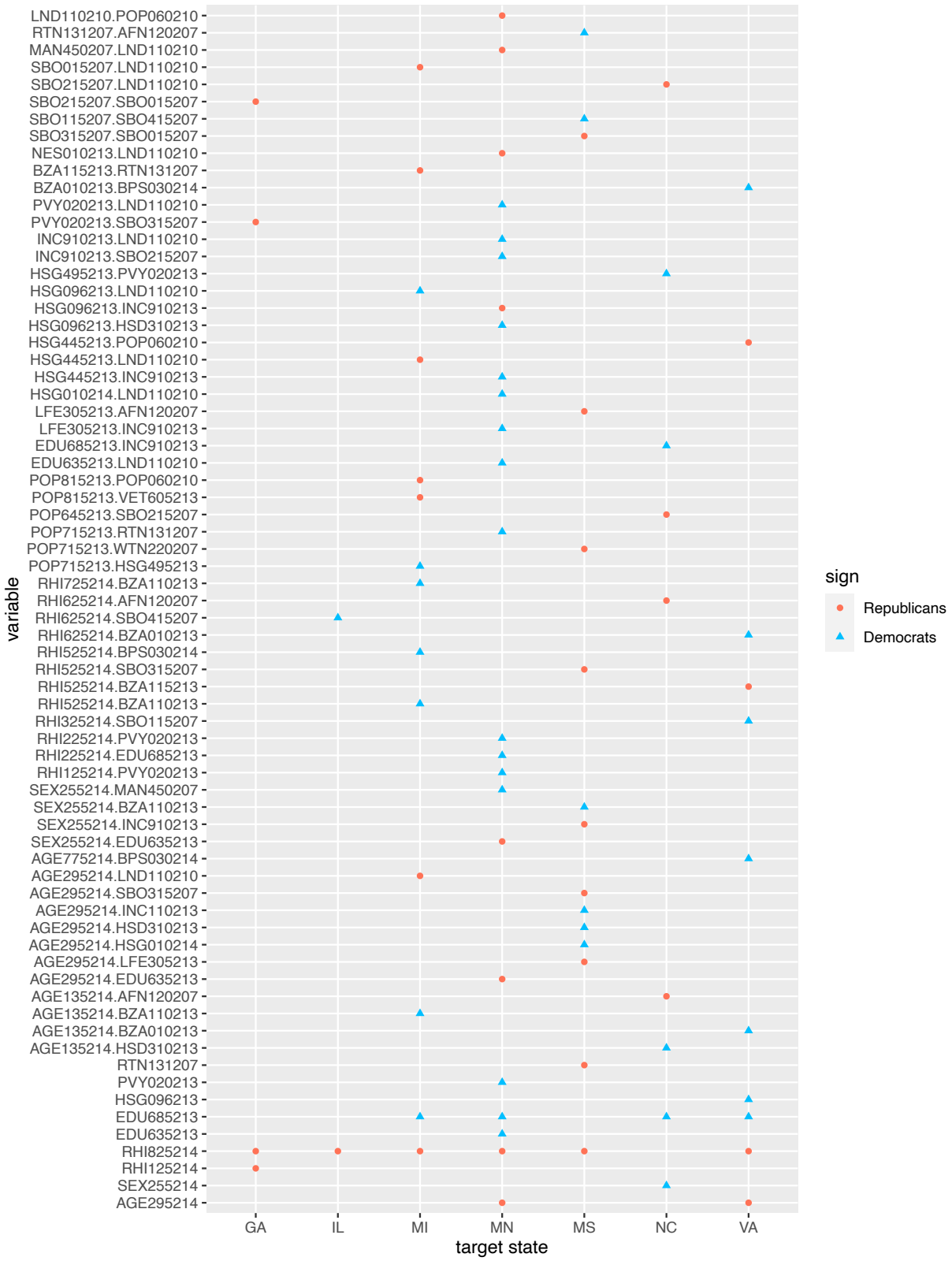


Figure 11: Variables that are significant under 90% level confidence for different targets in at least 5% of 500 replications. Provided by Algorithms 2 and 3.

| Variable name | Description |
|---|---|
| AGE135214 | Persons under 5 years, percent, 2014 |
| AGE295214 | Persons under 18 years, percent, 2014 |
| AGE775214 | Persons 65 years and over, percent, 2014 |
| SEX255214 | Female persons, percent, 2014 |
| RHI125214 | White alone, percent, 2014 |
| RHI225214 | Black or African American alone, percent, 2014 |
| RHI325214 | American Indian and Alaska Native alone, percent, 2014 |
| RHI525214 | Native Hawaiian and Other Pacific Islander alone, percent, 2014 |
| RHI625214 | Two or More Races, percent, 2014 |
| RHI725214 | Hispanic or Latino, percent, 2014 |
| RHI825214 | White alone, not Hispanic or Latino, percent, 2014 |
| POP715213 | Living in same house 1 year & over, percent, 2009-2013 |
| POP645213 | Foreign born persons, percent, 2009-2013 |
| POP815213 | Language other than English spoken at home, pct age 5+, 2009-2013 |
| EDU635213 | High school graduate or higher, percent of persons age 25+, 2009-2013 |
| EDU685213 | Bachelor's degree or higher, percent of persons age 25+, 2009-2013 |
| VET605213 | Veterans, 2009-2013 |
| LFE305213 | Mean travel time to work (minutes), workers age 16+, 2009-2013 |
| HSG010214 | Housing units, 2014 |
| HSG445213 | Homeownership rate, 2009-2013 |
| HSG096213 | Housing units in multi-unit structures, percent, 2009-2013 |
| HSG495213 | Median value of owner-occupied housing units, 2009-2013 |
| HSD310213 | Persons per household, 2009-2013 |
| INC910213 | Per capita money income in past 12 months (2013 dollars), 2009-2013 |
| INC110213 | Median household income, 2009-2013 |
| PVY020213 | Persons below poverty level, percent, 2009-2013 |
| BZA010213 | Private nonfarm establishments, 2013 |
| BZA110213 | Private nonfarm employment, 2013 |
| BZA115213 | Private nonfarm employment, percent change, 2012-2013 |
| NES010213 | Nonemployer establishments, 2013 |
| SBO315207 | Black-owned firms, percent, 2007 |
| SBO115207 | American Indian- and Alaska Native-owned firms, percent, 2007 |
| SBO215207 | Asian-owned firms, percent, 2007 |
| SBO415207 | Hispanic-owned firms, percent, 2007 |
| SBO015207 | Women-owned firms, percent, 2007 |
| MAN450207 | Manufacturers shipments, 2007 ($1,000) |
| WTN220207 | Merchant wholesaler sales, 2007 ($1,000) |
| RTN131207 | Retail sales per capita, 2007 |
| AFN120207 | Accommodation and food services sales, 2007 ($1,000) |
| BPS030214 | Building permits, 2014 |
| LND110210 | Land area in square miles, 2010 |
| POP060210 | Population per square mile, 2010 |

Table 2: Description of some variables in the original dataset.

$\psi'$ with corresponding predictors. Denote

$$\hat{L}(\boldsymbol{w}, \mathcal{D}) = -\frac{1}{n_{\mathcal{A}_h} + n_0} \sum_{k \in \{0\} \cup \mathcal{A}_h} (\boldsymbol{y}^{(k)})^T \boldsymbol{X}^{(k)} \boldsymbol{w} + \frac{1}{n_{\mathcal{A}_h} + n_0} \sum_{k \in \{0\} \cup \mathcal{A}_h} \sum_{i=1}^{n_k} \psi(\boldsymbol{w}^T \boldsymbol{x}_i^{(k)}),$$

$$\nabla \hat{L}(\boldsymbol{w}, \mathcal{D}) = -\frac{1}{n_{\mathcal{A}_h} + n_0} \sum_{k \in \{0\} \cup \mathcal{A}_h} (\boldsymbol{X}^{(k)})^T \boldsymbol{y}^{(k)} + \frac{1}{n_{\mathcal{A}_h} + n_0} \sum_{k \in \{0\} \cup \mathcal{A}_h} (\boldsymbol{X}^{(k)})^T \boldsymbol{\psi}'(\boldsymbol{w}^T \boldsymbol{x}^{(k)}),$$

$$\delta \hat{L}(\boldsymbol{u}, \mathcal{D}) = \hat{L}(\boldsymbol{w}^{\mathcal{A}_h} + \boldsymbol{u}, \mathcal{D}) - \hat{L}(\boldsymbol{w}^{\mathcal{A}_h}) - \nabla \hat{L}(\boldsymbol{w}^{\mathcal{A}_h})^T \boldsymbol{u}.$$

Denote $\partial \|\boldsymbol{w}\|_1$ as the subgradient of $\|\boldsymbol{w}\|_1$ w.r.t. $\boldsymbol{w} \in \mathbb{R}^p$, which falls between $-1$ and 1.

For any $\boldsymbol{w} \in \mathbb{R}^n$, denote $\boldsymbol{W}_{\boldsymbol{w}}^{(k)} = \text{diag}\left(\sqrt{\psi''((\boldsymbol{x}_1^{(k)})^T \boldsymbol{w})}, \ldots, \sqrt{\psi''((\boldsymbol{x}_{n_k}^{(k)})^T \boldsymbol{w})}\right)$ and $\boldsymbol{X}_{\boldsymbol{w}}^{(k)} = \boldsymbol{W}_{\boldsymbol{w}}^{(k)} \boldsymbol{X}^{(k)}$. $\boldsymbol{X}_{\boldsymbol{w},j}^{(k)}$ represents the $j$-th column of $\boldsymbol{X}_{\boldsymbol{w}}^{(k)}$ and $\boldsymbol{X}_{\boldsymbol{w},-j}^{(k)}$ represents the matrix $\boldsymbol{X}_{\boldsymbol{w}}^{(k)}$ without the $j$-th column. Denote $\boldsymbol{X}_{\boldsymbol{w},-j}^{(k)}$ as the submatrix without $j$-th column. $\boldsymbol{X}_{\boldsymbol{w},j}^{(k)}$ represents the $j$-th column of $\boldsymbol{X}_{\boldsymbol{w}}^{(k)}$ without the diagonal $(j,j)$ elements. Denote $\boldsymbol{x}_{\boldsymbol{w}}^{(k)} = \sqrt{\psi''((\boldsymbol{x}^{(k)})^T \boldsymbol{w})} \cdot \boldsymbol{x}^{(k)}$. $\boldsymbol{x}_{\boldsymbol{w},j}^{(k)}$ and $\boldsymbol{x}_{\boldsymbol{w},-j}^{(k)}$ represent the $j$-th component of $\boldsymbol{x}_{\boldsymbol{w}}^{(k)}$ and the vector without $j$-th component, respectively. Define

$$\boldsymbol{\gamma}_j^{(k)} = \underset{\boldsymbol{\gamma} \in \mathbb{R}^{p-1}}{\arg\min}\, \mathbb{E}\left[\boldsymbol{x}_{\boldsymbol{w}^{(k)},j}^{(k)} - (\boldsymbol{x}_{\boldsymbol{w}^{(k)},-j}^{(k)})^T \boldsymbol{\gamma}\right]^2 = \underset{\boldsymbol{\gamma} \in \mathbb{R}^{p-1}}{\arg\min}\, \mathbb{E}\left\{\psi''((\boldsymbol{w}^{(k)})^T \boldsymbol{x}^{(k)}) \cdot [\boldsymbol{x}_j^{(k)} - (\boldsymbol{x}_{-j}^{(k)})^T \boldsymbol{\gamma}]^2\right\}.$$

Also define $\boldsymbol{\eta}_{\boldsymbol{w}^{(k)},j}^{(k)} = \boldsymbol{X}_{\boldsymbol{w},j}^{(k)} - \boldsymbol{X}_{\boldsymbol{w},-j}^{(k)} \boldsymbol{\gamma}_j^{(k)}$ and $(\tau_j^{(k)})^2 = \mathbb{E}(\boldsymbol{\eta}_{\boldsymbol{w}^{(k)},j}^{(k)})^2$. And define

$$\boldsymbol{\gamma}_j^{\mathcal{A}} = \underset{\boldsymbol{\gamma} \in \mathbb{R}^{p-1}}{\arg\min} \left\{ \sum_{k \in \{0\} \cup \mathcal{A}_h} \mathbb{E}\left[\psi''((\boldsymbol{w}^{(k)})^T \boldsymbol{x}^{(k)}) \cdot [\boldsymbol{x}_j^{(k)} - (\boldsymbol{x}_{-j}^{(k)})^T \boldsymbol{\gamma}]^2\right] \right\}.$$

### S.2.1 Some lemmas

**Lemma 1.** *Under Assumptions 1 and 4,*

$$\|\boldsymbol{\delta}^{\mathcal{A}_h}\|_1 = \|\boldsymbol{w}^{\mathcal{A}_h} - \boldsymbol{\beta}\|_1 \leq C_1 h,$$

*where $\boldsymbol{w}^{\mathcal{A}_h}$ is defined by equation (2) and $C_1 := \sup_{k \in \{0\} \cup \mathcal{A}_h} \|\widetilde{\boldsymbol{\Sigma}}_h^{-1} \widetilde{\boldsymbol{\Sigma}}_h^{(k)}\|_1 < \infty$.*

**Lemma 2.** *Under Assumptions 1 and 2, there exists some positive constants $\kappa_1$, $\kappa_2$, $C_3$ and $C_4$ such that,*

$$\delta\hat{L}(\hat{\boldsymbol{u}}^{\mathcal{A}_h}, \mathcal{D}) \geq \kappa_1 \|\boldsymbol{u}\|_2^2 - \kappa_1\kappa_2 \sqrt{\frac{\log p}{n_{\mathcal{A}_h} + n_0}} \|\boldsymbol{u}\|_1 \|\boldsymbol{u}\|_2, \quad \forall \boldsymbol{u} : \|\boldsymbol{u}\|_2 \leq 1$$

*with probability at least $1 - C_3 \exp\{-C_4(n_{\mathcal{A}_h} + n_0)\}$.*

**Lemma 3.** *Under Assumptions 1-3 and Assumption 7.(i)-(iii),*

$$\sup_{k,j} \|\boldsymbol{\gamma}_j^{\mathcal{A}_h} - \boldsymbol{\gamma}_j^{(0)}\|_1 \lesssim h_1.$$

**Lemma 4.** *Assume Assumptions 1-4 and 7 hold. Let $\lambda_j \asymp \sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}} + (\frac{1}{\sqrt{s}}(\frac{\log p}{n_0})^{1/4} h^{1/2}) \wedge \frac{h}{\sqrt{s}}$ for any $j = 1, \ldots, p$. Suppose $h_1 \lesssim s^{-1/2}$. With probability at least $1 - K_{\mathcal{A}_h} n_0^{-1}$,*

$$\|\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}} - \boldsymbol{\gamma}_j^{\mathcal{A}}\|_2^2 \lesssim h_1 \left(\frac{\log p}{n_{\mathcal{A}_h} + n_0}\right)^{1/2} + \Re_1^2 + \Re_1 \frac{h_1}{\sqrt{s}},$$
$$\|\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}} - \boldsymbol{\gamma}_j^{\mathcal{A}}\|_1 \lesssim \sqrt{s}\Re_1 + s^{1/4} h_1^{1/2} \Re_1^{1/2} + h_1.$$

**Lemma 5.** *Let $\tilde{\lambda}_j \asymp \sqrt{\frac{\log p}{n_0}} + \Re_1$. Impose the same conditions assumed by Lemma 4. Suppose $h_1 \lesssim s_*^{-1/2}$. Then with probability at least $1 - K_{\mathcal{A}_h} n_0^{-1}$,*

$$\|\hat{\boldsymbol{\varrho}}_j - \boldsymbol{\varrho}_j\|_2^2 \lesssim h_1 \sqrt{\frac{\log p}{n_0}} + h_1\Re_1 + \Re_1^2,$$
$$\|\hat{\boldsymbol{\varrho}}_j - \boldsymbol{\varrho}_j\|_1 \lesssim h_1 + \Re_1.$$

The proof of Lemma 1 will be presented later in Section S.2.2.1. Lemma 2 can be derived in the same spirit as the proof of Proposition 2 in the full-length version of Negahban et al. (2009), so we omit the full proof and only highlight the sketch in Section S.2.2.2. Lemma 3 can be proved by following the same idea in the proof of Lemma 1, therefore we omit its proof as well.

Also, it is important to point out that all the constants involved in the proofs of Theorem 1-7 are independent with $\boldsymbol{\xi} = \{\boldsymbol{\beta}, \{\boldsymbol{w}^{(k)}\}_{k\in\mathcal{A}}\} \in \Xi(s,h)$, therefore we can take the supremum over $\boldsymbol{\xi} \in \Xi(s,h)$ in the final conclusion without changing the rate.

## S.2.2 Proof of lemmas

### S.2.2.1 Proof of Lemma 1

By definition,

$$\sum_{k\in\{0\}\cup\mathcal{A}} \alpha_k \mathbb{E}\left\{[\psi'((\boldsymbol{w}^{\mathcal{A}_h})^T\boldsymbol{x}^{(k)}) - \psi'((\boldsymbol{w}^{(k)})^T\boldsymbol{x}^{(k)})]\boldsymbol{x}^{(k)}\right\} = 0,$$

which implies

$$\begin{aligned}&\sum_{k\in\{0\}\cup\mathcal{A}} \alpha_k \mathbb{E}\left\{[\psi'((\boldsymbol{w}^{\mathcal{A}_h})^T\boldsymbol{x}^{(k)}) - \psi'(\boldsymbol{\beta}^T\boldsymbol{x}^{(k)})]\boldsymbol{x}^{(k)}\right\}\\ =&\sum_{k\in\mathcal{A}} \alpha_k \mathbb{E}\left\{[\psi'((\boldsymbol{w}^{(k)})^T\boldsymbol{x}^{(k)}) - \psi'(\boldsymbol{\beta}^T\boldsymbol{x}^{(k)})]\boldsymbol{x}^{(k)}\right\}.\end{aligned}$$

By Taylor expansion,

$$\begin{aligned}&\sum_{k\in\{0\}\cup\mathcal{A}} \alpha_k \mathbb{E}\left[\int_0^1 \psi''((\boldsymbol{w}^{\mathcal{A}_h})^T\boldsymbol{x}^{(k)} + t(\boldsymbol{w}^{\mathcal{A}_h} - \boldsymbol{\beta})^T\boldsymbol{x}^{(k)})\boldsymbol{x}^{(k)}(\boldsymbol{x}^{(k)})^T\right](\boldsymbol{w}^{\mathcal{A}_h} - \boldsymbol{\beta})\\ =&\sum_{k\in\mathcal{A}} \alpha_k \mathbb{E}\left[\int_0^1 \psi''((\boldsymbol{w}^{(k)})^T\boldsymbol{x}^{(k)} + t(\boldsymbol{w}^{(k)} - \boldsymbol{\beta})^T\boldsymbol{x}^{(k)})\boldsymbol{x}^{(k)}(\boldsymbol{x}^{(k)})^T\right](\boldsymbol{w}^{(k)} - \boldsymbol{\beta}).\end{aligned}$$

Therefore, by Assumption 4, $\|\boldsymbol{w}^{\mathcal{A}_h} - \boldsymbol{\beta}\|_1 \le \sum_{k\in\mathcal{A}} \alpha_k \|\widetilde{\boldsymbol{\Sigma}}_h^{-1}\widetilde{\boldsymbol{\Sigma}}_h^{(k)}\|_1 \cdot \|\boldsymbol{w}^{(k)} - \boldsymbol{\beta}\|_1 \le C_1 h$.

### S.2.2.2 Proof of Lemma 2

By the second-order Taylor expansion, for some $t_i^{(k)} \in [0,1]$,

$$\delta\hat{L}(\boldsymbol{u},\mathcal{D}) = \hat{L}(\boldsymbol{w}^{\mathcal{A}_h} + \boldsymbol{u},\mathcal{D}) - \hat{L}(\boldsymbol{w}^{\mathcal{A}_h}) - \nabla\hat{L}^{(k)}(\boldsymbol{w}^{\mathcal{A}_h})^T\boldsymbol{u}$$

$$= \frac{1}{n_{\mathcal{A}_h} + n_0} \sum_{k \in \{0\} \cup \mathcal{A}_h} \sum_{i=1}^{n_k} \left[ \psi((\boldsymbol{w}^{\mathcal{A}_h} + \boldsymbol{u})^T \boldsymbol{x}_i^{(k)}) - \psi((\boldsymbol{w}^{\mathcal{A}_h})^T \boldsymbol{x}_i^{(k)}) - \psi'((\boldsymbol{w}^{\mathcal{A}_h})^T \boldsymbol{x}_i^{(k)}) \boldsymbol{u}^T \boldsymbol{x}_i^{(k)} \right]$$

$$= \frac{1}{n_{\mathcal{A}_h} + n_0} \sum_{k \in \{0\} \cup \mathcal{A}_h} \sum_{i=1}^{n_k} \psi''((\boldsymbol{w}^{\mathcal{A}_h})^T \boldsymbol{x}_i^{(k)} + t_i^{(k)} \boldsymbol{u}^T \boldsymbol{x}_i^{(k)}) (\boldsymbol{u}^T \boldsymbol{x}_i^{(k)})^2,$$

which is the counterpart of equation (63) in the full-length version of Negahban et al. (2009). Due to the independence of between $\boldsymbol{x}_i^{(k)}$ for any $i$ and $k$, the arguments in Negahban et al. (2009) directly follow.

#### S.2.2.3 Proof of Lemma 4

Recall Theorem 1, under the assumptions, we have

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \lesssim \sqrt{\frac{s \log p}{n_{\mathcal{A}} + n_0}} + \left[ \left( \frac{\log p}{n_0} \right)^{1/4} h^{1/2} \right] \wedge h,$$

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \lesssim s \sqrt{\frac{\log p}{n_{\mathcal{A}} + n_0}} + h,$$

with probability at least $1 - n_0^{-1}$. By basic inequality,

$$\frac{1}{2(n_{\mathcal{A}_h} + n_0)} \sum_{k \in \{0\} \cup \mathcal{A}_h} \|\boldsymbol{X}_{\hat{\boldsymbol{\beta}},j}^{(k)} - \boldsymbol{X}_{\hat{\boldsymbol{\beta}},-j}^{(k)} \hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h}\|_2^2 + \lambda_j \|\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h}\|_1$$
$$\leq \frac{1}{2(n_{\mathcal{A}_h} + n_0)} \sum_{k \in \{0\} \cup \mathcal{A}_h} \|\boldsymbol{X}_{\hat{\boldsymbol{\beta}},j}^{(k)} - \boldsymbol{X}_{\hat{\boldsymbol{\beta}},-j}^{(k)} \boldsymbol{\gamma}_j^{\mathcal{A}_h}\|_2^2 + \lambda_j \|\boldsymbol{\gamma}_j^{\mathcal{A}_h}\|_1,$$

which implies

$$\frac{1}{2(n_{\mathcal{A}_h} + n_0)} \sum_{k \in \{0\} \cup \mathcal{A}_h} \|\boldsymbol{X}_{\hat{\boldsymbol{\beta}},-j}^{(k)} (\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h} - \boldsymbol{\gamma}_j^{\mathcal{A}_h})\|_2^2$$
$$\leq \frac{1}{n_{\mathcal{A}_h} + n_0} \sum_{k \in \{0\} \cup \mathcal{A}_h} \langle \boldsymbol{W}_{\hat{\boldsymbol{\beta}}}^{(k)} (\boldsymbol{W}_{\boldsymbol{\beta}}^{(k)})^{-1} \boldsymbol{\eta}_{\boldsymbol{\beta},j}^{(k)}, \boldsymbol{X}_{\hat{\boldsymbol{\beta}},-j}^{(k)} (\boldsymbol{\gamma}_j^{\mathcal{A}_h} - \hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h}) \rangle$$

$$+\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{k\in\{0\}\cup\mathcal{A}_h}\langle \boldsymbol{X}^{(k)}_{\hat{\boldsymbol{\beta}},-j}(\boldsymbol{\gamma}_j^{(k)}-\boldsymbol{\gamma}_j^{\mathcal{A}_h}),\boldsymbol{X}^{(k)}_{\hat{\boldsymbol{\beta}},-j}(\boldsymbol{\gamma}_j^{\mathcal{A}_h}-\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h})\rangle$$
$$+\lambda_j(\|\boldsymbol{\gamma}_j^{\mathcal{A}_h}\|_1-\|\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h}\|_1). \tag{S.2.12}$$

Denote $\boldsymbol{\Lambda}^{(k)}=\mathrm{diag}(\{\psi''((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\beta})\}_{i=1}^{n_k})$ and $\widehat{\boldsymbol{\Lambda}}^{(k)}=\mathrm{diag}(\{\psi''((\boldsymbol{x}_i^{(k)})^T\hat{\boldsymbol{\beta}})\}_{i=1}^{n_k})$.

$$\begin{aligned}&\left|\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{k\in\{0\}\cup\mathcal{A}_h}\langle \boldsymbol{X}^{(k)}_{\hat{\boldsymbol{\beta}},-j}(\boldsymbol{\gamma}_j^{(k)}-\boldsymbol{\gamma}_j^{\mathcal{A}_h}),\boldsymbol{X}^{(k)}_{\hat{\boldsymbol{\beta}},-j}(\boldsymbol{\gamma}_j^{\mathcal{A}_h}-\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h})\right|\\ &\le\frac{1}{n_{\mathcal{A}_h}+n_0}\left[\left\|\sum_{k\in\{0\}\cup\mathcal{A}_h}(\boldsymbol{X}^{(k)}_{-j})^T\boldsymbol{\Lambda}^{(k)}\boldsymbol{X}^{(k)}_{-j}(\boldsymbol{\gamma}_j^{(k)}-\boldsymbol{\gamma}_j^{\mathcal{A}_h})\right\|_\infty\right.\\ &\quad\left.+\left\|\sum_{k\in\{0\}\cup\mathcal{A}_h}(\boldsymbol{X}^{(k)}_{-j})^T(\widehat{\boldsymbol{\Lambda}}^{(k)}-\boldsymbol{\Lambda}^{(k)})\boldsymbol{X}^{(k)}_{-j}(\boldsymbol{\gamma}_j^{(k)}-\boldsymbol{\gamma}_j^{\mathcal{A}_h})\right\|_\infty\right]\cdot\|\boldsymbol{\gamma}_j^{\mathcal{A}_h}-\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h}\|_1.\end{aligned} \tag{S.2.13}$$

It's easy to see that each component of $\sum_{k\in\{0\}\cup\mathcal{A}_h}(\boldsymbol{X}^{(k)}_{-j})^T\boldsymbol{\Lambda}^{(k)}\boldsymbol{X}^{(k)}_{-j}(\boldsymbol{\gamma}_j^{(k)}-\boldsymbol{\gamma}_j^{\mathcal{A}_h})$ is a zero-mean, sub-exponential variable with variance $C(n_{\mathcal{A}}+n_0)h^2$. By union bounds,

$$\frac{1}{n_{\mathcal{A}_h}+n_0}\left\|\sum_{k\in\{0\}\cup\mathcal{A}_h}(\boldsymbol{X}^{(k)}_{-j})^T\boldsymbol{\Lambda}^{(k)}\boldsymbol{X}^{(k)}_{-j}(\boldsymbol{\gamma}_j^{(k)}-\boldsymbol{\gamma}_j^{\mathcal{A}_h})\right\|_\infty\lesssim\sqrt{\frac{\log p}{n_{\mathcal{A}}+n_0}}h_1,$$

with probability at least $1-p^{-1}$. In addition,

$$\begin{aligned}&\frac{1}{n_{\mathcal{A}_h}+n_0}\left\|\sum_{k\in\{0\}\cup\mathcal{A}_h}(\boldsymbol{X}^{(k)}_{-j})^T(\widehat{\boldsymbol{\Lambda}}^{(k)}-\boldsymbol{\Lambda}^{(k)})\boldsymbol{X}^{(k)}_{-j}(\boldsymbol{\gamma}_j^{(k)}-\boldsymbol{\gamma}_j^{\mathcal{A}_h})\right\|_\infty\\ &\le\frac{1}{n_{\mathcal{A}_h}+n_0}\left\|\sum_{k\in\{0\}\cup\mathcal{A}_h}(\boldsymbol{X}^{(k)}_{-j})^T(\widehat{\boldsymbol{\Lambda}}^{(k)}-\boldsymbol{\Lambda}^{(k)})\boldsymbol{X}^{(k)}_{-j}\right\|_{\max}\cdot\|\boldsymbol{\gamma}_j^{(k)}-\boldsymbol{\gamma}_j^{\mathcal{A}_h}\|_1.\end{aligned}$$

For $j_1$, $j_2\neq j$,

$$\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{k\in\{0\}\cup\mathcal{A}_h}\sum_{i=1}^{n_k}x_{i,j_1}^{(k)}x_{i,j_2}^{(k)}[\psi''((\boldsymbol{x}_i^{(k)})^T\hat{\boldsymbol{\beta}})-\psi''((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\beta})]$$

$$\leq \sqrt{\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{k\in\{0\}\cup\mathcal{A}_h}\sum_{i=1}^{n_k}(x_{i,j_1}^{(k)}x_{i,j_2}^{(k)})^2}\cdot\sqrt{\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{k\in\{0\}\cup\mathcal{A}_h}\sum_{i=1}^{n_k}\psi'''(\boldsymbol{a}_i^{(k)})[(\boldsymbol{x}_i^{(k)})^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})]^2}$$

$$\leq C\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2,$$

where the constant $C$ is the same for different $j_1$ and $j_2$, and $\boldsymbol{a}_i^{(k)}$ falls on the line segment between $(\boldsymbol{x}_i^{(k)})^T\hat{\boldsymbol{\beta}}$ and $(\boldsymbol{x}_i^{(k)})^T\boldsymbol{\beta}$. Therefore, the right-hand side of (S.2.13) can be upper bounded by

$$C\left(\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2+\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}\right)h_1\cdot\|\boldsymbol{\gamma}_j^{\mathcal{A}_h}-\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h}\|_1,$$

with probability at least $1-p^{-1}$.

On the other hand,

$$\begin{aligned}
&\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{k\in\{0\}\cup\mathcal{A}_h}\langle\boldsymbol{W}_{\hat{\boldsymbol{\beta}}}^{(k)}(\boldsymbol{W}_{\boldsymbol{\beta}}^{(k)})^{-1}\boldsymbol{\eta}_{\boldsymbol{\beta},j}^{(k)},\boldsymbol{X}_{\hat{\boldsymbol{\beta}},-j}^{(k)}(\boldsymbol{\gamma}_j^{\mathcal{A}_h}-\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h})\rangle\\
&\leq\frac{1}{n_{\mathcal{A}_h}+n_0}\left\|\sum_{k\in\{0\}\cup\mathcal{A}_h}\boldsymbol{X}_{\hat{\boldsymbol{\beta}},-j}^{(k)}\boldsymbol{\eta}_{\boldsymbol{\beta},j}^{(k)}\right\|_\infty\|\boldsymbol{\gamma}_j^{\mathcal{A}_h}-\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h}\|_1\\
&\quad+\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{k\in\{0\}\cup\mathcal{A}_h}\left[\frac{1}{4c_0}\left\|(\boldsymbol{W}_{\hat{\boldsymbol{\beta}}}^{(k)}(\boldsymbol{W}_{\boldsymbol{\beta}}^{(k)})^{-1}-\boldsymbol{I})\boldsymbol{\eta}_{\boldsymbol{\beta},j}^{(k)}\right\|_2^2\right.\\
&\qquad\left.+c_0\left\|\boldsymbol{X}_{\hat{\boldsymbol{\beta}},-j}^{(k)}(\boldsymbol{\gamma}_j^{\mathcal{A}_h}-\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h})\right\|_2^2\right],
\end{aligned}\tag{S.2.14}$$

where $c_0$ is a positive constant smaller than $1/4$. Note that

$$\begin{aligned}
&\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{k\in\{0\}\cup\mathcal{A}_h}\left\|(\boldsymbol{W}_{\hat{\boldsymbol{\beta}}}^{(k)}(\boldsymbol{W}_{\boldsymbol{\beta}}^{(k)})^{-1}-\boldsymbol{I})\boldsymbol{\eta}_{\boldsymbol{\beta},j}^{(k)}\right\|_2^2\\
&\leq\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{k\in\{0\}\cup\mathcal{A}_h}\sum_{i=1}^{n_k}\frac{\left(\sqrt{\psi''((\boldsymbol{x}_i^{(k)})^T\hat{\boldsymbol{\beta}})}-\sqrt{\psi''((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\beta})}\right)^2}{\psi''((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\beta})}\cdot|\eta_i^{(k)}|^2
\end{aligned}$$

$$
\begin{aligned}
&\lesssim \frac{1}{n_{\mathcal{A}_h}+n_0} \sum_{k\in\{0\}\cup\mathcal{A}_h} \sum_{i=1}^{n_k} [(\boldsymbol{x}_i^{(k)})^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})]^2, \\
&\lesssim \sup_{k\in\{0\}\cup\mathcal{A}_h} \|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2^2 \\
&\lesssim \Re_1^2,
\end{aligned}
$$

with probability at least $1-K_{\mathcal{A}_h}n_0^{-1}$, where we use Assumption 7.(i) to bound $\psi''((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\beta})$. The last two inequalities follows because of the same reason as that of (S.2.27). Plugging (S.2.13) and (S.2.14) and into (S.2.12),

$$
\begin{aligned}
&\frac{1}{4(n_{\mathcal{A}_h}+n_0)} \sum_{k\in\{0\}\cup\mathcal{A}_h} \|\boldsymbol{X}^{(k)}_{\hat{\boldsymbol{\beta}},-j}(\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h}-\boldsymbol{\gamma}_j^{\mathcal{A}_h})\|_2^2 \\
&\le C\left(\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}+\Re_1\right) h_1 \cdot \|\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h}-\boldsymbol{\gamma}_j^{\mathcal{A}_h}\|_1 + \Re_1^2 \\
&\quad + \frac{C}{n_{\mathcal{A}_h}+n_0} \left\| \sum_{k\in\{0\}\cup\mathcal{A}_h} \boldsymbol{X}^{(k)}_{\hat{\boldsymbol{\beta}},-j}\boldsymbol{\eta}^{(k)}_{\boldsymbol{\beta},j} \right\|_\infty \|\boldsymbol{\gamma}_j^{\mathcal{A}_h}-\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h}\|_1 + \lambda_j(\|\boldsymbol{\gamma}_j^{\mathcal{A}_h}\|_1 - \|\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h}\|_1)
\end{aligned}
$$

with probability at least $1-K_{\mathcal{A}_h}n_0^{-1}$. Note that $\frac{1}{n_{\mathcal{A}_h}+n_0}\left\|\sum_{k\in\{0\}\cup\mathcal{A}_h}\boldsymbol{X}^{(k)}_{\hat{\boldsymbol{\beta}},-j}\right\|_\infty \lesssim \sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}$ with probability at least $1-p^{-1}$. Therefore, if $\lambda_j \ge C'\Re_1/\sqrt{s}$ with some large $C'>0$, then since $h_1 \lesssim s^{-1/2}$, we have $\lambda_j > \frac{2C}{n_{\mathcal{A}_h}+n_0}\left\|\sum_{k\in\{0\}\cup\mathcal{A}_h}\boldsymbol{X}^{(k)}_{\hat{\boldsymbol{\beta}},-j}\right\|_\infty + C\left(\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}+\Re_1\right)h_1$, which implies

$$
\begin{aligned}
\frac{1}{4(n_{\mathcal{A}_h}+n_0)} \sum_{k\in\{0\}\cup\mathcal{A}_h} \|\boldsymbol{X}^{(k)}_{\hat{\boldsymbol{\beta}},-j}(\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h}-\boldsymbol{\gamma}_j^{\mathcal{A}_h})\|_2^2 &\le \frac{3}{2}\lambda_j\|(\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h}-\boldsymbol{\gamma}_j^{\mathcal{A}_h})_{S_j}\|_1 - \frac{1}{2}\lambda_j\|(\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h}-\boldsymbol{\gamma}_j^{\mathcal{A}_h})_{S_j^c}\|_1 \\
&\quad + \lambda_j\|(\boldsymbol{\gamma}_j^{\mathcal{A}_h})_{S_j^c}\|_1 + C\Re_1^2, \qquad \text{(S.2.15)}
\end{aligned}
$$

with probability at least $1-K_{\mathcal{A}_h}n_0^{-1}$. Therefore,

$$
\|\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h}-\boldsymbol{\gamma}_j^{\mathcal{A}_h}\|_1 \le 4\sqrt{s}\|\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h}-\boldsymbol{\gamma}_j^{\mathcal{A}_h}\|_2 + C\sqrt{s}\Re_1 + h_1, \qquad \text{(S.2.16)}
$$

with probability at least $1-K_{\mathcal{A}_h}n_0^{-1}$. Thus, we have either $\|\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h}-\boldsymbol{\gamma}_j^{\mathcal{A}_h}\|_1 \le 2C\sqrt{s}\Re_1+2h_1 \lesssim \sqrt{s}$ or $\|\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h}-\boldsymbol{\gamma}_j^{\mathcal{A}_h}\|_1 \le 8\sqrt{s}\|\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h}-\boldsymbol{\gamma}_j^{\mathcal{A}_h}\|_2$. By similar arguments to get (S.2.27), we obtain

$$\begin{aligned}\frac{1}{4(n_{\mathcal{A}_h}+n_0)}\sum_{k\in\{0\}\cup\mathcal{A}_h}\|\boldsymbol{X}^{(k)}_{\hat{\boldsymbol{\beta}},-j}(\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h}-\boldsymbol{\gamma}_j^{\mathcal{A}_h})\|_2^2 &\gtrsim \frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{k\in\{0\}\cup\mathcal{A}_h}\|\boldsymbol{X}^{(k)}_{-j}(\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h}-\boldsymbol{\gamma}_j^{\mathcal{A}_h})\|_2^2\\ &\gtrsim \|\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h}-\boldsymbol{\gamma}_j^{\mathcal{A}_h}\|_2^2, \qquad \text{(S.2.17)}\end{aligned}$$

with probability at least $1-K_{\mathcal{A}_h}n_0^{-1}$. Combine (S.2.15), (S.2.16) and (S.2.17) to get the desired conclusions.

#### S.2.2.4 Proof of Lemma 5

Similar to (S.2.12), by basic inequality,

$$\begin{aligned}\frac{1}{2n_0}\left\|\boldsymbol{X}^{(0)}_{\hat{\boldsymbol{\beta}},-j}(\hat{\boldsymbol{\varrho}}_j-\boldsymbol{\varrho}_j)\right\|_2^2 \le{}& \frac{1}{n_0}\langle \boldsymbol{W}^{(0)}_{\hat{\boldsymbol{\beta}}}(\boldsymbol{W}^{(0)}_{\boldsymbol{\beta}})^{-1}\boldsymbol{\eta}^{(0)}_{\boldsymbol{\beta},j}, \boldsymbol{X}^{(0)}_{\hat{\boldsymbol{\beta}},-j}(\hat{\boldsymbol{\varrho}}_j-\boldsymbol{\varrho}_j)\rangle\\ &+\frac{1}{n_0}\langle \boldsymbol{X}^{(0)}_{\hat{\boldsymbol{\beta}},-j}(\hat{\boldsymbol{\varrho}}_j-\boldsymbol{\varrho}_j), \boldsymbol{X}^{(0)}_{\hat{\boldsymbol{\beta}},-j}(\boldsymbol{\gamma}_j^{\mathcal{A}_h}-\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h})\rangle\\ &+\tilde{\lambda}_j(\|\boldsymbol{\varrho}_j\|_1-\|\hat{\boldsymbol{\varrho}}_j\|_1).\end{aligned}$$

Similar to the analysis in the proof of Lemma 4,

$$\begin{aligned}\frac{1}{4n_0}\left\|\boldsymbol{X}^{(0)}_{\hat{\boldsymbol{\beta}},-j}(\hat{\boldsymbol{\varrho}}_j-\boldsymbol{\varrho}_j)\right\|_2^2 \le{}& C\left(\sqrt{\frac{\log p}{n_0}}+\Re_1\right)h_1\cdot\|\hat{\boldsymbol{\varrho}}_j-\boldsymbol{\varrho}_j\|_1 + C\|\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h}-\boldsymbol{\gamma}_j^{\mathcal{A}_h}\|_2^2\\ &+C\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2^2+\frac{C}{n_0}\left\|\boldsymbol{X}^{(0)}_{\boldsymbol{\beta},-j}\boldsymbol{\eta}^{(0)}_j\right\|_\infty\cdot\|\hat{\boldsymbol{\varrho}}_j-\boldsymbol{\varrho}_j\|_1\\ &+\tilde{\lambda}_j(\|\boldsymbol{\varrho}_j\|_1-\|\hat{\boldsymbol{\varrho}}_j\|_1)\end{aligned}$$

with probability at least $1-n_0^{-1}$. Note that $\frac{1}{n_0}\|\boldsymbol{X}^{(0)}_{\boldsymbol{\beta},-j}\boldsymbol{\eta}^{(0)}_j\|_\infty \lesssim \sqrt{\frac{\log p}{n_0}}$ with probability at least $1-p^{-1}$. Therefore for $\tilde{\lambda}_j \ge C'\left(\sqrt{\frac{\log p}{n_0}}+\Re_1\right)$ with large enough $C'>0$, we have

$\tilde{\lambda}_j \geq C\left(\sqrt{\frac{\log p}{n_0}} + \Re_1\right) h_1 + \frac{C}{n_0}\|\boldsymbol{X}^{(0)}_{\hat{\boldsymbol{\beta}},-j}\boldsymbol{\eta}^{(0)}_j\|_\infty$. Then with probability at least $1 - n_0^{-1}$,

$$\frac{1}{4n_0}\left\|\boldsymbol{X}^{(0)}_{\hat{\boldsymbol{\beta}},-j}(\hat{\boldsymbol{\varrho}}_j - \boldsymbol{\varrho}_j)\right\|_2^2 \leq 2\tilde{\lambda}_j\|\boldsymbol{\varrho}_j\|_1 - \frac{1}{2}\tilde{\lambda}_j\|\hat{\boldsymbol{\varrho}}_j - \boldsymbol{\varrho}_j\|_1 + C\|\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h} - \boldsymbol{\gamma}_j^{\mathcal{A}_h}\|_2^2 + C\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2,$$

which leads to

$$\begin{aligned}\|\hat{\boldsymbol{\varrho}}_j - \boldsymbol{\varrho}_j\|_1 &\leq 4\|\boldsymbol{\varrho}_j\|_1 + C\tilde{\lambda}_j^{-1}\|\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h} - \boldsymbol{\gamma}_j^{\mathcal{A}_h}\|_2^2 + C\tilde{\lambda}_j^{-1}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \\ &\lesssim h_1 + \Re_1 + \left(\sqrt{\frac{n_0}{\log p}}\Re_1^2\right) \wedge \Re_1,\end{aligned}$$

with probability at least $1 - K_{\mathcal{A}_h} n_0^{-1}$. Similar to the trick we used before, we can get

$$\|\hat{\boldsymbol{\varrho}}_j - \boldsymbol{\varrho}_j\|_2^2 \lesssim \tilde{\lambda}_j h_1 + \|\hat{\boldsymbol{\gamma}}_j^{\mathcal{A}_h} - \boldsymbol{\gamma}_j^{\mathcal{A}_h}\|_2^2 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \lesssim h_1\sqrt{\frac{\log p}{n_0}} + h_1\Re_1 + \Re_1^2$$

with probability at least $1 - K_{\mathcal{A}_h} n_0^{-1}$, which completes the proof.

## S.2.3 Proof of theorems

### S.2.3.1 Proof of Theorem 1

**Transferring step:** Define $\hat{\boldsymbol{u}}^{\mathcal{A}_h} = \hat{\boldsymbol{w}}^{\mathcal{A}_h} - \boldsymbol{w}^{\mathcal{A}_h}$ and $\mathcal{D} = \{(\boldsymbol{X}^{(k)}, \boldsymbol{y}^{(k)})\}_{k\in\{0\}\cup\mathcal{A}_h}$. We first claim that when $\lambda_{\boldsymbol{w}} \geq 2\|\nabla L(\boldsymbol{w}^{\mathcal{A}_h}, \mathcal{D})\|_\infty$, with probability at least $1 - C_3\exp\{-C_4(n_{\mathcal{A}_h} + n_0)\}$, it holds that

$$\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_2 \leq 8\kappa_2 C_1 h\sqrt{\frac{\log p}{n_{\mathcal{A}_h} + n_0}} + 3\frac{\sqrt{s}}{\kappa_1}\lambda_{\boldsymbol{w}} + 2\sqrt{\frac{C_1}{\kappa_1}h\lambda_{\boldsymbol{w}}}. \tag{S.2.18}$$

To see this, first by the definition of $\hat{\boldsymbol{w}}^{\mathcal{A}_h}$, Hölder inequality and Lemma 1, we have

$$\begin{aligned}\delta\hat{L}(\hat{\boldsymbol{u}}^{\mathcal{A}_h}, \mathcal{D}) &\leq \lambda_{\boldsymbol{w}}(\|\boldsymbol{w}_S^{\mathcal{A}_h}\|_1 + \|\boldsymbol{w}_{S^c}^{\mathcal{A}_h}\|_1) - \lambda_{\boldsymbol{w}}(\|\hat{\boldsymbol{w}}_S^{\mathcal{A}_h}\|_1 + \|\hat{\boldsymbol{w}}_{S^c}^{\mathcal{A}_h}\|_1) + \nabla\hat{L}(\boldsymbol{w}, \mathcal{D})^T\hat{\boldsymbol{u}}^{\mathcal{A}_h} \\ &\leq \lambda_{\boldsymbol{w}}(\|\boldsymbol{w}_S^{\mathcal{A}_h}\|_1 + \|\boldsymbol{w}_{S^c}^{\mathcal{A}_h}\|_1) - \lambda_{\boldsymbol{w}}(\|\hat{\boldsymbol{w}}_S^{\mathcal{A}_h}\|_1 + \|\hat{\boldsymbol{w}}_{S^c}^{\mathcal{A}_h}\|_1) + \frac{1}{2}\lambda_{\boldsymbol{w}}\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1\end{aligned}$$

$$\leq \frac{3}{2}\lambda_{\boldsymbol{w}}\|\hat{\boldsymbol{u}}_S^{\mathcal{A}_h}\|_1 - \frac{1}{2}\lambda_{\boldsymbol{w}}\|\hat{\boldsymbol{u}}_{S^c}^{\mathcal{A}_h}\|_1 + 2\lambda_{\boldsymbol{w}}\|\boldsymbol{w}_{S^c}^{\mathcal{A}_h}\|_1$$
$$\leq \frac{3}{2}\lambda_{\boldsymbol{w}}\|\hat{\boldsymbol{u}}_S^{\mathcal{A}_h}\|_1 - \frac{1}{2}\lambda_{\boldsymbol{w}}\|\hat{\boldsymbol{u}}_{S^c}^{\mathcal{A}_h}\|_1 + 2\lambda_{\boldsymbol{w}}C_1 h. \tag{S.2.19}$$

If the claim does not hold, consider $\mathbb{C} = \left\{\boldsymbol{u} : \frac{3}{2}\|\boldsymbol{u}_S\|_1 - \frac{1}{2}\|\boldsymbol{u}_{S^c}\|_1 + 2C_1 h \geq 0\right\}$. Due to (S.2.19) and the convexity of $\hat{L}$, $\hat{\boldsymbol{u}}^{\mathcal{A}_h} \in \mathbb{C}$. Then for any $t \in (0,1)$, it's easy to see that

$$\frac{1}{2}\|t\hat{\boldsymbol{u}}_{S^c}^{\mathcal{A}_h}\|_1 = t \cdot \frac{1}{2}\|\hat{\boldsymbol{u}}_{S^c}^{\mathcal{A}_h}\|_1 \leq t \cdot \left(\frac{3}{2}\|\hat{\boldsymbol{u}}_S^{\mathcal{A}_h}\|_1 + 2C_{\boldsymbol{w}}h\right) \leq \frac{3}{2}\|t\hat{\boldsymbol{u}}_S^{\mathcal{A}_h}\|_1 + 2C_1 h,$$

implying that $t\hat{\boldsymbol{u}}^{\mathcal{A}_h} \in \mathbb{C}$. We could find some $t$ satisfying that $\|t\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_2 > 8\kappa_2 C_1 h\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}} + 3\frac{\sqrt{s}}{\kappa_1}\lambda_{\boldsymbol{w}} + 2\sqrt{\frac{C_1}{\kappa_1}h\lambda_{\boldsymbol{w}}}$ and $\|t\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_2 \leq 1$. Denote $\tilde{\boldsymbol{u}}^{\mathcal{A}_h} = t\hat{\boldsymbol{u}}^{\mathcal{A}_h}$ and $F(\boldsymbol{u}) = \hat{L}(\boldsymbol{w}^{\mathcal{A}_h} + \boldsymbol{u}, \mathcal{D}) - \hat{L}(\boldsymbol{w}^{\mathcal{A}_h}) + \lambda_{\boldsymbol{w}}(\|\boldsymbol{w}^{\mathcal{A}_h} + \boldsymbol{u}\|_1 - \|\boldsymbol{w}^{\mathcal{A}_h}\|_1)$. Since $F(\boldsymbol{0}) = 0$ and $F(\hat{\boldsymbol{u}}^{\mathcal{A}_h}) \leq 0$, by convexity,

$$F(\tilde{\boldsymbol{u}}^{\mathcal{A}_h}) = F(t\hat{\boldsymbol{u}}^{\mathcal{A}_h} + (1-t)\boldsymbol{0}) \leq tF(\hat{\boldsymbol{u}}^{\mathcal{A}_h}) \leq 0. \tag{S.2.20}$$

However, by Lemma 2 and the same trick of (S.2.19),

$$F(\tilde{\boldsymbol{u}}^{\mathcal{A}_h}) \geq \delta\hat{L}(\hat{\boldsymbol{u}}^{\mathcal{A}_h}, \mathcal{D}) + \nabla\hat{L}(\boldsymbol{w}^{\mathcal{A}_h})^T\tilde{\boldsymbol{u}}^{\mathcal{A}_h} - \lambda_{\boldsymbol{w}}\|\boldsymbol{w}^{\mathcal{A}_h}\|_1 + \lambda_{\boldsymbol{w}}\|\boldsymbol{w}^{\mathcal{A}_h} + \tilde{\boldsymbol{u}}^{\mathcal{A}_h}\|_1$$
$$\geq \kappa_1\|\tilde{\boldsymbol{u}}^{\mathcal{A}_h}\|_2^2 - \kappa_1\kappa_2\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}\|\tilde{\boldsymbol{u}}^{\mathcal{A}_h}\|_1\|\tilde{\boldsymbol{u}}^{\mathcal{A}_h}\|_2 - \frac{3}{2}\lambda_{\boldsymbol{w}}\|\tilde{\boldsymbol{u}}_S^{\mathcal{A}_h}\|_1 + \frac{1}{2}\lambda_{\boldsymbol{w}}\|\tilde{\boldsymbol{u}}_{S^c}^{\mathcal{A}_h}\|_1 - 2\lambda_{\boldsymbol{w}}C_1 h$$

$$\geq \kappa_1\|\tilde{\boldsymbol{u}}^{\mathcal{A}_h}\|_2^2 - \kappa_1\kappa_2\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}\|\tilde{\boldsymbol{u}}^{\mathcal{A}_h}\|_1\|\tilde{\boldsymbol{u}}^{\mathcal{A}_h}\|_2 - \frac{3}{2}\lambda_{\boldsymbol{w}}\|\tilde{\boldsymbol{u}}_S^{\mathcal{A}_h}\|_1 - 2\lambda_{\boldsymbol{w}}C_1 h.$$

Note that since $\tilde{\boldsymbol{u}}^{\mathcal{A}_h} \in \mathbb{C}$, it holds that

$$\frac{1}{2}\|\tilde{\boldsymbol{u}}^{\mathcal{A}_h}\|_1 \leq 2\|\tilde{\boldsymbol{u}}_S^{\mathcal{A}_h}\|_1 + 2C_{\boldsymbol{w}}h \leq 2\sqrt{s}\|\tilde{\boldsymbol{u}}^{\mathcal{A}_h}\|_2 + 2C_1 h.$$

When $n_{\mathcal{A}_h} + n_0 > 16\kappa_2^2 s\log p$, we have $2k_2\sqrt{\frac{s\log p}{n_{\mathcal{A}_h}+n_0}} \leq \frac{1}{2}$. Then it follows

$$F(\tilde{\boldsymbol{u}}^{\mathcal{A}_h}) \geq \frac{1}{2}\kappa_1\|\tilde{\boldsymbol{u}}^{\mathcal{A}_h}\|_2^2 - \left[2\kappa_1\kappa_2\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}C_{\boldsymbol{w}}h + \frac{3}{2}\lambda_{\boldsymbol{w}}\sqrt{s}\right]\|\tilde{\boldsymbol{u}}^{\mathcal{A}_h}\|_2 - 2\lambda_{\boldsymbol{w}}C_1 h > 0,$$

which conflicts with (S.2.20). Therefore our claim at the beginning holds.

Next, let's prove $\|\nabla \hat{L}(\boldsymbol{w}^{\mathcal{A}_h})\|_\infty \lesssim \sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}$ with probability at least $1-(n_{\mathcal{A}_h}+n_0)^{-1}$. To see this, notice that

$$\begin{aligned}\nabla \hat{L}(\boldsymbol{w}^{\mathcal{A}_h}) &= \frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{k\in\{0\}\cup\mathcal{A}_h}(\boldsymbol{X}^{(k)})^T[-\boldsymbol{y}^{(k)}+\boldsymbol{\psi}'(\boldsymbol{X}^{(k)}\boldsymbol{w}^{\mathcal{A}_h})] \\ &= \frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{k\in\{0\}\cup\mathcal{A}_h}(\boldsymbol{X}^{(k)})^T[-\boldsymbol{y}^{(k)}+\boldsymbol{\psi}'(\boldsymbol{X}^{(k)}\boldsymbol{\beta})] \\ &\quad + \frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{k\in\{0\}\cup\mathcal{A}_h}(\boldsymbol{X}^{(k)})^T[-\boldsymbol{\psi}'(\boldsymbol{X}^{(k)}\boldsymbol{\beta})+\boldsymbol{\psi}'(\boldsymbol{X}^{(k)}\boldsymbol{w}^{\mathcal{A}_h})]. \quad \text{(S.2.21)}\end{aligned}$$

Following a similar idea in the proof of Lemma 6 in Negahban et al. (2009), under Assumptions 1-3 and the fact $n_{\mathcal{A}_h} \geq Cs\log p$, we can show that

$$\frac{1}{n_{\mathcal{A}_h}+n_0}\left\|\sum_{k\in\{0\}\cup\mathcal{A}_h}(\boldsymbol{X}^{(k)})^T[-\boldsymbol{y}^{(k)}+\boldsymbol{\psi}'(\boldsymbol{X}^{(k)}\boldsymbol{\beta})]\right\|_\infty \lesssim \sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}},$$

with probability at least $1-(n_{\mathcal{A}_h}+n_0)^{-1}$.

The remaining step is to bound the infinity norm of the second term in (S.2.21). Denote $V_{ij}^{(k)} = x_{ij}^{(k)}[-\psi'((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\beta})+\psi'((\boldsymbol{x}_i^{(k)})^T\boldsymbol{w}^{\mathcal{A}_h})]$. Under Assumption 3, by mean value theorem and Lemma 1, there exists $v_i^{(k)} \in (0,1)$ such that

$$\begin{aligned}&\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{k\in\{0\}\cup\mathcal{A}_h}\sum_{i=1}^{n_k}V_{ij}^{(k)} \\ &= \frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{k\in\{0\}\cup\mathcal{A}_h}\sum_{i=1}^{n_k}\psi''((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\beta}+v_i^{(k)}(\boldsymbol{x}_i^{(k)})^T(\boldsymbol{w}^{\mathcal{A}_h}-\boldsymbol{\beta}))x_{ij}^{(k)}(\boldsymbol{x}_i^{(k)})^T(\boldsymbol{w}^{\mathcal{A}_h}-\boldsymbol{\beta}).\end{aligned}$$

$\psi''((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\beta}+v_i^{(k)}(\boldsymbol{x}_i^{(k)})^T(\boldsymbol{w}^{\mathcal{A}_h}-\boldsymbol{\beta}))x_{ij}^{(k)}$ is $M_\psi^2\kappa_u^2$-subGaussian due to the almost sure boundedness of $\psi''$ in Assumption 3. And $(\boldsymbol{x}_i^{(k)})^T(\boldsymbol{w}^{\mathcal{A}_h}-\boldsymbol{\beta})$ is a $4C_1^2h^2$-subGaussian due to Lemma 1. Then the multiplication is a $4C_1^2M_\psi^2\kappa_u^2h^2$-subexponential variable. By definition

of $\boldsymbol{w}^{\mathcal{A}_h}$, $\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{k\in\{0\}\cup\mathcal{A}_h}\sum_{i=1}^{n_k}V_{ij}^{(k)}$ has zero mean. Notice that the infinity norm of the second term in (S.2.21) equals to $\frac{1}{n_{\mathcal{A}_h}+n_0}\sup_{j=1,\ldots,p}|\sum_{k\in\{0\}\cup\mathcal{A}_h}\sum_{i=1}^{n_k}V_{ij}^{(k)}|$, by tail bounds of subexponential variables and union bounds, we have

$$\frac{1}{n_{\mathcal{A}_h}+n_0}\sup_{j=1,\ldots,p}\left|\sum_{k\in\{0\}\cup\mathcal{A}_h}\sum_{i=1}^{n_k}V_{ij}^{(k)}\right|\lesssim C_1M_\psi\kappa_u h\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}},$$

with probability at least $1-(n_{\mathcal{A}_h}+n_0)^{-1}$. Therefore $\|\nabla\hat{L}(\boldsymbol{w}^{\mathcal{A}_h})\|_\infty\lesssim\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}$ holds with probability at least $1-(n_{\mathcal{A}_h}+n_0)^{-1}$. Plugging the rate into (S.2.18), we have

$$\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_2\lesssim h\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}+\sqrt{\frac{s\log p}{n_{\mathcal{A}_h}+n_0}}+\left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/4}\sqrt{h},\tag{S.2.22}$$

with probability at least $1-(n_{\mathcal{A}_h}+n_0)^{-1}$, when $\lambda_{\boldsymbol{w}}\asymp C_{\boldsymbol{w}}\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}$ with $C_{\boldsymbol{w}}>0$ sufficiently large. Since $\hat{\boldsymbol{u}}^{\mathcal{A}_h}\in\mathbb{C}$, (S.2.22) encloses

$$\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1\lesssim s\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}+\left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/4}\sqrt{sh}+h\left(1+\sqrt{\frac{s\log p}{n_{\mathcal{A}_h}+n_0}}\right),\tag{S.2.23}$$

with probability at least $1-(n_{\mathcal{A}_h}+n_0)^{-1}$.

**Debiasing step:** Denote $\mathcal{D}^{(0)}=(\boldsymbol{X}^{(0)},\boldsymbol{y}^{(0)})$, $\hat{L}^{(0)}(\boldsymbol{w},\mathcal{D}^{(0)})=-\frac{1}{n_0}(\boldsymbol{y}^{(0)})^T\boldsymbol{X}^{(0)}\boldsymbol{w}$ $+\frac{1}{n_0}\sum_{i=1}^{n_0}\psi((\boldsymbol{x}_i^{(0)})^T\boldsymbol{w})$, $\nabla\hat{L}^{(0)}(\boldsymbol{w},\mathcal{D}^{(0)})=-\frac{1}{n_0}(\boldsymbol{X}^{(0)})^T\boldsymbol{y}^{(0)}+\frac{1}{n_0}(\boldsymbol{X}^{(0)})^T\boldsymbol{\psi}'(\boldsymbol{X}^{(0)}\boldsymbol{w})$, $\boldsymbol{\delta}^{\mathcal{A}_h}=\boldsymbol{\beta}-\boldsymbol{w}^{\mathcal{A}_h}$, $\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{w}}^{\mathcal{A}_h}+\hat{\boldsymbol{\delta}}^{\mathcal{A}_h}$, $\hat{\boldsymbol{v}}^{\mathcal{A}_h}=\hat{\boldsymbol{\delta}}^{\mathcal{A}_h}-\boldsymbol{\delta}^{\mathcal{A}_h}$, and $\delta\hat{L}^{(0)}(\boldsymbol{\delta},\mathcal{D})=\hat{L}^{(0)}(\hat{\boldsymbol{w}}^{\mathcal{A}_h}+\boldsymbol{\delta},\mathcal{D}^{(0)})-\hat{L}^{(0)}(\hat{\boldsymbol{w}}^{\mathcal{A}_h}+\boldsymbol{\delta}^{\mathcal{A}_h},\mathcal{D}^{(0)})-\nabla\hat{L}^{(0)}(\hat{\boldsymbol{w}}^{\mathcal{A}_h}+\boldsymbol{\delta}^{\mathcal{A}_h},\mathcal{D}^{(0)})^T\hat{\boldsymbol{v}}^{\mathcal{A}_h}$.

Similar to (S.2.19), when $\lambda_{\boldsymbol{\delta}}\geq2\|\nabla\hat{L}^{(0)}(\boldsymbol{\beta},\mathcal{D}^{(0)})\|_\infty$, we have

$$\begin{aligned}\delta\hat{L}^{(0)}(\hat{\boldsymbol{\delta}}^{\mathcal{A}_h},\mathcal{D})&\leq\lambda_{\boldsymbol{\delta}}(\|\boldsymbol{\delta}^{\mathcal{A}_h}\|_1-\|\hat{\boldsymbol{\delta}}^{\mathcal{A}_h}\|_1)-\nabla\hat{L}^{(0)}(\hat{\boldsymbol{w}}^{\mathcal{A}_h}+\boldsymbol{\delta}^{\mathcal{A}_h},\mathcal{D}^{(0)})^T\hat{\boldsymbol{v}}^{\mathcal{A}_h}\\&\leq\lambda_{\boldsymbol{\delta}}(2\|\boldsymbol{\delta}^{\mathcal{A}_h}\|_1-\|\hat{\boldsymbol{v}}^{\mathcal{A}_h}\|_1)+\|\nabla\hat{L}^{(0)}(\boldsymbol{\beta},\mathcal{D}^{(0)})\|_\infty\|\hat{\boldsymbol{v}}^{\mathcal{A}_h}\|_1\\&\quad-\left[\nabla\hat{L}^{(0)}(\hat{\boldsymbol{w}}^{\mathcal{A}_h}+\boldsymbol{\delta}^{\mathcal{A}_h},\mathcal{D}^{(0)})-\nabla\hat{L}^{(0)}(\boldsymbol{\beta},\mathcal{D}^{(0)})\right]^T\hat{\boldsymbol{v}}^{\mathcal{A}_h}\end{aligned}$$

$$
\begin{aligned}
&\leq 2\lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^{\mathcal{A}_h}\|_1 - \frac{1}{2}\lambda_{\boldsymbol{\delta}}\|\hat{\boldsymbol{v}}^{\mathcal{A}_h}\|_1 \\
&\quad - \frac{1}{n_0}\left[\boldsymbol{\psi}'((\boldsymbol{X}^{(0)})^T(\hat{\boldsymbol{w}}^{\mathcal{A}_h}+\boldsymbol{\delta}^{\mathcal{A}_h})) - \boldsymbol{\psi}'((\boldsymbol{X}^{(0)})^T\boldsymbol{\beta})\right]^T\hat{\boldsymbol{v}}^{\mathcal{A}_h} \\
&\leq 2\lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^{\mathcal{A}_h}\|_1 - \frac{1}{2}\lambda_{\boldsymbol{\delta}}\|\hat{\boldsymbol{v}}^{\mathcal{A}_h}\|_1 + \frac{1}{4c_0}M_\psi^2\cdot\frac{1}{n_0}\|\boldsymbol{X}^{(0)}\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_2^2 + c_0\cdot\frac{1}{n_0}\|\boldsymbol{X}^{(0)}\hat{\boldsymbol{v}}^{\mathcal{A}_h}\|_2^2,
\end{aligned}
\tag{S.2.24}
$$

where $c_0 > 0$ is an enough small constant. The last inequality holds because

$$
\begin{aligned}
&- \frac{1}{n_0}\left[\boldsymbol{\psi}'((\boldsymbol{X}^{(0)})^T(\hat{\boldsymbol{w}}^{\mathcal{A}_h}+\boldsymbol{\delta}^{\mathcal{A}_h})) - \boldsymbol{\psi}'((\boldsymbol{X}^{(0)})^T\boldsymbol{\beta})\right]^T\hat{\boldsymbol{v}}^{\mathcal{A}_h} \\
&\quad = \frac{1}{n_0}(\hat{\boldsymbol{u}}^{\mathcal{A}_h})^T(\boldsymbol{X}^{(0)})^T\Lambda^{(0)}\boldsymbol{X}^{(0)}\hat{\boldsymbol{v}}^{\mathcal{A}_h} \\
&\leq \frac{1}{4c_0}M_\psi^2\cdot\frac{1}{n_0}\|\boldsymbol{X}^{(0)}\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_2^2 + c_0\cdot\frac{1}{n_0}\|\boldsymbol{X}^{(0)}\hat{\boldsymbol{v}}^{\mathcal{A}_h}\|_2^2,
\end{aligned}
$$

where $\Lambda^{(0)} = \mathrm{diag}(\{\psi''((\boldsymbol{x}_i^{(0)})^T\boldsymbol{\beta} + t_i(\boldsymbol{x}_i^{(0)})^T\hat{\boldsymbol{u}}^{\mathcal{A}_h})\}_{i=1}^{n_0})$ is a $n_0\times n_0$ diagonal matrix and by Assumption 3, $\|\Lambda^{(0)}\|_{\max}\leq M_\psi$. Denote $\tilde{\boldsymbol{v}}^{\mathcal{A}_h} = t\hat{\boldsymbol{v}}^{\mathcal{A}_h}$ and $F^{(0)}(\boldsymbol{v}) = \hat{L}^{(0)}(\hat{\boldsymbol{w}}^{\mathcal{A}_h}+\boldsymbol{\delta}^{\mathcal{A}_h}+\boldsymbol{v},\mathcal{D}^{(0)}) - \hat{L}^{(0)}(\hat{\boldsymbol{w}}^{\mathcal{A}_h}+\boldsymbol{\delta}^{\mathcal{A}_h},\mathcal{D}^{(0)}) + \lambda_{\boldsymbol{\delta}}(\|\boldsymbol{\delta}^{\mathcal{A}_h}+\boldsymbol{v}\|_1 - \|\boldsymbol{\delta}^{\mathcal{A}_h}\|_1)$. Since $F(\mathbf{0}) = 0$ and $F^{(0)}(\hat{\boldsymbol{v}}^{\mathcal{A}_h})\leq 0$, by convexity, for any $t\in(0,1]$,

$$
F^{(0)}(\tilde{\boldsymbol{v}}^{\mathcal{A}_h}) = F^{(0)}(t\hat{\boldsymbol{v}}^{\mathcal{A}_h} + (1-t)\mathbf{0}) \leq tF^{(0)}(\hat{\boldsymbol{u}}^{\mathcal{A}_h}) \leq 0. \tag{S.2.25}
$$

We set $t\in(0,1]$ such that $\|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_2\leq 1$, which allows us to apply Proposition 1 in Loh and Wainwright (2015) on $\tilde{\boldsymbol{v}}^{\mathcal{A}_h}$. By basic inequality (S.2.25) and the same arguments in (S.2.24),

$$
\begin{aligned}
c_1\|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_2^2 - c_2\cdot\frac{\log p}{n_0}\|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_1^2 &\leq F^{(0)}(\tilde{\boldsymbol{v}}^{\mathcal{A}_h}) - \nabla\hat{L}^{(0)}(\hat{\boldsymbol{w}}^{\mathcal{A}_h}+\boldsymbol{\delta}^{\mathcal{A}_h},\mathcal{D}^{(0)})^T\tilde{\boldsymbol{v}}^{\mathcal{A}_h} \\
&\leq 2\lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^{\mathcal{A}_h}\|_1 - \frac{1}{2}\lambda_{\boldsymbol{\delta}}\|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_1 + \frac{1}{4c_0}M_\psi^2\cdot\frac{1}{n_0}\|\boldsymbol{X}^{(0)}\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_2^2 \\
&\quad + c_0\cdot\frac{1}{n_0}\|\boldsymbol{X}^{(0)}\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_2^2.
\end{aligned}
\tag{S.2.26}
$$

In the discussion of transferring step, we showed that $\|\hat{\boldsymbol{u}}_{S^c}^{\mathcal{A}_h}\|_1 \leq 3\|\hat{\boldsymbol{u}}_S^{\mathcal{A}_h}\|_1 + 4C_1 h$. We leverage on this fact to bound $\frac{1}{n_0}\|\boldsymbol{X}^{(0)}\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_2^2$ by $\|\boldsymbol{X}^{(0)}\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_2^2$.

If $3\|\hat{\boldsymbol{u}}_S^{\mathcal{A}_h}\|_1 \geq 4C_1 h$, we will have $\|\hat{\boldsymbol{u}}_{S^c}^{\mathcal{A}_h}\|_1 \leq 6\|\hat{\boldsymbol{u}}_S^{\mathcal{A}_h}\|_1$. Then by Theorem 1.6 of Zhou (2009),

$$\frac{1}{n_0}\|\boldsymbol{X}^{(0)}\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_2^2 \lesssim \|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_2^2 \lesssim \frac{s\log p}{n_{\mathcal{A}}+n_0} + h\cdot\sqrt{\frac{\log p}{n_{\mathcal{A}}+n_0}}, \tag{S.2.27}$$

with probability at least $1 - C(n_{\mathcal{A}}+n_0)^{-1} - C\exp\{-n_0\}$.

If $3\|\hat{\boldsymbol{u}}_S^{\mathcal{A}_h}\|_1 < 4C_1 h$, we will have $\|\hat{\boldsymbol{u}}_{S^c}^{\mathcal{A}_h}\|_1 \leq 8C_1 h \leq \sqrt{s}$. Also $\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_2 \leq 1$ with probability $1 - C(n_{\mathcal{A}}+n_0)^{-1}$. Denote

$$\Pi_0(s) = \{\boldsymbol{u}\in\mathbb{R}^p : \|\boldsymbol{u}\|_2 \leq 1, \|\boldsymbol{u}\|_0 \leq s\},$$
$$\Pi_1(s) = \{\boldsymbol{u}\in\mathbb{R}^p : \|\boldsymbol{u}\|_2 \leq 1, \|\boldsymbol{u}\|_1 \leq \sqrt{s}\}.$$

By Lemma 3.1 of Plan and Vershynin (2013), $\Pi_1(s) \subseteq 2\overline{\text{conv}}(\Pi_0(s))$, where $\overline{\text{conv}}(\Pi_0(s))$ is the closure of convex hull of $\Pi_0(s)$. Then following a similar argument in the proof of Theorem 2.4 in Mendelson et al. (2008) leads to (S.2.27) as well.

Next we want to bound $\frac{1}{n_0}\|\boldsymbol{X}^{(0)}\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_2^2$ by $\|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_2^2$. By another basic inequality,

$$\begin{aligned}
0 &\leq \hat{L}^{(0)}(\hat{\boldsymbol{\beta}}, \mathcal{D}^{(0)}) - \hat{L}^{(0)}(\boldsymbol{\beta}, \mathcal{D}^{(0)}) - \nabla\hat{L}^{(0)}(\boldsymbol{\beta}, \mathcal{D}^{(0)})^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&\leq \lambda_{\boldsymbol{\delta}}(\|\boldsymbol{\beta} - \hat{\boldsymbol{w}}^{\mathcal{A}_h}\|_1 - \|\hat{\boldsymbol{\delta}}^{\mathcal{A}_h}\|_1) + \|\nabla\hat{L}^{(0)}(\boldsymbol{\beta}, \mathcal{D}^{(0)})\|_\infty\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \\
&\leq \lambda_{\boldsymbol{\delta}}(\|\boldsymbol{\beta} - \hat{\boldsymbol{w}}^{\mathcal{A}_h}\|_1 - \|\hat{\boldsymbol{\delta}}^{\mathcal{A}_h}\|_1) + \frac{1}{2}\lambda_{\boldsymbol{\delta}}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \\
&\leq \frac{3}{2}\lambda_{\boldsymbol{\delta}}\|\boldsymbol{\beta} - \hat{\boldsymbol{w}}^{\mathcal{A}_h}\|_1 - \frac{1}{2}\lambda_{\boldsymbol{\delta}}\|\hat{\boldsymbol{\delta}}^{\mathcal{A}_h}\|_1 \\
&\leq \frac{3}{2}\lambda_{\boldsymbol{\delta}}C_1 h + \frac{3}{2}\lambda_{\boldsymbol{\delta}}\|\hat{\boldsymbol{u}}^{\mathcal{A}}\|_1 - \frac{1}{2}\lambda_{\boldsymbol{\delta}}\|\hat{\boldsymbol{\delta}}^{\mathcal{A}_h}\|_1,
\end{aligned}$$

which implies

$$\|\hat{\boldsymbol{v}}^{\mathcal{A}_h}\|_1 \leq \|\hat{\boldsymbol{\delta}}^{\mathcal{A}_h}\|_1 + C_1 h \leq 3\|\hat{\boldsymbol{u}}^{\mathcal{A}}\|_1 + 4C_1 h.$$

Combined with (S.2.23), this leads to $\|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_1 \le \|\hat{\boldsymbol{v}}^{\mathcal{A}_h}\|_1 \le \sqrt{s}$ when $s\log p/(n_{\mathcal{A}_h}+n_0)$ and $h$ are small enough. Due to the strict convexity, $\delta\hat{L}^{(0)}(\hat{\boldsymbol{\delta}}^{\mathcal{A}_h},\mathcal{D}) > 0$, leading to $\|\hat{\boldsymbol{\delta}}^{\mathcal{A}_h}\|_1 \le 3\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1 + 3h$. Then,

$$\|\hat{\boldsymbol{v}}^{\mathcal{A}_h}\|_1 \le \|\boldsymbol{\beta} - \hat{\boldsymbol{w}}^{\mathcal{A}_h}\|_1 + \|\hat{\boldsymbol{\delta}}^{\mathcal{A}_h}\|_1 \le 4\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1 + 4h \le \sqrt{s}. \tag{S.2.28}$$

Similar to the analysis of the case $3\|\hat{\boldsymbol{u}}_S^{\mathcal{A}_h}\|_1 < 4C_1h$ above, we can get

$$c_0 \cdot \frac{1}{n_0}\|\boldsymbol{X}^{(0)}\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_2^2 \le c_0 \cdot C\|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_2^2,$$

with probability at least $1 - C\exp\{-n_0\}$. As long as $c_0C < c_1/2$, by (S.2.26), we have

$$\begin{aligned} c_1\|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_2^2 - c_2 \cdot \frac{\log p}{n_0}\|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_1^2 \le 2\lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^{\mathcal{A}_h}\|_1 - \frac{1}{2}\lambda_{\boldsymbol{\delta}}\|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_1 + C\frac{s\log p}{n_{\mathcal{A}_h}+n_0} + Ch\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}} \\ + c_1/2\|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_2^2 \end{aligned} \tag{S.2.29}$$

with probability at least $1 - C'n_0^{-1}$.

If $\lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^{\mathcal{A}_h}\|_1 \le C\frac{s\log p}{n_{\mathcal{A}_h}+n_0} + Ch\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}$:

$$\|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_1 \lesssim \left[\frac{s\log p}{n_{\mathcal{A}_h}+n_0} + h\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}\right] \cdot \sqrt{\frac{n_0}{\log p}} + \|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_2^2.$$

Since $\|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_2 \le 1$, by (S.2.29), it holds

$$\|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_2^2 \lesssim \frac{s\log p}{n_{\mathcal{A}_h}+n_0} + h\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}} \lesssim \frac{s\log p}{n_{\mathcal{A}_h}+n_0} + \left[h\sqrt{\frac{\log p}{n_0}}\right] \wedge h^2,$$

with probability at least $1 - C'n_0^{-1}$.

If $\lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^{\mathcal{A}_h}\|_1 > C\frac{s\log p}{n_{\mathcal{A}_h}+n_0} + Ch\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}$: $\|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_1 \lesssim h + \|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_2^2$, which leads to

$$\|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_2^2 \lesssim 2\lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^{\mathcal{A}_h}\|_1 - \frac{1}{2}\lambda_{\boldsymbol{\delta}}\|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_1,$$

implying $\|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_1 \leq 4\|\boldsymbol{\delta}^{\mathcal{A}_h}\|_1 \leq 4C_1 h$. Besides, by plugging this result into (S.2.29), we have

$$\|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_2^2 \lesssim \frac{s\log p}{n_{\mathcal{A}_h}+n_0} + \left[h\sqrt{\frac{\log p}{n_0}}\right] \wedge h^2, \tag{S.2.30}$$

with probability at least $1 - C'n_0^{-1}$.

When $s\log p/(n_{\mathcal{A}_h}+n_0)$ and $h$ is small enough, and because $h\sqrt{\frac{\log p}{n_0}} = o(1)$, the right hand side of (S.2.30) can be very small. This means that we showed $\|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_2 \leq c < 1$ with probability at least $1 - C'n_0^{-1}$. Note that this result holds for any $t \in (0,1]$ such that $\|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_2 \leq 1$. Now let's consider our interested vector $\hat{\boldsymbol{v}}^{\mathcal{A}_h}$. Suppose $\|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_2^2 \geq C\frac{s\log p}{n_{\mathcal{A}_h}+n_0} + C\left[h\sqrt{\frac{\log p}{n_0}}\right] \wedge h^2$ for some constant $C > 0$ with probability at least $C'n_0^{-1}$, then we can choose $t \in (0,1]$ such that $c < \|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_2 \leq 1$, which is contradicted with the fact $\|\tilde{\boldsymbol{v}}^{\mathcal{A}_h}\|_2 \leq c$ with probability at least $1 - C'n_0^{-1}$. Therefore

$$\|\hat{\boldsymbol{v}}^{\mathcal{A}_h}\|_2^2 \lesssim \frac{s\log p}{n_{\mathcal{A}_h}+n_0} + \left[h\sqrt{\frac{\log p}{n_0}}\right] \wedge h^2 \lesssim 1,$$

with probability at least $1 - C'n_0^{-1}$. Then we can go over the analysis procedure of $\tilde{\boldsymbol{v}}^{\mathcal{A}_h}$ again with $\hat{\boldsymbol{v}}^{\mathcal{A}_h}$ (on the high-probability event $\|\hat{\boldsymbol{v}}^{\mathcal{A}_h}\|_2 \leq 1$) to get the $\ell_1$-bound

$$\|\hat{\boldsymbol{v}}^{\mathcal{A}_h}\|_1 \lesssim s\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}} + h,$$

with probability at least $1 - C'n_0^{-1}$.

Finally, we connect the conclusions above with the upper bounds on $\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_2$ and $\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1$, which completes the proof.

#### S.2.3.2 Proof of Theorem 2

The analysis in the proof of Theorem B and Theorem 2 in Li et al. (2021), which proved the $\ell_2$ minimax rate under linear models, can be directly extended to the GLMs. We only

point out some key observations. The readers can refer to the supplements of their paper for the full proof.

The proofs of Theorem B and Theorem 2 in Li et al. (2021) leverage on Fano's method, which involves Kullback-Leibler divergence. For $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ in $\mathbb{R}^p$, consider two GLMs where $y|\boldsymbol{x} \sim \mathbb{P}(y|\boldsymbol{x}; \boldsymbol{w}_1) = \rho(y)\exp\{y\boldsymbol{x}^T\boldsymbol{w}_1 - \psi(\boldsymbol{x}^T\boldsymbol{w}_1)\}$ and $y|\boldsymbol{x} \sim \mathbb{P}(y|\boldsymbol{x}; \boldsymbol{w}_2) = \rho(y)\exp\{y\boldsymbol{x}^T\boldsymbol{w}_2 - \psi(\boldsymbol{x}^T\boldsymbol{w}_2)\}$. Denote the corresponding joint distribution of $(\boldsymbol{x}, y)$ as $f_{\boldsymbol{w}_1}$ and $f_{\boldsymbol{w}_2}$ and suppose marginal distributions of $\boldsymbol{x}$ are the same. Then by definition and Assumption 3,

$$\begin{aligned}\mathrm{KL}(f_{\boldsymbol{w}_1}||f_{\boldsymbol{w}_2}) &= \mathbb{E}_{\boldsymbol{x}\sim f_{\boldsymbol{w}_1}}\left[\psi''(\boldsymbol{x}^T(t\boldsymbol{w}_1 + (1-t)\boldsymbol{w}_2)) \cdot [\boldsymbol{x}^T(\boldsymbol{w}_1 - \boldsymbol{w}_2)]^2\right] \\ &\leq C\mathbb{E}_{\boldsymbol{x}\sim f_{\boldsymbol{w}_1}}[\boldsymbol{x}^T(\boldsymbol{w}_1 - \boldsymbol{w}_2)]^2],\end{aligned}$$

for some constant $C > 0$, if $\|\psi''\|_\infty < \infty$ or $\|\boldsymbol{x}\|_\infty \leq U$ a.s. and $\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_1 \leq C'$ with some $C' > 0$. By using this fact, all the analysis of $\ell_2$-estimation error in their proof works out for GLM.

About the $\ell_1$-estimation error, we can make slight changes to make the argument work again. In case (i) of the proof of Theorem 2 (or case (i) of Theorem B) in Li et al. (2021), replace the $\delta$-packing under $\ell_2$ norm with the $\delta$-packing under $\ell_1$ norm, and set $\delta_0 = c_0 s\sqrt{\frac{\log p}{n_{\mathcal{A}}+n_0}}$. In case (ii-1), we can do the same and set $\delta_0 = c_0 s\sqrt{\frac{\log p}{n_0}}$. In (ii-2), we suppose $s\sqrt{\log p/n_0} > h$. Set $\delta_0 = \bar{m}\sqrt{\frac{\log p}{n_0}} \asymp h$ where $\bar{m}$ is the integer part of $h\sqrt{n_0/\log p}$. In (ii-3), the same argument works and we set $\delta_0 = h$. So finally we can get the minimax $\ell_1$ rate $s\sqrt{\log p/(n_{\mathcal{A}} + n_0)} + [s\sqrt{\log p/n_0}] \wedge h$.

#### S.2.3.3 Proof of Theorem 3

Throughout this proof, we denote $\boldsymbol{w}^{\mathcal{A}_h}$ as any vector $\boldsymbol{w}$ satisfying $\|\boldsymbol{w} - \boldsymbol{\beta}\|_1 \leq h$. Such a $\boldsymbol{w}^{\mathcal{A}_h}$ indeed exists, e.g. $\boldsymbol{w}^{\mathcal{A}_h} = \sum_{k\in\{0\}\cup\mathcal{A}_h} \alpha_k \boldsymbol{w}^{(k)}$. Note that $\boldsymbol{w}^{\mathcal{A}_h}$ here does not necessarily

enjoy the moment condition (2), which will bring more bias. This is the price we have to pay for relaxing Assumption 4. Other notations are defined the same as in the proof of Theorem 1.

The main idea of the proof is similar to that in proof of Theorem 1. We only highlight the different parts here and do not dig into all the details.

First, the claim in (S.2.18) still holds here, i.e. when $\lambda_{\boldsymbol{w}} \geq 2\|\nabla \hat{L}(\boldsymbol{w}^{\mathcal{A}_h}, \mathcal{D})\|_\infty$, with probability at least $1 - C_3 \exp\{-C_4(n_{\mathcal{A}_h} + n_0)\}$, it holds that

$$\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_2 \leq 8\kappa_2 C_{\boldsymbol{w}} h \sqrt{\frac{\log p}{n_{\mathcal{A}_h} + n_0}} + 3\frac{\sqrt{s}}{\kappa_1}\lambda_{\boldsymbol{w}} + 2\sqrt{\frac{1}{\kappa_1} h \lambda_{\boldsymbol{w}}}. \tag{S.2.31}$$

Via the decomposition in (S.2.21), $\|\nabla L(\boldsymbol{w}^{\mathcal{A}_h}, \mathcal{D})\|_\infty$ can be bounded by two parts where the first part has rate $\mathcal{O}_p\left(\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}\right)$. Denote $V_{ij}^{(k)} = x_{ij}^{(k)}[-\psi'((\boldsymbol{x}_i^{(k)})^T \boldsymbol{w}^{(k)}) + \psi'((\boldsymbol{x}_i^{(k)})^T \boldsymbol{w}^{\mathcal{A}_h})]$.

$$\begin{aligned}
&\frac{1}{n_{\mathcal{A}_h} + n_0} \sum_{k \in \{0\} \cup \mathcal{A}_h} \sum_{i=1}^{n_k} V_{ij}^{(k)} \\
&= \frac{1}{n_{\mathcal{A}_h} + n_0} \sum_{k \in \{0\} \cup \mathcal{A}_h} \sum_{i=1}^{n_k} \psi''((\boldsymbol{x}_i^{(k)})^T \boldsymbol{w}^{(k)} + v_i^{(k)} (\boldsymbol{x}_i^{(k)})^T (\boldsymbol{w}^{\mathcal{A}_h} - \boldsymbol{w}^{(k)})) x_{ij}^{(k)} (\boldsymbol{x}_i^{(k)})^T (\boldsymbol{w}^{\mathcal{A}_h} - \boldsymbol{w}^{(k)}).
\end{aligned}$$

Similar as before, the multiplication $\psi''((\boldsymbol{x}_i^{(k)})^T \boldsymbol{w}^{(k)} + v_i^{(k)} (\boldsymbol{x}_i^{(k)})^T (\boldsymbol{w}^{\mathcal{A}_h} - \boldsymbol{w}^{(k)})) x_{ij}^{(k)} (\boldsymbol{x}_i^{(k)})^T (\boldsymbol{w}^{\mathcal{A}_h} - \boldsymbol{w}^{(k)})$ is a $C_1^2 M_\psi^2 \kappa_u^2 h^2$-subexponential variable. And by Cauchy-Schwarz inequality and sub-Gaussian properties (Vershynin, 2018),

$$\mathbb{E}|V_{ij}^{(k)}| \leq (\mathbb{E}|x_{ij}^{(k)}|^2)^{1/2} (\mathbb{E}[(\boldsymbol{x}_i^{(k)})^T (\boldsymbol{w}^{\mathcal{A}_h} - \boldsymbol{w}^{(k)})]^2)^{1/2} \lesssim \kappa_u h.$$

Therefore, by tail bounds of sub-exponential variables and union bounds, we have

$$\frac{1}{n_{\mathcal{A}_h} + n_0} \sup_{j=1,\ldots,p} \left| \sum_{k \in \{0\} \cup \mathcal{A}_h} \sum_{i=1}^{n_k} V_{ij}^{(k)} \right| \lesssim \kappa_u h + C_1 M_\psi \kappa_u h \sqrt{\frac{\log p}{n_{\mathcal{A}_h} + n_0}},$$

with probability at least $1-(n_{\mathcal{A}_h}+n_0)^{-1}\vee p^{-1}$, which implies that $\|\nabla\hat{L}(\boldsymbol{w}^{\mathcal{A}_h})\|_\infty \lesssim \sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}+h$ with probability at least $1-(n_{\mathcal{A}_h}+n_0)^{-1}\vee p^{-1}$. Let $\lambda_{\boldsymbol{w}}=C_{\boldsymbol{w}}\left(\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}+h\right)$ and $\lambda_{\boldsymbol{w}}\geq 2\|\nabla L(\boldsymbol{w}^{\mathcal{A}_h},\mathcal{D})\|_\infty$ in high probability. Plugging it into (S.2.31), we get

$$\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_2 \lesssim \sqrt{\frac{s\log p}{n_{\mathcal{A}_h}+n_0}}+\sqrt{s}h+\sqrt{h}\left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/4}, \tag{S.2.32}$$

with probability at least $1-(n_{\mathcal{A}_h}+n_0)^{-1}\vee p^{-1}$. Similarly, we can obtain the $\ell_1$-error bound

$$\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1 \lesssim s\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}+sh+\sqrt{sh}\left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/4}, \tag{S.2.33}$$

with probability at least $1-(n_{\mathcal{A}_h}+n_0)^{-1}\vee p^{-1}$. Next, consider the debiasing step. Similar as the proof of Theorem 1, let $\lambda_{\boldsymbol{\delta}}=C_{\boldsymbol{\delta}}\sqrt{\frac{\log p}{n_0}}$ which satisfies $\lambda_{\boldsymbol{\delta}}\geq 2\|\nabla L^{(0)}(\boldsymbol{\beta},\mathcal{D}^{(0)})\|_\infty$ in high probability to get

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2^2 \lesssim \frac{s\log p}{n_{\mathcal{A}_h}+n_0}+sh^2+\left[h\sqrt{\frac{\log p}{n_0}}\right]\wedge h^2,$$
$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 \lesssim s\sqrt{\frac{\log p}{n_{\mathcal{A}}+n_0}}+sh+\sqrt{sh}\cdot\left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/4},$$

with probability at least $1-Cn_0^{-1}$.

#### S.2.3.4 Proof of Theorem 4

**Lemma 6.** *Under Assumptions 1-3,* $\sup_{k\in\mathcal{A}} L_0(\boldsymbol{\beta}^{(k)})-L_0(\boldsymbol{\beta})\lesssim\|\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta}\|_2^2\lesssim h^2$.

*Proof of Lemma 6.* Note that the term involving $\mathbb{E}[\rho(y^{(0)})]$ is canceled when taking the difference, therefore we drop that term and consider $L_0(\boldsymbol{w})=-\mathbb{E}[\psi'(\boldsymbol{w}^T\boldsymbol{x}^{(0)})\boldsymbol{w}^T\boldsymbol{x}^{(0)}]+\mathbb{E}[\psi(\boldsymbol{w}^T\boldsymbol{x}^{(0)})]$. Since $\nabla L_0(\boldsymbol{\beta})=\mathbf{0}$ and $\nabla^2 L_0(\boldsymbol{w})=\mathbb{E}[\psi''(\boldsymbol{w}^T\boldsymbol{x}^{(0)})\boldsymbol{x}^{(0)}(\boldsymbol{x}^{(0)})^T]$. By mean-theorem, $\exists t^{(k)}\in(0,1)$, such that

$$L_0(\boldsymbol{\beta}^{(k)})-L_0(\boldsymbol{\beta})=(\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta})^T\mathbb{E}[\psi''(\boldsymbol{\beta}^T\boldsymbol{x}^{(0)}+t^{(k)}(\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta})^T\boldsymbol{x}^{(0)})\boldsymbol{x}^{(0)}(\boldsymbol{x}^{(0)})^T](\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta}).$$

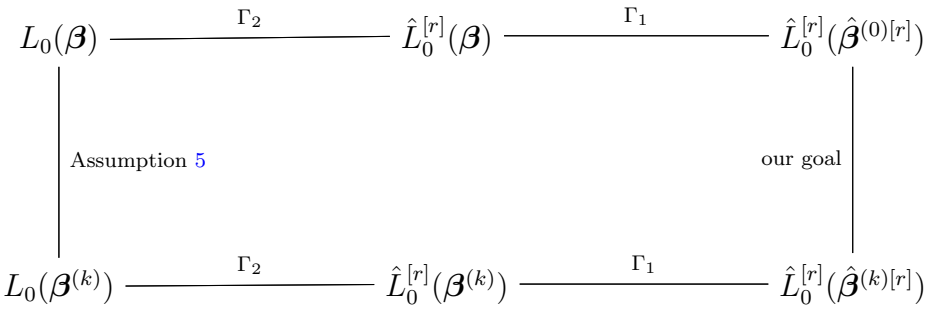


Figure 12: The idea behind Assumptions 4, 5 and Theorem 4.

Under Assumption 3.(i):

$$L_0(\boldsymbol{\beta}^{(k)}) - L_0(\boldsymbol{\beta}) \leq M\mathbb{E}[((\boldsymbol{x}^{(0)})^T(\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}))^2] \lesssim \|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}\|_2^2.$$

Under Assumption 3.(ii), by Cauchy-Schwarz inequality and the subGaussian moment bound (Vershynin, 2018):

$$L_0(\boldsymbol{\beta}^{(k)}) - L_0(\boldsymbol{\beta}) \leq \left[\mathbb{E}\left(\max_{z:|z|\leq 1} \psi''((\boldsymbol{x}^{(0)})^T\boldsymbol{\beta} + z)\right)^2\right]^{1/2} [\mathbb{E}((\boldsymbol{x}^{(0)})^T(\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}))^4]^{1/4} \lesssim \|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}\|_2^2.$$

The second half inequality automatically holds since $\boldsymbol{\beta}^{(k)}$ is a linear combination of $\boldsymbol{\beta}$ and $\boldsymbol{w}^{(k)}$. And it's easy to see that all the constants appearing in the inequalities are uniform for all $k \in \mathcal{A}$, which completes the proof. □

Next, we prove Theorem 4. We have

$$\begin{aligned}\hat{\sigma} &= \sqrt{\sum_{r=1}^{3} (\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(0)[r]}) - L_0(\boldsymbol{\beta}))^2/3} \leq \sqrt{\frac{2}{3}} \cdot \sum_{r=1}^{3} |\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(0)[r]}) - L_0(\boldsymbol{\beta})| \\ &\leq \sqrt{\frac{2}{3}} \cdot \sum_{r=1}^{3} \left[|\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(0)[r]}) - \hat{L}_0^{[r]}(\boldsymbol{\beta})| + |\hat{L}_0^{[r]}(\boldsymbol{\beta}) - L_0(\boldsymbol{\beta})|\right]\end{aligned}$$

$$\lesssim \zeta \left(\Gamma_1^{(0)} + \Gamma_2^{(0)}\right),$$

with probability at least $1 - g_1^{(0)}(\zeta) - g_2^{(0)}(\zeta)$. As Figure 12 shows, by Lemma 6, for $k \in \mathcal{A}$, there holds

$$\begin{aligned}
\sup_r |\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]}) - \hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(0)[r]})| &\le 2 \sup_r |\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]}) - \hat{L}_0^{[r]}(\boldsymbol{\beta}^{(k)[r]})| \\
&\quad + \sup_r |\hat{L}_0^{[r]}(\boldsymbol{\beta}^{(k)}) - \hat{L}_0^{[r]}(\boldsymbol{\beta}) - L_0(\boldsymbol{\beta}^{(k)}) + L_0(\boldsymbol{\beta})| \\
&\quad + |L_0(\boldsymbol{\beta}^{(k)}) - L_0(\boldsymbol{\beta})| \\
&\lesssim \zeta \left(\Gamma_1^{(k)} + \Gamma_2^{(k)} + h^2\right) \\
&\le C_0(\hat{\sigma} \vee 0.01), \qquad \text{(S.2.34)}
\end{aligned}$$

simultaneously with probability at least $1 - |\mathcal{A}| \max_{k\in\mathcal{A}}[g_1^{(k)}(\zeta) + g_2^{(k)}(\zeta)]$ for sufficiently small $\zeta > 0$ when $\min_{k\in\mathcal{A}} n_k$ and $n_0$ go to infinity since $\Gamma_1^{(k)} + \Gamma_2^{(k)} + h^2 = o(1)$. On the other hand, by Assumption 5 and the fact $\nabla L_0(\boldsymbol{\beta}) = \mathbf{0}$, for $k \in \mathcal{A}^c$,

$$\begin{aligned}
&\inf_r \hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]}) - \hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(0)[r]}) \\
&\ge |L_0(\boldsymbol{\beta}^{(k)}) - L_0(\boldsymbol{\beta})| - \Upsilon_1^{(k)} - \zeta\Gamma_1^{(k)} - \zeta\Gamma_2^{(k)} \\
&= \|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}\|_2^2 \cdot \lambda_{\min}\left(\int_0^1 \nabla^2 L_0(\boldsymbol{\beta} + t(\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}))dt\right) - \Upsilon_1^{(k)} - \zeta\Gamma_1^{(k)} - \zeta\Gamma_2^{(k)} \\
&\ge \|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}\|_2^2 \cdot \underline{\lambda} - \Upsilon_1^{(k)} - \zeta\Gamma_1^{(k)} - \zeta\Gamma_2^{(k)} \\
&\ge C_1^2 \left[(\Gamma_1^{(0)} + \Gamma_2^{(0)}) \vee 1\right] - \zeta\Gamma_1^{(k)} - \zeta\Gamma_2^{(k)} \\
&> C_0(\hat{\sigma} \vee 0.01),
\end{aligned}$$

simultaneously with probability at least $1 - |\mathcal{A}^c| \max_{k\in\mathcal{A}^c}[g_1^{(k)}(C_0^{-1}) + g_2^{(k)}(C_0^{-1})] - |\mathcal{A}^c| \max_{k\in\mathcal{A}^c}$

$[g_1^{(k)}(\zeta)+g_2^{(k)}(\zeta)]$. It entails

$$\mathbb{P}(\widehat{\mathcal{A}}\neq\mathcal{A}_h)\leq\mathbb{P}\left(\bigcup_{k\in\mathcal{A}}\left\{\hat{L}_0^{(k)}-\hat{L}_0^{(0)}>C_0(\hat{\sigma}\vee 0.01)\right\}\bigcup\bigcup_{k\in\mathcal{A}^c}\left\{\hat{L}_0^{(k)}-\hat{L}_0^{(0)}\leq C_0(\hat{\sigma}\vee 0.01)\right\}\right)$$

$$\begin{aligned}
&\leq\sum_{k\in\mathcal{A}}\mathbb{P}\left(\hat{L}_0^{(k)}-\hat{L}_0^{(0)}>C_0(\hat{\sigma}\vee 0.01)\right)+\sum_{k\in\mathcal{A}^c}\mathbb{P}\left(\hat{L}_0^{(k)}-\hat{L}_0^{(0)}\leq C_0(\hat{\sigma}\vee 0.01)\right)\\
&\leq\sum_{k\in\mathcal{A}}\mathbb{P}\left(\inf_r|\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]})-\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(0)[r]})|>C_0(\hat{\sigma}\vee 0.01)\right)\\
&\quad+\sum_{k\in\mathcal{A}^c}\mathbb{P}\left(\sup_r|\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]})-\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(0)[r]})|\leq C_0(\hat{\sigma}\vee 0.01)\right)\\
&\leq|\mathcal{A}|\max_{k\in\mathcal{A}}[g_1^{(k)}(\zeta)+g_2^{(k)}(\zeta)]+|\mathcal{A}^c|\max_{k\in\mathcal{A}^c}[g_1^{(k)}(C_0^{-1})+g_2^{(k)}(C_0^{-1})].
\end{aligned}$$

For any $\delta>0$, there exist constants $C'(\delta)$ and $\zeta'>0$ such that when $C_0=C'(\delta)$, $K\max_{k\in\mathcal{A}^c}[g_1^{(k)}(C_0^{-1})+g_2^{(k)}(C_0^{-1})]\leq\delta/2$, $K\max_{k\in\mathcal{A}}[g_1^{(k)}(\zeta')+g_2^{(k)}(\zeta')]<\delta/2$ and $C_1^2\left[(\Gamma_1^{(0)}+\Gamma_2^{(0)})\vee 1\right]>\zeta'\Gamma_1^{(k)}+\zeta'\Gamma_2^{(k)}$. On the other hand, there exists $N=N(\delta)>0$, such that when $\min_{k\in\{0\}\cup\mathcal{A}_h}n_k>N(\delta)$, $\zeta'\left(\Gamma_1^{(k)}+\Gamma_2^{(k)}+h^2\right)/[C'(\delta)\cdot(\hat{\sigma}\vee 0.01)]$ is sufficiently small to make (S.2.34) hold.

In summary, for any $\delta>0$, there exist constants $C'(\delta)$ and $N=N(\delta)>0$ such that when $C_0=C'(\delta)$ and $\min_{k\in\{0\}\cup\mathcal{A}_h}n_k>N(\delta)$, $\mathbb{P}(\widehat{\mathcal{A}}\neq\mathcal{A}_h)\leq\delta$.

#### S.2.3.5 Proof of Theorem 5

Denote $\Re_1=\sqrt{\frac{s\log p}{n_{\mathcal{A}}+n_0}}+\left[h^{1/2}\left(\frac{\log p}{n_0}\right)^{1/4}\right]\wedge h$. First we present the following proposition.

**Proposition 2.** *Assume Assumptions 1-4 and 7 hold (except (v)). Let* $\lambda_j\asymp\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}+\left[\frac{1}{\sqrt{s}}\cdot\left(\frac{\log p}{n_0}\right)^{1/4}h^{1/2}\right]\wedge\frac{h}{\sqrt{s}}$ *for any* $j=1,\ldots,p$, *and* $\tilde{\lambda}_j\asymp\sqrt{\frac{\log p}{n_0}}+\Re_1$. *Then with probability*

*at least* $1 - K_{\mathcal{A}_h} n_0^{-1}$,

$$\|\hat{\boldsymbol{\gamma}}_j^{(0)} - \boldsymbol{\gamma}_j^{(0)}\|_2^2 \lesssim h_1 \sqrt{\frac{\log p}{n_0}} + h_1 \Re_1 + \Re_1^2, \tag{S.2.35}$$

$$\|\hat{\boldsymbol{\gamma}}_j^{(0)} - \boldsymbol{\gamma}_j^{(0)}\|_1 \lesssim \sqrt{s}\Re_1 + s^{1/4}\Re_1^{1/2} h_1^{1/2} + h_1, \tag{S.2.36}$$

$$|\hat{\tau}_j^2 - \tau_j^2| \vee |\hat{\tau}_j^{-2} - \tau_j^{-2}| \lesssim \Re_1 + h_1^{1/2}\left(\frac{\log p}{n_0}\right)^{1/4} + h_1^{1/2}\Re_1^{1/2} + h_{\max}, \tag{S.2.37}$$

(S.2.35) and (S.2.36) can be proved by combining conclusions in Lemmas 4 and 5. The proof of (S.2.37) can be found in Section S.2.4.2.

Next we will leverage Proposition 2 to show (10) and (11). First, we notice that

$$\begin{aligned}
\hat{\boldsymbol{b}} &= \hat{\boldsymbol{\beta}} + \frac{1}{n_0}\widehat{\boldsymbol{\Theta}}(\boldsymbol{X}^{(0)})^T[\boldsymbol{Y}^{(0)} - \boldsymbol{\psi}'(\boldsymbol{X}^{(0)}\hat{\boldsymbol{\beta}})] \\
&= \hat{\boldsymbol{\beta}} + \frac{1}{n_0}\widehat{\boldsymbol{\Theta}}(\boldsymbol{X}^{(0)})^T[\boldsymbol{Y}^{(0)} - \boldsymbol{\psi}'(\boldsymbol{X}^{(0)}\boldsymbol{\beta})] + \frac{1}{n_0}\widehat{\boldsymbol{\Theta}}(\boldsymbol{X}^{(0)})^T[\boldsymbol{\psi}'(\boldsymbol{X}^{(0)}\hat{\boldsymbol{\beta}}) - \boldsymbol{\psi}'(\boldsymbol{X}^{(0)}\boldsymbol{\beta})] \\
&= \hat{\boldsymbol{\beta}} + \frac{1}{n_0}\widehat{\boldsymbol{\Theta}}(\boldsymbol{X}^{(0)})^T[\boldsymbol{Y}^{(0)} - \boldsymbol{\psi}'(\boldsymbol{X}^{(0)}\boldsymbol{\beta})] + \frac{1}{n_0}\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}^{(0)}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \\
&\quad + \frac{1}{n_0}\widehat{\boldsymbol{\Theta}}(\boldsymbol{X}^{(0)})^T \mathrm{diag}(\{\psi''(\tilde{u}_i^{(0)})\}) - \psi''((\boldsymbol{x}_i^{(0)})^T\boldsymbol{\beta})\}_{i=1}^{n_0})\boldsymbol{X}^{(0)}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}).
\end{aligned}$$

where $\tilde{u}_i^{(0)}$ falls on the line between $(\boldsymbol{x}_i^{(0)})^T\boldsymbol{\beta}$ and $(\boldsymbol{x}_i^{(0)})^T\hat{\boldsymbol{\beta}}$. Therefore for any $j = 1, \ldots, p$,

$$\begin{aligned}
\hat{b}_j - \beta_j = &\underbrace{\left[\boldsymbol{e}_j - (\widehat{\boldsymbol{\Theta}}_j)^T\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}^{(0)}\right]^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}_{(1)} + \underbrace{\frac{1}{n_0}(\widehat{\boldsymbol{\Theta}}_j)^T(\boldsymbol{X}^{(0)})^T[\boldsymbol{Y}^{(0)} - \boldsymbol{\psi}'(\boldsymbol{X}^{(0)}\boldsymbol{\beta})]}_{(2)} \\
&\underbrace{+ \frac{1}{n_0}(\widehat{\boldsymbol{\Theta}}_j)^T(\boldsymbol{X}^{(0)})^T \mathrm{diag}(\{\psi''(\tilde{u}_i^{(0)}) - \psi''((\boldsymbol{x}_i^{(0)})^T\boldsymbol{\beta})\}_{i=1}^{n_0})\boldsymbol{X}^{(0)}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}_{(3)},
\end{aligned}$$

For convenience, we write the $j$-th row of the matrix

$$\begin{pmatrix} 1 & -\hat{\gamma}_{1,2} & \dots & -\hat{\gamma}_{1,p} \\ -\hat{\gamma}_{2,1} & 1 & \dots & -\hat{\gamma}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\gamma}_{p,1} & -\hat{\gamma}_{p,2} & \dots & -\hat{\gamma}_{p,p} \end{pmatrix}$$

as $(\hat{\boldsymbol{\gamma}}_j^{(0)})^\dagger$ and the $j$-th row of matrix

$$\begin{pmatrix} 1 & -\gamma_{1,2} & \dots & -\gamma_{1,p} \\ -\gamma_{2,1} & 1 & \dots & -\gamma_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\gamma_{p,1} & -\gamma_{p,2} & \dots & -\gamma_{p,p} \end{pmatrix}$$

as $(\boldsymbol{\gamma}_j^{(0)})^\dagger$.

For (i), notice that

$$\begin{aligned}
&\left\| \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}^{(0)} \widehat{\boldsymbol{\Theta}}_j - \boldsymbol{e}_j \right\|_\infty \\
&= \left\| \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}^{(0)} \widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{(0)} \boldsymbol{\Theta}_j \right\|_\infty \\
&= \left\| \frac{1}{n_0} \sum_{i=1}^{n_0} \boldsymbol{x}_i^{(0)} \psi''((\boldsymbol{x}_i^{(0)})^T \boldsymbol{\beta}) (\boldsymbol{x}_i^{(0)})^T \frac{(\hat{\boldsymbol{\gamma}}_j^{(0)})^\dagger}{\hat{\tau}_j^2} - \mathbb{E}\left[ \boldsymbol{x}_i^{(0)} \psi''((\boldsymbol{x}_i^{(0)})^T \boldsymbol{\beta}) (\boldsymbol{x}_i^{(0)})^T \frac{(\boldsymbol{\gamma}_j^{(0)})^\dagger}{\tau_j^2} \right] \right\|_\infty \\
&= \left| \frac{1}{n_0} \sum_{i=1}^{n_0} \boldsymbol{x}_i^{(0)} \psi''((\boldsymbol{x}_i^{(0)})^T \boldsymbol{\beta}) (\boldsymbol{x}_i^{(0)})^T \frac{(\hat{\boldsymbol{\gamma}}_j^{(0)})^\dagger}{\hat{\tau}_j^2} - \frac{1}{n_0} \sum_{i=1}^{n_0} \boldsymbol{x}_i^{(0)} \psi''((\boldsymbol{x}_i^{(0)})^T \boldsymbol{\beta}) (\boldsymbol{x}_i^{(0)})^T \frac{(\boldsymbol{\gamma}_j^{(0)})^\dagger}{\tau_j^2} \right| \\
&\quad + \left| \frac{1}{n_0} \sum_{i=1}^{n_0} \boldsymbol{x}_i^{(0)} \psi''((\boldsymbol{x}_i^{(0)})^T \boldsymbol{\beta}) (\boldsymbol{x}_i^{(0)})^T \frac{(\boldsymbol{\gamma}_j^{(0)})^\dagger}{\tau_j^2} - \mathbb{E}\left[ \boldsymbol{x}_i^{(0)} \psi''((\boldsymbol{x}_i^{(0)})^T \boldsymbol{\beta}) (\boldsymbol{x}_i^{(0)})^T \frac{(\boldsymbol{\gamma}_j^{(0)})^\dagger}{\tau_j^2} \right] \right| \\
&\lesssim \|\hat{\boldsymbol{\gamma}}_j^{(0)} - \boldsymbol{\gamma}_j^{(0)}\|_2 + |\hat{\tau}_j^{-2} - \tau_j^{-2}| + n_0^{-1/2} \\
&\lesssim \Re_1 + h_1^{1/2} \left( \frac{\log p}{n_0} \right)^{1/4} + h_1^{1/2} \Re_1^{1/2} + h_{\max} + n_0^{-1/2},
\end{aligned}$$

implying that

$$
\begin{aligned}
(1) &\leq \left\| \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}^{(0)} \widehat{\boldsymbol{\Theta}}_j - \boldsymbol{e}_j \right\|_\infty \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \\
&\lesssim_p \left[ \Re_1 + h_1^{1/2} \left( \frac{\log p}{n_0} \right)^{1/4} + h_1^{1/2} \Re_1^{1/2} + h_{\max} + n_0^{-1/2} \right] \cdot \left( s \sqrt{\frac{\log p}{n_{\mathcal{A}_h} + n_0}} + h \right) \\
&= o(n_0^{-1/2}),
\end{aligned} \tag{S.2.38}
$$

under Assumption 7.

For (3),

$$
(3) \leq \frac{1}{n_0} \left\| \boldsymbol{X}^{(0)} \widehat{\boldsymbol{\Theta}}_j \right\|_\infty \left\| \mathrm{diag}(\{\psi''(\tilde{u}_i^{(0)}) - \psi''((\boldsymbol{x}_i^{(0)})^T \boldsymbol{\beta})\}_{i=1}^{n_0}) \boldsymbol{X}^{(0)} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right\|_1,
$$

where $|[\psi''(\tilde{u}_i^{(0)}) - \psi''((\boldsymbol{x}_i^{(0)})^T \boldsymbol{\beta})] \cdot (\boldsymbol{x}_i^{(0)})^T (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})| \leq \psi'''(t\tilde{u}_i^{(0)} + (1-t)(\boldsymbol{x}_i^{(0)})^T \boldsymbol{\beta}) \cdot |\tilde{u}_i^{(0)} - (\boldsymbol{x}_i^{(0)})^T \boldsymbol{\beta}||(\boldsymbol{x}_i^{(0)})^T (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})| \lesssim |(\boldsymbol{x}_i^{(0)})^T (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})|^2$. Therefore by similar arguments to get (S.2.27), we have

$$
(3) \lesssim_p (1 + \|\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j\|_1) \cdot \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \lesssim_p \Re_1^2 = o(n_0^{-1/2}) \tag{S.2.39}
$$

under Assumption 7.

For (2), we can see that

$$
(2) = \frac{1}{n_0} \widehat{\boldsymbol{\Theta}}_j^T (\boldsymbol{X}^{(0)})^T [\boldsymbol{Y}^{(0)} - \boldsymbol{\psi}'(\boldsymbol{X}^{(0)} \boldsymbol{\beta})] = \frac{1}{n_0} \sum_{i=1}^{n_0} \widehat{\boldsymbol{\Theta}}_j^T \boldsymbol{x}_i^{(0)} [y_i^{(0)} - \psi'((\boldsymbol{x}_i^{(0)})^T \boldsymbol{\beta})], \tag{S.2.40}
$$

where $\mathbb{E}\{(\widehat{\boldsymbol{\Theta}}_j)^T \boldsymbol{x}_i^{(0)} [y_i^{(0)} - \psi'((\boldsymbol{x}_i^{(0)})^T \boldsymbol{\beta})]\} = 0$ and

$$
\begin{aligned}
\mathbb{E}\{\widehat{\boldsymbol{\Theta}}_j^T \boldsymbol{x}_i^{(0)} [y_i^{(0)} - \psi'((\boldsymbol{x}_i^{(0)})^T \boldsymbol{\beta})]\}^2 &= \mathbb{E}\{(\widehat{\boldsymbol{\Theta}}_j^T \boldsymbol{x}_i^{(0)})^2 \psi''((\boldsymbol{x}_i^{(0)})^T \boldsymbol{\beta})\} \\
&= \boldsymbol{\Theta}_j^T \mathbb{E}\left[ \boldsymbol{x}^{(0)} (\boldsymbol{x}^{(0)})^T \psi''((\boldsymbol{x}^{(0)})^T \boldsymbol{\beta}) \right] \boldsymbol{\Theta}_j \\
&= \boldsymbol{\Theta}_j^T \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{(0)} \boldsymbol{\Theta}_j
\end{aligned}
$$

$$\lesssim \mathbb{E}\left(\boldsymbol{\Theta}_j^T \boldsymbol{x}^{(0)}\right)^2$$
$$< \infty,$$

by Assumption 7. Thus by Lindeberg's conditions,

$$\sqrt{n_0} \cdot (2) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Theta}_{j,j}). \tag{S.2.41}$$

On the other hand,

$$\left|\widehat{\boldsymbol{\Theta}}_j^T \widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} \widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_{j,j}\right| \leq \left|\widehat{\boldsymbol{\Theta}}_j^T (\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} \widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{(0)} \boldsymbol{\Theta}_j)\right| + \left|(\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j) \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{(0)} \boldsymbol{\Theta}_j\right|.$$

Note that

$$\begin{aligned}
&\left\|\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} \widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{(0)} \boldsymbol{\Theta}_j\right\|_\infty \\
&= \left\|\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} \frac{(\hat{\boldsymbol{\gamma}}_j^{(0)})^\dagger}{\hat{\tau}_j^2} - \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{(0)} \frac{(\boldsymbol{\gamma}_j^{(0)})^\dagger}{\tau_j^2}\right\|_\infty \\
&\leq \left\|\frac{1}{n_{\mathcal{A}_h} + n_0} \sum_{i,k} \left[\boldsymbol{x}_i^{(k)} \psi''(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i^{(k)}) (\boldsymbol{x}_i^{(k)})^T \frac{(\hat{\boldsymbol{\gamma}}_j^{(0)})^\dagger}{\hat{\tau}_j^2}\right] - \frac{1}{n_{\mathcal{A}_h} + n_0} \sum_{i,k} \left[\boldsymbol{x}_i^{(k)} \psi''(\boldsymbol{\beta}^T \boldsymbol{x}_i^{(k)}) (\boldsymbol{x}_i^{(k)})^T \frac{(\boldsymbol{\gamma}_j^{(0)})^\dagger}{\tau_j^2}\right]\right\|_\infty \\
&\quad + \frac{1}{n_{\mathcal{A}_h} + n_0} \left\|\sum_{i,k} \left[\boldsymbol{x}_i^{(k)} \psi''(\boldsymbol{\beta}^T \boldsymbol{x}_i^{(k)}) (\boldsymbol{x}_i^{(k)})^T \frac{(\boldsymbol{\gamma}_j^{(0)})^\dagger}{\tau_j^2}\right] - \sum_{k \in \{0\} \cup \mathcal{A}_h} n_k \mathbb{E}\left[\boldsymbol{x}_i^{(k)} \psi''(\boldsymbol{\beta}^T \boldsymbol{x}_i^{(k)}) (\boldsymbol{x}_i^{(k)})^T \frac{(\boldsymbol{\gamma}_j^{(0)})^\dagger}{\tau_j^2}\right]\right\|_\infty \\
&\quad + \left\|\sum_{k \in \{0\} \cup \mathcal{A}_h} \frac{n_k}{n_{\mathcal{A}_h} + n_0} \mathbb{E}\left[\boldsymbol{x}_i^{(k)} \psi''(\boldsymbol{\beta}^T \boldsymbol{x}_i^{(k)}) (\boldsymbol{x}_i^{(k)})^T \frac{(\boldsymbol{\gamma}_j^{(0)})^\dagger}{\tau_j^2} - \boldsymbol{x}_i^{(0)} \psi''(\boldsymbol{\beta}^T \boldsymbol{x}_i^{(0)}) (\boldsymbol{x}_i^{(0)})^T \frac{(\boldsymbol{\gamma}_j^{(0)})^\dagger}{\tau_j^2}\right]\right\|_\infty \\
&\lesssim \underbrace{\left\|\frac{1}{n_{\mathcal{A}_h} + n_0} \sum_{i,k} \boldsymbol{x}_i^{(k)} \left[\psi''(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i^{(k)}) - \psi''(\boldsymbol{\beta}^T \boldsymbol{x}_i^{(k)})\right] (\boldsymbol{x}_i^{(k)})^T \frac{(\hat{\boldsymbol{\gamma}}_j^{(0)})^\dagger}{\hat{\tau}_j^2}\right\|_\infty}_{(4)}
\end{aligned}$$

$$+\underbrace{\left\|\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{i,k}\boldsymbol{x}_i^{(k)}\psi''(\boldsymbol{\beta}^T\boldsymbol{x}_i^{(k)})(\boldsymbol{x}_i^{(k)})^T[(\hat{\boldsymbol{\gamma}}_j^{(0)})^\dagger-(\boldsymbol{\gamma}_j^{(0)})^\dagger]\hat{\tau}_j^{-2}\right\|_\infty}_{(5)}$$

$$+\underbrace{\left\|\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{i,k}\boldsymbol{x}_i^{(k)}\psi''(\boldsymbol{\beta}^T\boldsymbol{x}_i^{(k)})\cdot(\boldsymbol{x}_i^{(k)})^T(\hat{\boldsymbol{\gamma}}_j^{(0)})^\dagger(\hat{\tau}_j^2-\tau_j^2)\right\|_\infty}_{(6)}$$

$$+\underbrace{\frac{1}{n_{\mathcal{A}_h}+n_0}\left\|\sum_{i,k}\left[\boldsymbol{x}_i^{(k)}\psi''(\boldsymbol{\beta}^T\boldsymbol{x}_i^{(k)})(\boldsymbol{x}_i^{(k)})^T\frac{(\boldsymbol{\gamma}_j^{(0)})^\dagger}{\tau_j^2}\right]-\sum_{k\in\{0\}\cup\mathcal{A}_h}n_k\mathbb{E}\left[\boldsymbol{x}_i^{(k)}\psi''(\boldsymbol{\beta}^T\boldsymbol{x}_i^{(k)})(\boldsymbol{x}_i^{(k)})^T\frac{(\boldsymbol{\gamma}_j^{(0)})^\dagger}{\tau_j^2}\right]\right\|_\infty}_{(7)}$$

$$+\underbrace{\left\|\sum_{k\in\{0\}\cup\mathcal{A}_h}\frac{n_k}{n_{\mathcal{A}_h}+n_0}\mathbb{E}\left[\boldsymbol{x}_i^{(k)}\psi''(\boldsymbol{\beta}^T\boldsymbol{x}_i^{(k)})(\boldsymbol{x}_i^{(k)})^T\frac{(\boldsymbol{\gamma}_j^{(0)})^\dagger}{\tau_j^2}-\boldsymbol{x}_i^{(0)}\psi''(\boldsymbol{\beta}^T\boldsymbol{x}_i^{(0)})(\boldsymbol{x}_i^{(0)})^T\frac{(\boldsymbol{\gamma}_j^{(0)})^\dagger}{\tau_j^2}\right]\right\|_\infty}_{(8)}.$$

It's easy to see that

$$(4)\lesssim\left\|\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{i,k}\boldsymbol{x}_i^{(k)}\cdot\psi'''(t_i^{(k)}(\boldsymbol{x}_i^{(k)})^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})+(\boldsymbol{x}_i^{(k)})^T\boldsymbol{\beta})\cdot(\boldsymbol{x}_i^{(k)})^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\cdot(\boldsymbol{x}_i^{(k)})^T\frac{(\hat{\boldsymbol{\gamma}}_j^{(0)})^\dagger}{\hat{\tau}_j^2}\right\|_\infty$$

$$\lesssim\sqrt{\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{k,i}|(\boldsymbol{x}_i^{(k)})^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})|^2}$$

$$\lesssim\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2$$

$$\lesssim\Re_1,$$

with probability at least $1-n_0^{-1}$.

Similarly,

$$(5) \lesssim \|\hat{\boldsymbol{\gamma}}_j^{(0)} - \boldsymbol{\gamma}_j^{(0)}\|_2 + |\hat{\tau}_j^{-2} - \tau_j^{-2}| \lesssim \Re_1 + h_*^{1/2}\left(\frac{\log p}{n_0}\right)^{1/4} + h_1^{1/2}\Re_1^{1/2} + h_{\max}$$

$$(6) \lesssim |\hat{\tau}_j^{-2} - \tau_j^{-2}| \lesssim \Re_1 + h_*^{1/2}\left(\frac{\log p}{n_0}\right)^{1/4} + h_1^{1/2}\Re_1^{1/2} + h_{\max}$$

$$(7) \lesssim (n_{\mathcal{A}_h} + n_0)^{-1/2},$$

$$(8) \leq \sup_{k\in\mathcal{A}_h} \|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{(0)}\|_{\max} \cdot \left\|\frac{(\boldsymbol{\gamma}_j^{(0)})^\dagger}{\tau_j^2}\right\|_1 \lesssim h_{\max},$$

with probability at least $1 - K_{\mathcal{A}_h} n_0^{-1}$. Besides,

$$\|\widehat{\boldsymbol{\Theta}}_j\|_1 \leq \|\boldsymbol{\Theta}_j\|_1 + \|\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j\|_1 \lesssim \sqrt{s}.$$

with probability at least $1 - K_{\mathcal{A}_h} n_0^{-1}$. Therefore,

$$\begin{aligned}\left|\widehat{\boldsymbol{\Theta}}_j^T(\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{(0)}\boldsymbol{\Theta}_j)\right| &\leq \|\widehat{\boldsymbol{\Theta}}_j\|_1 \left\|\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{(0)}\boldsymbol{\Theta}_j\right\|_\infty \\ &\lesssim \sqrt{s}\left[\Re_1 + h_*^{1/2}\left(\frac{\log p}{n_0}\right)^{1/4} + h_1^{1/2}\Re_1^{1/2} + h_{\max}\right],\end{aligned}$$

with probability at least $1 - K_{\mathcal{A}_h} n_0^{-1}$. Similarly, we can obtain the same upper bound for the second term $\left|(\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j)\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{(0)}\boldsymbol{\Theta}_j\right|$ in (S.2.40). Then finally,

$$\left|\widehat{\boldsymbol{\Theta}}_j^T\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_{j,j}\right| \lesssim \sqrt{s}\left[\Re_1 + h_*^{1/2}\left(\frac{\log p}{n_0}\right)^{1/4} + h_1^{1/2}\Re_1^{1/2} + h_{\max}\right] = o(1),$$

with probability at least $1 - K_{\mathcal{A}_h} n_0^{-1}$. Then by Slutsky's lemma, equations (S.2.38), (S.2.41) and (S.2.39),

$$\frac{\sqrt{n_0}(\hat{b}_j - \beta_j)}{\sqrt{\widehat{\boldsymbol{\Theta}}_j^T\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\widehat{\boldsymbol{\Theta}}_j}} \xrightarrow{d} \mathcal{N}(0,1),$$

which completes the proof.

#### S.2.3.6 Proof of Theorem 6

Assume Assumption 3.(i) holds or Assumption 3.(ii) with $h \leq C'U^{-1}\bar{U}$ for some $C' > 0$ holds. When $\lambda_{\boldsymbol{\delta}} = C_{\boldsymbol{\delta}}\sqrt{\frac{\log p}{n_0}}$, where $C_{\boldsymbol{\delta}} > 0$ is a sufficiently large constant: By the definition of $\hat{\boldsymbol{\beta}}$, we have $\nabla L_{n_0}^{(0)}(\hat{\boldsymbol{\beta}}) + \lambda_{\boldsymbol{\delta}} \cdot \partial\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{w}}^{\mathcal{A}_h}\|_1 = \mathbf{0}$. Then by Hölder inequality,

$$\begin{aligned}\langle \nabla L_{n_0}^{(0)}(\hat{\boldsymbol{\beta}}) - \nabla L_{n_0}^{(0)}(\boldsymbol{\beta}), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle &= \langle -\lambda_{\boldsymbol{\delta}} \cdot \partial\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{w}}^{\mathcal{A}_h}\|_1 - \nabla L_{n_0}^{(0)}(\boldsymbol{\beta}), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle \\ &\leq \lambda_{\boldsymbol{\delta}}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 + \|\nabla L_{n_0}^{(0)}(\boldsymbol{\beta})\|_\infty \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1. \quad \text{(S.2.42)}\end{aligned}$$

Considering the fact that $\|\nabla L_{n_0}^{(0)}(\boldsymbol{\beta})\|_\infty \lesssim \sqrt{\frac{\log p}{n_0}}$ with probability at least $1 - n_0^{-1}$ (Lemma 6 in Negahban et al. (2009)) and the upper bound of $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$ we prove in Theorem 1, the desired upper bound of $\langle \nabla L_{n_0}^{(0)}(\hat{\boldsymbol{\beta}}) - \nabla L_{n_0}^{(0)}(\boldsymbol{\beta}), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle$ follows.

#### S.2.3.7 Proof of Theorem 7

Assume Assumption 3.(i) holds or Assumption 3.(ii) with $h \leq C'U^{-1}\bar{U}$ for some $C' > 0$ holds. If we take $\lambda_{\boldsymbol{\delta}} = C_{\boldsymbol{\delta}}\sqrt{\frac{\log p}{n_0}}$, where $C_{\boldsymbol{\delta}} > 0$ is a sufficiently large constant: Similar to (S.2.42), we can obtain

$$\langle \nabla L_{n_0}^{(0)}(\hat{\boldsymbol{\beta}}) - \nabla L_{n_0}^{(0)}(\boldsymbol{\beta}), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle \leq \lambda_{\boldsymbol{\delta}}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 + \|\nabla L_{n_0}^{(0)}(\boldsymbol{\beta})\|_\infty \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1.$$

To bound $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$, it suffices to combine (S.2.28) and the upper bound of $\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1$ in (S.2.33). Then the final upper bound follows.

### S.2.4 Proof of propositions

#### S.2.4.1 Proof of Proposition 1

The rate of $\Gamma_1$ can be derived from the following Lemma 7 and the union bound, together with the tail inequality (S.2.32). The rate of $\Gamma_2$ comes from the following Lemma 8.

**Lemma 7.** *Under the same assumptions as Theorem 4, we have the following conclusions:*

*(i) For logistic regression model:*

$$\sup_r |\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]}) - \hat{L}_0^{[r]}(\boldsymbol{\beta}^{(k)})| \leq C\left(1+\sqrt{\frac{1}{n_0}}\cdot\zeta\right)\cdot\sup_r \|\hat{\boldsymbol{\beta}}^{(k)[r]} - \boldsymbol{\beta}^{(k)}\|_2, k\in\mathcal{A},$$
$$\sup_r |\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]}) - \hat{L}_0^{[r]}(\boldsymbol{\beta}^{(k)})| \leq C\left(1+\sqrt{\frac{1}{n_0}}\cdot\zeta\right)\cdot\sup_r \|\hat{\boldsymbol{\beta}}^{(k)[r]} - \boldsymbol{\beta}^{(k)}\|_2, k\in\mathcal{A}^c,$$
$$\sup_r |\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(0)[r]}) - \hat{L}_0^{[r]}(\boldsymbol{\beta})| \leq C\sup_r \|\hat{\boldsymbol{\beta}}^{(0)[r]} - \boldsymbol{\beta}\|_2\cdot(1+\zeta),$$

*with probability at least* $1-\exp\{-\zeta^2\}$.

*(ii) For linear model:*

$$\sup_r |\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]}) - \hat{L}_0^{[r]}(\boldsymbol{\beta}^{(k)})| \leq C\left(1+\sqrt{\frac{1}{n_0}}\cdot\zeta\right)\cdot(\|\boldsymbol{w}^{(k)}\|_2\vee\|\boldsymbol{\beta}\|_2)\cdot\|\hat{\boldsymbol{\beta}}^{(k)[r]} - \boldsymbol{\beta}^{(k)}\|_2, k\in\mathcal{A},$$

$$\sup_r |\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]}) - \hat{L}_0^{[r]}(\boldsymbol{\beta}^{(k)})| \leq C\left(1+\sqrt{\frac{1}{n_0}}\cdot\zeta\right)\cdot\|\boldsymbol{\beta}^{(k)}\|_2\cdot\|\hat{\boldsymbol{\beta}}^{(k)[r]} - \boldsymbol{\beta}^{(k)}\|_2, k\in\mathcal{A}^c,$$

$$\sup_r |\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(0)[r]}) - \hat{L}_0^{[r]}(\boldsymbol{\beta})| \leq C\|\boldsymbol{\beta}\|_2\cdot\|\hat{\boldsymbol{\beta}}^{(0)[r]} - \boldsymbol{\beta}\|_2\cdot(1+\zeta),$$

*with probability at least* $1-\exp\{-\zeta^2\}$.

*(iii) For Poisson regression model with* $\sup_k \|\boldsymbol{x}^{(k)}\|_\infty \leq U$ *a.s.:*

$$\sup_r |\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]}) - \hat{L}_0^{[r]}(\boldsymbol{\beta}^{(k)})| \leq C\left(1+\sqrt{\frac{1}{n_0}}\cdot\zeta\right)\cdot\exp\left(U(\|\boldsymbol{w}^{(k)}\|_1\vee\|\boldsymbol{\beta}\|_1)\right)\|\hat{\boldsymbol{\beta}}^{(k)[r]} - \boldsymbol{\beta}^{(k)}\|_2, k\in\mathcal{A},$$

$$\sup_r |\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]}) - \hat{L}_0^{[r]}(\boldsymbol{\beta}^{(k)})| \leq C\left(1+\sqrt{\frac{1}{n_0}}\cdot\zeta\right)\cdot\exp\left(U\|\boldsymbol{\beta}^{(k)}\|_1\right)\|\hat{\boldsymbol{\beta}}^{(k)[r]} - \boldsymbol{\beta}^{(k)}\|_2, k\in\mathcal{A}^c,$$

$$|\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(0)[r]}) - \hat{L}_0^{[r]}(\boldsymbol{\beta})| \lesssim \exp\left(U\|\boldsymbol{\beta}\|_1\right) \cdot \|\hat{\boldsymbol{\beta}}^{(0)[r]} - \boldsymbol{\beta}\|_2 \cdot (1+\zeta),$$

*with probability at least* $1 - \exp\{-\zeta^2\}$.

**Remark 13.** *It's important to point out that based on Algorithm 2, the randomness of* $\hat{L}_0^{[r]}$, $\hat{\boldsymbol{\beta}}^{(k)[r]}$ $(k \neq 0)$, *and* $\hat{\boldsymbol{\beta}}^{(0)[r]}$ *is independent. Here* $\hat{\boldsymbol{\beta}}^{(k)[r]}$ *and* $\hat{\boldsymbol{\beta}}^{(0)[r]}$ *are regarded as fixed and we only consider the randomness from* $\hat{L}_0^{[r]}$.

*Proof of Lemma 7.* For convenience, we assume $n_0$ is divisible by 3. Note that the term involving $\sum_{i=1}^{n_0/3} \rho(y_i^{(0)[r]})$ is canceled when we take the difference between $\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]})$ and $\hat{L}_0^{[r]}(\boldsymbol{\beta}^{(k)})$. So in the following, we drop that term from the definition of $\hat{L}_0^{[r]}$ in equation (3). We only prove the bound for $|\hat{L}_0(\hat{\boldsymbol{\beta}}^{(k)[r]}) - \hat{L}_0(\boldsymbol{\beta}^{(k)})|$ when $k \in \mathcal{A}$. The cases that $k = 0$ or $k \in \mathcal{A}^c$ can be similarly discussed. Besides, according to the proof of Theorem 3, when $k \in \mathcal{A}$, we define $\boldsymbol{\beta}^{(k)} = \frac{2n_0/3}{2n_0/3+n_k}\boldsymbol{\beta} + \frac{n_k}{2n_0/3+n_k}\boldsymbol{w}^{(k)}$, which gives us the final results shown in Proposition 1.

(i) For logistic regression model, notice that

$$\begin{aligned}|\hat{L}_0(\hat{\boldsymbol{\beta}}^{(k)[r]}) - \hat{L}_0(\boldsymbol{\beta}^{(k)})| \leq \frac{1}{n_0/3}|(\boldsymbol{y}^{(0)[r]})^T \boldsymbol{X}^{(0)[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]} - \boldsymbol{\beta}^{(k)})| \\ + \frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}[\psi((\boldsymbol{x}_i^{(0)[r]})^T\hat{\boldsymbol{\beta}}^{(k)[r]}) - \psi((\boldsymbol{x}_i^{(0)[r]})^T\boldsymbol{\beta}^{(k)})]\right|.\end{aligned}$$

For the first term on the right-hand side, it holds that

$$\frac{1}{n_0/3}|(\boldsymbol{y}^{(0)[r]})^T \boldsymbol{X}^{(0)[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]} - \boldsymbol{\beta}^{(k)})| \leq \frac{1}{n_0/3}\sum_{i=1}^{n_0/3}|(\boldsymbol{x}_i^{(0)[r]})^T(\hat{\boldsymbol{\beta}}^{(k)[r]} - \boldsymbol{\beta}^{(k)})|,$$

where $\frac{1}{n_0/3}\sum_{i=1}^{n_0/3}|(\boldsymbol{x}_i^{(0)[r]})^T(\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k)})|$ is a $\frac{1}{n_0/3}\|\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k)}\|_2^2$ sub-Gaussian with mean less than $C\|\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k)}\|_2$, where $C > 0$ is a uniform constant, implying that

$$\frac{1}{n_0/3}|(\boldsymbol{y}^{(0)[r]})^T \boldsymbol{X}^{(0)[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]} - \boldsymbol{\beta}^{(k)})| \lesssim \left(1 + \sqrt{\frac{1}{n_0}} \cdot \zeta\right) \cdot \|\hat{\boldsymbol{\beta}}^{(k)[r]} - \boldsymbol{\beta}^{(k)}\|_2,$$

with probability at least $1-\exp\{-\zeta^2\}$. On the other hand, the second term can be bounded by $\frac{C}{n_0/3}\sum_{i=1}^{n_0/3}|(\boldsymbol{x}_i^{(0)[r]})^T(\hat{\boldsymbol{\beta}}^{(k)[r]}-\boldsymbol{\beta}^{(k)})|$, which can be similarly bounded as the first term, leading to the desired conclusion.

(ii) For linear model, note that $y_i^{(0)[r]}=\boldsymbol{\beta}^T\boldsymbol{x}_i^{(0)[r]}+\epsilon_i^{(0)[r]}$ and $\psi(u)=u^2/2$, leading to

$$\begin{aligned}|\hat{L}_0(\hat{\boldsymbol{\beta}}^{(k)[r]})-\hat{L}_0(\boldsymbol{\beta}^{(k)})| \le & \frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}\epsilon_i^{(0)[r]}(\boldsymbol{x}_i^{(0)[r]})^T(\hat{\boldsymbol{\beta}}^{(k)[r]}-\boldsymbol{\beta}^{(k)})\right| \\ & + \frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}(\boldsymbol{x}_i^{(0)[r]})^T\boldsymbol{\beta}\cdot(\boldsymbol{x}_i^{(0)[r]})^T(\hat{\boldsymbol{\beta}}^{(k)[r]}-\boldsymbol{\beta}^{(k)})\right| \\ & + \frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}(\boldsymbol{x}_i^{(0)[r]})^T(\hat{\boldsymbol{\beta}}^{(k)[r]}+\boldsymbol{\beta}^{(k)})(\hat{\boldsymbol{\beta}}^{(k)[r]}-\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}_i^{(0)[r]}\right|.\end{aligned}$$

It is easy to see that $\frac{1}{n_0/3}\sum_{i=1}^{n_0/3}\epsilon_i^{(0)}(\boldsymbol{x}_i^{(0)[r]})^T(\hat{\boldsymbol{\beta}}^{(k)[r]}-\boldsymbol{\beta}^{(k)})$ is $\frac{1}{n_0}\|\hat{\boldsymbol{\beta}}^{(k)[r]}-\boldsymbol{\beta}^{(k)}\|_2$-subGaussian with zero mean, while $\frac{1}{n_0/3}\sum_{i=1}^{n_0/3}(\boldsymbol{x}_i^{(0)[r]})^T\boldsymbol{\beta}\cdot(\boldsymbol{x}_i^{(0)[r]})^T(\hat{\boldsymbol{\beta}}^{(k)[r]}-\boldsymbol{\beta}^{(k)})$ and $\frac{1}{n_0/3}\sum_{i=1}^{n_0/3}(\boldsymbol{x}_i^{(0)[r]})^T(\hat{\boldsymbol{\beta}}^{(k)[r]}+\boldsymbol{\beta}^{(k)})(\hat{\boldsymbol{\beta}}^{(k)[r]}-\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}_i^{(0)[r]}$ are $\frac{1}{n_0/3}\|\boldsymbol{\beta}^{(k)}\|_2\|\hat{\boldsymbol{\beta}}^{(k)[r]}-\boldsymbol{\beta}^{(k)}\|_2$-subexponential with mean at most $C\|\boldsymbol{\beta}^{(k)}\|_2\|\hat{\boldsymbol{\beta}}^{(k)[r]}-\boldsymbol{\beta}^{(k)}\|_2$ (Vershynin, 2018), where $C>0$ is a uniform constant, then by tail bounds and union bounds, the conclusion follows.

(iii) For Poisson regression model with a.s. bounded covariates, it holds that

$$\begin{aligned}|\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]})-\hat{L}_0(\boldsymbol{\beta}^{(k)})| \le & \frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}y_i^{(0)[r]}(\boldsymbol{x}_i^{(0)[r]})^T(\hat{\boldsymbol{\beta}}^{(k)[r]}-\boldsymbol{\beta}^{(k)})\right| \\ & + \frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}[e^{(\boldsymbol{x}_i^{(0)[r]})^T\hat{\boldsymbol{\beta}}^{(k)[r]}}-e^{(\boldsymbol{x}_i^{(0)[r]})^T\boldsymbol{\beta}^{(k)}}]\right|.\end{aligned}$$

Conditioning on $\boldsymbol{X}^{(0)[r]}$, we know that $y_i^{(0)[r]}(\boldsymbol{x}_i^{(0)[r]})^T(\hat{\boldsymbol{\beta}}^{(k)[r]}-\boldsymbol{\beta}^{(k)}) \sim \text{Poisson}(e^{(\boldsymbol{x}_i^{(0)[r]})^T\boldsymbol{\beta}})\cdot (\boldsymbol{x}_i^{(0)[r]})^T(\hat{\boldsymbol{\beta}}^{(k)[r]}-\boldsymbol{\beta}^{(k)})$. By the fact that $\text{Poisson}(e^{(\boldsymbol{x}_i^{(0)[r]})^T\boldsymbol{\beta}})$ is a $e^{(\boldsymbol{x}_i^{(0)[r]})^T\boldsymbol{\beta}}$-subexponential given $\boldsymbol{x}_i^{(0)}$, we have

$$\frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3} y_i^{(0)}(\boldsymbol{x}_i^{(0)[r]})^T(\hat{\boldsymbol{\beta}}^{(k)[r]}-\boldsymbol{\beta}^{(k)})\right| \lesssim \frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3} e^{(\boldsymbol{x}_i^{(0)[r]})^T\boldsymbol{\beta}}(\boldsymbol{x}_i^{(0)[r]})^T(\hat{\boldsymbol{\beta}}^{(k)[r]}-\boldsymbol{\beta}^{(k)})\right|$$

$$+\frac{1}{\sqrt{n_0/3}}\max_i e^{(\boldsymbol{x}_i^{(0)[r]})^T\boldsymbol{\beta}}\sqrt{\sum_{i=1}^{n_0/3}|(\boldsymbol{x}_i^{(0)[r]})^T(\hat{\boldsymbol{\beta}}^{(k)[r]}-\boldsymbol{\beta}^{(k)})|^2}\cdot\zeta,$$

with probability at least $1-\exp\{-\zeta^2\}$. By Hölder inequality, the first term on the right hand side can be bounded by a $e^{2U\|\boldsymbol{\beta}\|_1}\|\hat{\boldsymbol{\beta}}^{(k)[r]}-\boldsymbol{\beta}^{(k)}\|_2^2$-subexponential with mean at most $Ce^{2U\|\boldsymbol{\beta}\|_1}\|\hat{\boldsymbol{\beta}}^{(k)[r]}-\boldsymbol{\beta}^{(k)}\|_2^2$, leading to

$$\frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3} e^{(\boldsymbol{x}_i^{(0)[r]})^T\boldsymbol{\beta}}(\boldsymbol{x}_i^{(0)[r]})^T(\hat{\boldsymbol{\beta}}^{(k)[r]}-\boldsymbol{\beta}^{(k)})\right| \lesssim \left(1+\sqrt{\frac{1}{n_0}}\cdot\zeta\right)\cdot e^{U\|\boldsymbol{\beta}\|_1}\|\hat{\boldsymbol{\beta}}^{(k)[r]}-\boldsymbol{\beta}^{(k)}\|_2,$$

with probability at least $1-\exp\{-\zeta^2\}$. On the other hand, by applying Bernstein inequality (Theorem 2.8.2 in Vershynin (2018)) as well as union bounds, we have

$$\frac{1}{\sqrt{n_0/3}}\max_i e^{(\boldsymbol{x}_i^{(0)[r]})^T\boldsymbol{\beta}}\sqrt{\sum_{i=1}^{n_0/3}|(\boldsymbol{x}_i^{(0)[r]})^T(\hat{\boldsymbol{\beta}}^{(k)[r]}-\boldsymbol{\beta}^{(k)})|^2} \lesssim_p \left(1+\sqrt{\frac{1}{n_0}}\cdot\zeta\right)$$

$$\cdot\, e^{U\|\boldsymbol{\beta}\|_1}\sup_k\|\hat{\boldsymbol{\beta}}^{(k)[r]}-\boldsymbol{\beta}^{(k)}\|_2,$$

with probability at least $1-\exp\{-\zeta^2\}$. Summarizing the conclusions before, we obtain the desired conclusion.

□

**Lemma 8.** *Under the same assumptions as Theorem 4, we have the following conclusions:*

*(i) For logistic regression model:*

$$|\hat{L}_0^{[r]}(\boldsymbol{\beta}) - L_0(\boldsymbol{\beta})| \vee |\hat{L}_0^{[r]}(\boldsymbol{\beta}^{(k)}) - \hat{L}_0^{[r]}(\boldsymbol{\beta}) - L_0(\boldsymbol{\beta}^{(k)}) + L_0(\boldsymbol{\beta})| \leq C\sqrt{\frac{1}{n_0}} \cdot \|\boldsymbol{w}^{(k)}\|_2 \cdot \zeta,$$

*with probability at least* $1 - \exp\{-\zeta^2\}$.

*(ii) For linear model:*

$$\begin{aligned}&|\hat{L}_0^{[r]}(\boldsymbol{\beta}) - L_0(\boldsymbol{\beta})| \vee |\hat{L}_0^{[r]}(\boldsymbol{\beta}^{(k)}) - \hat{L}_0(\boldsymbol{\beta}) - L_0(\boldsymbol{\beta}^{(k)}) + L_0(\boldsymbol{\beta})| \\ &\leq C\sqrt{\frac{1}{n_0}} \cdot (\|\boldsymbol{w}^{(k)}\|_2^2 \vee \|\boldsymbol{w}^{(k)}\|_2) \cdot \zeta,\end{aligned}$$

*with probability at least* $1 - \exp\{-\zeta^2\}$.

*(iii) For Poisson regression model with* $\sup_k \|\boldsymbol{x}^{(k)}\|_\infty \leq U$ *a.s.:*

$$\begin{aligned}&|\hat{L}_0^{[r]}(\boldsymbol{\beta}) - L_0(\boldsymbol{\beta})| \vee |\hat{L}_0^{[r]}(\boldsymbol{\beta}^{(k)}) - \hat{L}_0^{[r]}(\boldsymbol{\beta}) - L_0(\boldsymbol{\beta}^{(k)}) + L_0(\boldsymbol{\beta})| \\ &\leq C\sqrt{\frac{1}{n_0}} \exp\left(U\|\boldsymbol{w}^{(k)}\|_1\right) \left[1 + \|\boldsymbol{w}^{(k)}\|_2 + U\|\boldsymbol{w}^{(k)}\|_1\right] \cdot \zeta,\end{aligned}$$

*with probability at least* $1 - \zeta^{-2}$.

*Proof of Lemma 8.* Similar to the proofs of Lemmas 6 and 7, the terms involving $\sum_{i=1}^{n_0/3} \rho(y_i^{(0)[r]})$ and $\mathbb{E}[\rho(y^{(0)})]$ are canceled when taking the difference. Therefore without loss of generality, to prove the rate of $\sup_k |\hat{L}_0^{[r]}(\boldsymbol{\beta}^{(k)}) - \hat{L}_0^{[r]}(\boldsymbol{\beta}) - L_0(\boldsymbol{\beta}^{(k)}) + L_0(\boldsymbol{\beta})|$, throughout this proof, we discard these terms and consider

$$\begin{aligned}L_0(\boldsymbol{w}) &= -\mathbb{E}[\psi'(\boldsymbol{\beta}^T\boldsymbol{x}^{(0)})\boldsymbol{w}^T\boldsymbol{x}^{(0)}] + \mathbb{E}[\psi(\boldsymbol{w}^T\boldsymbol{x}^{(0)})], \\ \hat{L}_0^{[r]}(\boldsymbol{w}) &= -\frac{1}{n_0/3}\sum_{i=1}^{n_0/3} y_i^{(0)[r]}\boldsymbol{\beta}^T\boldsymbol{x}_i^{(0)[r]} + \frac{1}{n_0/3}\sum_{i=1}^{n_0/3}\psi(\boldsymbol{w}^T\boldsymbol{x}_i^{(0)[r]}).\end{aligned}$$

(i) For logistic regression model:

$$\begin{aligned}
&\sup_k \left|\hat{L}_0^{[r]}(\boldsymbol{\beta}^{(k)}) - L_0(\boldsymbol{\beta}^{(k)})\right| \\
&\leq \frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}[y_i^{(0)[r]} - \psi'(\boldsymbol{\beta}^T\boldsymbol{x}_i^{(0)[r]})](\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}_i^{(0)[r]}\right| \\
&\quad + \frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}\left[\psi'(\boldsymbol{\beta}^T\boldsymbol{x}_i^{(0)[r]})(\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}_i^{(0)[r]} - \mathbb{E}[\psi'(\boldsymbol{\beta}^T\boldsymbol{x}^{(0)})(\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}^{(0)}]\right]\right| \\
&\quad + \frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}\left[\psi((\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}_i^{(0)[r]}) - \mathbb{E}\psi((\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}_i^{(0)[r]})\right]\right|
\end{aligned}$$

Since $\frac{1}{n_0/3}\sum_{i=1}^{n_0/3}[y_i^{(0)[r]} - \psi'(\boldsymbol{\beta}^T\boldsymbol{x}_i^{(0)[r]})](\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}_i^{(0)[r]}$ is a zero-mean $\|\boldsymbol{\beta}^{(k)}\|_2^2$-subexponential variable, we have

$$\frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}[y_i^{(0)[r]} - \psi'(\boldsymbol{\beta}^T\boldsymbol{x}_i^{(0)[r]})](\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}_i^{(0)[r]}\right| \lesssim \sqrt{\frac{1}{n_0}}\|\boldsymbol{w}^{(k)}\|_2 \cdot \zeta, \tag{S.2.43}$$

with probability at least $1 - \exp\{-\zeta^2\}$. For the second term, since $\psi'$ is bounded, $\psi'(\boldsymbol{\beta}^T\boldsymbol{x}_i^{(0)[r]})(\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}_i^{(0)[r]}$ are i.i.d. $\|\boldsymbol{\beta}^{(k)}\|_2^2$-subexponential, leading to

$$\frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}\left[\psi'(\boldsymbol{\beta}^T\boldsymbol{x}_i^{(0)[r]})(\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}_i^{(0)[r]} - \mathbb{E}[\psi'(\boldsymbol{\beta}^T\boldsymbol{x}^{(0)})(\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}^{(0)}]\right]\right| \lesssim \sqrt{\frac{1}{n_0}}\|\boldsymbol{w}^{(k)}\|_2 \cdot \zeta, \tag{S.2.44}$$

with probability at least $1-\exp\{-\zeta^2\}$. For the last term, consider $g(u_1^{[r]}, \ldots, u_{n_0/3}^{[r]}) = \sum_{i=1}^{n_0/3}\psi(u_i^{[r]})$, where $u_i^{[r]} = (\boldsymbol{x}_i^{(0)[r]})^T\boldsymbol{\beta}^{(k)}$ is an i.i.d. $\|\boldsymbol{\beta}^{(k)}\|_2^2$-subGaussian. Since $\psi$ is 1-Lipschitz under $\ell_1$-norm, by Theorem 1 in Kontorovich (2014) and union bounds,

$$\frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}\left[\psi((\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}_i^{(0)[r]}) - \mathbb{E}\psi((\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}^{(0)})\right]\right| \lesssim \sqrt{\frac{1}{n_0}} \cdot \|\boldsymbol{w}^{(k)}\|_2 \cdot \zeta, \tag{S.2.45}$$

with probability at least $1-\exp\{-\zeta^2\}$. By (S.2.43), (S.2.44) and (S.2.45), the conclusion follows.

(ii) For linear model: recall that $y_i^{(0)} = (\boldsymbol{x}_i^{(0)[r]})^T\boldsymbol{\beta}^{(0)} + \epsilon_i^{(0)[r]}$, then

$$\begin{aligned}
&\left|\hat{L}_0^{[r]}(\boldsymbol{\beta}^{(k)}) - L_0(\boldsymbol{\beta}^{(k)})\right| \\
&\le \frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}\epsilon_i^{(0)[r]}(\boldsymbol{x}_i^{(0)[r]})^T\boldsymbol{\beta}\right| \\
&\quad + \frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}(\boldsymbol{x}_i^{(0)[r]})^T\boldsymbol{\beta}\cdot(\boldsymbol{x}_i^{(0)[r]})^T\boldsymbol{\beta}^{(k)} - \mathbb{E}[(\boldsymbol{x}^{(0)})^T\boldsymbol{\beta}\cdot(\boldsymbol{x}^{(0)})^T\boldsymbol{\beta}^{(k)}]\right| \\
&\quad + \frac{1}{2n_0/3}\left|\sum_{i=1}^{n_0/3}[(\boldsymbol{x}_i^{(0)[r]})^T\boldsymbol{\beta}^{(k)}]^2 - \mathbb{E}[(\boldsymbol{x}^{(0)})^T\boldsymbol{\beta}^{(k)}]^2\right|.
\end{aligned}$$

By subexponential tail bounds, we have

$$\frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}\epsilon_i^{(0)[r]}(\boldsymbol{x}_i^{(0)[r]})^T\boldsymbol{\beta}\right| \lesssim \sqrt{\frac{1}{n_0}}\cdot\|\boldsymbol{\beta}^{(k)}\|_2\cdot\zeta \lesssim \sqrt{\frac{1}{n_0}}\cdot\|\boldsymbol{w}^{(k)}\|_2\cdot\zeta,$$

$$\begin{aligned}
&\frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}(\boldsymbol{x}_i^{(0)[r]})^T\boldsymbol{\beta}^{(0)}\cdot(\boldsymbol{x}_i^{(0)[r]})^T\boldsymbol{\beta}^{(k)} - \mathbb{E}[(\boldsymbol{x}^{(0)})^T\boldsymbol{\beta}^{(0)}\cdot(\boldsymbol{x}^{(0)})^T\boldsymbol{\beta}^{(k)}]\right| \\
&\lesssim \sqrt{\frac{1}{n_0}}\cdot\|\boldsymbol{\beta}\|_2\sup_k\|\boldsymbol{\beta}^{(k)}\|_2\cdot\zeta \lesssim \sqrt{\frac{1}{n_0}}\cdot\|\boldsymbol{\beta}\|_2\sup_k\|\boldsymbol{w}^{(k)}\|_2,
\end{aligned}$$

$$\frac{1}{2n_0/3}\left|\sum_{i=1}^{n_0/3}[(\boldsymbol{x}_i^{(0)[r]})^T\boldsymbol{\beta}^{(k)}]^2 - \mathbb{E}[(\boldsymbol{x}^{(0)})^T\boldsymbol{\beta}^{(k)}]^2\right| \lesssim \sqrt{\frac{1}{n_0}}\cdot\|\boldsymbol{\beta}^{(k)}\|_2^2\cdot\zeta,$$

with probability at least $1-\exp\{-\zeta^2\}$, which leads to the desired conclusion.

(iii) For Poisson regression model: similar to the logistic regression model, it holds that

$$\begin{aligned}
&\left|\hat{L}_0^{[r]}(\boldsymbol{\beta}^{(k)}) - L_0(\boldsymbol{\beta}^{(k)})\right| \\
&\leq \frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}[y_i^{(0)[r]} - e^{\boldsymbol{\beta}^T\boldsymbol{x}_i^{(0)[r]}}](\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}_i^{(0)[r]}\right| \\
&\quad + \frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}\left[e^{\boldsymbol{\beta}^T\boldsymbol{x}_i^{(0)[r]}}(\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}_i^{(0)[r]} - \mathbb{E}[e^{\boldsymbol{\beta}^T\boldsymbol{x}^{(0)}}(\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}^{(0)}]\right]\right| \\
&\quad + \frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}\left[e^{(\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}_i^{(0)[r]}} - \mathbb{E}e^{(\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}^{(0)}}\right]\right|.
\end{aligned}$$

For the first term on the right-hand side, because $[y_i^{(0)[r]} - e^{\boldsymbol{\beta}^T\boldsymbol{x}_i^{(0)[r]}}](\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}_i^{(0)[r]}$ is an i.i.d. zero-mean $e^{U\|\boldsymbol{\beta}\|_1}\|\boldsymbol{\beta}^{(k)}\|_2$-subexponential variable, it follows

$$\frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}[y_i^{(0)[r]} - e^{\boldsymbol{\beta}^T\boldsymbol{x}_i^{(0)[r]}}](\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}_i^{(0)[r]}\right| \lesssim \sqrt{\frac{1}{n_0}}e^{U\|\boldsymbol{\beta}\|_1}\|\boldsymbol{\beta}^{(k)}\|_2 \cdot \zeta,$$

with probability at least $1-\exp\{-\zeta^2\}$. For the last term on the right-hand side, note that $\exp\{(\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}_i^{(0)[r]}\}$ is bounded by $\exp\{U\|\boldsymbol{\beta}^{(k)}\|_1\}$. Therefore by tail probability,

$$\frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}\left[e^{(\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}_i^{(0)[r]}} - \mathbb{E}e^{(\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}^{(0)}}\right]\right| \lesssim \exp\{U\|\boldsymbol{\beta}^{(k)}\|_1\} \cdot \sqrt{\frac{1}{n_0}} \cdot \zeta,$$

with probability at least $1-\exp\{-\zeta^2\}$. Denote $u_i^{(k)[r]} = (\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}_i^{(0)[r]}$. Finally, to bound the second term on the right-hand side, we follow the same idea to get

$$\frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}u_i^{(k)[r]}e^{u_i^{(0)[r]}} - \mathbb{E}u_i^{(k)[r]}e^{u_i^{(0)[r]}}\right| \lesssim U\|\boldsymbol{\beta}^{(k)}\|_1 \cdot \exp\{U\|\boldsymbol{\beta}^{(k)}\|_1\} \cdot \sqrt{\frac{1}{n_0}} \cdot \zeta,$$

with probability at least $1-\exp\{-\zeta^2\}$. By combining all the conclusions above, we obtain the desired bound.

The remaining task is to calculate the rate of $|\hat{L}_0^{[r]}(\boldsymbol{\beta}) - L_0(\boldsymbol{\beta})|$ under three scenarios. For logistic case, since $\rho = 0$, the calculation in (i) naturally follows and the same bound can be derived. For the linear case, we only have to show that

$$\left| \frac{1}{n_0/3} \sum_{i=1}^{n_0/3} (y_i^{(0)[r]})^2 - \mathbb{E}(y^{(0)})^2 \right| \lesssim \sqrt{\frac{1}{n_0}} \cdot (\|\boldsymbol{\beta}^{(k)}\|_2^2 \vee \|\boldsymbol{\beta}^{(k)}\|_2),$$

with probability at least $1 - \exp\{-n_0\}$. This can be easily checked by considering $y_i^{(0)[r]} = \boldsymbol{\beta}^T \boldsymbol{x}_i^{(0)[r]} + \epsilon_i^{(0)[r]}$ and applying subexponential tail bounds. For Poisson regression model, notice that

$$\mathrm{Var}(\log(y_i^{(0)[r]}!)) \leq \mathbb{E}[(y_i^{(0)[r]})^2 \log^2 y_i^{(0)[r]}] \leq \sqrt{\mathbb{E}[(y_i^{(0)[r]})^4]} \sqrt{\mathbb{E}(\log^4 y_i^{(0)[r]})}.$$

Due to moment bounds of subexponential variables and Jensen's inequality:

$$\begin{aligned} \mathbb{E}[(y_i^{(0)[r]})^4] &\lesssim \mathbb{E}_{\boldsymbol{x}} \left[ \mathbb{E}_{y^{(0)}|\boldsymbol{x}} (y_i^{(0)[r]})^4 \right] \leq \exp(4U\|\boldsymbol{\beta}\|_1), \\ \mathbb{E}(\log^4 y_i^{(0)[r]}) &\leq \log^4 \mathbb{E} y_i^{(0)[r]} \lesssim U^4 \|\boldsymbol{\beta}\|_1^4. \end{aligned}$$

Then by Chebyshev inequality and union bounds, it's straightforward to prove that

$$\left| \frac{1}{n_0/3} \sum_{i=1}^{n_0/3} y_i^{(0)[r]} - \mathbb{E} y^{(0)} \right| \lesssim_p \sqrt{\frac{1}{n_0}} U \|\boldsymbol{\beta}\|_1 \exp(U\|\boldsymbol{\beta}\|_1) \cdot \zeta,$$

with probability at least $1 - \zeta^{-2}$, which completes our proof.

□

#### S.2.4.2 Proof of Proposition 2

As analyzed in the proof of Theorem 5, it remains to show (S.2.37). Recall that $\hat{\tau}_j^2 = \widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}},j,j} - \widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}},j,-j} \hat{\boldsymbol{\gamma}}_j^{(0)} = (n_{\mathcal{A}_h} + n_0)^{-1} \sum_{k \in \{0\} \cup \mathcal{A}_h} (\boldsymbol{X}_j^{(k)})^T \mathrm{diag}(\{\psi''((\boldsymbol{x}_i^{(k)})^T \hat{\boldsymbol{\beta}})\}_{i=1}^{n_k}) \boldsymbol{X}_j^{(k)} - (n_{\mathcal{A}_h} +$

$n_0)^{-1}\sum_{k\in\{0\}\cup\mathcal{A}_h}(\boldsymbol{X}_j^{(k)})^T\mathrm{diag}(\{\psi''((\boldsymbol{x}_i^{(k)})^T\hat{\boldsymbol{\beta}})\}_{i=1}^{n_k})\boldsymbol{X}_{-j}^{(k)}\hat{\boldsymbol{\gamma}}_j^{(0)}$, $\tau_j^2 = \boldsymbol{\Sigma}_{\boldsymbol{\beta},j,j}^{(0)} - \boldsymbol{\Sigma}_{\boldsymbol{\beta},j,-j}^{(0)}\boldsymbol{\gamma}_j^{(0)} = \mathbb{E}[(x_j^{(0)})^2\psi''((\boldsymbol{x}^{(0)})^T\boldsymbol{\beta})] - \mathbb{E}[x_j^{(0)}\psi''((\boldsymbol{x}^{(0)})^T\boldsymbol{\beta})(x_{-j}^{(0)})^T]\boldsymbol{\gamma}_j^{(0)} = (\boldsymbol{\Theta}_{j,j})^{-1}$ stays away from zero because of Assumption 7.(iii). And we have

$$
\begin{aligned}
&\hat{\tau}_j^2 - \tau_j^2 \\
&= \underbrace{\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{i,k}(x_{ij}^{(k)})^2[\psi''((\boldsymbol{x}_i^{(k)})^T\hat{\boldsymbol{\beta}}) - \psi''((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\beta})]}_{(1)} \\
&\quad + \underbrace{\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{i,k}(x_{ij}^{(k)})^2\psi''((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\beta}) - \sum_{k\in\{0\}\cup\mathcal{A}_h}\frac{n_k}{n_{\mathcal{A}_h}+n_0}\mathbb{E}[(x_{ij}^{(k)})^2\psi''((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\beta})]}_{(2)} \\
&\quad + \underbrace{\sum_{k\in\{0\}\cup\mathcal{A}_h}\frac{n_k}{n_{\mathcal{A}_h}+n_0}\left\{\mathbb{E}[(x_{ij}^{(k)})^2\psi''((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\beta})] - \mathbb{E}[(x_{ij}^{(0)})^2\psi''((\boldsymbol{x}_i^{(0)})^T\boldsymbol{\beta})]\right\}}_{(3)} \\
&\quad + \underbrace{\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{k\in\{0\}\cup\mathcal{A}_h}(\boldsymbol{X}_j^{(k)})^T\mathrm{diag}(\{\psi''((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\beta}) - \psi''((\boldsymbol{x}_i^{(k)})^T\hat{\boldsymbol{\beta}})\}_{i=1}^{n_k})\boldsymbol{X}_{-j}^{(k)}\hat{\boldsymbol{\gamma}}_j^{(0)}}_{(4)} \\
&\quad - \underbrace{\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{k\in\{0\}\cup\mathcal{A}_h}(\boldsymbol{X}_j^{(k)})^T\mathrm{diag}(\{\psi''((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\beta})\}_{i=1}^{n_k})\boldsymbol{X}_{-j}^{(k)}(\hat{\boldsymbol{\gamma}}_j^{(0)} - \boldsymbol{\gamma}_j^{(0)})}_{(5)} \\
&\quad + \underbrace{\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{k\in\{0\}\cup\mathcal{A}_h}\left\{-(\boldsymbol{X}_j^{(k)})^T\mathrm{diag}(\{\psi''((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\beta})\}_{i=1}^{n_k})\boldsymbol{X}_{-j}^{(k)}\boldsymbol{\gamma}_j^{(0)} + \mathbb{E}[x_j^{(k)}\psi''((\boldsymbol{x}^{(k)})^T\boldsymbol{\beta})(x_{-j}^{(k)})^T]\boldsymbol{\gamma}_j^{(0)}\right\}}_{(6)} \\
&\quad + \underbrace{\sum_{k\in\{0\}\cup\mathcal{A}_h}\frac{n_k}{n_{\mathcal{A}_h}+n_0}\left\{-\mathbb{E}[x_j^{(k)}\psi''((\boldsymbol{x}^{(k)})^T\boldsymbol{\beta})(x_{-j}^{(k)})^T] + \mathbb{E}[x_j^{(0)}\psi''((\boldsymbol{x}^{(0)})^T\boldsymbol{\beta})(x_{-j}^{(0)})^T]\right\}\boldsymbol{\gamma}_j^{(0)}}_{(7)}.
\end{aligned}
$$

And we have the following control for each term:

$$
\begin{aligned}
|(1)| &\le \frac{1}{n_{\mathcal{A}_h}+n_0}\left|\sum_{i,k}(x_{ij}^{(k)})^2\psi'''((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\beta}+t(\boldsymbol{x}_i^{(k)})^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))(\boldsymbol{x}_i^{(k)})^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right| \\
&\le \sqrt{\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{i,k}(x_{ij}^{(k)})^4[\psi'''((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\beta}+t(\boldsymbol{x}_i^{(k)})^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))]^2}\cdot\sqrt{\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{i,k}[(\boldsymbol{x}_i^{(k)})^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))]^2} \\
&\lesssim \Re_1,
\end{aligned}
$$

with probability at least $1-n_0^{-1}$.

Since $\{(x_{ij}^{(k)})^2\psi''((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\beta})\}_{i,k}$ are independent sub-Gaussian variables with finite variance, by concentration inequality, $|(2)|\lesssim(n_{\mathcal{A}_h}+n_0)^{-1/2}$ with probability $1-\exp\{-(n_{\mathcal{A}_h}+n_0)\}$. Similarly, $|(6)|\lesssim(n_{\mathcal{A}_h}+n_0)^{-1/2}$ with probability $1-\exp\{-(n_{\mathcal{A}_h}+n_0)\}$.

$$
|(3)| \le \sup_{k\in\mathcal{A}_h}\left|\boldsymbol{\Sigma}_{\boldsymbol{\beta},j,j}^{(k)}-\boldsymbol{\Sigma}_{\boldsymbol{\beta},j,j}^{(0)}\right| \le h_{\max}.
$$

$$
\begin{aligned}
|(4)| &\le \left|\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{i,k}x_{ij}^{(k)}(\boldsymbol{x}_{-j}^{(k)})^T\hat{\boldsymbol{\gamma}}_j^{(0)}\cdot[\psi''((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\beta})-\psi''((\boldsymbol{x}_i^{(k)})^T\hat{\boldsymbol{\beta}})]\right| \\
&\le \left|\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{i,k}x_{ij}^{(k)}(\boldsymbol{x}_{-j}^{(k)})^T\boldsymbol{\gamma}_j^{(0)}\cdot\psi'''((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\beta}+t(\boldsymbol{x}_i^{(k)})^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))\cdot(\boldsymbol{x}_i^{(k)})^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right| \\
&\quad+\left|\frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{i,k}x_{ij}^{(k)}(\boldsymbol{x}_{-j}^{(k)})^T(\hat{\boldsymbol{\gamma}}_j^{(0)}-\boldsymbol{\gamma}_j^{(0)})\cdot\psi'''((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\beta}+t(\boldsymbol{x}_i^{(k)})^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))\cdot(\boldsymbol{x}_i^{(k)})^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right| \\
&\lesssim \Re_1,
\end{aligned}
$$

with probability at least $1-n_0^{-1}$.

$$
\begin{aligned}
|(5)| &\leq \left| \frac{1}{n_{\mathcal{A}_h} + n_0} \sum_{i,k} x_{ij}^{(k)} \psi''((\boldsymbol{x}_i^{(k)})^T \boldsymbol{\beta}) \cdot (\boldsymbol{x}_{i,-j}^{(k)})^T (\hat{\boldsymbol{\gamma}}_j^{(0)} - \boldsymbol{\gamma}_j^{(0)}) \right| \\
&\lesssim \sqrt{\frac{1}{n_{\mathcal{A}_h} + n_0} \sum_{i,k} [(\boldsymbol{x}_{i,-j}^{(k)})^T (\hat{\boldsymbol{\gamma}}_j^{(0)} - \boldsymbol{\gamma}_j^{(0)})]^2} \\
&\lesssim h_1^{1/2} \left( \frac{\log p}{n_0} \right)^{1/4} + h_1^{1/2} \mathfrak{R}_1^{1/2} + \mathfrak{R}_1
\end{aligned}
$$

with probability at least $1 - K_{\mathcal{A}_h} n_0^{-1}$ by (S.2.35).

$$
|(7)| \lesssim \sup_{k \in \mathcal{A}_h} \left| (\boldsymbol{\Sigma}_{\boldsymbol{\beta},-j,j}^{(k)} - \boldsymbol{\Sigma}_{\boldsymbol{\beta},-j,j}^{(0)}) \boldsymbol{\gamma}_j^{(0)} \right| \lesssim h_{\max}.
$$

Combine all the inequalities above to finish the proof of (S.2.37). Note that the bound of $|\hat{\tau}_j^{-2} - \tau_j^{-2}|$ follows because $\inf_j \tau_j^2 = \mathcal{O}(1)$.