# Transfer Learning under High-dimensional Generalized Linear Models

Ye Tian
Department of Statistics
Columbia University
and
Yang Feng
Department of Biostatistics, School of Global Public Health
New York University

**Abstract**

In this work, we study the transfer learning problem under high-dimensional generalized linear models (GLMs), which aim to improve the fit on *target* data by borrowing information from useful *source* data. Given which sources to transfer, we propose a transfer learning algorithm on GLM, and derive its $\ell_1/\ell_2$-estimation error bounds as well as a bound for a prediction error measure. The theoretical analysis shows that under certain conditions, when the target and source are sufficiently close to each other, these bounds could be improved over those of the classical penalized estimator using only target data. When we don't know which sources to transfer, an *algorithm-free* transferable source detection approach is introduced to detect informative sources. The detection consistency is proved under the high-dimensional GLM transfer learning setting. Extensive simulations and a real-data experiment verify the effectiveness of our algorithms. We summarize R codes for GLM transfer learning algorithms in a new R package `glmtrans`, which is available on CRAN.

*Keywords: Generalized linear models, transfer learning, high-dimensional data, Lasso, sparsity, negative transfer*

# 1 Introduction

With the fast development of science and technology, numerous machine learning methods have been applied in broad real-life applications successfully. Most of the statistical and machine learning algorithms work well only when sufficient training data with the same distribution as the test data is available. However, sometimes the collection of enough training data with the same distribution can be challenging and expensive in practice. Nevertheless, in many cases, data from a related domain and task are available, which is expected to be helpful with our target task. Human beings are good at transfer knowledge from other related domains. For example, many children under two years are viewing television on a daily basis. And they are found to be able to imitate actions presented on television using the corresponding real-world objects (Barr, 2010). The two-dimensional input from television helps them learn about objects in three-dimensional world. This scenario motivates *transfer learning*, which is often used to improve the learner from a specific domain and task by borrowing information from other related ones (Pan and Yang, 2009; Torrey and Shavlik, 2010; Weiss et al., 2016). We often call the main task and domain of predictors as *target*, and call the related tasks and domains as *sources*. The similarity shared between target and sources could happen on either the task or domain. Other transfer learning applications include the recommendation system (Bastani, 2020) (transferring information from clicking data to purchase data) and automatic driving (transferring information between types of terrain and different cities) among others. For more details and applications of transfer learning, refer to the survey papers by Pan and Yang (2009) and Weiss et al. (2016).

Nowadays, *high-dimensional* data become more frequently collected and analyzed in many research fields such as genomics and biomedical imaging. There are numerous ap-

proaches proposed to tackle this problem, including Tibshirani (1996); Fan and Li (2001); Zou and Hastie (2005); Zou (2006); Zhang (2010) among others. We notice that, the inherent difficulty of high-dimensional problems, that is, limited sample size with a huge number of variables, makes transfer learning potentially very useful. Recently, there are some works exploring transfer learning under the high-dimensional setting. Bastani (2020) studied the single-source case when the target data follows a high-dimensional linear model while the source sample size is much larger than the dimension $p$. A two-step transfer learning algorithm was developed and the estimation error bound was proved. Li et al. (2020a) further weakened the assumptions and explored the multi-source high-dimensional linear regression problem. The $\ell_2$-estimation error bound under $\ell_q$-regularization ($q \in [0, 1]$) was derived and proved to be minimax optimal under some conditions. In Li et al. (2020b), the analysis was extended to the Gaussian graphical models with false discovery rate control. Other related research on transfer learning with theoretical guarantee includes the nonparametric classification model (Cai and Wei, 2021; Reeve et al., 2021) and the analysis under general functional classes via transfer exponents (Hanneke and Kpotufe, 2020a,b) etc. Besides, during the past few years, there have been some related works studying parameter sharing under the regression setting. For instance, Chen et al. (2015) and Zheng et al. (2019) developed the so-called "data enriched model" for linear and logistic regression under a single-source setting, where the properties of the oracle tuned estimator with a quadratic penalty were studied. Gross and Tibshirani (2016) and Ollier and Viallon (2017) explored the so-called "data shared Lasso" under the multi-task learning setting, where the $\ell_1$ penalties of all contrasts are considered.

In this work, we contribute to the high-dimensional transfer learning framework in the following perspectives. First, we extend the results of Bastani (2020) and Li et al.

(2020a), by proposing multi-source transfer learning algorithms on generalized linear models (GLMs). Li et al. (2020a) studied the linear models only. Bastani (2020) solved the single-source problem on GLM, where only the target problem is high-dimensional and the source sample size is assumed to be much larger than the dimension. We consider the multi-source case. In addition, both target and source data are high-dimensional, which created more challenges. Besides, Bastani (2020) had more stringent model requirements. Concretely, they assume the second-order derivative of the cumulant function to be *strongly* convex, which excludes the Binomial families with unbounded predictors. We relax this condition and only impose the *strict* convexity of the cumulant function. Therefore the Binomial families with unbounded predictors are included. The theoretical analysis shows that under certain conditions, when the target and source are sufficiently close to each other, the estimation error bound could be improved over that of the classical penalized estimator using only target data. To the best of our knowledge, this is the first study of the multi-source transfer learning framework under the high-dimensional GLM setting. Second, as we mentioned, transferring sources that are close to the target can bring benefits. However, some sources might be far away from the target and including them into the model can damage the model performance. This phenomenon is often called *negative transfer* in literature (Torrey and Shavlik, 2010; Weiss et al., 2016). Pan and Yang (2009) proposed this questions as "when to transfer". We will show the impact of negative transfer in simulation studies in Section 4.1. To avoid this issue, we develop an *algorithm-free* transferable source detection algorithm, which can help identify informative sources. And with certain conditions satisfied, theoretical guarantees are provided under the GLM transfer learning framework. More discussions on negative transfer and detection algorithm can be found in Section 2.4.

The rest of this paper is organized as follows. Section 2 first introduces GLM basics and transfer learning settings under high-dimensional GLM, then presents the oracle algorithm and the transferable source detection algorithm. Section 3 provides the theoretical analysis on the algorithms, including $\ell_2$-estimation error bounds of the oracle algorithm and detection consistency property of the transferable source detection algorithm under certain conditions. We conduct extensive simulations and a real-data study in Section 4, and the results demonstrate the effectiveness of our GLM transfer learning algorithms. In Section 5, we review our contributions and shed light on some interesting future research directions. The proof of Theorem 1 is available in Appendix A. Some additional simulations results and the remaining proofs are relegated to supplementary materials.

# 2 Methodology

We first introduce some useful notations which will be used throughout the paper. We use bold capitalized letters (e.g. $\boldsymbol{X}$, $\boldsymbol{A}$) to denote matrices, and use bold little letters (e.g. $\boldsymbol{x}$, $\boldsymbol{y}$) to denote vectors. For a vector $\boldsymbol{x} = (x_1, \ldots, x_p)^T$, we denote its Euclidean norm or $\ell_2$-norm as $\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^p x_i^2}$. We also denote $\|\boldsymbol{x}\|_0 = \#\{j : x_j \neq 0\}$. For a matrix $\boldsymbol{A}_{p\times q} = [a_{ij}]_{p\times q}$, its 1-norm, 2-norm, $\infty$-norm and max-norm are defined as $\|\boldsymbol{A}\|_1 = \sup_j \sum_{i=1}^p |a_{ij}|$, $\|\boldsymbol{A}\|_2 = \max_{\boldsymbol{x}:\|\boldsymbol{x}\|_2=1} \|\boldsymbol{A}\boldsymbol{x}\|_2$, $\|\boldsymbol{A}\|_\infty = \sup_i \sum_{j=1}^q |a_{ij}|$ and $\|\boldsymbol{A}\|_{\max} = \sup_{i,j} |a_{ij}|$, respectively. For two non-zero real sequences $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, we use $a_n \ll b_n$, $b_n \gg a_n$ or $a_n = o(b_n)$ to represent $|a_n/b_n| \to 0$. And $a_n \lesssim b_n$ or $a_n = \mathcal{O}(b_n)$ means $\sup_n |a_n/b_n| < \infty$. Expression $a_n \asymp b_n$ means that $a_n/b_n$ converges to some positive constant as $n \to \infty$. For two random variable sequences $\{x_n\}_{n=1}^\infty$ and $\{y_n\}_{n=1}^\infty$, notation $x_n \lesssim_p y_n$ or $x_n = \mathcal{O}_p(y_n)$ means that for any $\epsilon > 0$, there exists a positive constant $M$ such that $\sup_n \mathbb{P}(|x_n/y_n| > M) \leq \epsilon$. And

for two real numbers $a$ and $b$, we use $a \vee b$ and $a \wedge b$ to represent $\max(a, b)$ and $\min(a, b)$, respectively. Without specific notes, the expectation $\mathbb{E}$, variance Var, and covariance Cov are calculated based on all randomness.

## 2.1 Generalized linear models (GLMs)

Given the predictors $\boldsymbol{x} \in \mathbb{R}^p$, if the response $y$ follows the generalized linear models (GLMs), then its conditional distribution takes the form

$$y|\boldsymbol{x} \sim \mathbb{P}(y|\boldsymbol{x}) = \rho(y) \exp\{y\boldsymbol{x}^T\boldsymbol{w} - \psi(\boldsymbol{x}^T\boldsymbol{w})\},$$

where $\boldsymbol{w} \in \mathbb{R}^p$ is the coefficient, $\rho$ and $\psi$ are some known univariate functions. $\psi'(\boldsymbol{x}^T\boldsymbol{w}) = \mathbb{E}(y|\boldsymbol{x})$ is called the *inverse link function* (McCullagh and Nelder, 1989). Another important property is that $\mathrm{Var}(y|\boldsymbol{x}) = \psi''(\boldsymbol{x}^T\boldsymbol{w})$, which follows from the fact that the distribution belongs to the exponential family. It is $\psi$ that characterizes different GLMs. For example, in linear model with Gaussian noise, we have a continuous response $y$ and $\psi(u) = \frac{1}{2}u^2$; in the logistic regression model, $y$ is binary and $\psi(u) = \log(1 + e^u)$; and in Poisson regression model, $y$ is a nonnegative integer and $\psi(u) = e^u$. For most GLMs, $\psi$ is strictly convex and infinitely differentiable.

## 2.2 Target data, source data, and transferring level

In this paper, we consider the following multi-source transfer learning problem. Suppose we have the *target* data $(\boldsymbol{X}^{(0)}, \boldsymbol{y}^{(0)}) = \{\boldsymbol{x}_i^{(0)}, y_i^{(0)}\}_{i=1}^{n_0}$ and *source* data $\{(\boldsymbol{X}^{(k)}, \boldsymbol{y}^{(k)})\}_{k=1}^K = \{\{(\boldsymbol{x}_i^{(k)}, y_i^{(k)})\}_{i=1}^{n_k}\}_{k=1}^K$, where $\boldsymbol{X}^{(k)} \in \mathbb{R}^{n_k \times p}$, $\boldsymbol{y}^{(k)} \in \mathbb{R}^{n_k}$ for $k = 0, \ldots, K$. The goal is to transfer useful information from source data to obtain a better model for the target data. We assume the responses in the target and source data all follow the generalized linear

model, that is,

$$y^{(k)}|\boldsymbol{x} \sim \mathbb{P}(y|\boldsymbol{x}) = \rho(y)\exp\{y\boldsymbol{x}^T\boldsymbol{w}^{(k)} - \psi(\boldsymbol{x}^T\boldsymbol{w}^{(k)})\}, \tag{1}$$

for $k = 0, \ldots, K$, with possibly different coefficient $\boldsymbol{w}^{(k)} \in \mathbb{R}^p$, the predictor $\boldsymbol{x} \in \mathbb{R}^p$, and some known univariate functions $\rho$ and $\psi$. Denote the target parameter as $\boldsymbol{\beta} = \boldsymbol{w}^{(0)}$. Suppose the target model is $\ell_0$-sparse, which satisfies $\|\boldsymbol{\beta}\|_0 = s \ll p$. This means that only $s$ of the $p$ variables contribute to the target response. Intuitively, if $\boldsymbol{w}^{(k)}$ is close to $\boldsymbol{\beta}$, the $k$-th source could be useful for transfer learning.

Define the $k$-th contrast $\boldsymbol{\delta}^{(k)} = \boldsymbol{\beta} - \boldsymbol{w}^{(k)}$ and we say $\|\boldsymbol{\delta}^{(k)}\|_1$ is the *transferring level* of source $k$. And we define the *level-$h$ transferring set* $\mathcal{A}_h = \{k : \|\boldsymbol{\delta}^{(k)}\|_1 \leq h\}$ as the set of sources which has transferring level lower than $h$. Note that in general, $h$ can be any positive values and different $h$ values define different $\mathcal{A}_h$ set. However, in our interested regime, $h$ should be reasonably small to guarantee that there are some benefits for transferring sources in $\mathcal{A}_h$. Denote $n_{\mathcal{A}_h} = \sum_{k\in\mathcal{A}_h} n_k$, $\alpha_k = \frac{n_k}{n_{\mathcal{A}}+n_0}$ for $k \in \{0\} \cup \mathcal{A}_h$ and $K_{\mathcal{A}_h} = |\mathcal{A}_h|$.

Note that in (1), we assume GLMs in the target and all sources share the same function $\psi$. After a careful examination of our proofs for theoretical properties in Section 3, we find that these theoretical results still hold even when the target and each source have their own function $\psi$, as long as these GLMs satisfy Assumptions 1 and 3 (to be presented in Section 3.1). It means that transferring information across different GLM families is possible. For simplicity, in the following discussion we assume GLMs in the target and all sources have the same function $\psi$, i.e. they belong to the same GLM family.

## 2.3 Two-step GLM transfer learning

We first introduce a transfer learning algorithm on GLMs, which can be applied to transfer all sources in a given index set $\mathcal{A}$. The algorithm we propose follow from the idea in

Bastani (2020) and Li et al. (2020a), which we call a *two-step transfer learning algorithm*. The main idea is to first transfer the information from transferable sources to obtain a rough estimator, then correct the bias in the second step using the target data. The algorithm with a given set $\mathcal{A}$ ($\mathcal{A}$-Trans-GLM) is presented in Algorithm 1. We fit a GLM Lasso model with target data and all source data in $\mathcal{A}$ first, then fit the contrast in the second step by another $\ell_1$-regularization. The transferring step could be understood as to find the solution of the following equation w.r.t. $\boldsymbol{w}^{\mathcal{A}} \in \mathbb{R}^p$:

$$\sum_{k\in\{0\}\cup\mathcal{A}} \left[ (\boldsymbol{X}^{(k)})^T \boldsymbol{y}^{(k)} - \sum_{i=1}^{n_k} \psi'((\boldsymbol{w}^{\mathcal{A}})^T \boldsymbol{x}_i^{(k)}) \boldsymbol{x}_i^{(k)} \right] = \mathbf{0}_p,$$

which converges to the solution of its population version under certain conditions

$$\sum_{k\in\{0\}\cup\mathcal{A}} \alpha_k \mathbb{E}\left\{ [\psi'((\boldsymbol{w}^{\mathcal{A}})^T \boldsymbol{x}^{(k)}) - \psi'((\boldsymbol{w}^{(k)})^T \boldsymbol{x}^{(k)})] \boldsymbol{x}^{(k)} \right\} = \mathbf{0}_p, \tag{2}$$

where $\alpha_k = \frac{n_k}{n_{\mathcal{A}}+n_0}$. Notice that for the linear case, $\boldsymbol{w}^{\mathcal{A}}$ can be explicitly expressed as a linear transformation of the true parameter $\boldsymbol{w}^{(k)}$, i.e.

$$\boldsymbol{w}^{\mathcal{A}} = \boldsymbol{\Sigma}^{-1} \sum_{k\in\{0\}\cup\mathcal{A}} \alpha_k \boldsymbol{\Sigma}^{(k)} \boldsymbol{w}^{(k)},$$

where $\boldsymbol{\Sigma}^{(k)} = \mathbb{E}[\boldsymbol{x}^{(k)}(\boldsymbol{x}^{(k)})^T]$ and $\boldsymbol{\Sigma} = \sum_{k\in\{0\}\cup\mathcal{A}} \alpha_k \mathbb{E}[\boldsymbol{x}^{(k)}(\boldsymbol{x}^{(k)})^T]$ (Li et al., 2020a).

To help readers better understand the algorithm, we draw a schematic in Section S.1.1 of the supplementary material. We refer interested readers who wants to get more intuitions to that.

## 2.4 Transferable source detection

As we described in Section 2.3, Algorithm 1 can be applied when we know which sources to transfer. However, in practice, we often don't know this information as a priori. Transfer-

**Algorithm 1:** $\mathcal{A}$-Trans-GLM

**Input:** target data $(\boldsymbol{X}^{(0)}, \boldsymbol{y}^{(0)})$, source data $\{(\boldsymbol{X}^{(k)}, \boldsymbol{y}^{(k)})\}_{k=1}^{K} = \{\{(\boldsymbol{x}_i^{(k)}, y_i^{(k)})\}_{i=1}^{n_k}\}_{k=1}^{K}$, penalty parameters $\lambda_{\boldsymbol{w}}$ and $\lambda_{\boldsymbol{\delta}}$, transferring set $\mathcal{A}$

**Output:** the estimated coefficient vector $\hat{\boldsymbol{\beta}}$

1 **Transferring step**: Compute

$$\hat{\boldsymbol{w}}^{\mathcal{A}} \leftarrow \arg\min_{\boldsymbol{w}} \left\{ \frac{1}{n_{\mathcal{A}}+n_0} \sum_{k\in\{0\}\cup\mathcal{A}} \left[ -(\boldsymbol{y}^{(k)})^T \boldsymbol{X}^{(k)} \boldsymbol{w} + \sum_{i=1}^{n_k} \psi(\boldsymbol{w}^T \boldsymbol{x}_i^{(k)}) \right] + \lambda_{\boldsymbol{w}} \|\boldsymbol{w}\|_1 \right\}$$

2 **Debiasing step**: Compute

$$\hat{\boldsymbol{\delta}}^{\mathcal{A}} \leftarrow \arg\min_{\boldsymbol{\delta}} \left\{ -\frac{1}{n_0} (\boldsymbol{y}^{(0)})^T \boldsymbol{X}^{(0)} (\hat{\boldsymbol{w}}^{\mathcal{A}} + \boldsymbol{\delta}) + \frac{1}{n_0} \sum_{i=1}^{n_0} \psi((\hat{\boldsymbol{w}}^{\mathcal{A}} + \boldsymbol{\delta})^T \boldsymbol{x}_i^{(0)}) + \lambda_{\boldsymbol{\delta}} \|\boldsymbol{\delta}\|_1 \right\}$$

3 Let $\hat{\boldsymbol{\beta}} \leftarrow \hat{\boldsymbol{w}}^{\mathcal{A}} + \hat{\boldsymbol{\delta}}^{\mathcal{A}}$

4 Output $\hat{\boldsymbol{\beta}}$

ring specific sources might not improve the performance compared to the fitted model using target data only. And sometimes, this can even lead to a worse performance. In transfer learning, we say *negative transfer* happens when the source data leads to a reduced performance of learning on the target task (Pan and Yang, 2009; Torrey and Shavlik, 2010; Weiss et al., 2016). How to avoid negative transfer has become an increasingly popular research topic.

We propose a simple, *algorithm-free*, and *data-driven* method to determine an informative transferring set $\widehat{\mathcal{A}}$. We call this approach as transferable source *detection* algorithm and refer it as Trans-GLM. In Section 3.2, we will show that $\widehat{\mathcal{A}} = \mathcal{A}_h$ for some specific $h$, if certain conditions are met. We will also see that under these conditions, transferring with $\widehat{\mathcal{A}}$ will lead to a faster convergence rate compared to Lasso fitted on the target data,

when target sample size $n_0$ falls into some regime. That is why we call this algorithm as the *transferable* source detection algorithm.

The main idea of this detection algorithm can be described as follows. First, we divide the target data into three folds, that is, $(\boldsymbol{X}^{(0)}, \boldsymbol{y}^{(0)}) = \{(\boldsymbol{X}^{(0)[r]}, \boldsymbol{y}^{(0)[r]})\}_{r=1}^{3}$ [1]. Second, run the transferring step on each source data and every two folds of target data. Then, for a given loss function, we calculate its value on the left-out fold of target data and get the average cross-validation loss $\hat{L}_0^{(k)}$ for each source. As benchmarks, we also fit Lasso on every choice of two folds of target data and calculate the loss on the remaining fold. The average cross-validation loss $\hat{L}_0^{(0)}$ is calculated as the loss of target. Finally, the difference between $\hat{L}_0^{(k)}$ and $\hat{L}_0^{(0)}$ will be calculated and compared with some threshold, and sources with difference less than the threshold will be recruited into $\widehat{\mathcal{A}}$.

Under the GLM setting, a natural loss function is the negative log-likelihood. For convenience, suppose $n_0$ is divisible by 3. According to (1), for any coefficient estimator $\boldsymbol{w}$, the average of negative log-likelihood on the $r$-th fold of target data $(\boldsymbol{X}^{(0)[r]}, \boldsymbol{y}^{(0)[r]})$ is

$$\hat{L}_0^{[r]}(\boldsymbol{w}) = -\frac{1}{n_0/3}\sum_{i=1}^{n_0/3} \log \rho(y_i^{(0)[r]}) - \frac{1}{n_0/3}(\boldsymbol{y}^{(0)[r]})^T \boldsymbol{X}^{(0)}\boldsymbol{w} + \frac{1}{n_0/3}\sum_{i=1}^{n_0/3} \psi(\boldsymbol{w}^T \boldsymbol{x}_i^{(0)[r]}). \quad (3)$$

The detailed algorithm is presented as Algorithm 2. It's important to point out that Algorithm 2 **does not** require the input of $h$. We will show that $\widehat{\mathcal{A}} = \mathcal{A}_h$ for some specific $h$, if certain conditions hold, in Section 3.2. Furthermore, under these conditions, transferring with $\widehat{\mathcal{A}}$ will lead to a faster convergence rate compared to Lasso fitted on the target data, when target sample size $n_0$ falls into some regime.

[1] We choose three folds only for convenience. It can actually be replaced by any finite number of folds.

**Algorithm 2:** Trans-GLM

**Input:** target data $(\boldsymbol{X}^{(0)}, \boldsymbol{y}^{(0)})$, all source data $\{(\boldsymbol{X}^{(k)}, \boldsymbol{y}^{(k)})\}_{k=1}^{K} = \{\{(\boldsymbol{x}_i^{(k)}, \boldsymbol{y}_i^{(k)})\}_{i=1}^{n_k}\}_{k=1}^{K}$, a constant $C_0 > 0$, penalty parameters $\{\{\lambda^{(k)[r]}\}_{k=0}^{K}\}_{r=1}^{3}$

**Output:** the estimated coefficient vector $\hat{\boldsymbol{\beta}}$, and the determined transferring set $\widehat{\mathcal{A}}$

**Transferable source detection**: Randomly divide $(\boldsymbol{X}^{(0)}, \boldsymbol{y}^{(0)})$ into three sets of equal size, that is $(\boldsymbol{X}^{(0)}, \boldsymbol{y}^{(0)}) = \{(\boldsymbol{X}^{(0)[i]}, \boldsymbol{y}^{(0)[i]})\}_{i=1}^{3}$

**for** $r = 1$ **to** 3 **do**

  $\hat{\boldsymbol{\beta}}^{(0)[r]} \leftarrow$ fit the Lasso on $\{(\boldsymbol{X}^{(0)[i]}, \boldsymbol{y}^{(0)[i]})\}_{i=1}^{3} \backslash (\boldsymbol{X}^{(0)[r]}, \boldsymbol{y}^{(0)[r]})$ with penalty parameter $\lambda^{(0)[r]}$

  $\hat{\boldsymbol{\beta}}^{(k)[r]} \leftarrow$ run step 1 in Algorithm 1 with $(\{(\boldsymbol{X}^{(0)[i]}, \boldsymbol{y}^{(0)[i]})\}_{i=1}^{3} \backslash (\boldsymbol{X}^{(0)[r]}, \boldsymbol{y}^{(0)[r]})) \cup (\boldsymbol{X}^{(k)}, \boldsymbol{y}^{(k)})$ and penalty parameter $\lambda^{(k)[r]}$ for all $k \neq 0$

  Calculate the loss function $\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]})$ on $(\boldsymbol{X}^{(0)[r]}, \boldsymbol{y}^{(0)[r]})$ for $k = 1, \ldots, K$

**end**

$\hat{L}_0^{(k)} \leftarrow \sum_{r=1}^{3} \hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]})/3$, $\hat{L}_0^{(0)} \leftarrow \sum_{r=1}^{3} \hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(0)[r]})/3$, $\hat{\sigma} = \sqrt{\sum_{r=1}^{3}(\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]}) - \hat{L}_0^{(0)})^2/3}$

$\widehat{\mathcal{A}} \leftarrow \{k \neq 0 : \hat{L}_0^{(k)} - \hat{L}_0^{(0)} \leq C_0(\hat{\sigma} \vee 0.01)\}$

**$\widehat{\mathcal{A}}$-Trans-GLM**: $\hat{\boldsymbol{\beta}} \leftarrow$ run Algorithm 1 using $\{(\boldsymbol{X}^{(k)}, \boldsymbol{y}^{(k)})\}_{k \in \{0\} \cup \widehat{\mathcal{A}}}$

Output $\hat{\boldsymbol{\beta}}$ and $\widehat{\mathcal{A}}$

# 3 Theory

In this section, we present a theoretical study on the two proposed algorithms. Section 3.1 provides a detailed analysis of Algorithm 1 with transferring set $\mathcal{A}_h$, which we denote as $\mathcal{A}_h$-Trans-GLM. Section 3.2 introduces certain conditions, under which we show that the

transferring set $\widehat{\mathcal{A}}$ detected by Algorithm 2 (Trans-GLM) is equal to $\mathcal{A}_h$ for some $h$ with high probability. The proof of Theorem 1 can be found in Appendix A. For all the other proofs, refer to supplementary materials.

## 3.1 Theory on $\mathcal{A}_h$-Trans-GLM

We first impose some common assumptions about GLM.

**Assumption 1.** *$\psi$ is infinitely differentiable and strictly convex. We call a second-order differentiable function $\psi$ strictly convex if $\psi''(x) > 0$.*

**Assumption 2.** *For any $\boldsymbol{a} \in \mathbb{R}^p$, $\boldsymbol{a}^T \boldsymbol{x}_i^{(k)}$'s are i.i.d. $\kappa_u \|\boldsymbol{a}\|_2^2$-subGaussian variables with zero mean for all $k = 0, \ldots, K$, where $\kappa_u$ is a positive constant. Denote the covariance matrix of $\boldsymbol{x}^{(k)}$ as $\boldsymbol{\Sigma}^{(k)}$, with $\inf_k \lambda_{\min}(\boldsymbol{\Sigma}^{(k)}) \geq \kappa_l > 0$, where $\kappa_l$ is a positive constant.*

**Assumption 3.** *At least one of the following assumptions hold: ($M_\psi$, $U$ and $\bar{U}$ are some positive constants)*

*(i)* $\sup_k \psi''(\boldsymbol{x}^{(k)}) \leq M_\psi < \infty$ *a.s.;*

*(ii)* $\sup_k \|\boldsymbol{x}^{(k)}\|_\infty \leq U < \infty$ *a.s.,* $\sup_k \sup_{|z| \leq \bar{U}} \psi''((\boldsymbol{x}^{(k)})^T \boldsymbol{w}^{(k)} + z) \leq M_\psi < \infty$ *a.s.*

Assumption 1 imposed the *strict convexity* and differentiability of $\psi$, which is satisfied by many popular distribution families, such as Gaussian, binomial, and Poisson distributions. Note that we do not require $\psi$ to be *strongly convex* (that is, $\exists C > 0$, such that $\psi''(x) > C$), which relaxes Assumption 4 in Bastani (2020). It is easy to verify that $\psi$ in logistic regression is in general not strongly convex with unbounded predictors. Assumption 2 requires the predictors in each source to be subGaussian with a well-behaved correlation

structure. Assumption 3 is motivated by Assumption (GLM 2) in the full-length version of Negahban et al. (2009), which is imposed to restrict $\psi''$ in a bounded region in some sense. Note that linear regression and logistic regression satisfy condition (i), while Poisson regression with coordinate-wise bounded predictors and $\ell_1$-bounded coefficients satisfies condition (ii).

Besides these common conditions on GLM, as discussed in Section 2.3, to guarantee the success of $\mathcal{A}_h$-Trans-GLM, we have to make sure that the estimator from the transferring step is close enough to $\boldsymbol{\beta}$. We first introduce the following assumption, which guarantees $\boldsymbol{w}^{\mathcal{A}_h}$ defined in (2) with $\mathcal{A} = \mathcal{A}_h$ is close to $\boldsymbol{\beta}$.

**Assumption 4.** *Denote* $\widetilde{\boldsymbol{\Sigma}}_h = \sum_{k\in\{0\}\cup\mathcal{A}_h} \alpha_k \mathbb{E}\Big[\int_0^1 \psi''((\boldsymbol{x}^{(k)})^T\boldsymbol{w}^{(k)} + t(\boldsymbol{x}^{(k)})^T(\boldsymbol{w}^{\mathcal{A}_h} - \boldsymbol{w}^{(k)}))dt \cdot \boldsymbol{x}^{(k)}(\boldsymbol{x}^{(k)})^T\Big]$ *and* $\widetilde{\boldsymbol{\Sigma}}_h^{(k)} = \mathbb{E}\Big[\int_0^1 \psi''((\boldsymbol{x}^{(k)})^T\boldsymbol{\beta} + t(\boldsymbol{x}^{(k)})^T(\boldsymbol{w}^{(k)} - \boldsymbol{\beta}))dt \cdot \boldsymbol{x}^{(k)}(\boldsymbol{x}^{(k)})^T\Big]$. *It holds that* $\sup_{k\in\{0\}\cup\mathcal{A}_h} \|\widetilde{\boldsymbol{\Sigma}}_h^{-1}\widetilde{\boldsymbol{\Sigma}}_h^{(k)}\|_1 < \infty$.

In the linear case, this assumption can be further simplified as a restriction on heterogeneity between target predictors and source predictors. More discussions can be found in Condition 4 of Li et al. (2020a). Now, we first present the following main result for $\mathcal{A}_h$-Trans-GLM algorithm with Assumption 4.

**Theorem 1** ($\ell_1/\ell_2$-estimation error of $\mathcal{A}_h$-Trans-GLM with Assumption 4)**.** *Assume Assumptions 1, 2 and 4 hold. Suppose* $n_0 \geq C\log p$ *and* $n_{\mathcal{A}_h} \geq Cs\log p$, *where* $C > 0$ *is a sufficiently large constant. We take* $\lambda_{\boldsymbol{w}} = C_{\boldsymbol{w}}\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}$, *where* $C_{\boldsymbol{w}} > 0$ *is a sufficiently large constant.*

*(i) 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:*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \lesssim \left(\frac{s\log p}{n_0}\right)^{1/4}\left(\frac{s\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/4} + \left(\frac{s\log p}{n_0}\right)^{1/4}\left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/8} h^{1/4}$$

$$+\left(\frac{\log p}{n_0}\right)^{1/4} h^{1/2}, \tag{4}$$

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 \lesssim s\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}+\left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/4}\sqrt{sh}+h\left(1+\sqrt{\frac{s\log p}{n_{\mathcal{A}_h}+n_0}}\right) \tag{5}$$

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

*(ii) Assume Assumption 3.(i) holds or Assumption 3.(ii) with* $n_{\mathcal{A}_h} \geq CU^2\bar{U}^{-2}s^2\log p$, $h \leq C'U^{-1}\bar{U}$ *for some* $C, C' > 0$ *holds. If we take* $\lambda_{\boldsymbol{\delta}} = C_{\boldsymbol{\delta}}\left[s\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}+\left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/4}\sqrt{sh}+h\right]$, *where* $C_{\boldsymbol{\delta}} > 0$ *is a sufficiently large constant:*

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2 \lesssim \left(\frac{s\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/2}+\left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/4}h^{1/2}+\left[\left(\frac{\log p}{n_0}\right)^{1/4}h^{1/2}+\left(\frac{s\log p}{n_0}\right)^{1/2}\right]\wedge h, \tag{6}$$

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

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

**Remark 1.** *When* $h \ll s\sqrt{\frac{\log p}{n_0}}$ *and* $n_{\mathcal{A}_h} \gg n_0$, *the upper bounds in (4) and (5) are better than the classical Lasso* $\ell_2$*-bound* $\mathcal{O}_p\left(\sqrt{\frac{s\log p}{n_0}}\right)$ *and* $\ell_1$*-bound* $\mathcal{O}_p\left(s\sqrt{\frac{\log p}{n_0}}\right)$ *using only target data. This requirement of* $h$ *matches the results under the linear case in Li et al. (2020a). The bound in (6) is better than the bound in (4) if* $h \ll \left(\frac{s\log p}{n_0}\right)^{1/3}\left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{2/3}$. *This result is intuitive because the scale of* $\lambda_{\boldsymbol{\delta}}$ *is larger in (ii), which leads to a smaller adjustment on the estimator* $\hat{\boldsymbol{w}}^{\mathcal{A}_h}$ *in the debiasing step. When* $h$ *is small, the debiasing step could be too aggressive when taking the same* $\lambda_{\boldsymbol{\delta}}$ *as (i). In the extreme case where* $h = 0$, *the choice of* $\lambda_{\boldsymbol{\delta}}$ *in condition (ii) will lead to a bound* $\mathcal{O}_p\left(\sqrt{\frac{s\log p}{n_{\mathcal{A}_h}+n_0}}\right)$, *which is the oracle rate, while using the results in (i) will lead to a suboptimal rate* $\mathcal{O}_p\left(\left(\frac{s\log p}{n_0}\right)^{1/4}\left(\frac{s\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/4}\right)$.

Next, we present a similar upper bound, which is weaker than the bound above but without requiring Assumption 4.

**Theorem 2** ($\ell_1/\ell_2$-estimation error of $\mathcal{A}_h$-Trans-GLM without Assumption 4)**.** *Assume Assumptions 1 and 2 hold. Suppose $n_0 \geq C \log p$ and $n_{\mathcal{A}_h} \geq Cs \log p$, where $C > 0$ is a sufficiently large constant. We take $\lambda_{\boldsymbol{w}} = C_{\boldsymbol{w}} \left( \sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}} + h \right)$, where $C_{\boldsymbol{w}} > 0$ is a sufficiently large constant.*

*(i) 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, then*

$$\begin{aligned}
\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 &\lesssim \left( \frac{s \log p}{n_0} \right)^{1/4} \left( \frac{s \log p}{n_{\mathcal{A}_h} + n_0} \right)^{1/4} + \left( \frac{s \log p}{n_0} \right)^{1/4} s^{1/4} h^{1/2} \\
&\quad + \left( \frac{s \log p}{n_0} \right)^{1/4} (sh)^{1/4} \left( \frac{\log p}{n_{\mathcal{A}_h} + n_0} \right)^{1/8}, \\
\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 &\lesssim s \sqrt{\frac{\log p}{n_{\mathcal{A}_h} + n_0}} + \left( \frac{\log p}{n_{\mathcal{A}_h} + n_0} \right)^{1/4} \sqrt{sh} + sh,
\end{aligned}$$

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

*(ii) Assume Assumption 3.(i) holds or Assumption 3.(ii) with $n_{\mathcal{A}_h} \geq C U^2 \bar{U}^{-2} s^2 \log p$, $h \leq C' \bar{U} U^{-1} s^{-1}$ for some $C$, $C' > 0$ holds. If we take $\lambda_{\boldsymbol{\delta}} = C_{\boldsymbol{\delta}} \Big[ s \sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}} + \left( \frac{\log p}{n_{\mathcal{A}_h}+n_0} \right)^{1/4} \sqrt{sh} + h \Big]$, where $C_{\boldsymbol{\delta}} > 0$ is a sufficiently large constant:*

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

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

**Remark 2.** *When $h \ll \sqrt{\frac{\log p}{n_0}}$ and $n_{\mathcal{A}_h} \gg n_0$, the upper bounds in (i) and (ii) are better than the classical Lasso bound $\mathcal{O}_p\left(\sqrt{\frac{\log p}{n_0}}\right)$ with target data. In addition, in this regime, the $\ell_2$-bound in (ii) is better than the $\ell_2$-bound in (i).*

Comparing the results in Theorems 1 and 2, we know that with Assumption 4, we could get a better $\ell_2$-estimation error bound.

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 3** (Bound of a prediction error measure for $\mathcal{A}_h$-Trans-GLM with Assumption 4)**.** *We impose the same conditions in Theorem 1. We take $\lambda_{\boldsymbol{w}} = C_{\boldsymbol{w}}\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}$, where $C_{\boldsymbol{w}} > 0$ is a sufficiently large constant.*

*(i) Assume Assumption 3.(i) holds or Assumption 3.(ii) with $h \le C'U^{-1}\bar{U}$ 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:*

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

$$+ h\left(\frac{\log p}{n_0}\right)^{1/2}\left(1+\sqrt{\frac{s\log p}{n_{\mathcal{A}_h}+n_0}}\right)$$

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

*(ii) Assume Assumption 3.(i) holds or Assumption 3.(ii) with* $n_{\mathcal{A}_h} \geq CU^2\bar{U}^{-2}s^2\log p$, $h \leq C'\bar{U}U^{-1}s^{-1}$ *for some* $C$, $C' > 0$ *holds. If we take* $\lambda_{\boldsymbol{\delta}} = C_{\boldsymbol{\delta}}\Big[s\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}} + \left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/4}\sqrt{sh}+h\Big]$, *where* $C_{\boldsymbol{\delta}} > 0$ *is a sufficiently large constant:*

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

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

**Remark 3.** *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 4** (Bound of a prediction error measure for $\mathcal{A}_h$-Trans-GLM without Assumption 4)**.** *We impose the same conditions in Theorem 2. We take* $\lambda_{\boldsymbol{w}} = C_{\boldsymbol{w}}\left(\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}+h\right)$, *where* $C_{\boldsymbol{w}} > 0$ *is a sufficiently large constant.*

*(i) 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:*

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

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

*(ii) Assume Assumption 3.(i) holds or Assumption 3.(ii) with* $n_{\mathcal{A}_h} \geq CU^2\bar{U}^{-2}s^2\log p$, $h \leq C'\bar{U}U^{-1}s^{-1}$ *for some* $C' > 0$ *for some* $C$, $C' > 0$ *holds. If we take* $\lambda_{\boldsymbol{\delta}} = C_{\boldsymbol{\delta}}\left[s\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}} + \left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/4}\sqrt{sh}+h\right]$, *where* $C_{\boldsymbol{\delta}} > 0$ *is a sufficiently large constant.*

$$\langle \nabla L_{n_0}^{(0)}(\hat{\boldsymbol{\beta}}) - \nabla L_{n_0}^{(0)}(\boldsymbol{\beta}), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle \lesssim \frac{s^2\log p}{n_{\mathcal{A}_h}+n_0} + \left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/2} sh + sh^2 + s^2h\left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/2} + s^{3/2}h^{3/2}\left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/4},$$

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

**Remark 4.** *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).*

## 3.2 Theory on the transferable source detection algorithm

In this section, we will show that under certain conditions our transferable set detection algorithm (Trans-GLM) can recover level-$h$ transferring set $\mathcal{A}_h$ for some specific $h$ with high probability, that is, $\widehat{\mathcal{A}} = \mathcal{A}_h$ with high probability. Under these conditions, transferring with $\widehat{\mathcal{A}}$ will lead to a faster convergence rate compared to Lasso fitted on the target data, when the target sample size $n_0$ falls into some regime. But as we described in Section 2.4, Algorithm 2 does not require any explicit input of $h$.

The corresponding population version of $\hat{L}_0^{[r]}(\boldsymbol{w})$ defined in (3) is

$$\begin{aligned} L_0(\boldsymbol{w}) &= -\mathbb{E}[\log\rho(y^{(0)})] - \mathbb{E}[y^{(0)}\boldsymbol{w}^T\boldsymbol{x}^{(0)}] + \mathbb{E}[\psi(\boldsymbol{w}^T\boldsymbol{x}^{(0)})] \\ &= -\mathbb{E}[\log\rho(y^{(0)})] - \mathbb{E}[\psi'(\boldsymbol{w}^T\boldsymbol{x}^{(0)})\boldsymbol{w}^T\boldsymbol{x}^{(0)}] + \mathbb{E}[\psi(\boldsymbol{w}^T\boldsymbol{x}^{(0)})]. \end{aligned}$$

Based on Assumption 6, similar to (2), for $\{k\}$-Trans-GLM (Algorithm 1 with $\mathcal{A} = \{k\}$) used in Algorithm 2, consider the following population version of estimators from the

transferring step with respect to target data and $k$-th source data, which is the solution $\boldsymbol{\beta}^{(k)}$ of equation

$$\sum_{j\in\{0,k\}} \alpha_j^{(k)} \mathbb{E}\left\{[\psi'((\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}^{(k)}) - \psi'((\boldsymbol{w}^{(k)})^T\boldsymbol{x}^{(k)})]\boldsymbol{x}^{(k)}\right\},$$

where $\alpha_0^{(k)} = \frac{2n_0/3}{2n_0/3+n_k}$ and $\alpha_k^{(k)} = \frac{n_k}{2n_0/3+n_k}$. Define $\boldsymbol{\beta}^{(0)} = \boldsymbol{\beta}$. Next, let's impose a general assumption to ensure the identifiability of some $\mathcal{A}_h$ by Trans-GLM.

**Assumption 5** (Identifiability of $\mathcal{A}_h$)**.** *Denote $\mathcal{A}_h^c = \{1,\ldots,K\}\backslash\mathcal{A}_h$. Suppose for some $h$, we have*

$$\mathbb{P}\left(\sup_r |\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]}) - \hat{L}_0^{[r]}(\boldsymbol{\beta}^{(k)})| > \Upsilon_1^{(k)} + \zeta\Gamma_1^{(k)}\right) \lesssim g_1^{(k)}(\zeta),$$
$$\mathbb{P}\left(\sup_r |\hat{L}_0^{[r]}(\boldsymbol{\beta}^{(k)}) - L_0(\boldsymbol{\beta}^{(k)})| > \zeta\Gamma_2^{(k)}\right) \lesssim g_2^{(k)}(\zeta),$$

*where $g_1^{(k)}(\zeta)$, $g_2^{(k)}(\zeta) \to 0$ as $\zeta \to \infty$. Assume $\inf_{k\in\mathcal{A}_h^c} \lambda_{\min}(\mathbb{E}[\int_0^1 \psi''((1-t)(\boldsymbol{x}^{(0)})^T\boldsymbol{\beta} + t(\boldsymbol{x}^{(0)})^T\boldsymbol{\beta}^{(k)})dt \cdot \boldsymbol{x}^{(0)}(\boldsymbol{x}^{(0)})^T]) := \underline{\lambda} > 0$, and*

$$\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}\|_2 \geq \underline{\lambda}^{-1/2}\left[C_1\left(\sqrt{\Gamma_1^{(0)}} \vee \sqrt{\Gamma_2^{(0)}} \vee 1\right) + \sqrt{2\Upsilon_1^{(k)}}\right], \forall k \in \mathcal{A}_h^c \quad (7)$$

$$\Upsilon_1^{(k)} + \Gamma_1^{(k)} + \Gamma_2^{(k)} + h^2 = o(1), \forall k \in \mathcal{A}_h; \quad \Gamma_1^{(k)} = o(1), \Gamma_2^{(k)} = o(1), \forall k \in \mathcal{A}_h^c, \quad (8)$$

*where $C_1 > 0$ is sufficiently large.*

**Remark 5.** *Here we use generic notations $\Upsilon_1^{(k)}$, $\Gamma_1^{(k)}$, $\Gamma_2^{(k)}$, $g_1^{(k)}(\zeta)$ and $g_2^{(k)}(\zeta)$. We show their explicit forms under linear, logistic, and Poisson regression models in Proposition 1 in Appendix S.1.2.*

**Remark 6.** *Condition (7) guarantees that for the sources not in $\mathcal{A}_h$, there is a sufficiently large gap between the population-level coefficient from the transferring step and the true*

*coefficient of target data. Condition (8) guarantees the variations of* $\sup_r |\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]}) - \hat{L}_0^{[r]}(\boldsymbol{\beta}^{(k)})|$ *and* $\sup_r |\hat{L}_0^{[r]}(\boldsymbol{\beta}^{(k)}) - L_0(\boldsymbol{\beta}^{(k)})|$ *are shrinking as the sample sizes go to infinity.*

Based on the aforementioned assumptions, we have the following detection consistency property.

**Theorem 5** (Detection consistency of $\mathcal{A}_h$)**.** *For Algorithm 2 (Trans-GLM), with Assumption 5 satisfied for some* $h$*, for any* $\delta > 0$*, there exist constants* $C'(\delta)$ *and* $N = N(\delta) > 0$ *such that when* $C_0 = C'(\delta)$ *and* $\min_{k \in \{0\} \cup \mathcal{A}_h} n_k > N(\delta)$*,*

$$\mathbb{P}(\widehat{\mathcal{A}} = \mathcal{A}_h) \geq 1 - \delta.$$

*Then Algorithm 2 has the same high-probability upper bounds of* $\ell_1/\ell_2$*-estimation error and the prediction error measure as those in Theorems 1-4 with same conditions, respectively.*

**Remark 7.** *We would like to emphasize again that Algorithm 2 does not require the explicit input of* $h$*. Theorem 5 tells us that the transferring set* $\widehat{\mathcal{A}}$ *suggested by Trans-GLM will be* $\mathcal{A}_h$ *for some* $h$*, under certain conditions.*

By recalling the procedure of Algorithm 2, we notice that it relies on the negative log-likelihood as the similarity metric between source and target data. As we know, the accurate estimation of coefficients or log-likelihood for GLM under the high-dimensional setting depends on the sparse structure. Therefore, to provide an explicit form of Assumption 6, we will impose a "weak" sparsity assumption on both $\boldsymbol{w}^{(k)}$ and $\boldsymbol{\beta}^{(k)}$ with $k \in \mathcal{A}_h^c$ for some $h$. Note that the source data in $\mathcal{A}_h$ automatically satisfy the sparsity condition due to the definition of $\mathcal{A}_h$.

Now we introduce the following assumption to present explicit forms of Assumption 5.

**Assumption 6.** *For some $h$ and any $k \in \mathcal{A}_h^c$, we assume the following decomposition of $\boldsymbol{w}^{(k)}$ and $\boldsymbol{\beta}^{(k)}$ to hold with some $s'$ and $h' > 0$:*

*(i)* $\boldsymbol{w}^{(k)} = \boldsymbol{\xi}^{(k)} + \boldsymbol{\tau}^{(k)}$, *where* $\|\boldsymbol{\xi}^{(k)}\|_0 \le s'$ *and* $\|\boldsymbol{\tau}^{(k)}\|_1 \le h'$;

*(ii)* $\boldsymbol{\beta}^{(k)} = \boldsymbol{\varrho}^{(k)} + \boldsymbol{\varpi}^{(k)}$, *where* $\|\boldsymbol{\varrho}^{(k)}\|_0 \le s'$ *and* $\|\boldsymbol{\varpi}^{(k)}\|_1 \le h'$.

**Corollary 1.** *Assume Assumptions 1, 2 and* $\inf_{k\in\mathcal{A}_h^c} \lambda_{\min}\big(\mathbb{E}[\int_0^1 \psi''((1-t)(\boldsymbol{x}^{(0)})^T\boldsymbol{\beta}+t(\boldsymbol{x}^{(0)})^T\boldsymbol{\beta}^{(k)})dt\cdot \boldsymbol{x}^{(0)}(\boldsymbol{x}^{(0)})^T]\big) := \underline{\lambda} > 0$ *hold. Also assume* $\sup_{k\in\mathcal{A}_h^c}\|\boldsymbol{\beta}^{(k)}\|_\infty < \infty$, $\sup_k \|\boldsymbol{w}^{(k)}\|_\infty < \infty$. *Then we have the following sufficient conditions to make Assumption 5 hold for logistic, linear and Poisson regression models. Denote* $\Omega = \sqrt{h'}\left(\frac{\log p}{\min_{k\in\mathcal{A}_h} n_k+n_0}\right)^{1/4} + \left(\frac{s'\log p}{\min_{k\in\mathcal{A}_h} n_k+n_0}\right)^{1/4}[(s\vee s')^{1/4}+\sqrt{h'}] + \left(\frac{\log p}{\min_{k\in\mathcal{A}_h} n_k+n_0}\right)^{1/8}(h')^{1/4}[(s\vee s')^{1/8}+(h')^{1/4}]$.

*(i) For logistic regression models, we require*

$$\inf_{k\in\mathcal{A}_h} n_k \gg s\log p, \quad n_0 \gg \big\{[s\vee s'+(h')^2]\vee\Omega^2\big\}\cdot\log K,$$
$$\inf_{k\in\mathcal{A}_h^c}\|\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta}\|_2 \gtrsim \left(\frac{s\log p}{n_0}\right)^{1/4}\vee 1+\Omega, \quad h \ll s^{-1/2}$$

*(ii) For linear models, we require*

$$\inf_{k\in\mathcal{A}_h} n_k \gg s^2\log p, \quad n_0 \gg \big\{[(s\vee s')^2+(h')^4]\vee[(s\vee s'+(h')^2)\Omega^2]\big\}\cdot\log K,$$

$$\inf_{k\in\mathcal{A}_h^c}\|\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta}\|_2 \gtrsim \left(\frac{s^2\log p}{n_0}\right)^{1/4}\vee 1+\Big[(s')^{1/4}+\sqrt{h'}\Big]\Omega, \quad h \ll s^{-1}.$$

*(iii) For Poisson regression models, we require*

$$\inf_{k\in\mathcal{A}_h} n_k \gg s^2\log p, \quad n_0 \gg \big[(s\vee s'+h')\vee\Omega^2\big]\cdot\log K, \quad U(s\vee s'+h\vee h')\lesssim 1$$
$$\inf_{k\in\mathcal{A}_h^c}\|\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta}\|_2 \gtrsim \left(\frac{s\log p}{n_0}\right)^{1/4}\vee 1+\Big[(s')^{1/4}+\sqrt{h'}\Big]\Omega, \quad h \ll s^{-1}.$$

Under Assumptions 1, 2, and the sufficient conditions derived in Corollary 1, by Theorem 5, we can conclude that $\widehat{\mathcal{A}} = \mathcal{A}_h$ for some $h$. Note that we don't impose Assumption 4 here. Remark 2 indicates that, for $\mathcal{A}_h$-Trans-GLM to have a faster convergence rate than Lasso on target data, we need $h \ll \sqrt{\frac{\log p}{n_0}}$ and $n_{\mathcal{A}_h} \gg n_0$. Suppose $s' \asymp s$, $h' \lesssim s^{1/2}$. Then for logistic regression models, when $s \log K \ll n_0 \ll s \log p$, the sufficient condition implies $h \ll s^{-1/2} \ll \sqrt{\frac{\log p}{n_0}}$. For linear models, when $s^2 \log K \ll n_0 \ll s^2 \log p$, $h \ll s^{-1} \ll \sqrt{\frac{\log p}{n_0}}$. And for Poisson models, when $s \log K \ll n_0 \ll s^2 \log p$, $h \ll s^{-1} \ll \sqrt{\frac{\log p}{n_0}}$. This implies that when target sample size $n_0$ is within certain regimes and the source data points are much more than target data points, Trans-GLM can lead to a better $\ell_2$-estimation error bound than the classical Lasso on target data.

# 4 Numerical Experiments

In this section, we demonstrate the power of our GLM transfer learning algorithms via extensive simulation studies and a real-data application. In the simulation part, we consider different $h$, and generate coefficient for sources from $\mathcal{A}_h$ and $\mathcal{A}_h^c$ accordingly. *Trans-GLM* (Algorithm 2) is compared with *naïve-Lasso*, $\mathcal{A}_h$*-Trans-GLM* (Algorithm 1 with $\mathcal{A} = \mathcal{A}_h$) and *Pooled-Trans-GLM* (Algorithm 1 with all sources). In real-data study, besides naïve-Lasso, Pooled-Trans-GLM, and Trans-GLM, other methods are explored for comparison, including support vector machines (SVM), decision trees (Tree), random forests (RF) and Adaboost algorithm with trees (Boosting). We run these benchmark methods twice. First we fit the models by only using target data, while at the second time, we run the approaches on target and all source data, which is called as a pooled version. We use the original method name to denote the corresponding method implemented on target data, and add a

prefix "Pooled" to denote the corresponding method implemented on target and all source data. For example, Pooled-SVM represents SVM fitted on all data from target and sources.

All experiments 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 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. All the other parameters are kept the same as default settings.

## 4.1 Simulations

We investigate the performance of Algorithms 1 and 2 under different settings. The number of sources is fixed as $K = 10$. We set the target sample size $n_0 = 200$ and source sample sample size $n_k = 200$ for all $k \neq 0$. The dimension $p = 2000$ for both target and source data. For the target, the coefficient is set to be $\boldsymbol{\beta} = (0.5 \cdot \mathbf{1}_s, \mathbf{0}_{p-s})^T$, where $\mathbf{1}_s$ has all $s$

elements 1 and $\mathbf{0}_{p-s}$ has all $(p-s)$ elements 0. We fix $s = 20$. Denote $\mathcal{R}_p^{(k)}$ as $p$ independent Rademacher variables (being $-1$ or 1 with equal probability) for any $k$. $\mathcal{R}_p^{(k)}$ and $\mathcal{R}_p^{(k')}$ are independent for any $k \neq k'$. Consider $h = 10$, 20 and 30. For any source data $k$ in $\mathcal{A}_h$, we set $\boldsymbol{w}^{(k)} = \boldsymbol{\beta} + h/p \cdot \mathcal{R}_p^{(k)}$. For linear and logistic regression models, predictors from target $\boldsymbol{x}_i^{(0)} \overset{i.i.d.}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = [\Sigma_{jj'}]_{p\times p}$ where $\Sigma_{jj'} = 0.9^{|j-j'|}$, for all $i = 1, \cdots, n_0$. For the source, we generate $p$-dimensional predictors from independent $t$-distribution with degree of freedom 4. For the target and sources of Poisson regression model, we generate predictors from the same Gaussian distribution and $t$-distribution respectively, and truncate each predictor at $\pm 0.5$.

To generate the coefficient $\boldsymbol{w}^{(k)}$ for $k \notin \mathcal{A}_h$, we randomly generate $S^{(k)}$ of size $s$ from $\{2s+1, \ldots, p\}$. Then, the $j$-th component of coefficient $\boldsymbol{w}^{(k)}$ is set to be

$$w_j^{(k)} = \begin{cases} 0.5 + 2h/p \cdot (-1)^{r_j^{(k)}}, & j \in \{s+1, \ldots, 2s\} \cup S^{(k)}, \\ 2h/p \cdot (-1)^{r_j^{(k)}}, & \text{otherwise}, \end{cases}$$

where $r_j^{(k)}$ is a Rademacher variable. We also add an intercept 0.5. The generating process of each source data is independent. We fit naïve-Lasso on only target data. $\mathcal{A}_h$-Trans-GLM and Pooled-Trans-GLM denote the results of Algorithm 1 on source data in $\mathcal{A}_h$ and target data or all sources and target data, respectively. Trans-GLM runs Algorithm 2 by first identifying some informative source set $\widehat{\mathcal{A}}$, then Algorithm 1 is applied to fit the model on sources in $\widehat{\mathcal{A}}$. We vary the values of $|\mathcal{A}_h|$ and $h$, and repeat each setting for 200 times. The average $\ell_2$-estimation errors are summarized in Figure 1.

From Figure 1, it can be observed that in all three models, $\mathcal{A}_h$-Trans-GLM always achieves the best performance as expected since it transfers information from sources in $\mathcal{A}_h$. Trans-GLM mimics the behavior of $\mathcal{A}_h$-Trans-GLM very well, implying that the transferable source detection algorithm can successfully recover $\mathcal{A}_h$. When $K_{\mathcal{A}_h}$ is small, Pooled-

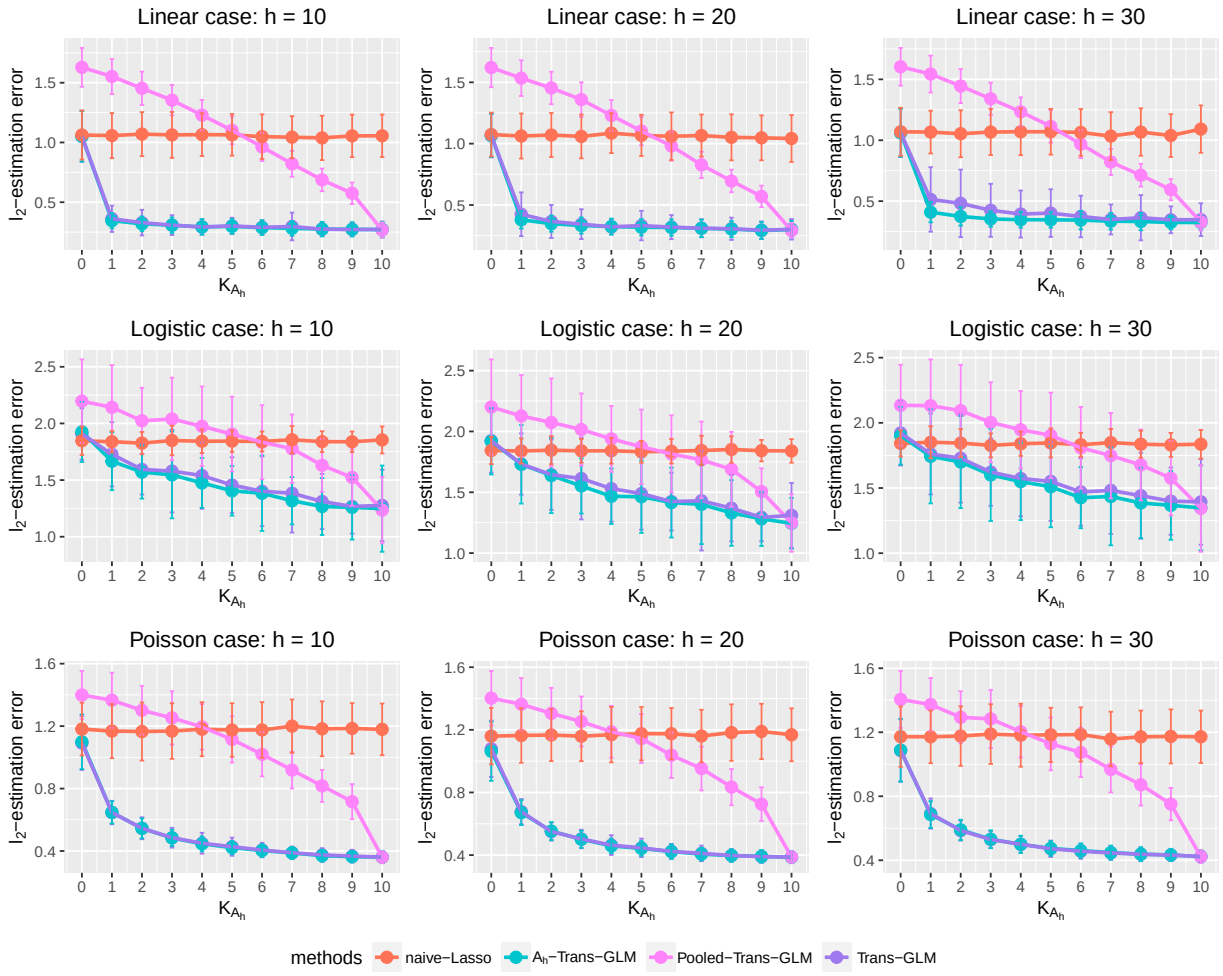


Figure 1: The average $\ell_2$-estimation error of various models with different settings of $h$ and $K_{\mathcal{A}_h}$ when $K = 10$. $n_k = 200$ for all $k = 0, \ldots, K$, $p = 2000$, $s = 20$. Error bars denote the standard deviations.

Trans-GLM performs worse than naïve-Lasso because of the negative transfer. As $K_{\mathcal{A}_h}$ increases, the performance of Pooled-Trans-GLM improves and finally matches those of $\mathcal{A}_h$-Trans-GLM and Trans-GLM when $K_{\mathcal{A}_h} = K = 10$.

Due to space limit, we leave additional simulation results in the supplementary materials, where we have more studies on oracle algorithms and explore different $(n, p, s)$ settings.

## 4.2 A real-data study

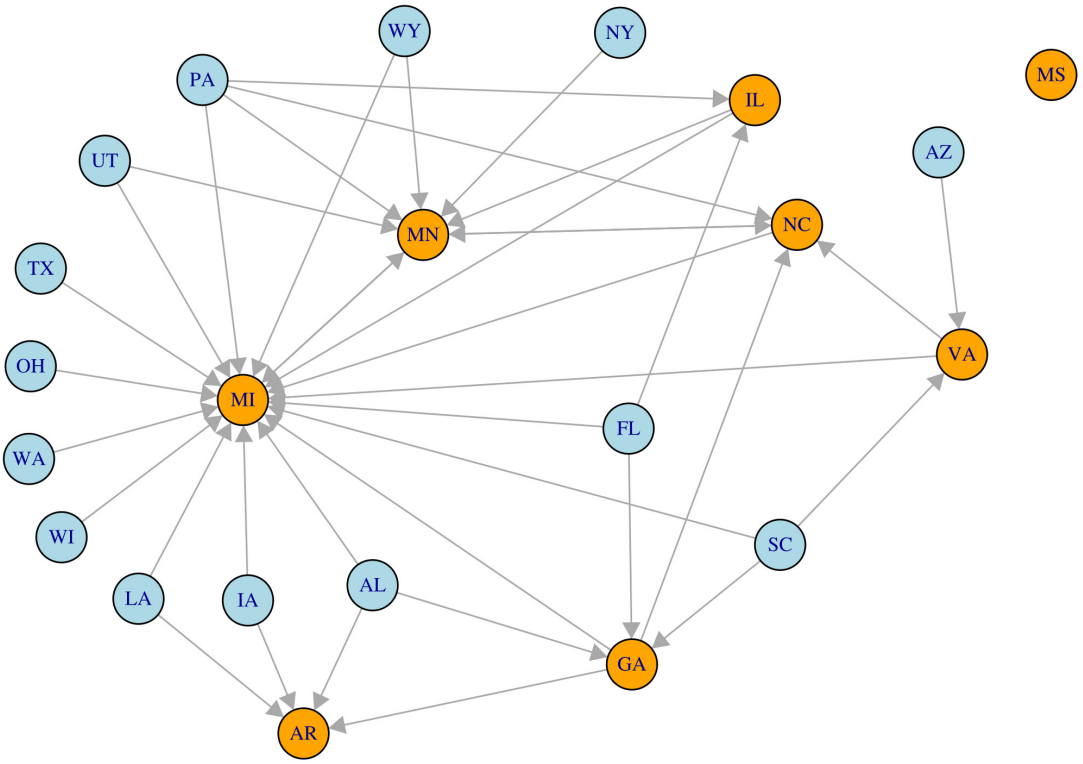


Figure 2: The transferability between different states for Trans-GLM. Edges connect groups with the highest 25 transferring frequencies among 500 replications. Orange nodes and blue nodes denote target states and source states, respectively. We do not plot the isolated source states with no direct edge to other nodes in the graph.

In this section, we study 2020 US presidential election results of each county. We only consider the win or lose between two main parties, Democrat and Republican, in each county. The 2020 county-level election result is available at https://github.com/tonmcg/US_County_Level_Election_Results_08-20. The response is the election result of each county. If Democrat wins, we denote this county as class 1, otherwise we denote it as class 0. And we also collect the county-level information as the predictors, including the population and race proportion, from https://www.kaggle.com/benhamner/2016-us-election.

The goal is to explore the relationship between states in the election by using our GLM transfer learning algorithm. We are interested in swing states with large number of counties. Therefore, among 50 states (Alaska excluded), we select the states where the proportion of counties voting Democrat falls between 10% and 90% and have at least 75 counties as target states. They include Arkansas (AR), Georgia (GA), Illinois (IL), Michigan (MI), Minnesota (MN), Mississippi (MS), North Carolina (NC), and Virginia (VA).

The original data includes 3111 counties and 52 county-level predictors. We also consider the pairwise interaction terms between predictors. After pre-processing, there are 3111 counties and 1128 predictors in the final data, belonging to 50 US states.

We would like to investigate which states have closer relationship with these target states by our transferable set detection algorithm. We set these target states as target dataset in turns and the remaining 49 states as source datasets. Each time we randomly sample 80% target data as training data and the remaining 20% data is used for testing. Then we run Trans-GLM (Algorithm 2) and see which states are in the estimated transferring set $\widehat{\mathcal{A}}$. We repeat the simulation for 500 times and count the transferring frequency between each two states. The 25 (directed) state pairs with the highest transferring frequencies are visualized in Figure 2. Each orange node represents a target state we mentioned above and

blue nodes are source states. States with the top 25 transferring frequencies are connected with a directed edge.

| Methods | Target states | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | AR | GA | IL | MI | MN | MS | NC | VA |
| naïve-Lasso | $3.95_{3.72}$ | $6.94_{3.56}$ | $6.43_{3.85}$ | $11.44_{2.13}$ | $11.14_{3.31}$ | $7.09_{4.56}$ | $11.46_{6.17}$ | $7.05_{3.98}$ |
| Pooled-Trans-GLM | $1.54_{2.70}$ | $4.14_{2.74}$ | $3.01_{3.13}$ | $5.73_{4.03}$ | $8.05_{5.03}$ | $6.68_{4.69}$ | $7.16_{4.02}$ | $6.01_{3.48}$ |
| Trans-GLM | $\mathbf{1.46}_{2.66}$ | $3.94_{2.68}$ | $\mathbf{2.97}_{3.17}$ | $\mathbf{5.28}_{4.01}$ | $\mathbf{7.63}_{4.87}$ | $6.97_{4.73}$ | $\mathbf{6.86}_{4.18}$ | $\mathbf{5.93}_{3.52}$ |
| SVM | $6.58_{2.40}$ | $14.00_{3.35}$ | $4.81_{3.46}$ | $11.64_{3.53}$ | $13.01_{2.74}$ | $13.83_{5.78}$ | $10.50_{5.00}$ | $8.39_{3.79}$ |
| Pooled-SVM | $7.54_{4.95}$ | $13.58_{2.85}$ | $7.53_{3.40}$ | $8.91_{4.66}$ | $10.74_{3.50}$ | $26.92_{5.71}$ | $18.34_{4.66}$ | $22.88_{3.88}$ |
| Tree | $4.11_{4.68}$ | $7.33_{3.43}$ | $5.72_{3.92}$ | $11.32_{5.59}$ | $11.76_{5.22}$ | $6.37_{4.41}$ | $12.68_{5.62}$ | $12.68_{5.67}$ |
| Pooled-Tree | $5.44_{5.50}$ | $7.35_{3.46}$ | $5.03_{3.30}$ | $6.66_{4.08}$ | $11.60_{4.26}$ | $14.72_{5.70}$ | $12.97_{5.35}$ | $11.30_{4.79}$ |
| RF | $1.72_{2.44}$ | $\mathbf{3.54}_{2.59}$ | $4.87_{3.34}$ | $6.61_{4.09}$ | $11.09_{4.54}$ | $\mathbf{4.42}_{3.53}$ | $8.00_{4.69}$ | $6.98_{3.89}$ |
| Pooled-RF | $5.25_{4.91}$ | $4.49_{2.30}$ | $5.39_{2.99}$ | $6.88_{4.09}$ | $10.70_{3.45}$ | $12.34_{5.70}$ | $7.54_{4.27}$ | $7.61_{3.87}$ |
| Boosting | $3.01_{3.49}$ | $5.62_{3.02}$ | $5.35_{3.98}$ | $8.41_{4.71}$ | $11.28_{5.18}$ | $5.80_{4.23}$ | $8.57_{5.44}$ | $6.24_{4.60}$ |
| Pooled-Boosting | $3.50_{4.05}$ | $4.09_{2.50}$ | $4.59_{3.16}$ | $7.00_{4.13}$ | $10.00_{4.10}$ | $9.03_{5.59}$ | $7.26_{4.42}$ | $6.96_{3.59}$ |

Table 1: The average test error rate of various methods with different targets among 500 replications. Numbers in the subscript indicate the standard deviations.

From Figure 2, we observe that Michigan has a strong relationship with other states, since that there are a lot of information transferable when predicting the county level results in Michigan. Another interesting finding is states that are geographically close to each other may share more similarities. For instance, see the connection between Indiana and Michigan, Ohio and Michigan, North Carolina and Virginia, South Carolina and Georgia, Alabama and Georgia, etc. In addition, one can observe that states in the Rust Belt[2] also

[2]The Rust Belt denotes a region of the Northeastern and Midwestern United States, which has been experiencing industrial decline starting around 1980. It includes parts of New York, Pennsylvania, Ohio, West Virginia, Kentucky, Indiana, Michigan, Illinois, Wisconsin and Minnesota. For more details, see the

share more similarities. As evidences, see the edges between Pennsylvania and Minnesota, Illinois and Minnesota, Wisconsin and Michigan, New York and Minnesota, etc.

To further verify the effectiveness of our GLM transfer learning framework on this dataset and make our findings more convincing, we calculate the average test misclassification error rates for each of eight target states. For comparison, we compare the performance of Trans-GLM, Pooled-Trans-GLM with naïve-Lasso and SVM, trees, random forests (RF), boosting trees (Boosting) as well as their pooled version (models are fitted on the pooled target and source datasets). Average test errors and the standard deviations of various methods are summarized in Table 1.

Table 1 shows that for six out of eight scenarios, Trans-GLM performs the best among all approaches, verifying the effectiveness of our GLM transfer learning algorithm. Besides, Pooled-Trans-GLM can always improve the performance of naïve-Lasso, while for other methods, pooling the data can sometimes lead to a worse performance than that of the model fitted on only the target data. This marks the success of our two-step transfer learning framework and the importance of the debiasing step. Combining the results with Figure 2, it can be seen that the performance improvement of Trans-GLM (compared to naïve-Lasso) for the target states which have more connections (share more similarities with other states) are larger. For example, Trans-GLM improves the performance of naïve-Lasso a lot on Michigan, Minnesota and North Carolina, while it performs similarly to naïve-Lasso on Mississippi.

Wikipedia and Crandall (1993).

# 5 Discussions

In this work, we introduce the GLM transfer learning problem. To the best of our knowledge, this is the first paper to study high-dimensional GLM under a transfer learning framework, which can be seen as an extension to Bastani (2020) and Li et al. (2020a). We propose GLM transfer learning algorithms, and derive bounds for $\ell_1/\ell_2$-estimation error and a prediction error measure with fast and slow rates under different conditions. In addition, to avoid the negative transfer, an algorithm-free transferable source detection algorithm is developed and its theoretical properties are presented in detail. Finally, we demonstrate the effectiveness of our algorithms via simulations and a real-data study.

There are several promising future avenues that are worth further research. The first interesting problem is how to extend the current framework and theoretical analysis to other models, such as multinomial regression and the Cox model. Second, it might be possible to revise the current algorithm to calculate the confidence interval of coefficient estimates, which is important for statistical inference. The ideas introduced in Zhang and Zhang (2014); Van de Geer et al. (2014) might help. Another promising direction is to explore similar frameworks for other machine learning models, including support vector machines, decision trees, and random forests. Besides, in the line of high-dimensional transfer learning research, most of the previous work, e.g. Bastani (2020); Li et al. (2020a,b), including this one, assume that target and all sources share the same predictor domain, which however can be hard to satisfy in practice. Relaxing this condition could motivate more useful applications in practice.

# Appendix A 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

$$\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_2 \leq 8\kappa_2 C_1 h\sqrt{\frac{\log p}{n_{\mathcal{A}_h} + n_0}} + 3\frac{\sqrt{s}}{\kappa_1}\lambda_{\boldsymbol{w}} + 2\sqrt{\frac{C_1}{\kappa_1}h\lambda_{\boldsymbol{w}}}. \tag{9}$$

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

$$\begin{aligned}
\delta\hat{L}(\hat{\boldsymbol{u}}^{\mathcal{A}_h}, \mathcal{D}) &\leq \lambda_{\boldsymbol{w}}(\|\boldsymbol{w}_S^{\mathcal{A}_h}\|_1 + \|\boldsymbol{w}_{S^c}^{\mathcal{A}_h}\|_1) - \lambda_{\boldsymbol{w}}(\|\hat{\boldsymbol{w}}_S^{\mathcal{A}_h}\|_1 + \|\hat{\boldsymbol{w}}_{S^c}^{\mathcal{A}_h}\|_1) + \nabla\hat{L}(\boldsymbol{w}, \mathcal{D})^T\hat{\boldsymbol{u}}^{\mathcal{A}_h} \\
&\leq \lambda_{\boldsymbol{w}}(\|\boldsymbol{w}_S^{\mathcal{A}_h}\|_1 + \|\boldsymbol{w}_{S^c}^{\mathcal{A}_h}\|_1) - \lambda_{\boldsymbol{w}}(\|\hat{\boldsymbol{w}}_S^{\mathcal{A}_h}\|_1 + \|\hat{\boldsymbol{w}}_{S^c}^{\mathcal{A}_h}\|_1) + \frac{1}{2}\lambda_{\boldsymbol{w}}\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1 \\
&\leq \frac{3}{2}\lambda_{\boldsymbol{w}}\|\hat{\boldsymbol{u}}_S^{\mathcal{A}_h}\|_1 - \frac{1}{2}\lambda_{\boldsymbol{w}}\|\hat{\boldsymbol{u}}_{S^c}^{\mathcal{A}_h}\|_1 + 2\lambda_{\boldsymbol{w}}\|\boldsymbol{w}_{S^c}^{\mathcal{A}_h}\|_1 \\
&\leq \frac{3}{2}\lambda_{\boldsymbol{w}}\|\hat{\boldsymbol{u}}_S^{\mathcal{A}_h}\|_1 - \frac{1}{2}\lambda_{\boldsymbol{w}}\|\hat{\boldsymbol{u}}_{S^c}^{\mathcal{A}_h}\|_1 + 2\lambda_{\boldsymbol{w}}C_1 h.
\end{aligned} \tag{10}$$

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 (10) 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 $F(\boldsymbol{u}) = L(\boldsymbol{w}^{\mathcal{A}_h} + \boldsymbol{u}, \mathcal{D}) - L(\boldsymbol{w}^{\mathcal{A}_h}) + \lambda_{\boldsymbol{w}}(\|\boldsymbol{w}^{\mathcal{A}_h} + \boldsymbol{u}\|_1 - \|\boldsymbol{w}^{\mathcal{A}_h}\|_1)$. Since $F(\boldsymbol{0}) = 0$ and $F(\hat{\boldsymbol{u}}^{\mathcal{A}_h}) \leq 0$, by convexity,

$$F(\tilde{\boldsymbol{u}}^{\mathcal{A}_h}) = F(t\hat{\boldsymbol{u}}^{\mathcal{A}_h} + (1-t)\boldsymbol{0}) \leq tF(\hat{\boldsymbol{u}}^{\mathcal{A}_h}) \leq 0. \tag{11}$$

However, by Lemma 2 and the same trick of (10),

$$F(\tilde{\boldsymbol{u}}^{\mathcal{A}_h}) \geq \delta\hat{L}(\hat{\boldsymbol{u}}^{\mathcal{A}_h}, \mathcal{D}) + \nabla L(\boldsymbol{w}^{\mathcal{A}_h})^T\tilde{\boldsymbol{u}}^{\mathcal{A}_h} - \lambda_{\boldsymbol{w}}\|\boldsymbol{w}^{\mathcal{A}_h}\|_1 + \lambda_{\boldsymbol{w}}\|\boldsymbol{w}^{\mathcal{A}_h} + \tilde{\boldsymbol{u}}^{\mathcal{A}_h}\|_1$$

$$\geq \kappa_1 \|\tilde{\boldsymbol{u}}^{\mathcal{A}_h}\|_2^2 - \kappa_1\kappa_2 \sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}} \|\tilde{\boldsymbol{u}}^{\mathcal{A}_h}\|_1 \|\tilde{\boldsymbol{u}}^{\mathcal{A}_h}\|_2 - \frac{3}{2}\lambda_{\boldsymbol{w}} \|\tilde{\boldsymbol{u}}_S^{\mathcal{A}_h}\|_1 + \frac{1}{2}\lambda_{\boldsymbol{w}} \|\tilde{\boldsymbol{u}}_{S^c}^{\mathcal{A}_h}\|_1 - 2\lambda_{\boldsymbol{w}} C_1 h$$

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

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 (11). 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

$$\begin{aligned}
\nabla \hat{L}(\boldsymbol{w}^{\mathcal{A}_h}) &= \frac{1}{n_{\mathcal{A}_h}+n_0} \sum_{k\in\{0\}\cup\mathcal{A}_h} (\boldsymbol{X}^{(k)})^T[-\boldsymbol{y}^{(k)} + \boldsymbol{\psi}'(\boldsymbol{X}^{(k)}\boldsymbol{w}^{\mathcal{A}_h})] \\
&= \frac{1}{n_{\mathcal{A}_h}+n_0} \sum_{k\in\{0\}\cup\mathcal{A}_h} (\boldsymbol{X}^{(k)})^T[-\boldsymbol{y}^{(k)} + \boldsymbol{\psi}'(\boldsymbol{X}^{(k)}\boldsymbol{w}^{(k)})] \\
&\quad + \frac{1}{n_{\mathcal{A}_h}+n_0} \sum_{k\in\{0\}\cup\mathcal{A}_h} (\boldsymbol{X}^{(k)})^T[-\boldsymbol{\psi}'(\boldsymbol{X}^{(k)}\boldsymbol{w}^{(k)}) + \boldsymbol{\psi}'(\boldsymbol{X}^{(k)}\boldsymbol{w}^{\mathcal{A}_h})]. \qquad (12)
\end{aligned}$$

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

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

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 (12). Denote $V_{ij}^{(k)} = x_{ij}^{(k)}[-\psi'((\boldsymbol{x}_i^{(k)})^T\boldsymbol{w}^{(k)}) + \psi'((\boldsymbol{x}_i^{(k)})^T\boldsymbol{w}^{\mathcal{A}_h})]$. Under Assumption 3, by mean value theorem and Lemma 1, there exists $v_i^{(k)} \in (0,1)$ such that

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

$\psi''((\boldsymbol{x}_i^{(k)})^T\boldsymbol{w}^{(k)}+v_i^{(k)}(\boldsymbol{x}_i^{(k)})^T(\boldsymbol{w}^{\mathcal{A}_h}-\boldsymbol{w}^{(k)}))x_{ij}^{(k)}$ is $M_\psi^2\kappa_u^2$-subGaussian due to the almost sure boundedness of $\psi''$ in Assumption 3. And $(\boldsymbol{x}_i^{(k)})^T(\boldsymbol{w}^{\mathcal{A}_h}-\boldsymbol{w}^{(k)})$ 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 (12) equals to $\frac{1}{n_{\mathcal{A}_h}+n_0}\sup_{j=1,\dots,p}|\sum_{k\in\{0\}\cup\mathcal{A}_h}\sum_{i=1}^{n_k}V_{ij}^{(k)}|$, by tail bounds of subexponential variables and union bounds, we have

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

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 (9), we have

$$\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_2\lesssim h\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}}+\sqrt{\frac{s\log p}{n_{\mathcal{A}_h}+n_0}}+\left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/4}\sqrt{h},\tag{13}$$

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 $\hat{\boldsymbol{u}}^{\mathcal{A}_h} \in \mathbb{C}$, (13) encloses

$$\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1 \lesssim s\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}} + \left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/4}\sqrt{sh} + h\left(1+\sqrt{\frac{s\log p}{n_{\mathcal{A}_h}+n_0}}\right), \tag{14}$$

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

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

(i) Similar to (10), when $\lambda_{\boldsymbol{\delta}} \geq 2\|\nabla\hat{L}^{(0)}(\boldsymbol{\beta}, \mathcal{D}^{(0)})\|_\infty$, we have

$$\begin{aligned}
\delta\hat{L}^{(0)}(\hat{\boldsymbol{\delta}}^{\mathcal{A}_h}, \mathcal{D}) &\leq \lambda_{\boldsymbol{\delta}}(\|\boldsymbol{\beta}-\hat{\boldsymbol{w}}^{\mathcal{A}_h}\|_1 - \|\hat{\boldsymbol{\delta}}^{\mathcal{A}_h}\|_1) - \nabla\hat{L}^{(0)}(\boldsymbol{\beta}, \mathcal{D}^{(0)})^T\hat{\boldsymbol{v}}^{\mathcal{A}_h} \\
&\leq \lambda_{\boldsymbol{\delta}}\|\boldsymbol{\beta}-\hat{\boldsymbol{w}}^{\mathcal{A}_h}\|_1 - \lambda_{\boldsymbol{\delta}}\|\hat{\boldsymbol{\delta}}^{\mathcal{A}_h}\|_1 + \frac{1}{2}\lambda_{\boldsymbol{\delta}}\|\hat{\boldsymbol{v}}^{\mathcal{A}_h}\|_1 \\
&\leq \frac{3}{2}\lambda_{\boldsymbol{\delta}}\|\boldsymbol{\beta}-\hat{\boldsymbol{w}}^{\mathcal{A}_h}\|_1 - \frac{1}{2}\lambda_{\boldsymbol{\delta}}\|\hat{\boldsymbol{\delta}}^{\mathcal{A}_h}\|_1 \\
&\leq \frac{3}{2}\lambda_{\boldsymbol{\delta}}\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1 + \frac{3}{2}\lambda_{\boldsymbol{\delta}}h - \frac{1}{2}\lambda_{\boldsymbol{\delta}}\|\hat{\boldsymbol{\delta}}^{\mathcal{A}_h}\|_1.
\end{aligned} \tag{15}$$

Due to the strict convexity, $\delta\hat{L}^{(0)}(\hat{\boldsymbol{\delta}}^{\mathcal{A}_h}, \mathcal{D}) > 0$, leading to $\|\hat{\boldsymbol{\delta}}^{\mathcal{A}_h}\|_1 \leq 3\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1 + 3h$. Then,

$$\|\hat{\boldsymbol{v}}^{\mathcal{A}_h}\|_1 \leq \|\boldsymbol{\beta}-\hat{\boldsymbol{w}}^{\mathcal{A}_h}\|_1 + \|\hat{\boldsymbol{\delta}}^{\mathcal{A}_h}\|_1 \leq 4\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1 + 4h. \tag{16}$$

On the other hand, by Proposition 2 in Negahban et al. (2009), there are constants $\kappa_1'$ and $\kappa_2'$ such that

$$\delta\hat{L}^{(0)}(\hat{\boldsymbol{\delta}}^{\mathcal{A}_h}, \mathcal{D}) \geq \kappa_1'\|\hat{\boldsymbol{v}}^{\mathcal{A}_h}\|_2^2 - \kappa_2'\sqrt{\frac{\log p}{n_0}}\|\hat{\boldsymbol{v}}^{\mathcal{A}_h}\|_1\|\hat{\boldsymbol{v}}^{\mathcal{A}_h}\|_2, \tag{17}$$

with probability at least $1-C_3\exp\{-C_4 n_0\}$. Combining (15), (16) and (17), we get

$$\kappa_1'\|\hat{\boldsymbol{v}}^{\mathcal{A}_h}\|_2^2 - \kappa_2'\sqrt{\frac{\log p}{n_0}}(4\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1+4h)\|\hat{\boldsymbol{v}}^{\mathcal{A}_h}\|_2 \leq \frac{3}{2}\lambda_{\boldsymbol{\delta}}\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1 + \frac{3}{2}\lambda_{\boldsymbol{\delta}}h,$$

with probability at least $1-C_3\exp\{-C_4 n_0\}$. By Lemma 6 in Negahban et al. (2009), since $n_0 \geq C\log p$, $\|\nabla \hat{L}^{(0)}(\boldsymbol{\beta}, \mathcal{D}^{(0)})\|_\infty \lesssim \sqrt{\frac{\log p}{n_0}}$ with probability at least $1-n_0^{-1}$. Due to (14), let $\lambda_{\boldsymbol{\delta}} = C_{\boldsymbol{\delta}}\sqrt{\frac{\log p}{n_0}} \geq 2\|\nabla \hat{L}^{(0)}(\boldsymbol{\beta}, \mathcal{D}^{(0)})\|_\infty$ with high probability, leading to

$$\begin{aligned}\|\hat{\boldsymbol{v}}^{\mathcal{A}_h}\|_2 &\lesssim \sqrt{\frac{\log p}{n_0}}(\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1 + h) + \sqrt{\lambda_{\boldsymbol{\delta}}\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1 + \lambda_{\boldsymbol{\delta}} h} \\ &\lesssim \left(\frac{s\log p}{n_0}\right)^{1/4}\left(\frac{s\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/4} + \left(\frac{s\log p}{n_0}\right)^{1/4}\left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/8} h^{1/4} + \left(\frac{\log p}{n_0}\right)^{1/4} h^{1/2},\end{aligned}$$

with probability at least $1-n_0^{-1}$. $\ell_1$-error bound directly comes from (14) and (16).

(ii) Consider $\nabla \hat{L}^{(0)}(\hat{\boldsymbol{w}}^{\mathcal{A}_h} + \boldsymbol{\delta}^{\mathcal{A}_h}, \mathcal{D}^{(0)}) = W_1 - W_2$, where $W_1 = -\frac{1}{n_0}(\boldsymbol{X}^{(0)})^T\boldsymbol{y}^{(0)} + \frac{1}{n_0}(\boldsymbol{X}^{(0)})^T \boldsymbol{\psi}'(\boldsymbol{X}^{(0)}\boldsymbol{\beta})$ and $W_2 = \frac{1}{n_0}(\boldsymbol{X}^{(0)})^T[\boldsymbol{\psi}'(\boldsymbol{X}^{(0)}(\hat{\boldsymbol{w}}^{\mathcal{A}_h} + \boldsymbol{\delta}^{\mathcal{A}_h})) - \boldsymbol{\psi}'(\boldsymbol{X}^{(0)}\boldsymbol{\beta})]$. By Lemma 6 in Negahban et al. (2009), since $n_0 \geq C\log p$, we have $\|W_1\|_\infty \lesssim \sqrt{\frac{\log p}{n_0}}$, with probability at least $1-n_0^{-1}$. Denote $\tilde{V}_{ij}^{(0)} = x_{ij}^{(0)}[\psi'((\boldsymbol{x}_i^{(0)})^T(\hat{\boldsymbol{w}}^{\mathcal{A}_h} + \boldsymbol{\delta}^{\mathcal{A}_h})) - \psi'((\boldsymbol{x}_i^{(0)})^T\boldsymbol{\beta})]$. Under Assumption 3, by mean value theorem and Lemma 1, $\exists v_i^{(k)} \in (0,1)$ such that

$$\begin{aligned}\|W_2\|_\infty &= \sup_{1\leq j\leq p}\left|\frac{1}{n_0}\sum_{i=1}^{n_0} V_{ij}^{(k)}\right| \\ &= \sup_{1\leq j\leq p}\left|\frac{1}{n_0}\sum_{i=1}^{n_0}\psi''((\boldsymbol{x}_i^{(0)})^T\boldsymbol{\beta} + v_i^{(0)}(\boldsymbol{x}_i^{(0)})^T\hat{\boldsymbol{u}}^{\mathcal{A}_h})x_{ij}^{(0)}(\boldsymbol{x}_i^{(0)})^T\hat{\boldsymbol{u}}^{\mathcal{A}_h}\right| \\ &\leq \sup_{1\leq j\leq p}\left\|\frac{1}{n_0}\sum_{i=1}^{n_0}\psi''((\boldsymbol{x}_i^{(0)})^T\boldsymbol{\beta} + v_i^{(0)}(\boldsymbol{x}_i^{(0)})^T\hat{\boldsymbol{u}}^{\mathcal{A}_h})x_{ij}^{(0)}\boldsymbol{x}_i^{(0)}\right\|_\infty \cdot \|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1 \\ &\lesssim s\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}} + \left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/4}\sqrt{sh} + h,\end{aligned}$$

with probability at least $1-(n_{\mathcal{A}_h}+n_0)^{-1}\vee p^{-1}$, where the last inequality is derived from the fact that each coordinate of $\psi''((\boldsymbol{x}_i^{(0)})^T\boldsymbol{\beta} + v_i^{(0)}(\boldsymbol{x}_i^{(0)})^T\hat{\boldsymbol{u}}^{\mathcal{A}_h})x_{ij}^{(0)}\boldsymbol{x}_i^{(0)}$ is a subexponential variable for $i = 1,\ldots,n_0$ and the mean of $\{\psi''((\boldsymbol{x}_i^{(0)})^T\boldsymbol{\beta} + v_i^{(0)}(\boldsymbol{x}_i^{(0)})^T\hat{\boldsymbol{u}}^{\mathcal{A}_h})x_{ij}^{(0)}\boldsymbol{x}_{ij'}^{(0)}\}_{i,j,j'}$

is uniformly bounded by $\mathcal{O}(1)$. The application of Bernstein inequality and union bounds encloses

$$\|\nabla \hat{L}^{(0)}(\hat{\boldsymbol{w}}^{\mathcal{A}_h} + \boldsymbol{\delta}^{\mathcal{A}_h}, \mathcal{D}^{(0)})\|_\infty \lesssim_p s\sqrt{\frac{\log p}{n_{\mathcal{A}_h} + n_0}} + \left(\frac{\log p}{n_{\mathcal{A}_h} + n_0}\right)^{1/4} \sqrt{sh} + h,$$

with probability at least $1 - n_0^{-1} \vee p^{-1}$. Let $\lambda_{\boldsymbol{\delta}} = C_{\boldsymbol{\delta}} \left[ s\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}} + \left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/4} \sqrt{sh} + h \right]$ satisfying $\lambda_{\boldsymbol{\delta}} \geq 2\|\nabla \hat{L}^{(0)}(\hat{\boldsymbol{w}}^{\mathcal{A}_h} + \boldsymbol{\delta}^{\mathcal{A}_h}, \mathcal{D}^{(0)})\|_\infty$ in high probability to get

$$\hat{L}^{(0)}(\hat{\boldsymbol{w}}^{\mathcal{A}_h} + \hat{\boldsymbol{\delta}}^{\mathcal{A}_h}, \mathcal{D}^{(0)}) - \hat{L}^{(0)}(\hat{\boldsymbol{w}}^{\mathcal{A}_h} + \boldsymbol{\delta}^{\mathcal{A}_h}, \mathcal{D}^{(0)}) - \nabla \hat{L}^{(0)}(\hat{\boldsymbol{w}}^{\mathcal{A}_h} + \boldsymbol{\delta}^{\mathcal{A}_h}, \mathcal{D}^{(0)})^T(\hat{\boldsymbol{\delta}}^{\mathcal{A}_h} - \boldsymbol{\delta}^{\mathcal{A}_h})$$

$$\leq \lambda_{\boldsymbol{\delta}}(\|\boldsymbol{\delta}^{\mathcal{A}_h}\|_1 - \|\hat{\boldsymbol{\delta}}^{\mathcal{A}_h}\|_1) + \|\nabla \hat{L}^{(0)}(\hat{\boldsymbol{w}}^{\mathcal{A}_h} + \boldsymbol{\delta}^{\mathcal{A}_h}, \mathcal{D}^{(0)})\|_\infty \|\hat{\boldsymbol{\delta}}^{\mathcal{A}_h} - \boldsymbol{\delta}^{\mathcal{A}_h}\|_1$$
$$\leq 2\lambda_{\boldsymbol{\delta}}\|\boldsymbol{\delta}^{\mathcal{A}_h}\|_1 - \frac{1}{2}\lambda_{\boldsymbol{\delta}}\|\hat{\boldsymbol{\delta}}^{\mathcal{A}_h} - \boldsymbol{\delta}^{\mathcal{A}_h}\|_1.$$

with probability at least $1 - n_0^{-1} \vee p^{-1}$. By convexity of $\hat{L}^{(0)}$, $\|\hat{\boldsymbol{\delta}}^{\mathcal{A}_h} - \boldsymbol{\delta}^{\mathcal{A}_h}\|_1 \leq 4\|\boldsymbol{\delta}^{\mathcal{A}_h}\|_1 \leq 4C_1 h$. Therefore, by (13),

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \leq \|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_2 + \|\hat{\boldsymbol{\delta}}^{\mathcal{A}_h} - \boldsymbol{\delta}^{\mathcal{A}_h}\|_2 \lesssim_p \sqrt{\frac{s \log p}{n_{\mathcal{A}_h} + n_0}} + \left(\frac{\log p}{n_{\mathcal{A}_h} + n_0}\right)^{1/4} \sqrt{h} + h,$$

with probability at least $1 - n_0^{-1} \vee p^{-1}$. For the $\ell_1$-error bound, it can be easily derived from the fact that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \leq \|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1 + \|\hat{\boldsymbol{\delta}}^{\mathcal{A}_h} - \boldsymbol{\delta}^{\mathcal{A}_h}\|_1 \leq \|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1 + 4C_1 h$ with probability at least $1 - n_0^{-1} \vee p^{-1}$ and the inequality (14), which completes our proof.

# References


Barr, R. (2010). Transfer of learning between 2d and 3d sources during infancy: Informing theory and practice. *Developmental review*, 30(2):128–154.

Bastani, H. (2020). Predicting with proxies: Transfer learning in high dimension. *Management Science*.

Cai, T. T. and Wei, H. (2021). Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *Annals of Statistics*, 49(1):100–128.

Chen, A., Owen, A. B., Shi, M., et al. (2015). Data enriched linear regression. *Electronic journal of statistics*, 9(1):1078–1112.

Crandall, R. W. (1993). *The continuing decline of manufacturing in the Rust Belt*. Brookings Institution.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Gross, S. M. and Tibshirani, R. (2016). Data shared lasso: A novel tool to discover uplift. *Computational statistics & data analysis*, 101:226–235.

Hanneke, S. and Kpotufe, S. (2020a). A no-free-lunch theorem for multitask learning. *arXiv preprint arXiv:2006.15785*.

Hanneke, S. and Kpotufe, S. (2020b). On the value of target data in transfer learning. *arXiv preprint arXiv:2002.04747*.

Kontorovich, A. (2014). Concentration in unbounded metric spaces and algorithmic stability. In *International Conference on Machine Learning*, pages 28–36. PMLR.

Li, S., Cai, T. T., and Li, H. (2020a). Transfer learning for high-dimensional linear regression: Prediction, estimation, and minimax optimality. *arXiv preprint arXiv:2006.10593*.

Li, S., Cai, T. T., and Li, H. (2020b). Transfer learning in large-scale gaussian graphical models with false discovery rate control. *arXiv preprint arXiv:2010.11037*.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research*, 16(1):559–616.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, volume 37. CRC Press.

Negahban, S., Yu, B., Wainwright, M. J., and Ravikumar, P. K. (2009). A unified framework for high-dimensional analysis of $m$-estimators with decomposable regularizers. In *Advances in neural information processing systems*, pages 1348–1356. Citeseer.

Ollier, E. and Viallon, V. (2017). Regression modelling on stratified data with the lasso. *Biometrika*, 104(1):83–96.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Reeve, H. W., Cannings, T. I., and Samworth, R. J. (2021). Adaptive transfer learning. *Annals of Statistics*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Torrey, L. and Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global.

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

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, 3(1):1–40.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 217–242.

Zheng, C., Dasgupta, S., Xie, Y., Haris, A., and Chen, Y. Q. (2019). On data enriched logistic regression. *arXiv preprint arXiv:1911.06380*.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.

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

## S.1 More details

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

To better illustrate our oracle algorithm, we draw a schematic in Figure 3. 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 $\hat{\boldsymbol{\beta}}_{\mathcal{A}\text{-Trans-GLM}}$.

### S.1.2 Theory

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

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

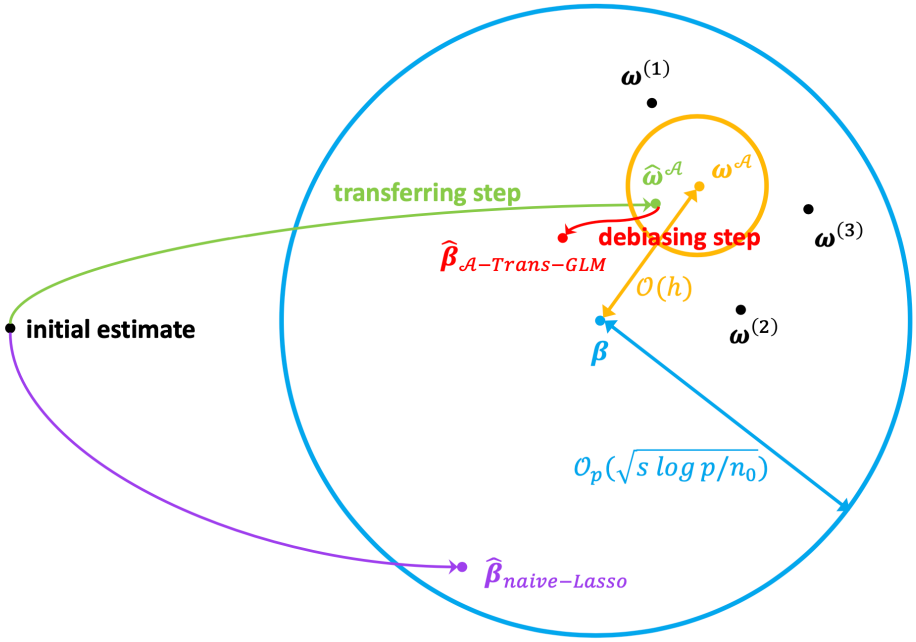


Figure 3: A schematic of $\mathcal{A}$-Trans-GLM (Algorithm 1). $\mathcal{A} = \{1, 2, 3\}$.

*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* $h \leq U^{-1}\bar{U}$ *and* $U \sup_{k \in \mathcal{A}^c}\{\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}\|_1 \vee \|\boldsymbol{\beta}^{(k)} - \boldsymbol{w}^{(k)}\|_1\} \leq \bar{U}$. *With* $\lambda^{(k)} = C\left(\sqrt{\frac{\log p}{n_k + n_0}} + h\right)$ *when* $k \in \mathcal{A}_h$, $\lambda^{(k)} = C\sqrt{\frac{\log p}{n_k + n_0}} \cdot W_k$ *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* $\Gamma_1$ *and* $\Gamma_2$ *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:*

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

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

$$\begin{aligned}
\Gamma_1^{(0)} &= \sqrt{\frac{s\log p}{n_0}} \cdot \exp\left(U\|\boldsymbol{\beta}\|_1\right), \quad \Gamma_2^{(0)} = \exp\left(U\|\boldsymbol{\beta}\|_1\right) \cdot \frac{1 + \|\boldsymbol{\beta}\|_2 + U\|\boldsymbol{\beta}\|_1}{\sqrt{n_0}}, \\
\Upsilon_1^{(k)} &= \Omega_k \cdot \exp\left\{U\|\boldsymbol{w}^{(k)}\|_1 \cdot \mathbb{1}(k \in \mathcal{A}_h) + U\|\boldsymbol{\beta}^{(k)}\|_1 \cdot \mathbb{1}(k \in \mathcal{A}_h^c)\right\}, \\
\Gamma_1^{(k)} &= \sqrt{\frac{1}{n_0}}\Omega_k \cdot \exp\left\{U\|\boldsymbol{w}^{(k)}\|_1 \cdot \mathbb{1}(k \in \mathcal{A}_h) + U\|\boldsymbol{\beta}^{(k)}\|_1 \cdot \mathbb{1}(k \in \mathcal{A}_h^c)\right\}, \\
\Gamma_2^{(k)} &= \sqrt{\frac{1}{n_0}} \Big[\exp\left(U\|\boldsymbol{w}^{(k)}\|_1\right)\left(1 + \|\boldsymbol{w}^{(k)}\|_2 + U\|\boldsymbol{w}^{(k)}\|_1\right) \cdot \mathbb{1}(k \in \mathcal{A}_h) \\
&\qquad + \exp\left(U\|\boldsymbol{\beta}^{(k)}\|_1\right)\left(1 + \|\boldsymbol{\beta}^{(k)}\|_2 + U\|\boldsymbol{\beta}^{(k)}\|_1\right) \cdot \mathbb{1}(k \in \mathcal{A}_h^c)\Big], \\
g_1^{(k)}(\zeta) &= g_2^{(k)}(\zeta) = \exp(-\zeta^2), k \neq 0; \quad g_1^{(0)}(\zeta) = \exp(-\zeta^2), g_2^{(0)}(\zeta) = \zeta^{-2}.
\end{aligned}$$

### S.1.3 Additional simulation results

#### S.1.3.1 Transfer learning on $\mathcal{A}_h$

In this section, we study the performance of oracle algorithm (Algorithm 1) under different $h$ and $(\{n_k\}_{k=0}^K, p, s)$ settings. The following three settings of $(\{n_k\}_{k=0}^K, p, s)$ are considered:

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

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

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

Consider the same setting we use in Section 4.1, that is, for the target, the coefficient is set to be $\boldsymbol{\beta} = (0.5 \cdot \mathbf{1}_s, \mathbf{0}_{p-s})^T$, where $\mathbf{1}_s$ has all $s$ elements 1 and $\mathbf{0}_{p-s}$ has all $(p-s)$ elements 0. Denote $\boldsymbol{\mathcal{R}}_p^{(k)}$ as $p$ independent Rademacher variables (being $-1$ or 1 with equal probability) for any $k$. $\boldsymbol{\mathcal{R}}_p^{(k)}$ is independent with $\boldsymbol{\mathcal{R}}_p^{(k')}$ for any $k \neq k'$. For any source data $k$ in $\mathcal{A}_h$, we set $\boldsymbol{w}^{(k)} = \boldsymbol{\beta} + h/p \cdot \boldsymbol{\mathcal{R}}_p^{(k)}$. For linear and logistic regression models, predictors from target $\boldsymbol{x}_i^{(0)} \overset{i.i.d.}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = [\Sigma_{jj'}]_{p\times p}$ where $\Sigma_{jj'} = 0.9^{|j-j'|}$, for all $i = 1, \cdots, n$. And for $k \in \mathcal{A}_h$, we generate $p$-dimensional predictors from independent $t$-distribution with degree of freedom 4. For Poisson regression model, we generate predictors from the same Gaussian and $t$ distribution and truncate each predictor at $\pm 0.5$. We train naïve-Lasso and $\mathcal{A}_h$-Trans-GLM models under different settings of $h$ and $K_{\mathcal{A}_h}$, then calculate the $\ell_2$-estimation error of $\boldsymbol{\beta}$. All the experiments are replicated for 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 4, 5 and 6.

From Figure 4-6, it can be seen that the oracle algorithm outperforms naïve-Lasso for most combinations of $h$ and $K$. As more and more source data become available,

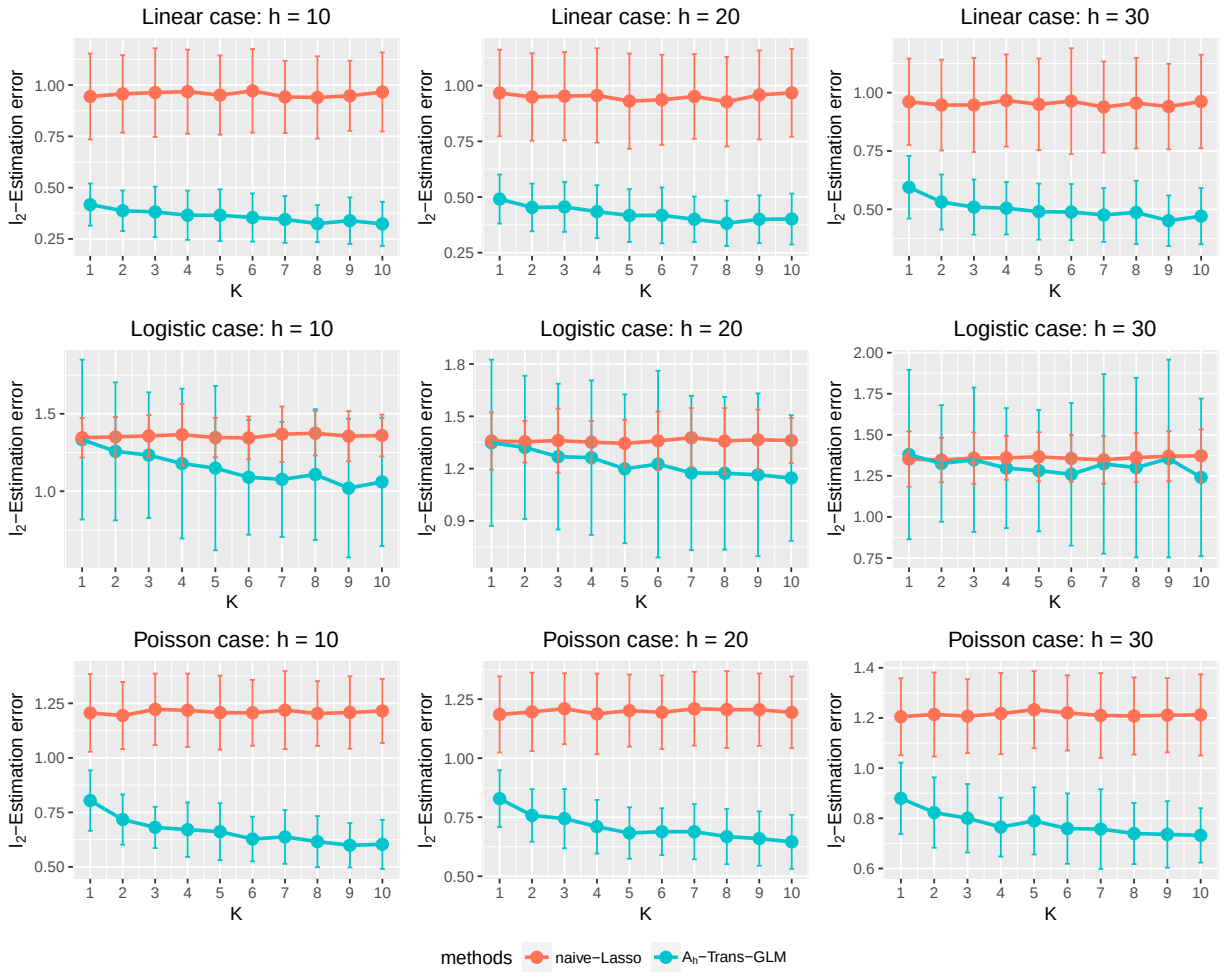


Figure 4: The average $\ell_2$-estimation of $\mathcal{A}_h$-Trans-GLM and naïve-Lasso under linear, logistic and Poisson regression models with different settings of $h$ and $K_{\mathcal{A}_h}$. $n_k = 100$ for all $k = 0, \ldots, K$, $p = 500$, $s = 10$. Error bars denote the standard deviations.

the performance of $\mathcal{A}_h$-Trans-GLM improves significantly (except the logistic case when $h = 30$). This matches our theoretical analysis because the $\ell_2$-estimation error bounds in Theorems 1 and 2 become sharper as $n_{\mathcal{A}_h}$ grows. When $h$ increases, the performance of

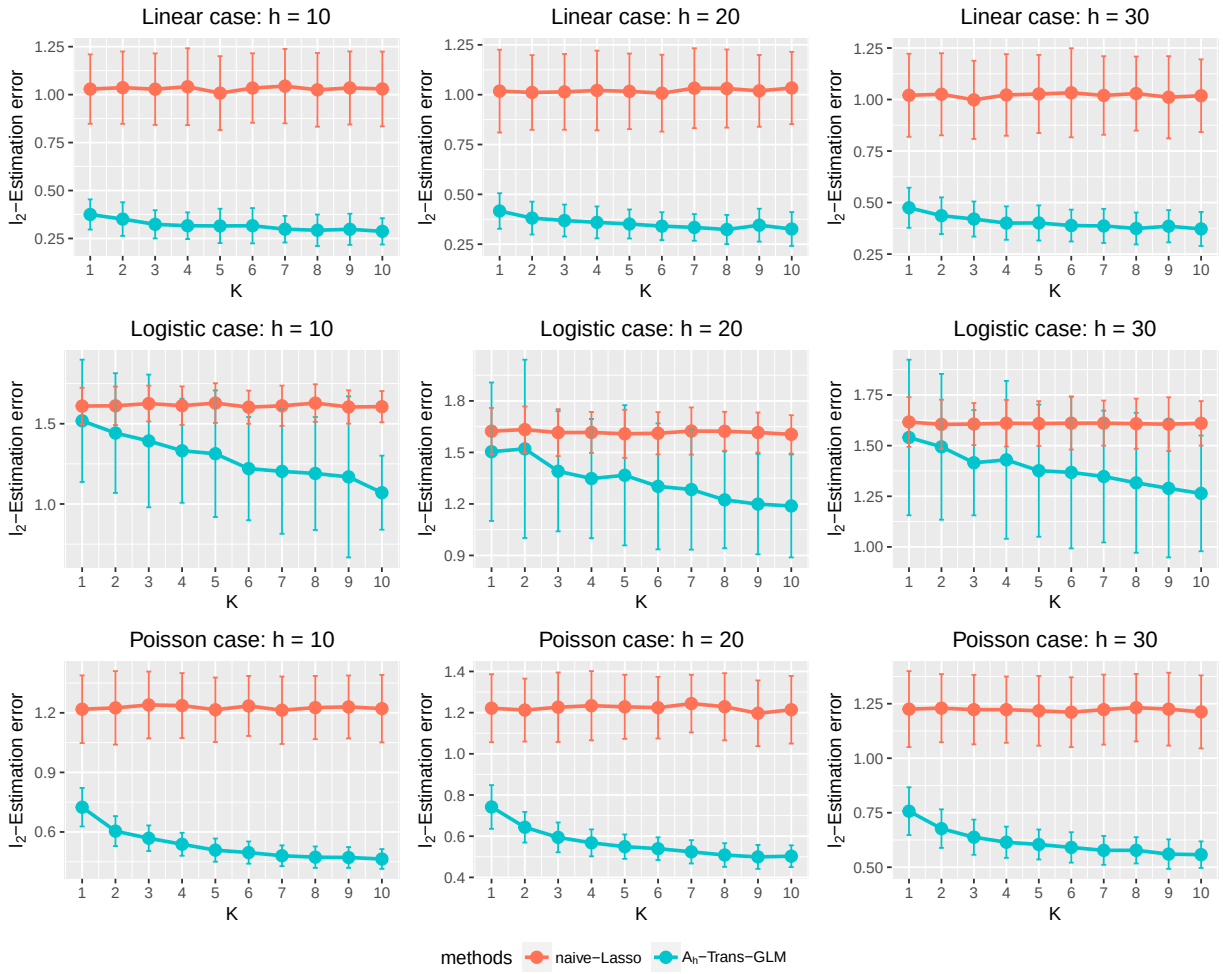


Figure 5: The average $\ell_2$-estimation of $\mathcal{A}_h$-Trans-GLM and naïve-Lasso under linear, logistic and Poisson regression models with different settings of $h$ and $K_{\mathcal{A}_h}$. $n_k = 150$ for all $k = 0, \ldots, K$, $p = 1000$, $s = 15$. Error bars denote the standard deviations.

the oracle algorithm becomes worse.

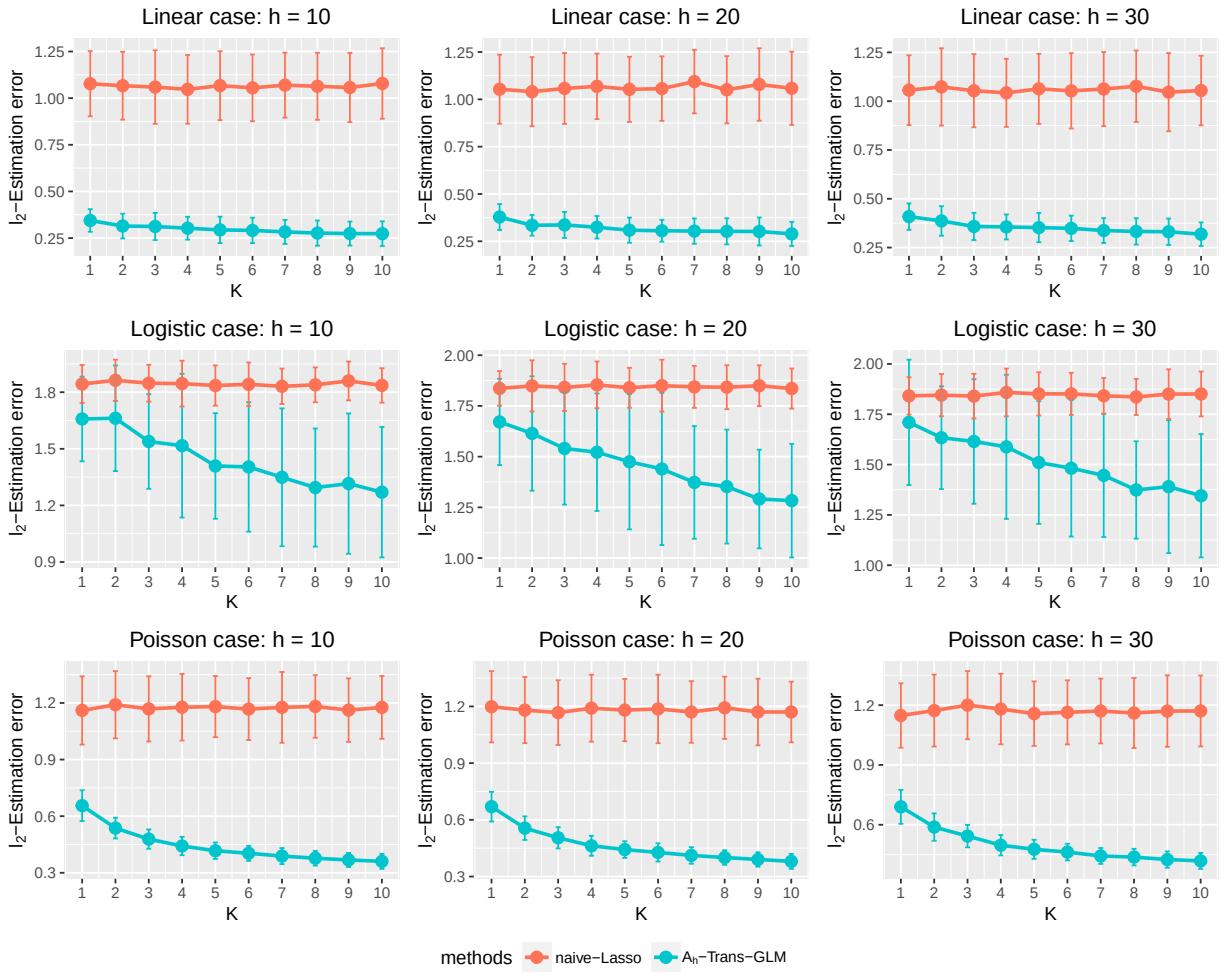


Figure 6: The average $\ell_2$-estimation of $\mathcal{A}_h$-Trans-GLM and naïve-Lasso under linear, logistic and Poisson regression models with different settings of $h$ and $K_{\mathcal{A}_h}$. $n_k = 200$ for all $k = 0, \ldots, K$, $p = 2000$, $s = 20$. Error bars denote the standard deviations.

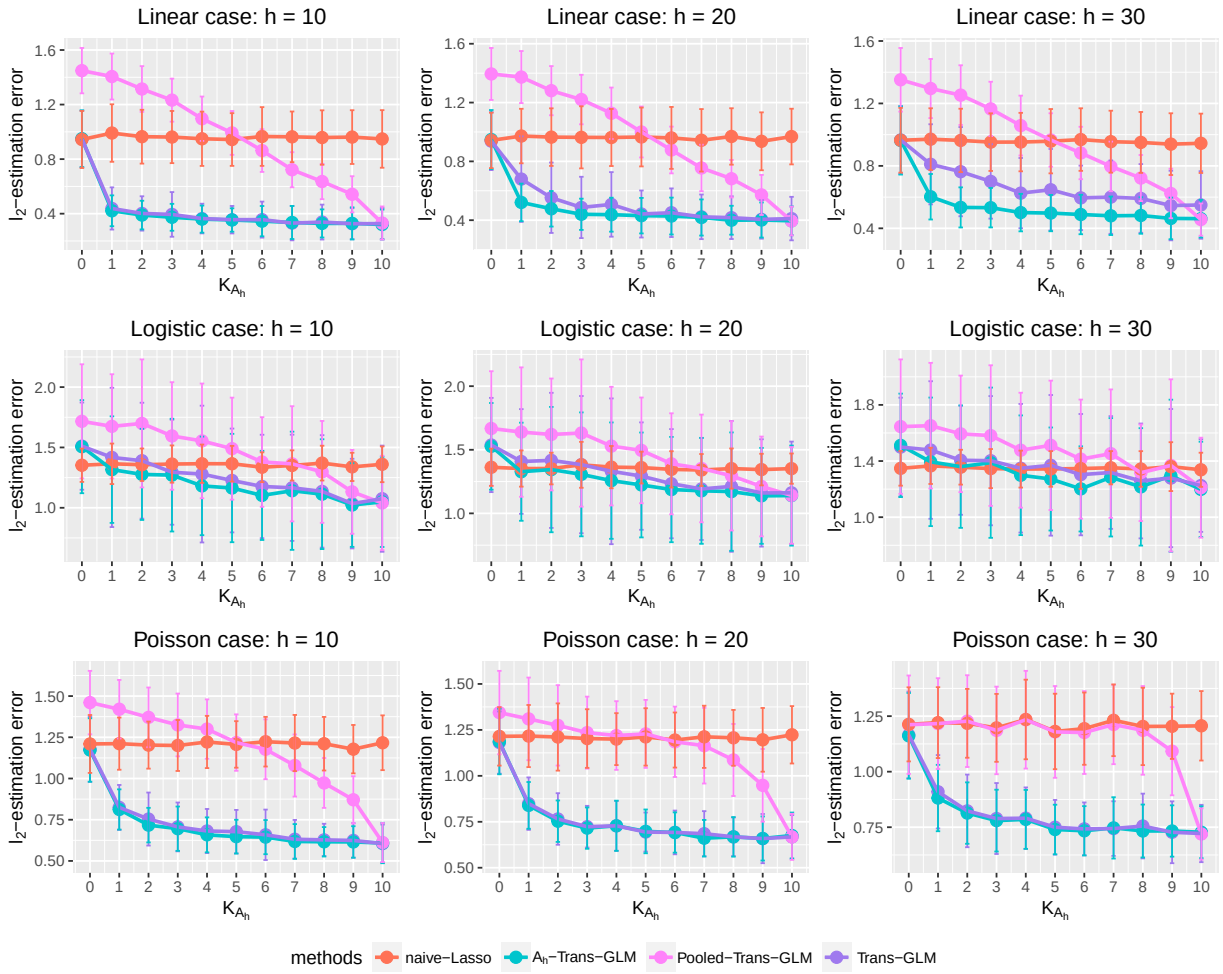


Figure 7: The average $\ell_2$-estimation error of various models with different settings of $h$ and $K_{\mathcal{A}_h}$ when $K = 10$. $n_k = 100$ for all $k = 0, \ldots, K$, $p = 500$, $s = 10$. Error bars denote the standard deviations.

### S.1.3.2 Transferable source detection

In this section we consider the case that some sources are not in the level-$h$ transferring set $\mathcal{A}$. The model settings are the same as those in Section 4.1. We also consider different

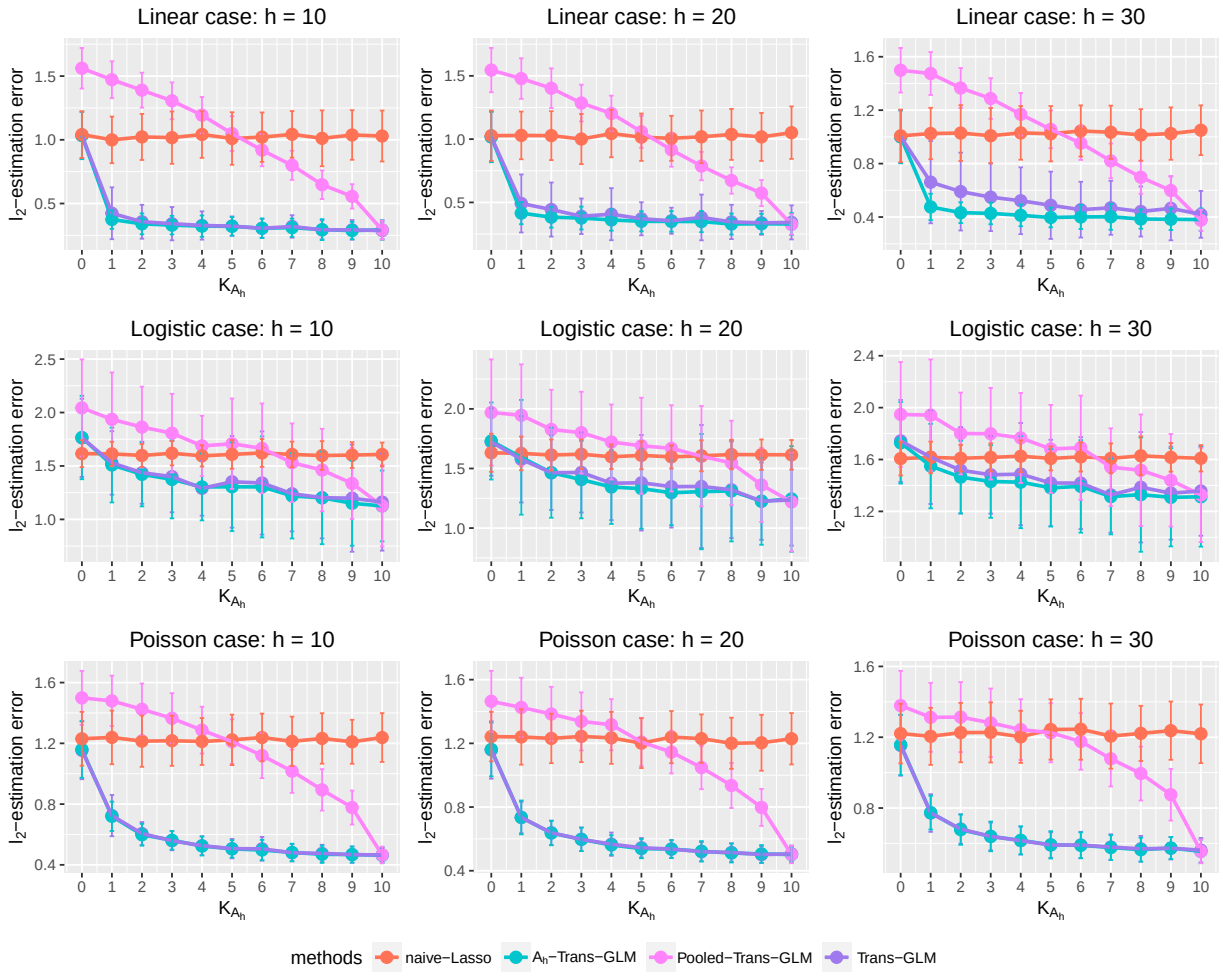


Figure 8: The average $\ell_2$-estimation error of various models with different settings of $h$ and $K_{\mathcal{A}_h}$ when $K = 10$. $n_k = 150$ for all $k = 0, \ldots, K$, $p = 1000$, $s = 15$. Error bars denote the standard deviations.

$(\{n_k\}_{k=0}^K, p, s)$ settings as those in Section S.1.3.1. Note that the case where $n_k = 200$, $p = 2000$ and $s = 20$ has been presented in Section 4.1. 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 7 and 8.

From Figures 7 and 8, we can see 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 $h = 30$, there is a small gap between the estimation error of $\mathcal{A}_h$-Trans-GLM and Trans-GLM, meaning that when $h$ increases, Trans-GLM might begin missing sources in $\mathcal{A}$ or wrongly including sources in $\mathcal{A}$.

## 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 $\psi'$ with corresponding predictors. Denote

$$\begin{aligned}
\hat{L}(\boldsymbol{w}, \mathcal{D}) &= -\frac{1}{n_{\mathcal{A}_h} + n_0} \sum_{k\in\{0\}\cup\mathcal{A}_h} (\boldsymbol{y}^{(k)})^T \boldsymbol{X}^{(k)} \boldsymbol{w} + \frac{1}{n_{\mathcal{A}_h} + n_0} \sum_{k\in\{0\}\cup\mathcal{A}_h} \sum_{i=1}^{n_k} \psi(\boldsymbol{w}^T \boldsymbol{x}_i^{(k)}), \\
\nabla\hat{L}(\boldsymbol{w}, \mathcal{D}) &= -\frac{1}{n_{\mathcal{A}_h} + n_0} \sum_{k\in\{0\}\cup\mathcal{A}_h} (\boldsymbol{X}^{(k)})^T \boldsymbol{y}^{(k)} + \frac{1}{n_{\mathcal{A}_h} + n_0} \sum_{k\in\{0\}\cup\mathcal{A}_h} (\boldsymbol{X}^{(k)})^T \boldsymbol{\psi}'(\boldsymbol{w}^T \boldsymbol{x}^{(k)}), \\
\delta\hat{L}(\boldsymbol{u}, \mathcal{D}) &= \hat{L}(\boldsymbol{w}^{\mathcal{A}_h} + \boldsymbol{u}, \mathcal{D}) - \hat{L}(\boldsymbol{w}^{\mathcal{A}_h}) - \nabla\hat{L}(\boldsymbol{w}^{\mathcal{A}_h})^T \boldsymbol{u}.
\end{aligned}$$

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.

**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* $C_1 \coloneqq \sup_{k\in\{0\}\cup\mathcal{A}_h} \|\widetilde{\boldsymbol{\Sigma}}_h^{-1} \widetilde{\boldsymbol{\Sigma}}_h^{(k)}\|_1 < \infty$.

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

The proof of Lemma 1 will be presented later in Section S.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.

### S.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, $\|\boldsymbol{w}^{\mathcal{A}_h} - \boldsymbol{\beta}\|_1 \leq \sum_{k\in\mathcal{A}} \alpha_k\|\widetilde{\boldsymbol{\Sigma}}_h^{-1}\widetilde{\boldsymbol{\Sigma}}_h^{(k)}\|_1 \cdot \|\boldsymbol{w}^{(k)} - \boldsymbol{\beta}\|_1 \leq C_1 h$.

### S.2.2 Proof of Lemma 2

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

$$\begin{aligned}\delta\hat{L}(\boldsymbol{u},\mathcal{D}) &= \hat{L}(\boldsymbol{w}^{\mathcal{A}_h}+\boldsymbol{u},\mathcal{D}) - \hat{L}(\boldsymbol{w}^{\mathcal{A}_h}) - \nabla\hat{L}^{(k)}(\boldsymbol{w}^{\mathcal{A}_h})^T\boldsymbol{u} \\ &= \frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{k\in\{0\}\cup\mathcal{A}_h}\sum_{i=1}^{n_k}\Big[\psi((\boldsymbol{w}^{\mathcal{A}_h}+\boldsymbol{u})^T\boldsymbol{x}_i^{(k)}) - \psi((\boldsymbol{w}^{\mathcal{A}_h})^T\boldsymbol{x}_i^{(k)}) - \psi'((\boldsymbol{w}^{\mathcal{A}_h})^T\boldsymbol{x}_i^{(k)})\boldsymbol{u}^T\boldsymbol{x}_i^{(k)}\Big] \\ &= \frac{1}{n_{\mathcal{A}_h}+n_0}\sum_{k\in\{0\}\cup\mathcal{A}_h}\sum_{i=1}^{n_k}\psi''((\boldsymbol{w}^{\mathcal{A}_h})^T\boldsymbol{x}_i^{(k)} + t_i^{(k)}\boldsymbol{u}^T\boldsymbol{x}_i^{(k)})(\boldsymbol{u}^T\boldsymbol{x}_i^{(k)})^2,\end{aligned}$$

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.2.3 Proof of Theorem 2

Throughout this proof, we denote $\boldsymbol{w}^{\mathcal{A}_h}$ as any vector $\boldsymbol{w}$ satisfying $\|\boldsymbol{w}-\boldsymbol{\beta}\|_1 \le h$. 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 (9) still holds here, i.e. when $\lambda_{\boldsymbol{w}} \ge 2\|\nabla\hat{L}(\boldsymbol{w}^{\mathcal{A}_h},\mathcal{D})\|_\infty$, with probability at least $1 - C_3\exp\{-C_4(n_{\mathcal{A}_h}+n_0)\}$, it holds that

$$\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_2 \le 8\kappa_2 C_{\boldsymbol{w}} h\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}} + 3\frac{\sqrt{s}}{\kappa_1}\lambda_{\boldsymbol{w}} + 2\sqrt{\frac{1}{\kappa_1}h\lambda_{\boldsymbol{w}}}. \tag{S.2.18}$$

Via the decomposition in (12), $\|\nabla L(\boldsymbol{w}^{\mathcal{A}_h}, \mathcal{D})\|_\infty$ 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 $V_{ij}^{(k)} = x_{ij}^{(k)}[-\psi'((\boldsymbol{x}_i^{(k)})^T\boldsymbol{w}^{(k)}) + \psi'((\boldsymbol{x}_i^{(k)})^T\boldsymbol{w}^{\mathcal{A}_h})]$.

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

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 $C_1^2M_\psi^2\kappa_u^2h^2$-subexponential variable. And by Cauchy-Schwarz inequality and sub-Gaussian properties (Vershynin, 2018),

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

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

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

with probability at least $1-(n_{\mathcal{A}_h}+n_0)^{-1}\vee p^{-1}$, 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 $\lambda_{\boldsymbol{w}} \ge 2\|\nabla L(\boldsymbol{w}^{\mathcal{A}_h}, \mathcal{D})\|_\infty$ in high probability. Plugging it into (S.2.18), we get

$$\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_2 \lesssim \sqrt{\frac{s\log p}{n_{\mathcal{A}_h}+n_0}} + \sqrt{s}h + \sqrt{h}\left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/4}, \tag{S.2.19}$$

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

$$\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1 \lesssim s\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}} + sh + \sqrt{sh}\left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/4}, \tag{S.2.20}$$

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

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2 &\lesssim \sqrt{\frac{\log p}{n_0}}(\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1 + h) + \sqrt{\lambda_{\boldsymbol{\delta}}\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1 + \lambda_{\boldsymbol{\delta}} h} \\ &\lesssim \left(\frac{s\log p}{n_0}\right)^{1/4}\left(\frac{s\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/4} + \left(\frac{s\log p}{n_0}\right)^{1/4} s^{1/4}h^{1/2} \\ &\quad + \left(\frac{s\log p}{n_0}\right)^{1/4}(sh)^{1/4}\left(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\right)^{1/8} + \left(\frac{\log p}{n_0}\right)^{1/4} h^{1/2},\end{aligned}$$

with probability at least $1-n_0^{-1}\vee p^{-1}$. And $\ell_1$-error bound comes from (S.2.20) and (14). On the other hand, if we set $\lambda_{\boldsymbol{\delta}} = C_{\boldsymbol{\delta}}\Big[s\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}} + \Big(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\Big)^{1/4}\sqrt{sh} + h\Big]$ satisfying $\lambda_{\boldsymbol{\delta}} \geq 2\|\nabla L^{(0)}(\hat{\boldsymbol{w}}^{\mathcal{A}_h} + \boldsymbol{\delta}^{\mathcal{A}_h}, \mathcal{D}^{(0)})\|_\infty$ in high probability, we could obtain

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

with probability at least $1-n_0^{-1}\vee p^{-1}$. For the $\ell_1$-error bound, similar to the proof of Theorem 1, it can be easily derived from the fact that $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 \leq \|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1 + \|\hat{\boldsymbol{\delta}}^{\mathcal{A}_h} - \boldsymbol{\delta}^{\mathcal{A}_h}\|_1 \leq \|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1 + 4C_1 h$ with probability at least $1-n_0^{-1}\vee p^{-1}$ and the inequality (14), which completes our proof.

### S.2.4 Proof of Theorem 3

(i) 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,

$$\langle \nabla L_{n_0}^{(0)}(\hat{\boldsymbol{\beta}}) - \nabla L_{n_0}^{(0)}(\boldsymbol{\beta}), \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rangle = \langle -\lambda_{\boldsymbol{\delta}}\cdot\partial\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{w}}^{\mathcal{A}_h}\|_1 - \nabla L_{n_0}^{(0)}(\boldsymbol{\beta}), \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rangle$$

$$\leq \lambda_{\boldsymbol{\delta}} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 + \|\nabla L_{n_0}^{(0)}(\boldsymbol{\beta})\|_\infty \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1. \quad \text{(S.2.21)}$$

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.

(ii) Assume Assumption 3.(i) holds or Assumption 3.(ii) with $n_{\mathcal{A}_h} \geq CU^2\bar{U}^{-2}s^2 \log p$, $h \leq C'\bar{U}U^{-1}s^{-1}$ for some $C$, $C' > 0$ holds. If we take $\lambda_{\boldsymbol{\delta}} = C_{\boldsymbol{\delta}}\Big[s\sqrt{\frac{\log p}{n_{\mathcal{A}_h}+n_0}} + \Big(\frac{\log p}{n_{\mathcal{A}_h}+n_0}\Big)^{1/4}\sqrt{sh} + h\Big]$, where $C_{\boldsymbol{\delta}} > 0$ is a sufficiently large constant: We have

$$\langle \nabla L_{n_0}^{(0)}(\hat{\boldsymbol{\beta}}) - \nabla L_{n_0}^{(0)}(\boldsymbol{\beta}), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle \leq \|\nabla L_{n_0}^{(0)}(\hat{\boldsymbol{\beta}}) - \nabla L_{n_0}^{(0)}(\boldsymbol{\beta})\|_\infty \cdot \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1. \quad \text{(S.2.22)}$$

And

$$\begin{aligned}
\|\nabla L_{n_0}^{(0)}(\hat{\boldsymbol{\beta}}) - \nabla L_{n_0}^{(0)}(\boldsymbol{\beta})\|_\infty &= \sup_{1\leq j\leq p} \left| \frac{1}{n_0} \sum_{i=1}^{n_0} [\psi'((\boldsymbol{x}_i^{(0)})^T\hat{\boldsymbol{\beta}}) - \psi'((\boldsymbol{x}_i^{(0)})^T\boldsymbol{\beta})] x_{ij}^{(0)} \right| \\
&= \sup_{1\leq j\leq p} \left| \frac{1}{n_0} \sum_{i=1}^{n_0} \psi''((\boldsymbol{x}_i^{(0)})^T\boldsymbol{\beta} + v_i^{(0)}(\boldsymbol{x}_i^{(0)})^T\hat{\boldsymbol{v}}^{\mathcal{A}_h}) x_{ij}^{(0)} (\boldsymbol{x}_i^{(0)})^T\hat{\boldsymbol{v}}^{\mathcal{A}_h} \right| \\
&\leq \sup_{1\leq j\leq p} \left\| \frac{1}{n_0} \sum_{i=1}^{n_0} \psi''((\boldsymbol{x}_i^{(0)})^T\boldsymbol{\beta} + v_i^{(0)}(\boldsymbol{x}_i^{(0)})^T\hat{\boldsymbol{v}}^{\mathcal{A}_h}) x_{ij}^{(0)} \boldsymbol{x}_i^{(0)} \right\|_\infty \cdot \|\hat{\boldsymbol{v}}^{\mathcal{A}_h}\|_1,
\end{aligned}$$

where $v_i^{(0)}$ is some constant falling between 0 and 1 and $\hat{\boldsymbol{v}}^{\mathcal{A}_h} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$. When Assumption 3.(i) holds, $\|\psi''\|_\infty$ is bounded, implying that each coordinate of $\psi''((\boldsymbol{x}_i^{(0)})^T\boldsymbol{\beta} + v_i^{(0)}(\boldsymbol{x}_i^{(0)})^T\hat{\boldsymbol{v}}^{\mathcal{A}_h}) x_{ij}^{(0)} \boldsymbol{x}_i^{(0)}$ is a subexponential variable for $i = 1, \ldots, n_0$ and the mean of $\{\psi''((\boldsymbol{x}_i^{(0)})^T\boldsymbol{\beta} + v_i^{(0)}(\boldsymbol{x}_i^{(0)})^T\hat{\boldsymbol{v}}^{\mathcal{A}_h}) x_{ij}^{(0)} \boldsymbol{x}_{ij'}^{(0)}\}_{i,j,j'}$ is uniformly bounded by $\mathcal{O}(1)$. The application of Bernstein inequality, the bound of $\|\hat{\boldsymbol{v}}^{\mathcal{A}_h}\|_1$ and union bounds encloses

$$\|\nabla L_{n_0}^{(0)}(\hat{\boldsymbol{\beta}}) - \nabla L_{n_0}^{(0)}(\boldsymbol{\beta})\|_\infty \lesssim s\sqrt{\frac{\log p}{n_{\mathcal{A}_h} + n_0}} + \left(\frac{\log p}{n_{\mathcal{A}_h} + n_0}\right)^{1/4} \sqrt{sh} + h, \quad \text{(S.2.23)}$$

with probability at least $1 - n_0^{-1} \vee p^{-1}$. On the other hand, when Assumption 3.(ii) holds, combine the condition $n_{\mathcal{A}_h} \geq CU^2\bar{U}^{-2}s^2 \log p$ and the $\ell_1$-bound of $\hat{\boldsymbol{v}}^{\mathcal{A}_h}$, we can get the same high probability bound in (S.2.23). Therefore, combine (S.2.23) and (S.2.22), we have

$$\|\nabla L_{n_0}^{(0)}(\hat{\boldsymbol{\beta}}) - \nabla L_{n_0}^{(0)}(\boldsymbol{\beta})\|_\infty \lesssim \left[ s\sqrt{\frac{\log p}{n_{\mathcal{A}_h} + n_0}} + \left(\frac{\log p}{n_{\mathcal{A}_h} + n_0}\right)^{1/4} \sqrt{sh} + h \right]^2,$$

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

### S.2.5 Proof of Theorem 4

(i) 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.2.21), 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 (16) and the upper bound of $\|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1$ in (S.2.20). Then the final upper bound follows.

(ii) Assume Assumption 3.(i) holds or Assumption 3.(ii) with $n_{\mathcal{A}_h} \geq CU^2\bar{U}^{-2}s^2 \log p$, $h \leq C'\bar{U}U^{-1}s^{-1}$ holds for some $C$, $C' > 0$. If we take $\lambda_{\boldsymbol{\delta}} = C_{\boldsymbol{\delta}}\left[ s\sqrt{\frac{\log p}{n_{\mathcal{A}_h} + n_0}} + \left(\frac{\log p}{n_{\mathcal{A}_h} + n_0}\right)^{1/4} \sqrt{sh} + h \right]$, where $C_{\boldsymbol{\delta}} > 0$ is a sufficiently large constant: In this case, by the proof of Theorem 2, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \leq \|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1 + \|\hat{\boldsymbol{\delta}}^{\mathcal{A}_h} - \boldsymbol{\delta}^{\mathcal{A}_h}\|_1 \leq \|\hat{\boldsymbol{u}}^{\mathcal{A}_h}\|_1 + 4C_1 h$ with probability at least $1 - n_0^{-1} \vee p^{-1}$, where $\hat{\boldsymbol{u}}^{\mathcal{A}_h} = \hat{\boldsymbol{w}}^{\mathcal{A}_h} - \boldsymbol{w}^{\mathcal{A}_h}$. Then the same discussion as in part (ii) of the proof of Theorem 3 leads to the final bound of $\langle \nabla L_{n_0}^{(0)}(\hat{\boldsymbol{\beta}}) - \nabla L_{n_0}^{(0)}(\boldsymbol{\beta}), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \rangle$.

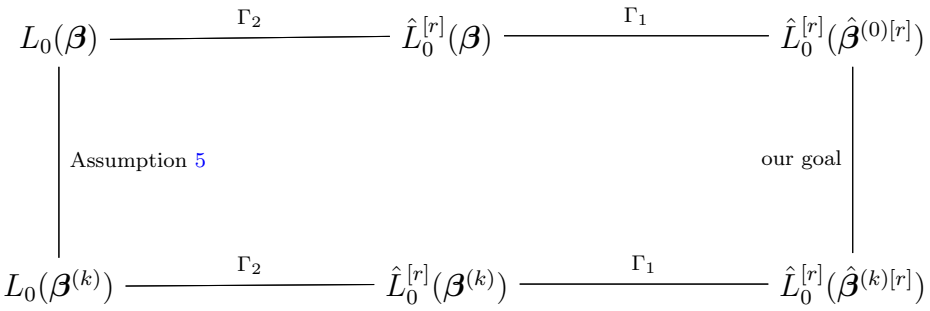


Figure 9: The idea behind Assumptions 4, 5 and Theorem 5.

### S.2.6 Proof of Theorem 5

**Lemma 3.** *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 3.* 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

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

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 5. We have

$$\begin{aligned}
\hat{\sigma} &= \sqrt{\sum_{r=1}^{3} (\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(0)[r]}) - L_0(\boldsymbol{\beta}))^2/3} \leq \sqrt{\frac{2}{3}} \cdot \sum_{r=1}^{3} |\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(0)[r]}) - L_0(\boldsymbol{\beta})| \\
&\leq \sqrt{\frac{2}{3}} \cdot \sum_{r=1}^{3} \left[ |\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(0)[r]}) - \hat{L}_0^{[r]}(\boldsymbol{\beta})| + |\hat{L}_0^{[r]}(\boldsymbol{\beta}) - L_0(\boldsymbol{\beta})| \right] \\
&\lesssim \zeta \left( \Gamma_1^{(0)} + \Gamma_2^{(0)} \right),
\end{aligned}$$

with probability at least $1 - g_1^{(0)}(\zeta) - g_2^{(0)}(\zeta)$. As Figure 9 shows, by Lemma 3, for $k \in \mathcal{A}$, there holds

$$\begin{aligned}
\sup_r |\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]}) - \hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(0)[r]})| &\leq 2 \sup_r |\hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]}) - \hat{L}_0^{[r]}(\boldsymbol{\beta}^{(k)[r]})| \\
&\quad + \sup_r |\hat{L}_0^{[r]}(\boldsymbol{\beta}^{(k)}) - \hat{L}_0^{[r]}(\boldsymbol{\beta}) - L_0(\boldsymbol{\beta}^{(k)}) + L_0(\boldsymbol{\beta})| \\
&\quad + |L_0(\boldsymbol{\beta}^{(k)}) - L_0(\boldsymbol{\beta})| \\
&\lesssim \zeta \left( \Gamma_1^{(k)} + \Gamma_2^{(k)} + h^2 \right) \\
&\leq C_0(\hat{\sigma} \vee 0.01), \qquad \text{(S.2.24)}
\end{aligned}$$

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 $k \in \mathcal{A}^c$,

$$\inf_r \hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(k)[r]}) - \hat{L}_0^{[r]}(\hat{\boldsymbol{\beta}}^{(0)[r]})$$

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

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

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

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.2.24) 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.2.7 Proof of Proposition 1

The rate of $\Gamma_1$ can be derived from the following Lemma 4 and the union bound, together with the tail inequality (S.2.19). The rate of $\Gamma_2$ comes from the following Lemma 5.

**Lemma 4.** *Under the same assumptions as Theorem 5, 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* $\sup_k \|\boldsymbol{x}^{(k)}\|_\infty \le U$ *a.s.:*

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

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

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

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

**Remark 8.** *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 4.* 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 2, 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

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

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

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

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

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

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 5.** *Under the same assumptions as Theorem 5, 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.:*

$$|\hat{L}_0^{[r]}(\boldsymbol{\beta}) - L_0(\boldsymbol{\beta})| \vee |\hat{L}_0^{[r]}(\boldsymbol{\beta}^{(k)}) - \hat{L}_0^{[r]}(\boldsymbol{\beta}) - L_0(\boldsymbol{\beta}^{(k)}) + L_0(\boldsymbol{\beta})|$$

$$\leq C\sqrt{\frac{1}{n_0}}\exp\left(U\|\boldsymbol{w}^{(k)}\|_1\right)\left[1+\|\boldsymbol{w}^{(k)}\|_2+U\|\boldsymbol{w}^{(k)}\|_1\right]\cdot\zeta,$$

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

*Proof of Lemma 5.* Similar to the proofs of Lemmas 3 and 4, 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

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

(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

$$\frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}[y_i^{(0)[r]}-\psi'(\boldsymbol{\beta}^T\boldsymbol{x}_i^{(0)[r]})](\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}_i^{(0)[r]}\right|\lesssim\sqrt{\frac{1}{n_0}}\|\boldsymbol{w}^{(k)}\|_2\cdot\zeta,\tag{S.2.25}$$

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

$$\frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}\left[\psi'(\boldsymbol{\beta}^T\boldsymbol{x}_i^{(0)[r]})(\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}_i^{(0)[r]}-\mathbb{E}[\psi'(\boldsymbol{\beta}^T\boldsymbol{x}^{(0)})(\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}^{(0)}]\right]\right|\lesssim\sqrt{\frac{1}{n_0}}\|\boldsymbol{w}^{(k)}\|_2\cdot\zeta, \tag{S.2.26}$$

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,

$$\frac{1}{n_0/3}\left|\sum_{i=1}^{n_0/3}\left[\psi((\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}_i^{(0)[r]})-\mathbb{E}\psi((\boldsymbol{\beta}^{(k)})^T\boldsymbol{x}^{(0)})\right]\right|\lesssim\sqrt{\frac{1}{n_0}}\cdot\|\boldsymbol{w}^{(k)}\|_2\cdot\zeta, \tag{S.2.27}$$

with probability at least $1-\exp\{-\zeta^2\}$. By (S.2.25), (S.2.26) and (S.2.27), 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

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

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

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

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

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

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

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

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

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

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.

□