Estimation and Inference for High-Dimensional Generalized Linear Models with Knowledge Transfer

Sai Li, Linjun Zhang, T. Tony Cai and Hongzhe Li

Abstract

Transfer learning provides a powerful tool for incorporating related data into a target study of interest. In epidemiology and medical studies, the classification of a target disease could borrow information across diseases and populations. In this work, we consider transfer learning for high-dimensional generalized linear models (GLMs). A novel algorithm, TransGLM, that incorporates data from the target study as well as the auxiliary studies is proposed. Minimax rate of convergence for estimation is established and the proposed estimator is shown to be rate-optimal.

Statistical inference for the target regression coefficients is also studied. Asymptotic normality for a debiased estimator is established and confidence intervals are constructed. Numerical studies show significant improvements in estimation and inference accuracy. Proposed methods are applied to a real data study concerning the classification of colorectal cancer using microbiomes, and are shown to enhance the classification accuracy in comparison to the single-task methods.

Keywords: Generalized linear models, meta learning, multi-task learning, debiased estimator.

Sai Li is a postdoctoral researcher, Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *sai.li@pennmedicine.upenn.edu*). Linjun Zhang is an assistant professor at Rutgers University(). T. Tony Cai is Daniel H. Silberberg Professor of Statistics, Department of Statistics, the Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (E-mail:*tcai@wharton.upenn.edu*). Hongzhe Li is Perelman Professor of Biostatistics, Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *hongzhe@upenn.edu*).

1 Introduction

Generalized linear models (GLMs) are widely used in many areas of statistical applications (Hastie et al., 2009). In genetic applications and other medical studies, the number of covariates can be quite large and high-dimensional GLMs are frequently adopted for classifying diseases and health-related outcomes. In the age of big data, the availability of public datasets makes it possible to improve the learning performance of a new study by incorporating information from the existing ones. This is the goal of transfer learning, which aims to incorporate the knowledge from different but related studies to enhance the accuracy of the target study of interest (Torrey and Shavlik, 2010). Transfer learning has been successfully applied in a range of different fields, including pattern recognition, natural language processing, and drug discovery (Pan and Yang, 2009; Turki et al., 2017; Bastani, 2018). In particular, transfer learning for the GLMs has been used in image classification and disease diagnosis (Hosny et al., 2018; Sevakula et al., 2018). However, little is known about their statistical guarantees.

In this paper, we study transfer learning for high-dimensional GLMs in the setting where the data are available from a target study and multiple auxiliary studies. In the target study, we observe n_0 *i.i.d.* samples $\boldsymbol{x}_i^{(0)} \in \mathbb{R}^p$ and $y_i^{(0)} \in \mathcal{Y} \subseteq \mathbb{R}$, $i = 1, \ldots, n_0$ drawn from a GLM with parameter $\boldsymbol{\beta} \in \mathbb{R}^p$. The negative log-likelihood for the target data is

$$L^{(0)}(\boldsymbol{\beta}) = \sum_{i=1}^{n_0} \{\psi((\boldsymbol{x}_i^{(0)})^\top \boldsymbol{\beta}) - y_i^{(0)} \cdot (\boldsymbol{x}_i^{(0)})^\top \boldsymbol{\beta}\}, \tag{1}$$

where ψ is the link function that satisfies certain smoothness conditions. Additionally, we have observations from K different auxiliary studies. For $k = 1, \ldots, K$, let $(\boldsymbol{x}_i^{(k)}, y_i^{(k)})$, $i = 1, \ldots, n_k$, denote the observations from the k-th study drawn from a GLM with the parameter $\boldsymbol{w}^{(k)} \in \mathbb{R}^p$ and link function ψ. The similarity between the k-th study and the target study is captured by the contrast vector $\boldsymbol{\delta}^{(k)} = \boldsymbol{w}^{(k)} - \boldsymbol{\beta}$. The smaller the magnitude of $\boldsymbol{\delta}^{(k)}$, the higher the similarity. Let h denote the similarity level such that $\max_{1\leq k\leq K}\|\boldsymbol{\delta}^{(k)}\|_q \leq h$ for some fixed $q \in [0,1]$. Specifically, $q=0$ corresponds to the exact sparse contrast vectors and when $q>0$, $\boldsymbol{\delta}^{(k)}$ can have many nonzero coefficients but their magnitude decays relatively fast. The range of q in consideration is flexible in applications and our proposed method can adapt to q.

The goal is to optimally estimate and make inference for the target parameter $\boldsymbol{\beta} \in \mathbb{R}^p$ based on the available data from the target and auxiliary studies.

1.1 Related work

In the conventional setting where only data from the target study is available, estimation for high-dimensional GLMs has been well-studied. Van de Geer (2008) uses ℓ_1-penalty and derives an oracle inequality and estimation error rates. Negahban et al. (2012) studies M-estimators and proves estimation error rates under the restricted strong convexity condition. Huang and Zhang (2012) considers convex loss functions with weighted Lasso penalties. van de Geer et al. (2014) proposes a debiasing procedure for inference by computing the correction score via another Lasso on the Hessian matrix. Cai et al. (2020a) introduces a debiasing procedure for the GLMs with binary outcomes via quadratic optimization. The idea of debiasing has also been generalized to tackle high-dimensional proportional hazards models (Fang et al., 2017), mixed-effects models (Bradic et al., 2020; Li et al., 2019), and for multiple testing (Zhang and Cheng, 2017; Dezeure et al., 2017; Javanmard and Javadi, 2019; Ma et al., 2020).

Transfer learning has been studied in different models. Cai and Wei (2021) considers nonparametric classification and establishes the minimax optimal rate and proposes an adaptive classifier. Tripuraneni et al. (2020a) proposes an algorithm in linear models that assumes all the auxiliary studies and the target study share a common, low-dimensional linear representation. Transfer learning in general functional classes has been studied in Tripuraneni et al. (2020b) and Hanneke and Kpotufe (2020). Li et al. (2020a) proposes methods for transfer learning in high-dimensional linear models and establishes the minimax optimal rate. Li et al. (2020b) introduces a method for estimation and edge detection in high-dimensional Gaussian graphical models with knowledge transfer. However, the methods established in the aforementioned two papers cannot be directly used as the link functions in GLMs are nonlinear in general. Liang et al. (2020) studies high-dimensional classification with auxiliary outcomes in the setting that the same set of individuals are used to generate different outcomes, which is different from our setting.

A related but different problem is multi-task learning (Zhang and Yang, 2017), where the goal is to jointly estimate all the parameters for multiple tasks. Multi-task learning has been studied in various settings, including linear regression (Agarwal et al., 2012; Dondelinger et al., 2020) and graphical models (Chen et al., 2010; Danaher et al., 2014). An optimal multi-task procedure does not necessarily yield an optimal estimator for the target task in transfer learning.

1.2 Our contributions

A novel algorithm is developed for estimation and inference in high-dimensional GLMs with knowledge transfer. The proposed method estimates the target parameter and contrast vectors jointly via constrained ℓ_1-minimization. Minimax rate of convergence is established and the proposed estimator is shown to attain the optimal rate under mild conditions. The optimal rate for transfer learning is faster than the corresponding rate in the single-task setting under mild similarity conditions between the auxiliary and target tasks.

A debiasing method is introduced in the transfer learning setting. The debiased estimator of an individual coefficient is shown to be asymptotically normal and is then used for constructing confidence intervals. It is shown that this debiased estimator has a smaller magnitude of remaining bias in comparison to the one in the single-task setting. As a result, the asymptotic normality holds under weaker sparsity conditions on $\boldsymbol{\beta}$ in transfer learning when the auxiliary studies are sufficiently informative. Consequently, inference for a given coefficient β_j is no longer restricted to the "ultra-sparse" regime for $\boldsymbol{\beta}$. This reveals the benefit of transfer learning for statistical inference.

1.3 Organization

The rest of the paper is organized as follows. In Section 2, we introduce a transfer learning algorithm using a constrained ℓ_1-minimization approach for estimation in GLMs. Section 3 provides the theoretical guarantees for our proposal and establishes the minimax lower bound. In Section 4, we introduce a debiasing procedure for inference of β_j and prove its asymptotic normality. To guarantee positive transfer, an aggregation procedure is developed in Section 5. Section 6 considers the numerical performance of our proposed algorithms in comparison to some existing methods. The results provide empirical evidence of the gain of transfer learning. The proposed methods are applied to analyze a microbiome data set for classifying colorectal cancer in Section 7. The results demonstrate the advantage of transfer learning. The proofs and additional numerical results are given in the supplementary materials (Li et al., 2021).

1.4 Notation

For two sequences of positive numbers $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ if $a_n \leq cb_n$ for some universal constant $c \in (0, \infty)$, and $a_n \gtrsim b_n$ if $a_n \geq c'b_n$ for some universal constant $c' \in (0, \infty)$. We say $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. We use $c, C, c_0, c_1, c_2, \cdots$, and so on to denote universal constants. Their specific values may vary from place to place. For an integer $k > 0$, $[k]$ denotes the set $\{1, 2, ..., k\}$. For a vector $\boldsymbol{v} \in \mathbb{R}^d$ and a subset $S \subseteq [d]$, we use $\boldsymbol{v}_S$ to denote the restriction of vector $\boldsymbol{v}$ to the index set S. We write $\text{supp}(\boldsymbol{v}) := \{j \in [d] : v_j \neq 0\}$. Let $\|\boldsymbol{v}\|_p = (\sum_{j=1}^d |v_j|^p)^{1/p}$ for $0 < p \leq \infty$, and let $\|\boldsymbol{v}\|_0$ denote the number of non-zero coordinates of $\boldsymbol{v}$. For a function $f : \mathbb{R} \to \mathbb{R}$, $\|f\|_\infty$ denotes the the essential supremum of $|f|$ and $\dot{f}$ and $\ddot{f}$ denote the first and second derivatives respectively. The sub-Gaussian norm of a random variable $u \in \mathbb{R}$ is $\|u\|_{\psi_2} = \sup_{l \geq 1} l^{-1/2} \mathbb{E}^{1/l}[|u|^l]$ and the sub-Gaussian norm of a random vector $\boldsymbol{U} \in \mathbb{R}^n$ is $\|\boldsymbol{U}\|_{\psi_2} = \sup_{\|\boldsymbol{v}\|_2 = 1, \boldsymbol{v} \in \mathbb{R}^n} \|\langle \boldsymbol{U}, \boldsymbol{v} \rangle\|_{\psi_2}$. Let z_α be the $(1-\alpha)$-th quantile of the standard normal distribution.

2 Transfer learning via constrained ℓ_1-minimization

In this section, we introduce our proposed algorithm via constrained ℓ_1-minimization. We begin with preliminaries and model setup in Section 2.1. The rationale behind the proposed method is described in Section 2.2 and the algorithm for estimating $\boldsymbol{\beta}$ is introduced in Section 2.3. The theoretical guarantees, including both the upper and matching lower bounds, are provided in Section 3.

2.1 Model setup

Formally, the target model can be written as

$$f_{\boldsymbol{\beta}}(y_i^{(0)} | \boldsymbol{x}_i^{(0)}) = h(y, \sigma^{(0)}) \exp\left(\frac{(\boldsymbol{x}_i^{(0)})^\top \boldsymbol{\beta} \cdot y_i^{(0)} - \psi((\boldsymbol{x}_i^{(0)})^\top \boldsymbol{\beta})}{c(\sigma^{(0)})} \right), \tag{2}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the target parameter of interest, $c(\sigma^{(0)})$ is a nuisance scale parameter and $\psi(\cdot)$ is the cumulant generating function of y given $\boldsymbol{x}$. The GLM is, first of all, a generalization of the linear model: setting $\psi(\mu) = \mu^2/2$ and $c(\sigma) = \sigma^2$ in (2) recovers the (Gaussian) linear model. Model (2) also includes other popular models such as logistic, multinomial and Poisson regression models. In the high-dimensional regime where p can be much larger than the sample size n_0, $\boldsymbol{\beta}$ is often assumed to be sparse such that the number of nonzero elements of $\boldsymbol{\beta}$, denoted by s, is much smaller than p. With *i.i.d.* samples $\{(y_i^{(0)}, \boldsymbol{x}_i^{(0)})\}_{i=1}^{n_0}$ drawn from the model (2), the general approach is to minimize the negative log-likelihood function (1) with some sparsity-induced penalty.

In the context of transfer learning, we additionally observe $\{(y_i^{(k)}, \boldsymbol{x}_i^{(k)})\}_{i=1}^{n_k}$, $k = 1, ..., K$, generated from the auxiliary models

$$f_{\boldsymbol{w}^{(k)}}(y_i^{(k)}|\boldsymbol{x}_i^{(k)}) = h(y_i^{(k)}, \sigma^{(k)}) \exp\left(\frac{(\boldsymbol{x}_i^{(k)})^\top \boldsymbol{w}^{(k)} \cdot y_i^{(k)} - \psi((\boldsymbol{x}_i^{(k)})^\top \boldsymbol{w}^{(k)})}{c(\sigma^{(k)})}\right), \qquad (3)$$

where $\boldsymbol{w}^{(k)} \in \mathbb{R}^p$ is the coefficient vector for the k-th study satisfying $\boldsymbol{w}^{(k)} = \boldsymbol{\beta} + \boldsymbol{\delta}^{(k)}$. For convenience, we define $\boldsymbol{\delta}^{(0)} = 0$. As described in Section 1, we assume $\max_{1\leq k\leq K} \|\delta^{(k)}\|_q \leq h$ for some constant $q \in [0, 1]$. We will introduce the estimator for $\boldsymbol{\beta}$ in the sequel.

2.2 Rationale from moment equations

To estimate $\boldsymbol{\beta}$ and $\{\boldsymbol{\delta}^{(k)}\}_{k=1}^K$, we start with the moment equations. Let $\dot{\psi}(\mu) = \partial\psi(\mu)/\partial\mu$. The function $\dot{\psi}(\mu)$ is nonlinear in general. For instance, $\dot{\psi}(\mu) = 1/(1+\exp(-\mu))$ for logistic regression. The score functions based on the likelihood functions (2) and (3) satisfy

$$\mathbb{E}\left[\boldsymbol{x}_i^{(k)}\big\{y_i^{(k)} - \dot{\psi}\big((\boldsymbol{x}_i^{(k)})^\top(\boldsymbol{\beta} + \boldsymbol{\delta}^{(k)})\big)\big\}\right] = 0, \;\; k = 0, \ldots, K. \qquad (4)$$

These $(K+1)\times p$ moment equations guarantee the identifiability of the unknown parameters $\boldsymbol{\beta}$ and $\{\boldsymbol{\delta}^{(k)}\}_{k=1}^K$. As $\boldsymbol{\beta}$ and $\{\boldsymbol{\delta}^{(k)}\}_{k=1}^K$ are assumed to be (approximately) sparse, we will consider a sparsity-induced estimator based on the moment equations.

As opposed to transfer learning for linear models, we see from (4) that there is no way to separate the estimation of $\boldsymbol{\beta}$ and $\{\boldsymbol{\delta}^{(k)}\}_{k=1}^K$ in GLMs. This brings additional challenges in devising the algorithm and in the theoretical analysis. We propose a constrained optimization algorithm for jointly estimating the target parameter $\boldsymbol{\beta}$ and contrast vectors $\{\boldsymbol{\delta}^{(k)}\}_{k=1}^{K}$. For a parameter vector $\boldsymbol{b} \in \mathbb{R}^p$, we denote the empirical score function by $\dot{L}^{(k)}(\boldsymbol{b}) = \sum_{i=1}^{n_k} \boldsymbol{x}_i^{(k)}(y_i^{(k)} - \dot{\psi}((\boldsymbol{x}_i^{(k)})^\top \boldsymbol{b}))$. We consider

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}^{(1)}, ..., \hat{\boldsymbol{\delta}}^{(K)}) = \underset{\boldsymbol{\beta}, \{\boldsymbol{\delta}^{(k)}\}_{k=1}^{K}}{\arg\min} \left\{ \lambda_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 + \sum_{k=1}^{K} \lambda_k \|\boldsymbol{\delta}^{(k)}\|_1 \right\} \tag{5}$$
$$\text{subject to} \begin{cases} \left\| \dot{L}^{(k)}(\boldsymbol{\beta} + \boldsymbol{\delta}^{(k)}) \right\|_\infty \leq \lambda_k, \text{ for } 0 \leq k \leq K \\ \left\| \dot{L}^{(0)}(\boldsymbol{\beta}) + \sum_{k=1}^{K} \dot{L}^{(k)}(\boldsymbol{\beta} + \boldsymbol{\delta}^{(k)}) \right\|_\infty \leq \lambda_{\boldsymbol{\beta}}, \end{cases}$$

where $\lambda_{\boldsymbol{\beta}}$ and λ_k, $1 \leq k \leq K$ are the tuning parameters and will be specified later. The objective function in (5) encourages sparse solutions. Notice that there are $(K+2) \times p$ constraints in (5) while there are $(K+1) \times p$ unknown parameters. All these constraints are essential. Specifically, the constraint $\|\dot{L}^{(0)}(\boldsymbol{\beta})\|_\infty \leq \lambda_0$ is inherited from the target model, imposing that $\boldsymbol{\beta}$ should be identified as the true parameter for the target model. The constraint $\|\dot{L}^{(k)}(\boldsymbol{\beta}+\boldsymbol{\delta}^{(k)})\|_\infty \leq \lambda_k$ comes from the score functions from k-th auxiliary study, imposing that $\boldsymbol{\delta}^{(k)}$ should be identified as $\boldsymbol{w}^{(k)} - \boldsymbol{\beta}$. The last constraint in (5) aggregates the moment equations for all the studies in use. It ensures that the estimation of $\boldsymbol{\beta}$ borrows information across auxiliary studies. Specifically, imagining $\{\boldsymbol{\delta}^{(k)}\}_{k=1}^{K}$ are known, the last constraint ensures that $\boldsymbol{\beta}$ is estimated based on N independent samples and hence can lead to a faster convergence rate. We formalize the transfer learning algorithm in Section 2.3.

2.3 Estimation of the target parameter

Our proposed algorithm for estimating $\boldsymbol{\beta}$ is a detailed version of (5). In Step 1, an initial estimator of $\boldsymbol{\beta}$ is constructed by minimizing an ℓ_1-penalized negative likelihood based only on the target data. In Step 2, we modify (5) by adding one more constraint using the initial estimator. We now introduce the detail algorithm and then provide further comments on the algorithm. Let $\boldsymbol{x}_i^{(k)}$ be the i-th row of $\boldsymbol{X}^{(k)}$ and $y_i^{(k)}$ be the i-th element of $\boldsymbol{y}^{(k)}$, $k = 0, \ldots, K$.

Algorithm 1: TransGLM, transfer learning via constrained ℓ_1-minimization

Input : Target data $(\boldsymbol{X}^{(0)}, \boldsymbol{y}^{(0)})$, auxiliary data $\{(\boldsymbol{X}^{(k)}, \boldsymbol{y}^{(k)})\}_{k=1}^{K}$, tuning parameter

$$\lambda_k = c_2\sqrt{\frac{\log p}{n_0 \wedge n_k}},\ 0 \le k \le K,\ \text{and } \lambda_{\boldsymbol{\beta}} \text{ as in (8)}.$$

Output: $\hat{\boldsymbol{\beta}}$.

Step 1: Compute an initial estimator

$$\hat{\boldsymbol{\beta}}^{(init)} = \arg\min_{\boldsymbol{b}\in\mathbb{R}^p}\left\{L^{(0)}(\boldsymbol{b}) + \lambda_0\|\boldsymbol{b}\|_1\right\}.$$

Step 2:

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}^{(1)}, ..., \hat{\boldsymbol{\delta}}^{(K)}) = \arg\min_{\boldsymbol{\beta}, \|\boldsymbol{\delta}^{(k)}\|_2 \le C}\left\{\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_1 + \sum_{k=1}^{K}\lambda_k\|\boldsymbol{\delta}^{(k)}\|_1\right\} \tag{6}$$

$$\text{subject to}\begin{cases}\left\|\dot{L}^{(k)}(\boldsymbol{\beta}+\boldsymbol{\delta}^{(k)})\right\|_\infty \le \lambda_k, \forall\ 0 \le k \le K\\ \left\|\dot{L}^{(0)}(\boldsymbol{\beta}) + \sum_{k=1}^{K}\dot{L}^{(k)}(\boldsymbol{\beta}+\boldsymbol{\delta}^{(k)})\right\|_\infty \le \lambda_{\boldsymbol{\beta}}\\ \left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(init)}\right\|_1 \le \lambda_0^{-1}.\end{cases} \tag{7}$$

In comparison to (5), the last ℓ_1-constraint $\left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(init)}\right\|_1 \le \lambda_0^{-1}$ in (7) is needed for technical convenience when $\dot{\psi}(\mu)$ is nonlinear. This constraint is mild as $\lambda_0 = o(1)$. This condition can be removed if the target parameter satisfies $\|\boldsymbol{\beta}\|_1 \le c\lambda_0^{-1}$ for some positive constant c. Computationally, the joint optimization in (6) is still a convex programming.

The proposed algorithm can also be used for multi-task GLM learning, where the goal is to jointly estimate $\boldsymbol{\beta}$ and $\{\boldsymbol{w}^{(k)}\}_{k=1}^{K}$ (Zhang and Yang, 2017). Specifically, after fitting $\boldsymbol{\beta}$ and

$\boldsymbol{\delta}^{(k)}$ with the proposed algorithm, one can estimate $\boldsymbol{w}^{(k)}$ with $\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\delta}}^{(k)}$. The corresponding convergence rate is implied by the results in Section 3.

3 Theoretical guarantees for estimation

In this section, we establish the minimax optimal rate and show that the proposed estimator is rate optimal. Define the population Hessian matrices as

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \mathbb{E}[\boldsymbol{x}_i^{(0)}(\boldsymbol{x}_i^{(0)})^\top \ddot{\psi}((\boldsymbol{x}_i^{(0)})^\top \boldsymbol{\beta})],\ \boldsymbol{\Sigma}_{\boldsymbol{w}^{(k)}} = \mathbb{E}\left[\boldsymbol{x}_i^{(k)}(\boldsymbol{x}_i^{(k)})^\top \ddot{\psi}((\boldsymbol{x}_i^{(k)})^\top \boldsymbol{w}^{(k)})\right],\ k = 1, \ldots, K.$$

We introduce two regularity conditions.

Condition 3.1 (Sub-Gaussian designs and positive definite Hessians)**.** *For $k = 0, \ldots, K$, $\boldsymbol{x}_i^{(k)}$ are independent distributed with mean zero and covariance $\boldsymbol{\Sigma}^{(k)}$ such that $\Lambda_{\max}(\boldsymbol{\Sigma}^{(k)}) \leq c_1$. For $k = 0, \ldots, K$, the population Hessian matrices $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{w}^{(k)}}$ satisfy that $\Lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \geq c_2 > 0$ and $\Lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{w}^{(k)}}) \geq c_2 > 0$. For $k = 0, \ldots, K$, $\boldsymbol{x}_i^{(k)}$ have finite sub-Gaussian norms.*

Condition 3.2 (Sub-Gaussian random errors)**.** *For any $k = 0, \ldots, K$, the random errors $y_i^{(k)} - \dot{\psi}((\boldsymbol{x}_i^{(k)})^\top \boldsymbol{w}^{(k)})$ are independent and have finite sub-Gaussian norms.*

Condition 3.3 (Lipschitz condition for ψ)**.** *The derivatives $\dot{\psi}(a)$ and $\ddot{\psi}(a)$ exist for $a \in \mathbb{R}$. Moreover, $\ddot{\psi}(a)$ is uniformly bounded and $|\log \ddot{\psi}(a+b) - \log \ddot{\psi}(a)| \leq C|b|$ for all $a, b \in \mathbb{R}$.*

Condition 3.1 assumes independent sub-Gaussian designs with positive definite covariance matrices. The positive definiteness of Hessian $\boldsymbol{\Sigma}_{\boldsymbol{w}^{(k)}}$ essentially requires that $\ddot{\psi}((\boldsymbol{x}_i^{(k)})^\top \boldsymbol{w}^{(k)})$ is bounded away from zero with high probability and it is mild for sub-Gaussian designs. The covariance matrix $\boldsymbol{\Sigma}^{(k)}$ for different studies can be different, i.e., the distributions of the covariates in different tasks are allowed to be heterogeneous. Condition 3.2 requires the random noises to be sub-Gaussian, which is typical in high-dimensional analysis for fast convergence rates. Condition 3.3 is a Lipschitz condition on the link function. Conditions 3.1, 3.2, and 3.3 are common in the study of the GLMs, see Huang and Zhang (2012); Negahban et al. (2012); Cai et al. (2020b) and the reference therein. It holds for linear, logistic, and multinomial models. Beyond the GLMs, some other models for binary outcomes can also applicable, such as model (1.1) in Cai et al. (2020a). The Poisson or log-linear models have heavy-tailed distributions and may not satisfy Condition 3.2. We comment that our method is still applicable but the convergence rate may not be as sharp as what we will establish in Theorem 3.1.

We now analyze the convergence rate of the estimator obtained in Algorithm 1. Formally, the parameter space we consider is

$$\Theta_q(s,h)=\left\{(\boldsymbol{\beta},\boldsymbol{\delta}^{(1)},\ldots,\boldsymbol{\delta}^{(K)}):\|\boldsymbol{\beta}\|_0\le s,\max_{1\le k\le K}\|\boldsymbol{\delta}^{(k)}\|_q\le h\right\},$$

where $q\in[0,1]$ enforces either a hard ($q=0$) or soft ($q\in(0,1]$) form of sparsity on the contrast vectors. Let $n_{\min}=\min_{0\le k\le K}n_k$ and $N=\sum_{k=0}^K n_k$. In our theoretical analysis, we take the tuning parameter $\lambda_{\boldsymbol{\beta}}$ as

$$\lambda_{\boldsymbol{\beta}}=\begin{cases} c_2N(\sqrt{\frac{\log p}{N}}+\sqrt{\frac{h\log p}{n_0 s}}) \text{ if } q=0\\ c_2N(\sqrt{\frac{\log p}{N}}+h^{\frac{q}{2}}(\frac{\log p}{n_0})^{\frac{1}{2}-\frac{q}{4}}/\sqrt{s}) \text{ if } q\in(0,1].\end{cases} \tag{8}$$

This tuning parameter $\lambda_{\boldsymbol{\beta}}$ depends on the sparsity parameter s and h. This is mainly for establishing a desirable ℓ_1-error bound for the proposed estimator, which is needed in the debiasing step for statistical inference. As we will prove in Remark 3.1, if $\boldsymbol{\beta}$ is sufficiently sparse, then it suffices to choose $\lambda_{\boldsymbol{\beta}}=c_1\sqrt{N\log p}$, which is independent of h and s. In practice, the tuning parameters can be chosen by cross-validation. Next, we define the following quantity that will be used to characterize the rate of convergence.

$$T_{n_0,q}=\begin{cases}\frac{h\log p}{n_0} & \text{if } q=0\\ h^q(\frac{\log p}{n_0})^{1-q/2} & \text{if } q\in(0,1].\end{cases}$$

We are now ready to present the theoretical guarantees for the output $\hat{\boldsymbol{\beta}}$ of Algorithm 1.

Theorem 3.1 (Convergence rate of $\hat{\boldsymbol{\beta}}$)**.** *Let $q \in [0,1]$ be a fixed constant. Assume Conditions 3.1, 3.2, and 3.3 and the true parameters are in $\Theta_q(s,h)$. Suppose $s\log p \leq c_1 n_0 \wedge \sqrt{N}$, $T_{n_0,q} \leq c_1$, $s\log p T_{n_0,q} \leq c_1$, and $Kn_0 \leq c_1 N$, for some small enough constant c_1. Taking $\lambda_{\boldsymbol{\beta}}$ as in (8), then with probability at least $1-\exp(-c_2\min\{\log p, n_{\min}\})$, it holds that*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \leq c_3(\frac{s\log p}{N} + T_{n_0,q} \wedge h^2) \tag{9}$$

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \leq c_4 s\sqrt{\frac{\log p}{N}} + \sqrt{sT_{n_0,q}}. \tag{10}$$

Remark 3.1. *Assume the conditions of Theorem 3.1 and $(s\log p)^2 \leq c_1 n_0$. Then for $\lambda_{\boldsymbol{\beta}} = c_2\sqrt{N\log p}$, expression (9) still holds with probability at least $1-\exp(-c_2\min\{\log p, n_{\min}\})$.*

Theorem 3.1 establishes the convergence rate of $\hat{\boldsymbol{\beta}}$ under mild regularity conditions for any fixed $q \in [0,1]$. We first highlight the gain of transfer learning over the single-task GLM estimation. We know that the minimax optimal rate for single-task GLM is $s\log p/n_0$. Theorem 3.1 implies that when $N \gg n_0$ and $T_{n_0,q} \wedge h^2 \ll s\log p/n_0$, $\hat{\boldsymbol{\beta}}$ would admit a faster convergence rate than the single-task minimax rate. This result implies that a significant amount of knowledge can be transferred from auxiliary tasks to the target task when the similarities between target and auxiliary studies are high. In fact, $T_{n_0,q} \wedge h^2$ is the minimax error rate for estimating a p-dimensional vector with sample size n_0 and ℓ_q-sparsity h. This term comes from the estimation of contrast vectors. The condition $T_{n_0,q} \wedge h^2 \ll s\log p/n_0$ is guaranteed by $h \ll s$ when $q = 0$ and by $h \ll s\sqrt{\log p/n_0}$ when $q = 1$. Hence, when the similarity between auxiliary studies and the target study is high, the estimation performance can be improved by transfer learning. When $q = 1$, (9) recovers the convergence rate of Oracle Trans-Lasso in linear models (Li et al., 2020a). We also remark that the ℓ_1-error in Theorem 3.1 is useful for conducting statistical inference for the target parameters. We will illustrate this further in Section 4.

We now provide some discussion on the regularity conditions in Theorem 3.1. The condition $s \log p \leq n_0$ is standard for single-task sparse regression. The condition $s \log p \leq \sqrt{N}$ is mild in the regime of interest $N \gg n_0$. As h is relatively small, bounded $T_{n_0,q}$ is not hard to satisfy in applications. Finally, the condition $s \log p T_{n_0,q} \leq c_1$ guarantees the consistency of $\hat{\boldsymbol{\beta}}$ in ℓ_1-norm.

Moreover, we establish the following lower bound result showing that our proposed algorithm makes full use of the auxiliary information as the convergence rate obtained in 3.1 is in fact minimax rate-optimal.

Theorem 3.2 (Minimax lower bound)**.** *Suppose $\hat{\boldsymbol{\beta}}$ is an estimator based on n_0 i.i.d. samples $\{(\boldsymbol{x}_i^{(0)}, y_i^{(0)})\}_{i=1}^{n_0}$ drawn from model (2), and auxiliary samples $\{(\boldsymbol{x}_i^{(k)}, y_i^{(k)})\}_{i=1}^{n_k}$ drawn from model (3) for $1 \leq k \leq K$. For $T_{n_0,q} \wedge h^2 \leq s \log p / n_0 = o(1)$, we have*

$$\mathbb{P}\left(\inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta} \in \Theta_q(s,h)} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \gtrsim \frac{s \log p}{N} + T_{n_0,q} \wedge h^2 \right) \geq \frac{1}{2}.$$

4 Inference for the target parameters

In this section, we consider statistical inference of β_j for a given $j \in [p]$ in the transfer learning setting. A debiasing method for the proposed estimator is introduced in Section 4.1 and its asymptotic normality is established in Section 4.2.

4.1 A debiased estimator

We introduce a debiased estimator for β_j based on $\hat{\boldsymbol{\beta}}$, the output of Algorithm 1. We will use the target data for debiasing. Specifically, following the general debiasing recipe (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014), define

$$\hat{\beta}_j^{(db)} = \hat{\beta}_j + \frac{\sum_{i=1}^{n_0} (\boldsymbol{x}_i^{(0)})^\top \hat{\boldsymbol{\gamma}}_j \{y_i^{(0)} - \dot{\psi}((\boldsymbol{x}_i^{(0)})^\top \hat{\boldsymbol{\beta}})\}}{n_0}, \tag{11}$$

where $\hat{\boldsymbol{\gamma}}_j \in \mathbb{R}^p$ is the correction score approximating the j-th column of the inverse Hessian $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}$. To obtain $\hat{\boldsymbol{\gamma}}_j \in \mathbb{R}^p$, we estimate $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ by $\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = \frac{1}{n_0} \sum_{i=1}^{n_0} \ddot{\psi}((\boldsymbol{x}_i^{(0)})^\top \hat{\boldsymbol{\beta}}) \boldsymbol{x}_i^{(0)} (\boldsymbol{x}_i^{(0)})^\top$, and then solve $\hat{\boldsymbol{\gamma}}_j$ by the following constrained optimization

$$\begin{aligned} \hat{\boldsymbol{\gamma}}_j = & \operatorname*{arg\,min}_{\boldsymbol{\gamma} \in \mathbb{R}^p} \|\boldsymbol{\gamma}\|_1 \\ & \text{subject to} \begin{cases} \left\| \widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} \boldsymbol{\gamma} - \boldsymbol{e}_j \right\|_\infty \leq c_1 \sqrt{\frac{\log p}{n_0}}. \\ \max_{1 \leq i \leq n_0} |(\boldsymbol{x}_i^{(0)})^\top \boldsymbol{\gamma}| \leq c_2 \sqrt{\log n_0}, \end{cases} \end{aligned} \tag{12}$$

where c_1 and c_2 are two tuning parameters. In (12), the correction score $\hat{\boldsymbol{\gamma}}_j$ is obtained via a constrained ℓ_1-optimization based on the target Hessian matrix. The two constraints are linear and therefore the optimization is convex and computationally efficient. The first constraint guarantees that $\hat{\boldsymbol{\gamma}}_j$ approximates the j-th column of $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}$. The population Hessian matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ is approximated by an empirical estimator based on the design of the target model and $\hat{\boldsymbol{\beta}}$. The second constraint is on the magnitude of $|(\boldsymbol{x}_i^{(0)})^\top \hat{\boldsymbol{\gamma}}_j|$. This constraint is employed in justifying the Lyapunov central limit theorem for the sum of independent noises. Additionally, we would like to point out that while the ℓ_1-minimization in (12) encourages a sparse solution, the probabilistic limit of $\hat{\boldsymbol{\gamma}}_j$ is not necessarily sparse. Indeed, we will see that the optimization in (12) is effective no matter the j-th column of the true inverse Hessian $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}$ is sparse or not. In other words, any feasible solution to (12) is a proper correction score for the debiasing task. A similar constraint has been studied in Zhu and Bradic (2018) for hypothesis testing in single-task high-dimensional linear models. Here we extend this idea for constructing confidence intervals in high-dimensional GLMs, and further to the transfer-learning setting.

Our proposed debiasing scheme can also be used in single-task GLMs, in which case one can replace $\hat{\boldsymbol{\beta}}$ with, say, the single-task generalized Lasso estimator (Van de Geer, 2008). In comparison, the Lasso-based debiasing for the GLMs (van de Geer et al., 2014) requires $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}$ to be sparse. Another method, Cai et al. (2020a), computes the correction score under the same constraints as in (12) but the objective function is a quadratic function of $\boldsymbol{\gamma}$. The theoretical benefits of the current method will be demonstrated in detail in the next subsection.

Next, we provide a variance estimator for the debiased estimator (11). In GLMs, the variance estimation necessitates to estimate $\sigma_i^2 = \mathrm{Var}(y_i^{(0)}|(\boldsymbol{x}_i^{(0)})^\top \boldsymbol{\beta})$ for each individual $1 \leq i \leq n_0$. Our variance estimator is given as follows. For linear models, let $\hat{\sigma}_i^2 = \sum_{i=1}^{n_0} \|y_i^{(0)} - (\boldsymbol{x}_i^{(0)})^\top \hat{\boldsymbol{\beta}}\|_2^2/n_0$. For models with $c(\sigma) = 1$ in (2), which includes logistic, multinomial, Poisson, and log-linear models, let $\hat{\sigma}_i^2 = \ddot{\psi}((\boldsymbol{x}_i^{(0)})^\top \hat{\boldsymbol{\beta}})$. We now define the variance estimate of $\hat{\beta}_j^{(db)}$:

$$\widehat{V}_j = \frac{1}{n_0} \sum_{i=1}^{n_0} \{(\boldsymbol{x}_i^{(0)})^\top \hat{\boldsymbol{\gamma}}_j\}^2 \hat{\sigma}_i^2. \tag{13}$$

We establish the asymptotic distribution of $\hat{\beta}_j^{(db)}$ for some $1 \leq j \leq p$ and show the variance estimator $\widehat{V}_j$ is consistent in the next subsection.

4.2 Asymptotic normality

We next study the asymptotic distribution of $\hat{\beta}_j^{(db)}$ for some $1 \leq j \leq p$. We first show that the limiting distribution of $\hat{\beta}_j^{(db)}$ is normal in linear models, and present the result beyond linear models afterward.

In the following lemma, we prove that, with high probability, the variance estimator $\widehat{V}_j$ in (13) converges to its limit and its limit is lower bounded by a positive constant.

Lemma 4.1 (Asymptotic property of the variance estimator in linear models)**.** *Assume the conditions of Theorem 3.1 and $\dot{\psi}(\mu) = \mu$. For $\widehat{V}_j$ defined in (13), $V_j = \frac{1}{n_0}\sum_{i=1}^{n_0}\{(\boldsymbol{x}_i^{(0)})^\top \hat{\boldsymbol{\gamma}}_j\}^2\sigma_i^2$, and some positive constant c_0, it holds that*

$$|\widehat{V}_j - V_j| = o_P(1) \quad \textit{and} \quad V_j \geq c_0 - o_P(1).$$

By Lemma 4.1, V_j is the probabilistic limit of $\widehat{V}_j$ and it is only a function of $\{\boldsymbol{x}_i^{(0)}\}_{i=1}^{n_0}$ in linear models. In fact, V_j is the asymptotic variance of $\hat{\beta}_j^{(db)}$ conditioning on $\{\boldsymbol{x}_i^{(0)}\}_{i=1}^{n_0}$ in linear models.

Theorem 4.1 (Asymptotic normality of $\hat{\beta}_j^{(db)}$ for linear models)**.** *For any fixed $1 \leq j \leq p$, under the same conditions as those in Theorem 3.1 and $\dot{\psi}(\mu) = \mu$. It holds that*

$$\hat{\beta}_j^{(db)} - \beta_j = rem_j + z_j,$$

where

$$rem_j = O_P\left(\frac{s\log p}{\sqrt{Nn_0}} + T_{n_0,q}^{1/2}\sqrt{\frac{s\log p}{n_0}}\right)$$

and

$$\sqrt{\frac{n_0}{\widehat{V}_j}}z_j \xrightarrow{D} N(0,1).$$

In Theorem 4.1, we decompose the limiting distribution of $\hat{\beta}_j^{(db)}$ into two parts: an asymptotically normal part z_j and a remaining bias part rem_j. To have the asymptotic normality, one needs the asymptotically normal part to dominate the bias term, that is, $rem_j = o_P(n_0^{-1/2})$. This leads to the following sparsity conditions for asymptotic normality, which are

$$s\log p \ll \sqrt{N} \text{ and } s\log pT_{n_0,q} \ll 1. \tag{14}$$

In the single-task setting, the minimax optimal rate in Cai and Guo (2017) implies that it is necessary to require $s\log p \ll \sqrt{n_0}$. We see that the requirement in (14) is much weaker when we have a large amount of auxiliary data ($N \gg n_0$) and these data share the similarity with our target ($\sqrt{n_0} T_{n_0,q} \ll 1$). The condition $\sqrt{n_0} T_{n_0,q} \ll 1$ holds when $h = o(\sqrt{n_0 / \log p})$ if $q = 0$ and when $h\sqrt{\log p} = o(1)$ for $q = 1$. In words, when the similarity of the auxiliary studies are sufficiently large, i.e., when h is sufficiently small, the asymptotic normality of $\hat{\beta}_j^{(db)}$ requires weaker sparsity conditions than the debiased estimator in the single-task setting. Additionally, while we require a much weaker condition, the length of the proposed confidence interval in the transfer learning setting has the same order ($n_0^{-1/2}$) as that in the single-task setting. In applications, these results imply more accurate coverage probabilities with the debiased transfer learning estimator without inflating the lengths of confidence intervals. To summarize, the confidence interval in (16) is asymptotically valid for linear models when the conditions of Theorem 4.1 and (14) hold.

We remark that the results of Theorem 4.1 do not require the sparsity of inverse Hessian $\mathbf{\Sigma}^{-1}$. When $\{\mathbf{\Sigma}^{-1}\}_{.,j}$ is sufficiently sparse, standard arguments can be leveraged to show that $\|\hat{\boldsymbol{\gamma}}_j - \{\mathbf{\Sigma}^{-1}\}_{.,j}\|_1 = o_P(1)$. That is, $\hat{\beta}_j^{(db)}$ can adapt to the sparsity of the inverse Hessian. The advantage of $\hat{\boldsymbol{\gamma}}_j$ is that it is robust to non-sparse inverse Hessian and can achieve semi-parametric efficiency (van de Geer et al., 2014) for sparse inverse Hessian. In comparison, the quadratic optimization-based debiasing (Javanmard and Montanari, 2014) does not assume sparse $\mathbf{\Sigma}^{-1}$ but the semi-parametric efficiency is not shown.

We now derive the asymptotic normality for the proposed $\hat{\beta}_j^{(db)}$ beyond linear models. In this case, $\hat{\boldsymbol{\gamma}}_j$ depends on $\hat{\boldsymbol{\beta}}$ and hence depends on $y_i^{(0)}$ given $\boldsymbol{x}_i^{(0)}$. This leads to technical difficulties in justifying the asymptotic normality in GLMs. For the GLMs, we first impose a high-level Condition 4.1 and prove the main theorem. We will later verify this condition in different settings.

Condition 4.1 (Independence of the correction score)**.** *There exists some* $\boldsymbol{\gamma}_j^o \in \mathbb{R}^p$ *such that conditioning on* $\boldsymbol{\gamma}_j^o$ *and* $\{\boldsymbol{x}_i^{(0)}\}_{i=1}^{n_0}$, $y_i^{(0)} - \dot{\psi}((\boldsymbol{x}_i^{(0)})^\top \boldsymbol{\beta})$ *are independent with mean zero.*

Assume that the correction score computed via (12) satisfies $\|\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^o\|_1 = o_P((\log p)^{-1/2})$.

Condition 4.1 essentially requires that the estimated $\hat{\boldsymbol{\gamma}}_j$ converges to a "deterministic" vector $\boldsymbol{\gamma}_j^o$ in ℓ_1-norm. Here, "deterministic" means that $\boldsymbol{\gamma}_j^o$ is independent of the random noises $y_i^{(0)} - \dot{\psi}((\boldsymbol{x}_i^{(0)})^\top \boldsymbol{\beta})$. We will demonstrate the realization of Condition 4.1 after presenting the following main theorem on the asymptotic normality for $\hat{\beta}_j^{(db)}$.

We first establish the consistency of the proposed variance estimator $\widehat{V}_j$ in (13).

Lemma 4.2 (Asymptotic property of the variance estimator in GLMs)**.** *Assume the conditions of Theorem 3.1 and Condition 4.1. For* $\widehat{V}_j$ *defined in (13),* $V_j^o = \frac{1}{n_0}\sum_{i=1}^{n_0}((\boldsymbol{x}_i^{(0)})^\top \boldsymbol{\gamma}_j^o)^2 \sigma_i^2$, *and some positive constant* c_0, *we have*

$$|\widehat{V}_j - V_j^o| = o_P(1) \quad \textit{and} \quad V_j^o \geq c_0 - o_P(1).$$

By Lemma 4.2, V_j^o is the probabilistic limit of $\widehat{V}_j$ and it is independent of the random noises by Condition 4.1 in GLMs. In fact, V_j^o is the variance of $\hat{\beta}_j^{(db)}$ conditioning on $\{\boldsymbol{x}_i^{(0)}\}_{i=1}^{n_0}$ and $\boldsymbol{\gamma}_j^o$. We mention that Lemma 4.2 can be viewed as a generalization of Lemma 4.1 beyond linear models. This is because, in the case that $\dot{\psi}(\mu) = \mu$, Condition 4.1 always holds with $\boldsymbol{\gamma}_j^o = \hat{\boldsymbol{\gamma}}_j$. Hence, Lemma 4.2 recovers Lemma 4.1 when $\dot{\psi}(\mu) = \mu$, i.e., in linear models.

Theorem 4.2 (Asymptotic normality for $\hat{\beta}_j^{(db)}$ in GLMs)**.** *Assume the conditions of Theorem 3.1 and Condition 4.1. It holds that*

$$\hat{\beta}_j^{(db)} - \beta_j = rem_j + z_j,$$

where

$$rem_j = O_P\left(\frac{s \log p \sqrt{\log n_0}}{\sqrt{N n_0}} + T_{n_0,q}^{1/2}\sqrt{\frac{s \log p \log n_0}{n_0}}\right)$$

and

$$\sqrt{\frac{n_0}{\widehat{V}_j}}\, z_j \xrightarrow{D} N(0,1).$$

In Theorem 4.2, we see that the remaining bias term rem_j has an extra $\sqrt{\log n_0}$ term comparing to the results for linear models (Theorem 4.1). This inflation comes from the uncertainty in the weights of the Hessian matrix, which is estimated based on $\hat{\boldsymbol{\beta}}$. This extra term also appears in Cai et al. (2020a) for the single-task debiased estimator. Implied by Theorem 4.2, the sparsity condition for asymptotic normality in GLMs is

$$s \log p \ll \sqrt{N/\log n_0} \text{ and } T_{n_0,q} \log n_0 s \log p \ll 1. \tag{15}$$

With the target study only, the analysis in Cai et al. (2020a) requires $s \log p \ll \sqrt{n_0/\log n_0}$ for the asymptotic normality. Again, this shows that transfer learning helps reduce the remaining bias when the auxiliary studies are sufficiently similar to the target one. We can conclude that the confidence interval in (16) is asymptotically valid for the GLMs when the conditions of Theorem 4.2 and (18) hold.

In the following, we verify Condition 4.1 in different cases.

Lemma 4.3 (Sufficient conditions for Condition 4.1)**.** *Condition 4.1 holds if one of the following three statements hold:*

(i) $\ddot{\psi}$ *is a positive constant.*

(ii) We first split n_0 *samples into two folds such that* $\hat{\boldsymbol{\beta}}$ *is independent of the debiasing samples* $\{\tilde{\boldsymbol{x}}_i^{(0)}, \tilde{y}_i^{(0)}\}_{i=1}^{\tilde{n}}$. *Replace* $\{\boldsymbol{x}_i^{(0)}, y_i^{(0)}\}_{i=1}^{n_0}$ *in (11) with* $\{\tilde{\boldsymbol{x}}_i^{(0)}, \tilde{y}_i^{(0)}\}_{i=1}^{\tilde{n}}$.

(iii) The j*-th column of* $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}$ *has at most* s_j *nonzero elements such that* $(s_j \log p)^2 = o(n_0)$ *and* $N \gtrsim n_0 \log n_0$.

To summarize, for linear models, Condition 4.1 holds for free. For the GLMs, Condition 4.1 can be guaranteed by a sample splitting argument or by the sparsity of $\{\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\}_{\cdot,j}$. The third statement demonstrates the benefit of the optimization (12) over the quadratic programming in Cai et al. (2020a). That is, when s_j is sufficiently small, the sample splitting technique can be avoided. In fact, sample splitting always leads to sub-optimal empirical performance, especially when the samples are limited. We will see from the numerical experiments that our proposal has reliable performance for both sparse and non-sparse inverse Hessian matrices.

We conclude this subsection by summarizing the algorithm of constructing a two-sided $(1-\alpha)$-level confidence interval for β_j in Algorithm 2.

```
Algorithm 2: $(1-\alpha)$-level confidence interval for $\beta_j$
  Input  : $\hat{\boldsymbol{\beta}}$ obtained in Algorithm 1, target data $(\boldsymbol{X}^{(0)}, \boldsymbol{y}^{(0)})$, tuning parameters $c_1$
           and $c_2$.
  Output: $I_j$.
  Compute $\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = \frac{1}{n_0}\sum_{i=1}^{n_0} \ddot{\psi}((\boldsymbol{x}_i^{(0)})^\top \hat{\boldsymbol{\beta}})\boldsymbol{x}_i^{(0)}(\boldsymbol{x}_i^{(0)})^\top$.
  Compute $\hat{\beta}_j^{(db)}$ via (11) with $\hat{\boldsymbol{\gamma}}_j$ computed via (12).
  Compute
      $I_j = [\hat{\beta}_j^{(db)} - z_{\alpha/2}\sqrt{\widehat{V}_j/n_0},\ \ \hat{\beta}_j^{(db)} + z_{\alpha/2}\sqrt{\widehat{V}_j/n_0}],$    (16)
  where $\widehat{V}_j$ is defined in (13).
```

5 Aggregated TransGLM with positive transfer warranty

As seen in the theoretical analysis, the performance of transfer learning always depends on the level of similarity, h, which is typically unknown. When h is large, incorporating the auxiliary studies into the analysis can potentially reduce the estimation and inference accuracy of the target parameter. To guard against such "negative transfer", we propose an additional aggregation step based on the likelihood.

Given a collection of initial estimators, an aggregation procedure (Rigollet and Tsybakov, 2011; Dai et al., 2012) selects the best or a convex combination of the initial estimators by minimizing certain empirical risk measures based on the observed data. Here our primary goal is to prevent negative transfer and we propose a simple step to aggregate two initial estimators, the estimator obtained by using the target samples only, and the estimator obtained using combined dataset. More specifically, we propose our final procedure, aggregated TransGLM, shorthanded as "aTransGLM", that aggregates the transfer learning estimator $\hat{\boldsymbol{\beta}}$ with the single-task GLM Lasso $\hat{\boldsymbol{\beta}}^{(init)}$, which is formally given below.

Algorithm 3: aTransGLM, an aggregated transfer learning algorithm

Input : $\hat{\boldsymbol{\beta}}^{(init)}$, $\hat{\boldsymbol{\beta}}$, and some samples from the target study which are independent of $(\hat{\boldsymbol{\beta}}^{(init)}, \hat{\boldsymbol{\beta}})$, denoted by $\{((\tilde{\boldsymbol{x}}_i^{(0)})^\top, \tilde{y}_i^{(0)})\}_{i=1}^{\tilde{n}}$ for $\tilde{n} \asymp n_0$.

Output: $\check{\boldsymbol{\beta}}$.

Step 1: Thresholding $\hat{\boldsymbol{\beta}}$:

$$\hat{\beta}_j^t = \hat{\beta}_j \mathbb{1}(|\hat{\beta}_j| \geq \lambda_{\boldsymbol{\beta}}/N). \tag{17}$$

Step 2: Aggregation based on the likelihood. For $\widehat{\boldsymbol{B}} = (\hat{\boldsymbol{\beta}}^{(init)}, \hat{\boldsymbol{\beta}}^t) \in \mathbb{R}^{p\times 2}$,

$$\hat{\boldsymbol{\eta}} = \mathop{\arg\min}_{\boldsymbol{\eta}\in \text{a positive simplex}} \sum_{i=1}^{\tilde{n}} \left\{ \tilde{y}_i^{(0)} \cdot (\tilde{\boldsymbol{x}}_i^{(0)})^\top \widehat{\boldsymbol{B}}\boldsymbol{\eta} - \psi((\tilde{\boldsymbol{x}}_i^{(0)})^\top \widehat{\boldsymbol{B}}\boldsymbol{\eta}) \right\}.$$

Output $\check{\boldsymbol{\beta}} = \widehat{\boldsymbol{B}}\hat{\boldsymbol{\eta}}$.

We show in the supplement (Li et al., 2021) that the truncated estimators $\hat{\boldsymbol{\beta}}^t$ has the same convergence rate as $\hat{\boldsymbol{\beta}}$ but $\hat{\boldsymbol{\beta}}^t$ has sparsity no larger than the order of s. This facilitates upper-bounding the ℓ_1-error of $\check{\boldsymbol{\beta}}$ and further prepares $\check{\boldsymbol{\beta}}$ for the downstream statistical inference. In Step 2 of Algorithm 3, the independent target samples can be obtained by a sample splitting of the target samples before the analysis. Hence, we consider $\tilde{n} \asymp n_0$. The computed $\hat{\boldsymbol{\eta}}$ is a weight vector to combine two initial estimators. We also comment that the optimization of $\hat{\boldsymbol{\eta}}$ can be replaced with the Q-aggregation (Dai et al., 2012) or its variations, which can achieve the same convergence rate but sharper constants. As an illustration, we focus on a more intuitive aggregation based on the likelihood as in Step 2.

Theorem 5.1 shows that the aggregated estimator $\check{\boldsymbol{\beta}}$ is guaranteed to be no worse than the single-task estimator with high probability, which demonstrates that it provides a positive transfer warranty.

Theorem 5.1 (Consequences of aggregation)**.** *Assuming Conditions 3.1, 3.2, and 3.3 hold. Let $q \in [0,1]$ be a fixed constant. Assume that the true parameters are in $\Theta_q(s,h)$, $s\log p \leq c_1 n_0 \wedge \sqrt{N}$, $T_{n_0,q} \leq c_1$, $s\log p T_{n_0,q} \leq c_1$, and $N \geq c_2 K n_0$, for some positive constants c_1 and c_2. Then with probability at least $1-\exp(-c_1\log p)-\exp(-c_1 n_{\min})-\exp(-c_1 t)$,*

$$\|\check{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2^2 \leq c_2\frac{s\log p}{N} + c_3 T_{n_0,q} \wedge h^2 \wedge \frac{s\log p}{n_0} + \frac{c_4 t}{n_0}$$

$$\|\check{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 \leq c_5\sqrt{s}\|\check{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2.$$

Theorem 5.1 essentially shows that the aggregated estimator $\check{\boldsymbol{\beta}}$ has no slower convergence rate than the those obtained by $\hat{\boldsymbol{\beta}}^{(init)}$ and $\hat{\boldsymbol{\beta}}^t$. Theorem 5.1 guarantees that $\|\check{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2^2 \lesssim s\log p/n_0$ with high probability as long as $s \neq 0$. Hence, the performance of $\check{\boldsymbol{\beta}}$ is robust to a large h, i.e. low similarity levels. We also obtained the convergence rate in ℓ_1-norm by utilizing the sparsity of $\hat{\boldsymbol{\beta}}^{(init)}$ and the sparsity of the thresholded estimator $\hat{\boldsymbol{\beta}}^t$. The cost of aggregation is of order $1/n_0$, which is negligible in most scenarios of interest. For example, when $q = 0$, as long as $h \geq 1$ and $s \geq 1$, the cost of aggregation is always dominated by the second term. Hence, in practice, it is almost no harm to perform an aggregation step.

The inference results based on $\check{\boldsymbol{\beta}}$ can be similarly proved. Let $\check{\boldsymbol{\beta}}_j^{(db)}$ be the debiased estimator in (11) with $\hat{\boldsymbol{\beta}}$ replaced by $\check{\boldsymbol{\beta}}$. The score $\check{\boldsymbol{\gamma}}_j$ for $\check{\boldsymbol{\beta}}_j^{(db)}$ is computed based on $\widehat{\boldsymbol{\Sigma}}_{\check{\boldsymbol{\beta}}}$ instead of $\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$. In the following, we establish the asymptotic normality of $\check{\boldsymbol{\beta}}_j^{(db)}$.

Theorem 5.2 (Asymptotic normality for $\check{\beta}_j^{(db)}$ in GLMs)**.** *Assume the conditions of Theorem 5.1 and Condition 4.1. It holds that*

$$\check{\beta}_j^{(db)} - \beta_j = rem_j + z_j,$$

where

$$rem_j = O_P\left(\frac{s\log p\sqrt{\log n_0}}{\sqrt{Nn_0}} + T_{n_0,q}^{1/2}\sqrt{\frac{s\log p\log n_0}{n_0}} \wedge \frac{s\log p\sqrt{\log n_0}}{n_0}\right) + o_P(n_0^{-1/2})$$

and for $\widehat{V}_j$ defined in (13),

$$\sqrt{\frac{n_0}{\widehat{V}_j}}z_j \xrightarrow{D} N(0,1).$$

Implied by Theorem 5.2, the sparsity condition for asymptotic normality is

$$\begin{cases} s\log p \ll \sqrt{N/\log n_0} \ \text{ and } \ T_{n_0,q}s\log p\log n_0 \ll 1 & \text{if } T_{n_0,q} \ll n_0^{-1/2} \\ s\log p \ll \sqrt{n_0/\log n_0} & \text{otherwise.} \end{cases} \tag{18}$$

The requirement in (18) is always no worse than the sparsity requirement in the single-task setting as a consequence of aggregation. The verification of Condition 4.1 can be similarly proved as in Lemma 4.3. In the next section, we evaluate the numerical performance of $\check{\boldsymbol{\beta}}$ and $\check{\boldsymbol{\beta}}_j^{(db)}$.

6 Simulation studies

We study the numerical performance of our proposal and other comparable methods. We set $n_0 = \cdots = n_K = 200$, $p = 500$, and $s = 10$. We set $\boldsymbol{\beta}_{1:s} = (0.8, 0.65, 0.50, \ldots, -0.55)^\top$ and $\beta_j = 0$ for $j > s$. For $k = 0, \ldots, K$, we generate $\boldsymbol{x}_i^{(k)} \sim N(0, \boldsymbol{\Sigma}^{(k)})$ independently. We consider two configurations of the covariance matrices.

(a) For $k = 0, \ldots, K$, we consider Toeplitz matrices $\{\boldsymbol{\Sigma}^{(k)}\}_{j,l} = (k/(K+2))^{|j-l|}$.

(b) We consider equi-correlated $\boldsymbol{\Sigma}^{(0)}_{j,k} = 0.3$ for $j \neq k$ and $\boldsymbol{\Sigma}^{(0)}_{j,j} = 1$. For each $k = 1, \ldots, K$, we generate a random matrix $\boldsymbol{A}^{(k)}$ where each entry equals 0.1 with probability 0.1 and equals 0 with probability 0.9. We set $\boldsymbol{\Sigma}^{(k)} = (\boldsymbol{A}^{(k)})^\top \boldsymbol{A}^{(k)} + I_p$, $k = 1, \ldots, K$.

In both (a) and (b), the design matrices are heterogeneous among studies. The target covariance matrix $\boldsymbol{\Sigma}^{(0)}$ is sparse in (a) but not in (b). Hence, (b) provides a challenging setting for statistical inference.

To accommodate the practical setting that some auxiliary studies can be very far from the target study, we define $\mathcal{A} \subseteq \{1, \ldots, K\}$ to be the set of informative studies. Specifically, we generate $\boldsymbol{\delta}^{(k)}$ in two ways.

(i) For $k \in \mathcal{A}$, let H_k be a random subset of $\{1, \ldots, p\}$ with $|H_k| = h \in \{2, 6, 10\}$. For $k \notin \mathcal{A}$, let H_k be a random subset of $\{1, \ldots, p\}$ with $|H_k| = 50$. For $k = 1, \ldots, K$, we set $\delta^{(k)}_j = 0.3$ for $j \in H_k$ and $\delta^{(k)}_j = 0$ otherwise.

(ii) For $k \in \mathcal{A}$, $\delta^{(k)}_j \sim N(0, (h/50)^2)$ for $j \leq 100$ and $h \in \{2, 6, 10\}$ and $\delta^{(k)}_j = 0$ otherwise. For $k \notin \mathcal{A}$, $\delta^{(k)}_j \sim N(0, 0.5^2)$ for $j \leq 100$ and $\delta^{(k)}_j = 0$ otherwise.

We see that in both (i) and (ii), $\{\delta^{(k)}\}_{k \in \mathcal{A}}$ are sparser than $\{\delta^{(k)}\}_{k \in \mathcal{A}^c}$. Moreover, $\{\delta^{(k)}\}_{k \in \mathcal{A}^c}$ are even denser than $\boldsymbol{\beta}$ and we treat studies in $\mathcal{A}^c$ as non-informative studies. In (i), $\boldsymbol{\delta}^{(k)}$ is exact sparse and in (ii), $\boldsymbol{\delta}^{(k)}$ are approximately sparse. We will consider four scenarios generated by (a) and (b) crossing (i) and (ii), denoted by (a-i), (a-ii), (b-i) and (b-ii), respectively. Each configuration is replicated with 300 independent experiments. In the main paper, we report two settings generated by (a-i) and (b-i). The results for (a-ii) and (b-ii) are analogous and are reported in the supplementary materials (Section E).

We compare five methods numerically. The first one is generalized Lasso based on the target study, denoted as "GLM Lasso". The second one is Algorithm 1, denoted by "TransGLM". The third method is Algorithm 1 based on target and informative auxiliary studies. That is, we apply Algorithm 1 with $\{1, \ldots, K\}$ replaced by $\mathcal{A}$. We denote this method by "oracle TransGLM" as it depends on the oracle $\mathcal{A}$. The fourth method is Algorithm 3, denoted by "aTransGLM". The last one is a simple aggregated estimator, denoted by "Simple-Agg". It first applies the GLM Lasso to each task and then aggregate these $K + 1$ estimators using the optimization in Section 5. This method can be viewed as a meta-analysis paradigm with adaptive weights. It is widely used in applications for its simplicity and we include it as another benchmark method. For the inference results, we construct confidence intervals with oracle TransGLM, aTransGLM, and the single-task method in van de Geer et al. (2014). The detailed implementation of different methods is illustrated in the supplementary materials.

6.1 Classification errors

In every experiment, we evaluate the classification errors in an independent target sample with a sample size 200. From Figure 1, we see that the performance of single-task GLM Lasso does not change as the informative sample size changes. The oracle TransGLM significantly reduces the classification errors in comparison to the GLM Lasso as the informative sample size increases. It is always no worse than the GLM Lasso because it never incorporates non-informative samples. The TransGLM method reduces classifications errors when a significant proportion of the auxiliary samples are informative. This is because it uses all the auxiliary studies and when few studies are informative, the errors can be large according to Section 4. The aTransGLM method also improves classification accuracy when the informative sample size is relatively large. On the other hand, the aggregation step in aTransGLM achieves robustness to negative transfer in the sense that the performance of aTransGLM is always no worse than the single-task GLM Lasso. When $|\mathcal{A}|$ is close to K, the TransGLM has slightly smaller errors than aTransGLM. This is because TransGLM does not split the samples for aggregation but aTransGLM does. However, robustness can be more important than the mild gain in accuracy and hence aTransGLM should be favorable over TransGLM in most practical applications. The “Simple-Agg” method has limited improvement when the informative samples are large and its performance is very sensitive to the levels of h. By comparing the plots at different levels of h, we see that the performances of Oracle TransGLM, TransGLM, and aTransGLM are getting slightly worse as h increases, which agrees with our theoretical analysis. The overall performance also demonstrates that our method is robust to heterogeneous design matrices. The estimation errors are reported in the supplementary materials (Section E).

6.2 Confidence intervals

We construct 95% two-sided confidence intervals for β_j, $j = 1, \ldots, p$. We compare our proposed debiased oracle TransGLM and debiased aTransGLM with the single-task inference method for the GLMs (van de Geer et al., 2014).

In Table 1, we report the results in setting a-i, where the inverse Hessian matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}$ is relatively sparse. All the methods have reliable coverages for $\beta_j = 0$. For $\beta_j = 0.5$, we see that the single-task method has coverage probabilities lower than the nominal level. This is mainly due to the large remaining bias of the single-task debiased estimators, which have been studied in Li (2020). The proposed debiased oracle TransGLM and debiased aTransGLM have improvements in coverage probabilities for $\beta_j \neq 0$ without inflating the length of confidence intervals. The increased coverage probabilities are due to the smaller remaining bias of the debiased transfer learning estimator, which agrees with our theoretical results. In Table 2, we report the inference results in b-i which gives a non-sparse $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}$. For the true signals, the debiased transfer learning estimators have significantly higher coverage probabilities than the single-task debiased method. This again demonstrates the smaller

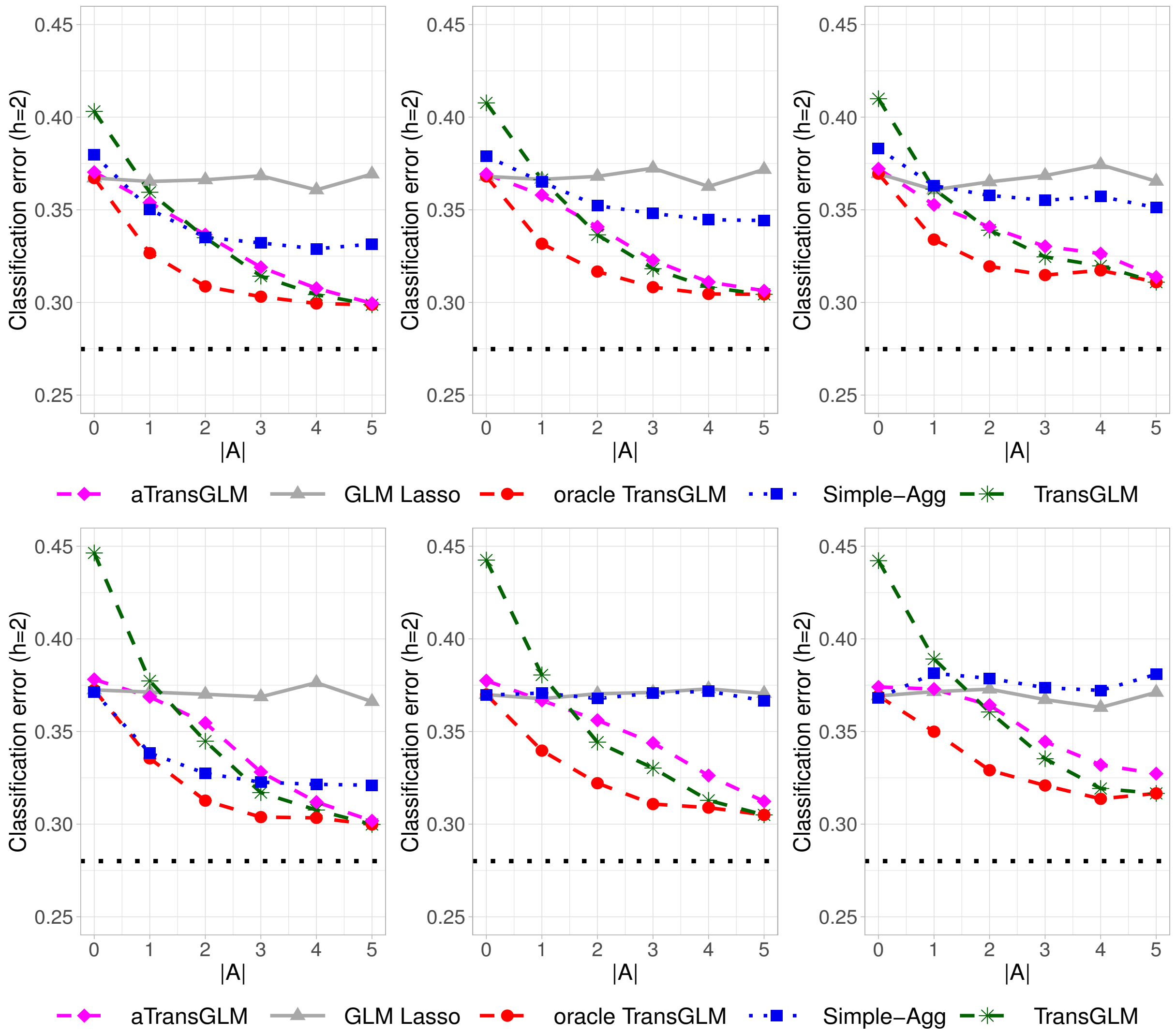

Figure 1: Classification errors in setting a-i (first row) and in setting b-i (second row). The dotted horizontal line is the average classification errors given by oracle $\boldsymbol{\beta}$.

remaining bias of the debiased transfer learning estimators.

Table 1: Average coverage probabilities (standard deviations) for $\beta_3 = 0.5$ and $\beta_{13} = 0$ in setting a-i.

h	$\lvert\mathcal{A}\rvert$	van de Geer et al. (2014)		Debiased oracle TransGLM		Debiased aTransGLM	
		0.5	0	0.5	0	0.5	0
2	1	0.937(0.154)	0.987(0.153)	0.947(0.154)	0.983(0.152)	0.930(0.150)	0.987(0.149)
2	2	0.920(0.156)	0.977(0.153)	0.930(0.156)	0.967(0.152)	0.933(0.152)	0.967(0.149)
2	3	0.897(0.155)	0.973(0.153)	0.913(0.156)	0.970(0.153)	0.900(0.152)	0.970(0.151)
2	4	0.950(0.155)	0.970(0.153)	0.967(0.157)	0.957(0.153)	0.963(0.154)	0.967(0.151)
2	5	0.917(0.155)	0.987(0.154)	0.927(0.156)	0.980(0.154)	0.930(0.156)	0.980(0.154)
6	1	0.943(0.155)	0.973(0.154)	0.947(0.152)	0.980(0.151)	0.947(0.151)	0.973(0.150)
6	2	0.933(0.157)	0.977(0.155)	0.947(0.152)	0.980(0.151)	0.937(0.150)	0.977(0.150)
6	3	0.933(0.156)	0.983(0.155)	0.937(0.152)	0.983(0.151)	0.933(0.150)	0.980(0.150)
6	4	0.910(0.156)	0.973(0.154)	0.917(0.153)	0.963(0.151)	0.927(0.151)	0.963(0.151)
6	5	0.933(0.156)	0.967(0.154)	0.947(0.153)	0.967(0.151)	0.957(0.153)	0.967(0.151)
10	1	0.950(0.156)	0.957(0.154)	0.937(0.152)	0.957(0.150)	0.937(0.152)	0.953(0.150)
10	2	0.953(0.157)	0.980(0.155)	0.967(0.152)	0.973(0.150)	0.957(0.150)	0.970(0.149)
10	3	0.920(0.158)	0.963(0.156)	0.923(0.152)	0.967(0.150)	0.923(0.151)	0.963(0.150)
10	4	0.943(0.157)	0.970(0.155)	0.963(0.151)	0.970(0.150)	0.957(0.152)	0.970(0.150)
10	5	0.913(0.156)	0.987(0.154)	0.933(0.152)	0.977(0.150)	0.933(0.153)	0.973(0.151)

Table 2: Average coverage probabilities (standard deviations) for $\beta_3 = 0.5$ and $\beta_{13} = 0$ in setting b-i.

h	$\lvert\mathcal{A}\rvert$	van de Geer et al. (2014)		Debiased oracle TransGLM		Debiased aTransGLM	
		0.5	0	0.5	0	0.5	0
2	1	0.893(0.178)	0.967(0.176)	0.917(0.174)	0.960(0.173)	0.910(0.173)	0.963(0.172)
2	2	0.883(0.176)	0.957(0.175)	0.913(0.175)	0.957(0.174)	0.907(0.172)	0.947(0.171)
2	3	0.903(0.176)	0.963(0.174)	0.927(0.176)	0.957(0.172)	0.913(0.171)	0.950(0.169)
2	4	0.933(0.176)	0.977(0.174)	0.963(0.175)	0.973(0.172)	0.960(0.172)	0.963(0.170)
2	5	0.927(0.176)	0.963(0.176)	0.953(0.176)	0.963(0.174)	0.953(0.176)	0.967(0.175)
6	1	0.913(0.179)	0.960(0.178)	0.920(0.173)	0.980(0.172)	0.917(0.177)	0.973(0.172)
6	2	0.920(0.177)	0.960(0.177)	0.927(0.172)	0.957(0.173)	0.920(0.172)	0.960(0.172)
6	3	0.903(0.176)	0.970(0.175)	0.913(0.172)	0.960(0.171)	0.903(0.171)	0.957(0.171)
6	4	0.920(0.177)	0.967(0.175)	0.937(0.172)	0.960(0.171)	0.933(0.171)	0.963(0.170)
6	5	0.920(0.175)	0.967(0.175)	0.927(0.171)	0.967(0.171)	0.927(0.172)	0.970(0.171)
10	1	0.883(0.177)	0.960(0.176)	0.890(0.172)	0.960(0.171)	0.880(0.172)	0.960(0.171)
10	2	0.900(0.176)	0.970(0.177)	0.910(0.171)	0.973(0.171)	0.910(0.171)	0.973(0.172)
10	3	0.903(0.177)	0.983(0.174)	0.917(0.172)	0.980(0.170)	0.913(0.172)	0.983(0.169)
10	4	0.910(0.178)	0.980(0.176)	0.940(0.171)	0.980(0.171)	0.930(0.171)	0.980(0.170)
10	5	0.890(0.177)	0.977(0.176)	0.917(0.172)	0.973(0.171)	0.917(0.172)	0.973(0.171)

7 Application to the colorectal cancer data

We apply our method to several human gut microbiome studies concerning colorectal cancer (CRC). These are case-control studies where the response indicates whether an individual has CRC and the covariates are the common genera and phyla of the microbiomes and three other covariates (age, gender, and BMI). The raw data is publicly available at https://zenodo.org/record/840333#.X6qTRS9h3u2 and has been studied in Duvallet et al. (2017). We analyze the data from three studies, referred to as Zackular, Zeller, and Baxter, where are collected in USA/Canada, France, and USA, respectively. These studies are all related to the CRC but are measured in different populations. Hence, it is likely that these studies share some similarities but the underlying true models may not be identical. Therefore, it is proper to apply transfer learning to these studies. The sample sizes of Zackular, Zeller, and Baxter studies are 83, 127, and 488, respectively. Some genera and phyla of the microbiomes are relatively rare and are removed from the analysis if their abundance are zero in more than 90% of the samples in each study. Altogether, 146 genera and phyla of the microbiomes and three covariates ($p = 149$) remain in the analysis. The covariates are standardized before analysis.

We consider Zackular, Baxter, and Zeller as the target study individually and use the other two studies as auxiliary studies. We first look at the classification errors given by our proposed transfer learning method and the single-task method, the GLM Lasso. The results based on leave-one-out prediction are reported in Table 3. Specifically, we iteratively use one sample from the target data as the test sample and the rest of the data as training samples. We see that the TransGLM, aTransGLM, and Simple-Agg all have smaller classification errors for the target Zackular. This demonstrates the improvement of transfer learning. Furthermore, we see that aTransGLM is robust in the sense that its classification error is always no larger than the single-task method. Both TransGLM and Simple-Agg are not as robust as aTransGLM. This demonstrates the benefit of aggregation. We also see the improvement of transfer learning in Zackular study is the most significant. One potential reason is that the sample size of Zackular study is the smallest and transfer learning has the potential to contribute more improvements. In the Baxter study, the target sample size is significantly larger than the overall auxiliary sample size. Hence, one would expect that transfer learning may not lead to significant improvements.

Table 3: Misclassification rates given by the single-task method (GLM Lasso), TransGLM, aTransGLM, and a simple aggregation method (Simple-Agg) described in Section 6 based on leave-one-out prediction for three studies.

Target	Sample Size	GLM Lasso	TransGLM	aTransGLM	Simple-Agg
Zackular	83	33.7%	26.7%	25.3%	26.7%
Zeller	127	29.1%	31.5%	27.6%	31.5%
Baxter	488	23.0%	21.3%	21.3%	24.6%

We also construct 95% confidence intervals for each regression coefficient in the target study. We calculate the confidence intervals using the single-task method (van de Geer et al., 2014) and our proposed debiased aTransGLM. In the Zackular study (Table 4), two covariates are significant at 95% confidence level using the single-task method and three covariates are significant at 95% confidence level using the debiased aTransGLM. Our findings agree with some existing studies on CRC. For example, BMI has been shown to be positively correlated with the risk of CRC in multiple studies (Zheng et al., 2018; Campbell et al., 2021). Clostridium group XVIII has been found negatively correlated with the occurrence of CRC (Baxter et al., 2014) and Enterobacter can potentially promote CRC (Yurdakul et al., 2015). The results for Zeller study and Baxter Study are reported in Table 3 and Table 4 in the supplementary files, respectively. In the Zeller study, ten covariates are selected using the single-task method with the 95% CI not including zero and 18 covariates are selected using the transfer learning method with the 95% CI not including zero. In the Baxter Study, 13 covariates are selected using the single-task method with the 95% CI not including zero and 16 covariates are selected using the transfer learning method with the 95% CI not including zero.

Table 4: Significant covariates based on the single-task method or the proposed method at 95% confidence level in the Zackular study. The p-values with * are significant at 95% confidence level.

no	Variables	van de Geer et al. (2014)		Debiased aTransGLM	
		CI	p-value	CI	p-value
1	BMI	0.595 ± 0.46	0.011*	0.536 ± 0.45	0.020*
2	Clostridium.XVIII	-0.681 ± 0.51	0.009*	-0.555 ± 0.47	0.021*
3	Enterobacter	0.432 ± 0.44	0.052	0.445 ± 0.44	0.047*

References

Agarwal, A., Negahban, S., and Wainwright, M. J. (2012). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 40(2):1171–1197.

Bastani, H. (2018). Predicting with proxies: Transfer learning in high dimension. *arXiv: 1812.11097*.

Baxter, N. T., Zackular, J. P., Chen, G. Y., and Schloss, P. D. (2014). Structure of the gut microbiome following colonization with human feces determines colonic tumor burden. *Microbiome*, 2(1):1–11.

Bradic, J., Claeskens, G., and Gueuning, T. (2020). Fixed effects testing in high-dimensional linear mixed models. *Journal of the American Statistical Association*, 115(532):1835–1850.

Cai, T. T. and Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646.

Cai, T. T., Guo, Z., and Ma, R. (2020a). Statistical inference for high-dimensional generalized linear models with binary outcomes. *Technical report*.

Cai, T. T., Wang, Y., and Zhang, L. (2020b). The cost of privacy in generalized linear models: Algorithms and minimax lower bounds. *arXiv preprint arXiv:2011.03900*.

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

Campbell, P. T., Lin, Y., Bien, S. A., Figueiredo, J. C., Harrison, T. A., Guinter, M. A., Berndt, S. I., Brenner, H., Chan, A. T., Chang-Claude, J., et al. (2021). Association of body mass index with colorectal cancer risk by genome-wide variants. *JNCI: Journal of the National Cancer Institute*, 113(1):38–47.

Chen, X., Kim, S., Lin, Q., Carbonell, J. G., and Xing, E. P. (2010). Graph-structured multi-task regression and an efficient optimization method for general fused lasso. *arXiv preprint arXiv:1005.3579*.

Dai, D., Rigollet, P., and Zhang, T. (2012). Deviation optimal learning using greedy q-aggregation. *The Annals of Statistics*, 40(3):1878–1905.

Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society. Series B (Statistical methodology)*, 76(2):373–397.

Dezeure, R., Bühlmann, P., and Zhang, C.-H. (2017). High-dimensional simultaneous inference with the bootstrap. *Test*, 26(4):685–719.

Dondelinger, F., Mukherjee, S., and Initiative, A. D. N. (2020). The joint lasso: high-dimensional regression for group structured data. *Biostatistics*, 21(2):219–235.

Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., and Alm, E. J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature communications*, 8(1):1–10.

Fang, E. X., Ning, Y., and Liu, H. (2017). Testing and confidence intervals for high dimensional proportional hazards models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 79(5):1415–1437.

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

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Hosny, K. M., Kassem, M. A., and Foaud, M. M. (2018). Skin cancer classification using deep learning and transfer learning. In *2018 9th Cairo International Biomedical Engineering Conference (CIBEC)*, pages 90–93. IEEE.

Huang, J. and Zhang, C.-H. (2012). Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. *The Journal of Machine Learning Research*, 13(1):1839–1864.

Javanmard, A. and Javadi, H. (2019). False discovery rate control via debiased lasso. *Electronic Journal of Statistics*, 13(1):1212–1253.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15:2869–2909.

Li, S. (2020). Debiasing the debiased lasso with bootstrap. *Electronic Journal of Statistics*, 14(1):2298–2337.

Li, S., Cai, T. T., and Li, H. (2019). Inference for high-dimensional linear mixed-effects models: A quasi-likelihood approach. *Journal of the American Statistical Association (in press)*.

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*.

Li, S., Zhang, L., Cai, T. T., and Li, H. (2021). Supplements to "estimation and inference in high-dimensional generalized linear models with knowledge transfer".

Liang, M., Zhong, X., and Park, J. (2020). Learning a high-dimensional classification rule using auxiliary outcomes. *arXiv preprint arXiv:2011.05493*.

Ma, R., Tony Cai, T., and Li, H. (2020). Global and simultaneous hypothesis testing for high-dimensional logistic regression models. *Journal of the American Statistical Association*, pages 1–15.

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

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

Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771.

Sevakula, R. K., Singh, V., Verma, N. K., Kumar, C., and Cui, Y. (2018). Transfer learning for molecular cancer classification using deep neural networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(6):2089–2100.

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.

Tripuraneni, N., Jin, C., and Jordan, M. I. (2020a). Provable meta-learning of linear representations. *arXiv preprint arXiv:2002.11684*.

Tripuraneni, N., Jordan, M. I., and Jin, C. (2020b). On the theory of transfer learning: The importance of task diversity. *arXiv preprint arXiv:2006.11650*.

Turki, T., Wei, Z., and Wang, J. T. (2017). Transfer learning approaches to improve drug sensitivity prediction in multiple myeloma patients. *IEEE Access*, 5:7381–7393.

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

Van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645.

Yurdakul, D., Yazgan-Karataş, A., and Şahin, F. (2015). Enterobacter strains might promote colon cancer. *Current Microbiology*, 71(3):403–411.

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*, 76(1):217–242.

Zhang, X. and Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, 112(518):757–768.

Zhang, Y. and Yang, Q. (2017). A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*.

Zheng, R., Du, M., Zhang, B., Xin, J., Chu, H., Ni, M., Zhang, Z., Gu, D., and Wang, M. (2018). Body mass index (bmi) trajectories and risk of colorectal cancer in the plco cohort. *British Journal of Cancer*, 119(1):130–132.

Zhu, Y. and Bradic, J. (2018). Significance testing in non-sparse high-dimensional linear models. *Electronic Journal of Statistics*, 12(2):3312–3364.