Guided Transfer Learning for High-Dimensional Linear Regression

Xinhao Qu

(School of Economics, Xiamen University, Xiamen 361005, China)

Abstract I study a transfer learning problem with adaptive l_1 penalization for dissimilarity between target and source parameters. Through zero-consistent data-driven guidance, the learning algorithms are free of detection for transferable sources and own doubly robust minimax convergence rate for arbitrary sources compared with Oracle Trans-Lasso (Li et al., 2022). Numerical experiments also show the computational efficiency and robustness of Guided Transfer Learning (GTL) for negative masses.

Keywords Adaptive Estimation; Transfer Learning; High-Dimensional Linear Regression

1 Introduction

Transfer learning conveys referable knowledge to information-lacking machines of the target, which has become a powerful tool in various tasks, including image recognition, natural language processing, recommendation system and so on (Daumé III, 2007; Pan & Yang, 2009; Pan & Yang, 2013; Weiss et al., 2016). Recently, Li et al. (2022) provided a minimax optimality insurance for knowledge-transfer through Oracle Trans-Lasso within high-dimensional parameter space under sparsity, whose formalization utilized an initial middle-point $\hat{\beta}^{\mathcal{A}}$ followed by a de-biasing procedure through l_1 regularization

$$\min_{\delta\in\mathbb{R}^p}\left\{\left\|y^{(0)}-X^{(0)}\left(\hat{\beta}^{\mathcal{A}}+\delta\right)\right\|_2^2+\lambda\|\delta\|_1\right\}$$

on target data $y^{(0)}, X^{(0)}$, where λ serves as the tuning.

However, originating from the underlying dissimilarity between the target and non-informative sources, the latent negativity remains to be one of the biggest challenges for transfer learning. Existing literatures suggest two main streams handling such issue, including interval-type source detections and adaptive estimations through aggregation. The former, introduced by Tian & Feng (2022), employed sample splitting technique to construct the upper and lower bound for transferable sources' likelihood functions. Although the sampling stableness assumption for transferable source $\mathcal{A}$'s identifiability was mild and could be further sharpened through Bootstrap, as shown by Qu (2024+), the other gap assumption in terms of $\|\beta^{(k)}-\beta^{(0)}\|_2$ for $k\notin\mathcal{A}$ contributing to identifiability is untestable and may be hardly true in practice, resulting in failure of consistent source detection as well as negative transfer.

The second adaptive Q-aggregation strategy, proposed by Li et al. (2022) as Trans-Lasso, also mildly relied on the existing gap condition for $\min_{k\notin\mathcal{A}}\sum_j|\Sigma^{(k)}_{j,\cdot}\beta^{(k)}-\Sigma^{(0)}_{j,\cdot}\beta^{(0)}|$ to ensure proper adaptivity and thus positive transfer. Apart from this, the avoidance of negativity requires $n_{\mathcal{A}}>>n_0$ and is never cheap through source-wise adaptation in practice. By shrinking candidate numbers, Trans-Lasso switched the computation burden from the ensemble process to the sure screening step under both large p and large n_k, which could be even more time-consuming (Li et al., 2020).

To this end, this paper intends to inherit the adaptive estimation methodology to effectively avoid negative transfer but in a more robust, computationally efficient, parameter-wise perspective. The framework is primely established under small source numbers K with diverging sample sizes n_k for each source k, which is typically close to the setting in image recognition tasks, where the sources are largely composed of several large-scale picture bases, such as ImageNet (Deng et al., 2009).

The parameter-wise adaptivity stems from a separable thinking while transferring as in Qu (2024+), who

divide the parameter set into causal (dissimilar) and nuisance (similar) part under confounding, and conduct shallow-layer weight-transfer plus down-stream fine-tuning procedures comparable to image recognition tasks after de-confounding. The gradient ascent recovery and recent progress on using explainable tools such as Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) for exploring the reasoning behind Convolutional Neural Networks (Haar et al., 2023) also corroborated the parameterized causal nuisance ratiocination, indicating the need of adaptive analysis parameter-wisely.

However, beyond hard separation of the entire parameter set, this paper also introduces a smooth adaptivity embedded in transferring under sparse high-dimensional regression, which has been thoroughly analyzed for single dataset (Zou, 2006; Huang et al., 2008; Chatterjee & Lahiri, 2013). Base on similar concept of zero-consistent guidance as in Huang et al. (2008), this paper shows the oracle property for transfer learning, meaning that GTL could consistently carry on knowledge transfer on transferable parameter sets for arbitrarily given sources. Implementing the cross-fitting technique (Fan et al., 2012), GTL presents doubly robust rate of convergence compared with Oracle Trans-Lasso, the minimax optimality is also comparable to Li et al. (2022)'s benchmark research. The synthetic and empirical analysis further merit GTL's computational efficiency and estimation and prediction accuracy.

In summary, the main contributions of GTL to previous literatures are: First, design transfer learning algorithm that is free of detection algorithms through adaptive weighting procedure from a parameter-wise perspective; Second, introduce oracle property for transfer learning under sparse dissimilarity between target and source parameters; Third, provide doubly robust convergence rate with minimax optimality under l_1 measure of dissimilarity for GTL; Last but not least, enhance the computational efficiency of GTL compared with Trans-Lasso aggregation.

The rest of paper is organized as below: Section 2 gives detailed procedures about implementations of GTL method and asymptotics related to oracle transfer property; followed by non-asymptotic analysis for convergence rate and minimax lower bound in section 4; the last two sections are dedicated to synthetic simulations and empirical applications to testify well-performance of GTL's estimation and prediction as well as its computational efficiency.

2 Methodology

Under the setting of sparse high-dimensional linear model, this section introduces Guided Trans-Lasso algorithms in detail as for knowledge transferring from source to target dataset that is free of informative source detection or data aggregation procedures.

2.1. Model

Without loss of generality, here singly focuses on limited set of large sources and assumes the total number $K = 1$. The data matrices are identically formulated between the target and the source under independent generating process:

$$y_i^{(k)} = {x_i^{(k)}}^\top \beta^{(k)} + \epsilon_i^{(k)}, \quad i = 1, \ldots, n_k$$

for $k \in \{0, 1\}$ with $k = 0$ representing the target generating process. Sample size $n_1 >> n_0$. Response $y_i^{(k)}$ are continuous, $x_i^{(k)} \in \mathbb{R}^p$ with high dimensional p greater than n_0 and n_1, $\epsilon_i^{(k)}$ denotes the random noise. The parameter space is specified as

$$\Theta_q(s, h) = \left\{ \left(\beta^{(0)}, \delta \right) : \|\beta^{(0)}\|_0 \leq s, \|\delta\|_q \leq h \right\},$$

where $\delta \equiv \beta^{(0)} - \beta^{(1)}$, its q-norm measures the dissimilarity between the target and source, and is controlled through h; $q = \{0, 1\}$ with $q = 0$ stands for sparse δ and $\beta^{(1)}$ cases while $q = 1$ measures the cumulative absolute difference between target and source parameters; s confines the sparsity level for target parameters.

2.2. Guided Trans-Lasso

Following the combining and de-biasing strategy in Oracle Trans-Lasso (Li et al., 2022), however, here I employ the distinctive level of transferability for each parameter as the data-driven guidance, whose measure is of similar forms with Zou (2006). As in the following algorithm, the first step simply conducts Adaptive Lasso for the source under zero-consistent initial estimator (Huang et al., 2008), if $q = 0$, this is serving as an Oracle (Fan & Li, 2001) start await for transferring. Step 2 mimics the fusion penalty as for the de-biasing procedure in Oracle Trans-Lasso through regularizing the dissimilarity for target and source parameters. Importantly, the adaptive weight $1/|\tilde{\delta}_j|$ guides the de-biasing level, which should lead harsh penalization when the actual gap δ_j is small. To offer robustness, here similarly introduces positive constants γ_1 and γ_0 for refining the level of adaptiveness as in Zou (2006). λ_{n_1} and λ_{n_0}, on the other hand, restrict the general degree of regularization for sparse $\beta^{(1)}$ and δ if $q = 0$, which are commonly selected by cross-validation in practice.

Algorithm 1: Guided Trans-Lasso

Input: Target data $\{y_i^{(0)}, x_i^{(0)}\}_{i=1,\dots,n_0}$ and informative auxiliary samples $\{y_i^{(1)}, x_i^{(1)}\}_{i=1,\dots,n_1}$; initial zero-consistent estimators $\tilde{\beta}_j^{(1)}$, $\tilde{\delta}_j \equiv \check{\beta}_j^{(0)} - \check{\beta}_j^{(1)}$; tunings $\lambda_{n_1}, \lambda_{n_0}$ and $\gamma_1, \gamma_0 > 0$

Output: $\hat{\beta}^{(0)}$

1 Compute

$$\hat{\beta}^{(1)} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n_1} \sum_{i=1}^{n_1} \left(y_i^{(1)} - {x_i^{(1)}}^\top \beta \right)^2 + \lambda_{n_1} \sum_{j=1}^{p} \frac{|\beta_j|}{|\tilde{\beta}_j^{(1)}|^{\gamma_1}} \right\}$$

2 Compute

$$\hat{\delta} = \arg\min_{\delta \in \mathbb{R}^p} \left\{ \frac{1}{2n_0} \sum_{i=1}^{n_0} \left[y_i^{(0)} - {x_i^{(0)}}^\top \left(\hat{\beta}^{(1)} + \delta \right) \right]^2 + \lambda_{n_0} \sum_{j=1}^{p} \frac{|\delta_j|}{|\tilde{\delta}_j|^{\gamma_0}} \right\}$$

Let

$$\hat{\beta}^{(0)} = \hat{\beta}^{(1)} + \hat{\delta}$$

2.3. Initial Zero-Consistent Estimator

To ensure proper adaptivity in theory, this section employs the concept of zero-consistency (Huang et al., 2008) to argue for the oracle transfer property of $\hat{\delta}$. The main definition could be simply translated as the consistency for zero and non-zero sets, or more specifically,

Definition 1 *(zero-consistency)* Estimator $\tilde{\delta}$ is zero-consistent of rate r_n if $\exists\ \xi_l > 0, \forall \varepsilon > 0$, for sufficiently large n,

$$\max_{j \in \mathcal{A}} \left| \tilde{\delta}_j \right| = O_p(1/r_n), \quad \mathbb{P}\left(\min_{j \in \mathcal{A}^c} \left| \tilde{\delta}_j \right| \geq \xi_l\ l_{\mathcal{A}^c} \right) > 1 - \varepsilon,$$

where $\mathcal{A} \equiv \{j : \delta_j = 0\}$, $l_{\mathcal{A}^c} = \min\{|\delta_j| : j \in \mathcal{A}^c\}$ the lower bound for absolute value of non-zero parameters.

Based on the above definition, the oracle start still requires the following regular conditions. Define $\mathcal{B}^{(k)} \equiv \{j : \beta_j^{(k)} = 0\}$, $l_{\mathcal{B}^c}^{(k)} = \min\{|\beta_j^{(k)}| : j \notin \mathcal{B}^{(k)}\}$, $k \in \{0, 1\}$, conditions below promise oracle estimation of sparse δ if $q = 0$: For $k \in \{0, 1\}$,

(A1) $\epsilon_i^{(k)}$'s are i.i.d. sub-Gaussian with zero mean and s.t.d. $\sigma^{(k)}$;

(A2) Eigenvalues of $\Sigma_{\mathcal{B}^c}^{(k)}$ are bounded away from zero and infinity: $\exists\ M_1 < M_2$, s.t.

$$0 < M_1 < \Lambda_{\min}(\Sigma_{\mathcal{B}^c}^{(k)}) < \Lambda_{\max}(\Sigma_{\mathcal{B}^c}^{(k)}) < M_2 < \infty;$$

(A3) Initial estimators $\tilde{\beta}^{(1)}$ and $\tilde{\delta}$ are zero-consistent of rate $r_{n1}, r_{n0} \to \infty$;

(A4) $\lambda_{nk} \to \infty, \lambda_{nk}/n_k^{1/2} \to 0$ and

$$\frac{\sqrt{\lambda_{nk}}}{\sqrt{n_k} l_{\mathcal{B}^c}^{(k)}} \to 0, \quad \frac{\sqrt{n_k \log p}}{\lambda_{nk} r_{nk}} \to 0, \quad \frac{1}{r_{nk} l_{\mathcal{B}^c}^{(k)}} \to 0.$$

Lemma 1 Under *(A1)-(A4)* for $k = 1$, first step $\hat{\beta}^{(1)}$ has oracle property when $q = 0$:

$$\mathbb{P}\left(\hat{\beta}_{\mathcal{B}}^{(1)} = \beta_{\mathcal{B}}^{(1)}\right) \to 1,$$
$$\sqrt{n_1}\tau_{n_1}^{-1}\alpha_{n_1}^\top \left(\hat{\beta}_{\mathcal{B}^c}^{(1)} - \beta_{\mathcal{B}^c}^{(1)}\right) \xrightarrow{d} N(0,1),$$

where $\tau_{n_1} = \sigma^{(1)}\sqrt{\alpha_{n_1}^\top \Sigma_{\mathcal{B}^c}^{(1)^{-1}} \alpha_{n_1}}$ for any $\alpha_{n_1} \in \mathbb{R}^{|\mathcal{B}^{(1)^c}|}$ satisfying $\|\alpha_{n_1}\|_2 \leq 1$.

The first-stage $\hat{\beta}^{(1)}$ then contributes an oracle baseline for further de-biasing, however, the following theorem ensures regularity to such procedure.

Theorem 1 *(oracle transfer for $q = 0$)* Under *(A1)-(A4)* with bound l_2 norm of $x_{i,\mathcal{B}^{(1)^c}}^{(0)}$, $i = 1, ..., n_0$, sparse bias $\hat{\delta}$ is oracle through cross-fitted estimation for $\hat{\beta}^{(1)}$ and zero-consistent $\tilde{\delta}$ in **Algorithm 1**:

$$\mathbb{P}\left(\hat{\delta}_{\mathcal{A}} = \delta_{\mathcal{A}}\right) \to 1,$$
$$\sqrt{n_0}\tau_{n_0}^{-1}\alpha_{n_0}^\top \left(\hat{\delta}_{\mathcal{A}^c} - \delta_{\mathcal{A}^c}\right) \xrightarrow{d} N(0,1),$$

with $\tau_{n_0} = \sigma^{(0)}\sqrt{\alpha_{n_0}^\top \Sigma_{\mathcal{A}^c}^{(0)^{-1}} \alpha_{n_0}}$ with $\alpha_{n_0} \in \mathbb{R}^{|\mathcal{A}^c|}$, $\|\alpha_{n_0}\|_2 \leq 1$ analogously defined as in **Lemma 1**.

Notice that the convergence of $\hat{\delta}_{\mathcal{A}}$ ensures automatic detection of transferable parameter sets, which matches the oracle observation, while the asymptotic normality of set $\mathcal{A}^c$ promises estimation consistency for parameter set $\mathcal{A}^c \cap \mathcal{B}^{(1)^c}$ if combined with cross-fitted $\hat{\beta}^{(1)}$ in the first step. **Theorem 1** thus is similarly referred as the 'Oracle Transfer' property for GTL when $q = 0$. Then it is natural to consider the ultimate estimator for $\beta^{(0)}$, and the following argument intends to reveal performance enhancement of GTL compared with trivial non-transfer estimators, referred as the partial sharpness in convergence rate.

Corollary 1 *(partial sharpness for $q = 0$)* Under *(A1)-(A4)*, with bound l_2 norm of $x_{i,\mathcal{B}^{(1)^c}}^{(0)}$, $i = 1, ..., n_0$, and $n_0 = o(n_1)$, parameter set of $\hat{\beta}_{\mathcal{J}_2}^{(0)}$ has sharper convergence rate $O_p(n_1^{-1/2})$ through cross-fitted estimation for $\hat{\beta}^{(1)}$ and zero-consistent $\tilde{\delta}$ based on separation of the entire parameter set $\{1, ..., p\} = \mathcal{J}_1 \cup \mathcal{J}_2 \cup \mathcal{J}_3 \cup \mathcal{J}_4 \equiv \{\mathcal{A} \cap \mathcal{B}^{(1)}\} \cup \{\mathcal{A} \cap \mathcal{B}^{(1)^c}\} \cup \{\mathcal{A}^c \cap \mathcal{B}^{(1)}\} \cup \{\mathcal{A}^c \cap \mathcal{B}^{(1)^c}\}$.

Notice that for the other subset $\mathcal{J}_1 \cup \mathcal{J}_3 \cup \mathcal{J}_4$, the rate of convergence for GTL estimators is dominated by $O_p(n_0^{-1/2})$, which is of slower order $O_p(n_1^{-1/2})$ since $n_0 = o(n_1)$, and remains the same order before and after transferring.

2.4. Implementation

This section thus tends to answer the sequential question of how to effectively design such zero-consistent initial estimators in practice. For $\tilde{\beta}^{(1)}$'s zero-consistency, it is of typical example to apply Ordinary Least Square (OLS) when $p \leq n_0 << n_1$ as well as marginal correlation for high dimensions under partial orthogonality assumption as in Huang et al. (2008), which was described as weak correlations between covariates of zero and non-zero parameters:

$$\frac{1}{n_0}\sum_{i=1}^{n_0} x_{ij}^{(0)} x_{ik}^{(0)} = O\left(n_0^{-1/2}\right), \quad j \in \mathcal{B}^{(0)}, k \notin \mathcal{B}^{(0)}.$$

for target data.

For $\tilde{\delta}$, however, simply conduct respective marginal correlation for $\check{\beta}^{(0)}$ and $\check{\beta}^{(1)}$ is not enough to guarantee zero-consistency, since the inconsistency for non-zero sets may confuse the difference. Therefore, here I further employ the Adaptive Lasso estimators respectively on target and source data under zero-consistent initial estimators beforehand, in order to get oracle $\check{\beta}^{(0)}, \check{\beta}^{(1)}$. And the following theorem confirms $\tilde{\delta}$'s zero-consistency by **Lemma 1**.

Theorem 2 *(zero-consistent $\tilde{\delta}$)* Under *(A1)-(A4)* and $\check{\beta}^{(0)}, \check{\beta}^{(1)}$ with oracle property. For $n_0 = o(n_1)$, $|\mathcal{B}^{(0)^c} \cap \mathcal{B}^{(1)^c} \cap \mathcal{A}| = o(\sqrt{n_0}/r_{n_0})$, $\exists\ \xi_l > 0$, $\forall \varepsilon > 0$,

$$\mathbb{P}\left(r_{n_0} \max_{j \in \mathcal{A}} \left|\tilde{\delta}_j\right| > \varepsilon\right) \to 0, \quad \mathbb{P}\left(\min_{j \in \mathcal{A}^c} \left|\tilde{\delta}_j\right| > \xi_l\ l_{\mathcal{A}^c}\right) \to 1.$$

Note that the initial zero-consistent estimation procedure could be replaced with any regularization estimators owning oracle properties, such as Smoothly Clipped Absolute Deviation (SCAD) penalty as in Fan & Li (2001), or Minimax Concave Penalty (MCP) from Zhang (2010).

3 Non-Asymptotic Analysis

The formal discussion on sparse δ fails immediately when $q = 1$, this section thus intends to analyze the complementary case for cumulative l_1 measure, and is willing to be extended for $q \in (0, 1)$ if following similar argument as in Li et al. (2022)'s appendix C. At the first hand, we reveal the upper bound relying on the restricted strong convexity (RSC) technique in Raskutti et al. (2010) for both estimation and prediction issues. And the second part automatically considers the minimax lower bound for GTL estimators. The following assumption and lemma in addition promises theorems' validity:

(A5) For $k \in \{0, 1\}$, each row of $X^{(k)}$ is i.i.d. Gaussian with mean zero and covariance matrix $\Sigma^{(k)}$ and $\mathbb{E}[(y_i^{(k)})^2]$ is finite. For constant C_0, $\max_{k \in \{0,1\}} \mathbb{E}[\exp\{t\epsilon_i^{(k)}\}] \leq \exp\{t^2 C_0\}$ for all $t \in \mathbb{R}$.

3.1. Rate of Convergence

To fully show GTL's adaptive rate in transferring process, define the pre-specified guidance $\|w\|_\infty \equiv \max_{j \in \{1,\ldots,p\}} w_j$, where $w_j \equiv 1/|\tilde{\delta}_j|^{\gamma_0} \in \mathbb{R}^p$. The following theorem reveals simultaneous control of the convergence rate by both h and $\|w\|_\infty$. For informative sources as defined in Li et al. (2022), small h similarly serves as the dominating factor in the convergence rate for both prediction and estimation issues; while for huge dissimilarities among target and sources, the guidance $\|w\|_\infty$ should play the leading role with small values to ensure enhanced performance for parameter estimation and outcome prediction compared with applying pure Lasso on the target data, which owns the rate of $O_p(s \log p/n_0)$. γ_0 thus determines the sensitivity for negative transfers.

Theorem 3 *(convergence rate for $q = 1$)* Under *(A1)*,*(A2)*, and *(A5)*, take $\lambda_{n_1} = c_1\sqrt{\mathbb{E}[(y_i^{(1)})^2] \log p/n_1}$ and $\lambda_{n_0} = c_0\sqrt{\log p/n_0}$ for large c_1, c_0, with $C_\Sigma h \lesssim s\sqrt{\log p/n_0}$ and $n_0 \lesssim n_1$. If $s \log p/n_1 + C_\Sigma h\,(\log p/n_0)^{1/2} = o(1)$, then

$$\inf_{B \in \Theta_1(s,h)} \mathbb{P}\left(\frac{1}{n_0}\left\|X^{(0)}(\hat{\beta}^{(0)} - \beta^{(0)})\right\|_2^2 \vee \|\hat{\beta}^{(0)} - \beta^{(0)}\|_2^2 \lesssim \frac{s \log p}{n_0 + n_1} + \frac{s \log p}{n_0} \wedge \lambda_{n_0}\|w\|_\infty C_\Sigma h \wedge \bar{w}^2 C_\Sigma^2 h^2\right)$$
$$\geq 1 - \exp\left(-c_2 \log p\right),$$

where $C_\Sigma = 1 + \max_{j \leq p} \|e_j^\top \left(\Sigma^{(1)} - \Sigma^{(0)}\right)\left(\Sigma^{(1)}\right)^{-1}\|_1$, and $\bar{w} \equiv \frac{\|w\|_\infty}{\min_j\{w_j - 1/\kappa\}} > 0$ for κ being a large positive constant.

As is obviously shown in the theorem, the prediction and estimation error are jointly controlled by the estimation rate $s \log p/(n_0 + n_1)$ for sparse $\beta^{(0)}$ from combined sample of target and source, plus the original $O_p(s \log p/n_0)$ rate that is willing to be dominated by the following $\lambda_{n_0}\|w\|_\infty C_\Sigma h$ and $\bar{w}^2 C_\Sigma^2 h^2$ terms, where the formal serves for a doubly robust domination through both h and $\|w\|_\infty$, which is expected to variate in the opposite direction, and this is further demonstrated in section 3.3. However, both are contaminated by the heterogeneity within covariance matrices $\Sigma^{(0)}$ and $\Sigma^{(1)}$.

3.2. Minimax Optimality

In addition to the upper bound derived above, this section tries to claim that the convergence rate in **Theorem 3** is minimax optimal. And this depends on the following least dissimilarity condition, which calibrates the scenario to the case that there exists few large differences among p dimensions between the

target and source model while transferring.

(A6) The largest absolute difference of initial estimator $\left\|\check{\beta}^{(0)}-\check{\beta}^{(1)}\right\|_\infty$ should be great enough such that $\min_j\{w_j-1/\kappa\}^2\leq s$ for s being the sparsity level.

Then the theorem below reveals the minimax lower bound that owns the same order with the previous upper bound, resulting in the minimax optimality of GTL estimators. Detailed proofs could be found in appendix.

Theorem 4 *(minimax lower bound for $q=1$)* Under *(A1)*,*(A2)*,*(A5)*, and *(A6)*, for constant $0<C_{01}<1$, if $\max\{s\log p/\left(n_{\mathcal{A}}+n_0\right),h\sqrt{\log p/n_0}\}=o(1)$, $1\leq s\leq\frac{2}{3}p$ for large p, then

$$\inf_{\hat{\beta}^{(0)}}\sup_{\Theta_1(s,h)}\mathbb{P}\left(\|\hat{\beta}^{(0)}-\beta^{(0)}\|_2^2\gtrsim\frac{s\log p}{n_0+n_1}+\frac{s\log p}{n_0}\wedge\sqrt{\log p/n_0}\|w\|_\infty C_\Sigma h\wedge\bar{w}^2C_\Sigma^2h^2\right)\geq C_{01}.$$

3.3. Achievability of Double Robustness

To perform double robustness in the above convergence rate, the pre-assumed guidance $\|w\|_\infty$ needs to variate oppositely with respect to h. This section thus intends to show that zero-consistent $\tilde{\delta}$ that constructs $w\equiv 1/|\tilde{\delta}|^{\gamma_0}$ leads to such conclusion naturally.

Corollary 2 Based on oracle $\check{\beta}^{(0)}$ and $\check{\beta}^{(1)}$, $\tilde{\delta}_j-\delta_j=O_p(1/\sqrt{n_0})$.

From the above corollary, it is not hard to observe that if large h is due to few huge point differences among all dimensions, then $\|w\|_\infty$ is restricted small by these extreme $\tilde{\delta}_j$'s. However, if the volume of h is resulting from cumulatively mild differences for p dimensions, the infinity norm will not be able to capture such information and would thus lose the double robustness.

4 Simulation

This section contains extensive scenarios to verify GTL's enhanced performance for parameter estimation, following the data generating process in 2.1 as well as the partial orthogonality assumption for covariate matrix.

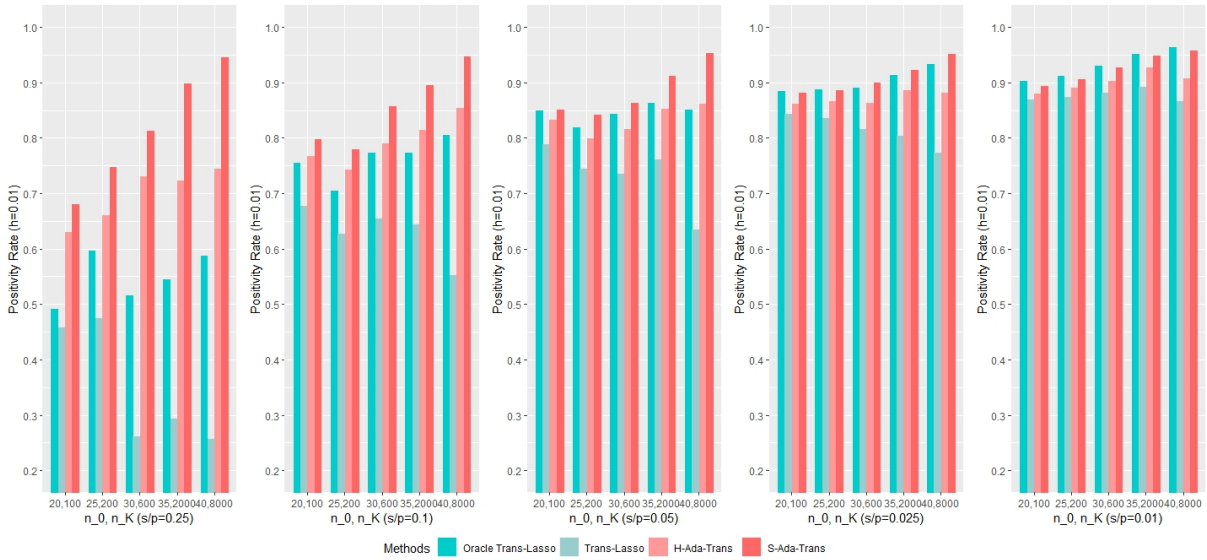
4.1. Enhanced Estimation Accuracy

The first setting is analogous to Li et al. (2022), where the source is suitably designed for Oracle Trans-Lasso. Specifically, with $(n_0,n_1)\in\{(20,100),(25,200),(30,600),(35,2000),(40,8000)\}$ divergent in a slack and explosive rate respectively; Without loss of generality, assume the target and source are calibrated with the same sparsity structure with degree $s/p\in\{0.01,0.025,0.05,0.1,0.25\}$ and the corresponding dimension $p\in\{20,50,100,200,500\}$; Combine with $h=0.01$ under $q=1$ to complete the specified parameter space. On the other hand, the covariate matrix for the target and source are both with mean 1 and $\Sigma_{ij}=0.5^{|i-j|},i,j\in\{2,...,p\}$ with first row and column being 0 except the diagonal. The noises are both set to be standard normal. For tunings, $\gamma_2=0.05$ for avoiding extremes, while λ's are chosen through cross-validation.

The following histograms are comparing GTL with the previous benchmark-Oracle Trans-Lasso method under such scenario. In addition, multiple types of adaptiveness are included for better illustration: Trans-Lasso for a source-wise perspective; H-Ada-Trans with a hard threshold for adaptivity, which employs

$$\hat{\delta}=\arg\min_{\delta\in\mathbb{R}^p}\left\{\frac{1}{2n_0}\sum_{i=1}^{n_0}\left[y_i^{(0)}-x_i^{(0)\top}\left(\hat{\beta}^{(1)}+\delta\right)\right]^2+\lambda_{n0}\sum_{j=1}^{p}w_j|\delta_j|\right\}$$

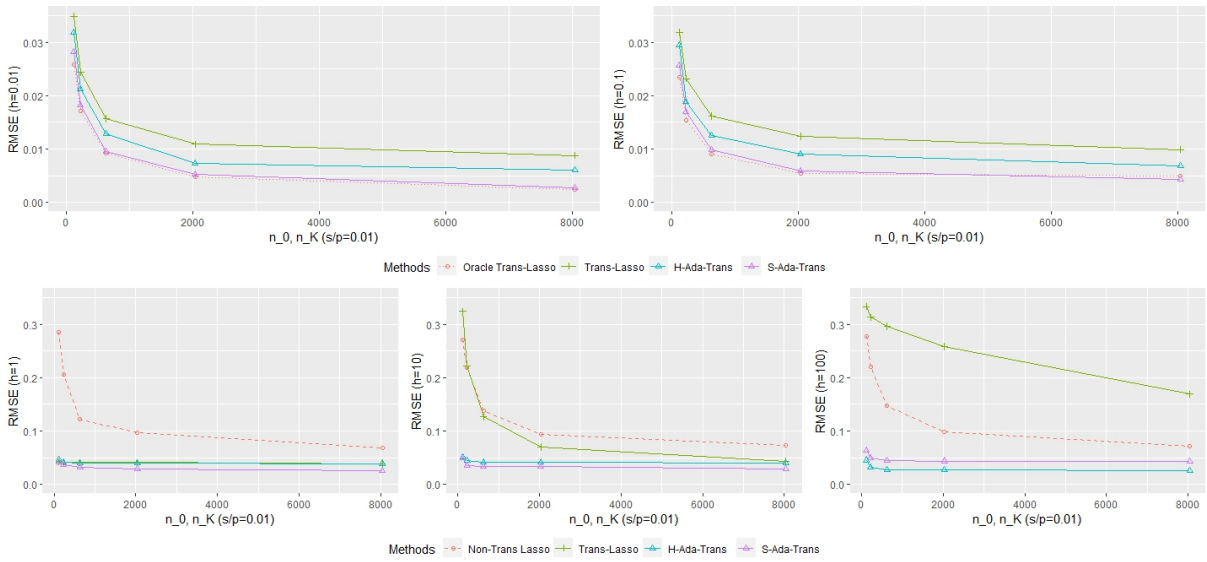
with $w_j\equiv I\{j\in\mathcal{M}\}$, where $\mathcal{M}$ is the index set for the ν largest values of $|\tilde{\beta}_j^{(0)}-\tilde{\beta}_j^{(1)}|$, and $\nu=1$ in simulation; as the counterpart for parameter-wise adaptivity, GTL is indicated as the S-Ada-Trans method for smooth adaptiveness.



Under various sample sizes and sparsity levels, the y-axis average positive rate among 100 numerical experiments denotes for the comparable ratio of the mentioned methods over Root Mean Square Error (RMSE) for parameter estimation on target data only -1. It is clear that firstly, all four transfer learning methods obtain positive rate > 0, indicating non-negative transfer with knowledge from the informative source. However, among five different sparsity structure, GTL shows uniform advantages over the other three while Oracle Trans-Lasso and Trans-Lasso perform better when there is extremely sparsity underlying. Additionally, there are fluctuations for positive rates for growing sample size of n_0, n_1 in Oracle Trans-Lasso and Trans-Lasso method, however, GTL presents almost strictly advances when the sample's information increases, and this is intuitive from the adaptive regularization of each dimension.

4.2. Consistency and Robustness

What's following is a relaxation for $h \in \{0.01, 0.1, 1, 10, 100\}$, which differs from the original settings in Li et al. (2022), and here the dissimilarity is concentrated in a point mass under l_1 measure in $\Theta_q(s, h)$ for calibration of $\nu = 1$ in H-Ada-Trans. The second row in the chart below also contains the red dashed line serving as the boundary for negative transfer when the gap h is large.

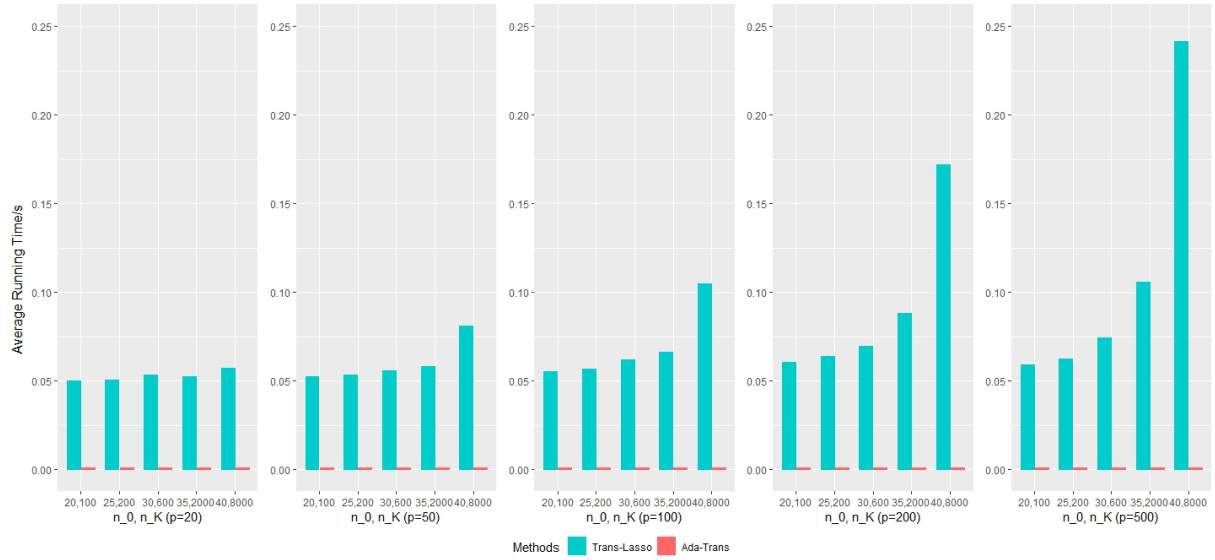


The green solid line, representing Trans-Lasso, shows a obvious shifting from positive to negative transfer

above the boundary as h increases from 1 to 100 while GTL method stays robust for 'non-informative' source. Even for source that is comparably transferable, GTL enjoys a faster rate of convergence over Trans-Lasso and is nearly oracle.

4.3. Computational Efficiency

Through embedding the adaptiveness within the de-biasing step while transferring, GTL shrinks the computation burden compared with Trans-Lasso. The following graph explicitly reveals such property in practice.



The average running time for each simulation explodes as sample size (n_0, n_1) goes larger especially under high dimensions if applying Trans-Lasso for delivering the knowledge, which is understandable since through sure screening for estimated rank indexes, Trans-Lasso in Li et al. (2022) was indeed switching the amount of computation from the ensemble procedure to the screening issue under large n_1 as well as large p, and would be never efficient through source-wise adaptivity. However, GTL serves from another perspetive and saves large amount of computation time even in case of large n_1 large p scenario. More details of simulations could be found in the appendix.

5 Empirical Analysis

To demonstrate GTL's performance in practice, this section introduces multiple empirical scenarios that fits into the framework of transfer learning for further applications of Guided Trans-Lasso, which is also compared with the benchmark method introduced by Li et al. (2022).

The first background relates to bioscience especially, where the database refers to Genotype-Tissue Expression (GTEx) (`https://gtexportal.org/`), a project offering resources for human gene expression study and regulations, as well as the relationship with genetic variation. A subset of which is picked by choosing MODULE_137 (`https://www.gsea-msigdb.org/gsea/msigdb/cards/MODULE 137.html`) genes and 13 human brains tissues (`https://gtexportal.org/home/samplingSitePage`) samples, forming the data matrix of size $2,642 \times 546$ that primely concerns on human brains' functioning process.

Also, similar to Li *et al.* (2022)'s transferring strategy, I set `JAM2`, a gene located at chromosome 21 that encodes a crucial protein for lymphocyte homing to secondary lymphoid organs (Johnson-Léger *et al.*, 2002), as the response variable, whose malfunction may lay threatening factors for primary familial brain calcification (Cen *et al.*, 2020; Schottlaender *et al.*, 2020). Then allocate each individual brain tissue as target

dataset, with average sample size of 203, and other 12 tissues as a whole being potential sources, what's following compares the relative prediction error and the average running time between various methods including Oracle Trans-Lasso, Trans-Lasso and GTL.

With x-axis denotes each target brain tissue, the relative prediction error in the y-axis describes the comparative error ratio of transferring methods over Lasso (Tibshirani, 1996) merely based on target sample. Dashed line represents the negative transfer boundary with relative prediction error being 1, which indicates efficacy of adaptivity for avoidance of negative transfer through comparison between Oracle Trans-Lasso and Trans-Lasso as well as GTL. However, Uniformly for all target tissues, GTL performs the most accurate prediction with the lowest relative error below Trans-Lasso. In addition, parallel to the estimation result, the computation time is much shrinked through parameter-wise adaptive GTL.

6 Conclusion

This paper analyzes a transfer learning scenario from large but limited sources arbitrarily, designing Guided Trans-Lasso algorithm that is free of detection procedures for informative sources or data ensemble processes, ensuring oracle transfer under sparse differences between target and source parameters, also providing doubly robust rate of convergence for estimation and prediction tasks with minimax optimality under cumulative measure of parameter differences. Simulation and empirical research further validates the efficacy of GTL estimation and prediction in various setting.

REFERENCE

[1] Cen, Z., Y. Chen, S. Chen, et al. (2020). Biallelic Loss-Of-Function Mutations in Jam2 Cause Primary Familial Brain Calcification. *Brain*, 143(2), 491-502.

[2] Chatterjee, A., and S. N. Lahiri (2013). Rates of Convergence of The Adaptive LASSO Estimators to The Oracle Distribution and Higher Order Refinements by The Bootstrap. *The Annals of Statistics*, 41(3), 1232-1259.

[3] Deng, J., W. Dong, R. Socher, L. Li, K. Li, and F. Li (2009). ImageNet: A Large-scale Hierarchical Image Database. *IEEE Conference on Computer Vision and Pattern Recognition*, 248-255.

[4] Duan, Y., and K. Wang (2023). Adaptive and Robust Multi-Task Learning. *The Annals of Statistics*, 51(5), 2015-2039.

[5] Fan, J., S. Guo, and N. Hao (2012). Variance Estimation Using Refitted Cross-Validation in Ultrahigh Dimensional Regression. *Journal of the Royal Statistical Society: Series B*, 74(1), 37-65.

[6] Fan, J., and R. Li (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96(456), 1348-1360.

[7] Haar L. V., T. Elvira, and O. Ochoa (2023). An Analysis of Explainability Methods for Convolutional Neural Networks. *Engineering Applications of Artificial Intelligence*, 117(A), 105606.

[8] Huang, J., S. Ma, and C. Zhang (2008). Adaptive Lasso for Sparse High-Dimensional Regression Models. *Statistica Sinica*, 18(4), 1603-1618.

[9] Johnson-Léger, C. A., M. Aurrand-Lions, N. Beltraminelli, N. Fasel, and B. A. Imhof (2002). Junctional Adhesion Molecule-2 (Jam-2) Promotes Lymphocyte Transendothelial Migration. *Blood, The Journal of the American Society of Hematology*, 100(7), 2479-2486.

[10] Li, S., T. T. Cai, and H. Li (2022). Transfer Learning for High-Dimensional Linear Regression: Prediction, Estimation and Minimax Optimality. *Journal of the Royal Statistical Society: Series B*, 84(1), 149-173.

[11] Li, X., R. Li, Z. Xia, and C. Xu (2020). Distributed Feature Screening via Component-wise Debiasing. *Journal of Machine Learning Research*, 21(24), 1-32.

[12] Qu, X. (2024+). Partial Transfer Learning Under High-Dimensional Confounding: Estimation, Prediction, and Efficiency. *Working Paper*.

[13] Raskutti, G., M. J. Wainwright, and B. Yu (2010). Restricted Eigenvalue Properties for Correlated Gaussian Designs. *Journal of Machine Learning Research*, 11(78), 2241-2259.

[14] Raskutti, G., M. J. Wainwright, and B. Yu (2011). Minimax Rates of Estimation for High-Dimensional Linear Regression Over l_q-Balls. *IEEE Transactions on Information Theory*, 57(10), 6976-6994.

[15] Rigollet, P., and A. Tsybakov (2011). Exponential Screening and Optimal Rates of Sparse Estimation. *The Annals of Statistics*, 39(2), 731-771.

[16] Schottlaender, L. V., R. Abeti, Z. Jaunmuktane, et al. (2020). Bi-Allelic Jam2 Variants Lead to Early-Onset Recessive Primary Familial Brain Calcification. *The American Journal of Human Genetics*, 106(3), 412-421.

[17] Tian, Y., and Y. Feng (2022). Transfer Learning Under High-Dimensional Generalized Linear Models. *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2022.2071278.

[18] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, 58, 267-288.

[19] Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476), 1418-1429.

[20] Zhang, C. (2010). Nearly Unbiased Variable Selection Under Minimax Concave Penalty. *The Annals of Statistics*, 38(2), 894-942.

APPENDIX

A Technical Proofs

A.1. Asymptotics for $q = 0$

Lemma 1 Under *(A1)-(A4)* for $k = 1$, first step $\hat{\beta}^{(1)}$ has oracle property when $q = 0$:

$$\mathbb{P}\left(\hat{\beta}^{(1)}_{\mathcal{B}} = \beta^{(1)}_{\mathcal{B}}\right) \to 1,$$

$$\sqrt{n_1}\tau_{n_1}^{-1}\alpha_{n_1}^{\top}\left(\hat{\beta}^{(1)}_{\mathcal{B}^c} - \beta^{(1)}_{\mathcal{B}^c}\right) \xrightarrow{d} N(0,1),$$

where $\tau_{n_1} = \sigma^{(1)}\sqrt{\alpha_{n_1}^{\top}\Sigma^{(1)^{-1}}_{\mathcal{B}^c}\alpha_{n_1}}$ for any $\alpha_{n_1} \in \mathbb{R}^{|\mathcal{B}^{(1)^c}|}$ satisfying $\|\alpha_{n_1}\|_2 \leq 1$.

Proof. This lemma could be verified the same way as for Theorem 1 and 2 in Huang et al. (2008). □

Theorem 1 *(oracle transfer)* Under *(A1)-(A4)* with bound l_2 norm of $x^{(0)}_{i,\mathcal{B}^{(1)^c}}$, $i = 1, ..., n_0$, sparse bias $\hat{\delta}$ is oracle through cross-fitted estimation for $\hat{\beta}^{(1)}$ and zero-consistent $\tilde{\delta}$ in **Algorithm 1**:

$$\mathbb{P}\left(\hat{\delta}_{\mathcal{A}} = \delta_{\mathcal{A}}\right) \to 1,$$

$$\sqrt{n_0}\tau_{n_0}^{-1}\alpha_{n_0}^{\top}\left(\hat{\delta}_{\mathcal{A}^c} - \delta_{\mathcal{A}^c}\right) \xrightarrow{d} N(0,1),$$

with $\tau_{n_0} = \sigma^{(0)}\sqrt{\alpha_{n_0}^{\top}\Sigma^{(0)^{-1}}_{\mathcal{A}^c}\alpha_{n_0}}$ with $\alpha_{n_0} \in \mathbb{R}^{|\mathcal{A}^c|}$, $\|\alpha_{n_0}\|_2 \leq 1$ analogously defined as in **Lemma 1**.

Proof. Rewrite the target Data Generating Process (DGP) as

$$\begin{aligned}
y_i^{(0)} &= x_i^{(0)^{\top}}\beta^{(1)} + x_i^{(0)^{\top}}\delta + \epsilon_i^{(0)} \\
&= x_i^{(0)^{\top}}\hat{\beta}^{(1)} + x_i^{(0)^{\top}}\left(\beta^{(1)} - \hat{\beta}^{(1)}\right) + x_i^{(0)^{\top}}\delta + \epsilon_i^{(0)} \\
&= x_i^{(0)^{\top}}\hat{\beta}^{(1)} + x^{(0)^{\top}}_{i,\mathcal{B}^{(1)}}\left(\beta^{(1)}_{\mathcal{B}} - \hat{\beta}^{(1)}_{\mathcal{B}}\right) + x^{(0)^{\top}}_{i,\mathcal{B}^{(1)^c}}\left(\beta^{(1)}_{\mathcal{B}^c} - \hat{\beta}^{(1)}_{\mathcal{B}^c}\right) + x_i^{(0)^{\top}}\delta + \epsilon_i^{(0)}
\end{aligned}$$

for $i = 1, ..., n_0$. Based on **Lemma 1**, $\exists\ N_1 > 0$, $\forall n_1 > N_1$, $\beta^{(1)}_{\mathcal{B}} = \hat{\beta}^{(1)}_{\mathcal{B}}$ with high probability. Also, since $\bar{x}^{(0)^{\top}}_{i,\mathcal{B}^{(1)^c}}\left(\beta^{(1)}_{\mathcal{B}^c} - \hat{\beta}^{(1)}_{\mathcal{B}^c}\right) = O_p(n_1^{-1/2})$ with $\bar{x}^{(0)}_{i,\mathcal{B}^{(1)^c}} = x^{(0)}_{i,\mathcal{B}^{(1)^c}}/\|x^{(0)}_{i,\mathcal{B}^{(1)^c}}\|_2$, as long as $\|x^{(0)}_{i,\mathcal{B}^{(1)^c}}\|_2 < C < \infty$ bounded by a constant C, $x^{(0)^{\top}}_{i,\mathcal{B}^{(1)^c}}\left(\beta^{(1)}_{\mathcal{B}^c} - \hat{\beta}^{(1)}_{\mathcal{B}^c}\right) = O_p(n_1^{-1/2})$ is approximately zero for large N_1 with high probability. Through cross-fitted estimation for $\hat{\beta}^{(1)}$ and zero-consistent $\tilde{\delta}$, the de-biasing step equals to applying Adaptive Lasso on

$$\tilde{y}_i^{(0)} = x_i^{(0)^{\top}}\delta + \epsilon_i^{(0)} \tag{1}$$

asymptotically for large n_1, where $\tilde{y}_i^{(0)} \equiv y_i^{(0)} - x_i^{(0)^{\top}}\hat{\beta}^{(1)}$. Therefore, inheriting the oracle start at step 1, the underlying DGP prior to step 2 matches the pre-assumed model (1) asymptotically before implementing Adaptive Lasso, and $\hat{\delta}$ is thus oracle promised by *(A1)-(A4)* and **Lemma 1**. □

Corollary 1 *(partial sharpness)* Under *(A1)-(A4)*, with bound l_2 norm of $x^{(0)}_{i,\mathcal{B}^{(1)^c}}$, $i = 1, ..., n_0$, and $n_0 = o(n_1)$, parameter set of $\hat{\beta}^{(0)}_{\mathcal{J}_2}$ has sharper convergence rate $O_p(n_1^{-1/2})$ through cross-fitted estimation for $\hat{\beta}^{(1)}$ and zero-consistent $\tilde{\delta}$.

Proof. Continue on **Lemma 1** and **Theorem 1**, we have

$$\mathbb{P}\left(\hat{\delta}_{\mathcal{J}_2} = \delta_{\mathcal{J}_2}\right) \to 1, \quad \sqrt{n_1}\tau_{n_1}^{-1}\alpha_{n_1}^{\top}\left(\hat{\beta}^{(1)}_{\mathcal{J}_2} - \beta^{(1)}_{\mathcal{J}_2}\right) \xrightarrow{d} N(0,1).$$

For independent sample estimation of $\hat{\delta}_{\mathcal{J}_2}$ and $\hat{\beta}^{(1)}_{\mathcal{J}_2}$ promised by cross-fitting technique,

$$\hat{\beta}^{(0)}_{\mathcal{J}_2} - \beta^{(0)}_{\mathcal{J}_2} = \hat{\delta}_{\mathcal{J}_2} - \delta_{\mathcal{J}_2} + \hat{\beta}^{(1)}_{\mathcal{J}_2} - \beta^{(1)}_{\mathcal{J}_2} = O_p(1/\sqrt{n_1})$$

for n_1 sufficiently large. This is sharper than $O_p(1/\sqrt{n_0})$, the convergence rate of pure oracle estimation on non-zero $\mathcal{J}_2$ based on target sample if applying **Lemma 1**. □

Theorem 2 *(zero-consistent $\tilde{\delta}$)* Under *(A1)-(A4)* and $\check{\beta}^{(0)}, \check{\beta}^{(1)}$ with oracle property. For $n_0 = o(n_1)$, $|\mathcal{B}^{(0)^c} \cap \mathcal{B}^{(1)^c} \cap \mathcal{A}| = o(\sqrt{n_0}/r_{n_0})$, $\exists\ \xi_l > 0$, $\forall \varepsilon > 0$,

$$\mathbb{P}\left(r_{n_0} \max_{j\in\mathcal{A}} \left|\tilde{\delta}_j\right| > \varepsilon\right) \to 0, \quad \mathbb{P}\left(\min_{j\in\mathcal{A}^c} \left|\tilde{\delta}_j\right| > \xi_l\ l_{\mathcal{A}^c}\right) \to 1.$$

Proof. By definition, $\tilde{\delta}_j \equiv \check{\beta}_j^{(0)} - \check{\beta}_j^{(1)}$, consider the following partition of $\{1, ..., p\} = \{\mathcal{B}^{(0)} \cap \mathcal{B}^{(1)}, \mathcal{B}^{(0)^c} \cap \mathcal{B}^{(1)}, \mathcal{B}^{(0)} \cap \mathcal{B}^{(1)^c}, \mathcal{B}^{(0)^c} \cap \mathcal{B}^{(1)^c} \cap \mathcal{A}, \mathcal{B}^{(0)^c} \cap \mathcal{B}^{(1)^c} \cap \mathcal{A}^c\}$ and denote these index sets as $\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3, \mathcal{I}_4, \mathcal{I}_5$, such that $\mathcal{I}_1 \cup \mathcal{I}_4 = \mathcal{A}$ and $\mathcal{I}_2 \cup \mathcal{I}_3 \cup \mathcal{I}_5 = \mathcal{A}^c$.

For $j \in \mathcal{I}_1$, due to oracle property of $\check{\beta}_j^{(0)}$ and $\check{\beta}_j^{(1)}$ by **Lemma 1** under *(A1)-(A4)*,

$$\mathbb{P}\left(\check{\beta}_{\mathcal{B}}^{(0)} = \beta_{\mathcal{B}}^{(0)}\right) \to 1, \quad \mathbb{P}\left(\check{\beta}_{\mathcal{B}}^{(1)} = \beta_{\mathcal{B}}^{(1)}\right) \to 1,$$

i.e. $\exists\ N$, $\forall n_0, n_1 > N$, $\forall \epsilon > 0$,

$$\mathbb{P}\left(\check{\beta}_{\mathcal{B}}^{(0)} \neq \beta_{\mathcal{B}}^{(0)}\right) < \epsilon, \quad \mathbb{P}\left(\check{\beta}_{\mathcal{B}}^{(1)} \neq \beta_{\mathcal{B}}^{(1)}\right) < \epsilon.$$

Thus for $\varepsilon > 0$,

$$\begin{aligned}
\mathbb{P}\left(r_{n_0} \max_{j\in\mathcal{I}_1} \left|\tilde{\delta}_j\right| > \varepsilon\right) &\leq \mathbb{P}\left(r_{n_0} \sum_{j\in\mathcal{I}_1} \left|\tilde{\delta}_j\right| > \varepsilon\right) = \mathbb{P}\left(r_{n_0} \sum_{j\in\mathcal{I}_1} \left|\check{\beta}_j^{(0)} - \check{\beta}_j^{(1)}\right| > \varepsilon\right) \\
&\leq \mathbb{P}\left(r_{n_0} \sum_{j\in\mathcal{I}_1} \left|\check{\beta}_j^{(0)} - \beta_j^{(0)}\right| + r_{n_0} \sum_{j\in\mathcal{I}_1} \left|\check{\beta}_j^{(1)} - \beta_j^{(1)}\right| > \varepsilon\right) \\
&\leq \mathbb{P}\left(r_{n_0} \sum_{j\in\mathcal{I}_1} \left|\check{\beta}_j^{(0)} - \beta_j^{(0)}\right| > \varepsilon/2\right) + \mathbb{P}\left(r_{n_0} \sum_{j\in\mathcal{I}_1} \left|\check{\beta}_j^{(1)} - \beta_j^{(1)}\right| > \varepsilon/2\right) \\
&\leq \mathbb{P}\left(\sum_{j\in\mathcal{I}_1} \left|\check{\beta}_j^{(0)} - \beta_j^{(0)}\right| > 0\right) + \mathbb{P}\left(\sum_{j\in\mathcal{I}_1} \left|\check{\beta}_j^{(1)} - \beta_j^{(1)}\right| > 0\right) < 2\epsilon
\end{aligned}$$

For $j \in \mathcal{I}_4$, since

$$\sqrt{n_0}\tau_{n_0}^{-1}\alpha_{n_0}^\top \left(\check{\beta}_{\mathcal{B}^c}^{(0)} - \beta_{\mathcal{B}^c}^{(0)}\right) \xrightarrow{d} N(0,1), \quad \sqrt{n_1}\tau_{n_1}^{-1}\alpha_{n_1}^\top \left(\check{\beta}_{\mathcal{B}^c}^{(1)} - \beta_{\mathcal{B}^c}^{(1)}\right) \xrightarrow{d} N(0,1),$$

due to **Lemma 1**,

$$r_{n_0} \max_{j\in\mathcal{I}_4} \left|\tilde{\delta}_j\right| \leq r_{n_0} \sum_{j\in\mathcal{I}_4} \left|\check{\beta}_j^{(0)} - \beta_j^{(0)}\right| + r_{n_0} \sum_{j\in\mathcal{I}_4} \left|\check{\beta}_j^{(1)} - \beta_j^{(1)}\right| = O_p\left(r_{n_0} \left|\mathcal{I}_4\right| n_0^{-1/2}\right)$$

for $n_0 = o(n_1)$. Therefore, for $n_0, n_1 > N$, $\forall \epsilon > 0$,

$$\mathbb{P}\left(r_{n_0} \max_{j\in\mathcal{A}} \left|\tilde{\delta}_j\right| > \varepsilon\right) \leq \mathbb{P}\left(r_{n_0} \max_{j\in\mathcal{I}_1} \left|\tilde{\delta}_j\right| > \varepsilon/2\right) + \mathbb{P}\left(r_{n_0} \max_{j\in\mathcal{I}_4} \left|\tilde{\delta}_j\right| > \varepsilon/2\right) \leq C\epsilon$$

as long as $r_{n_0}|\mathcal{I}_4|/\sqrt{n_0} \to 0$ as $n_0 \to \infty$, where $C > 0$ is a constant.

On the other hand, for $j \in \mathcal{I}_2$, $\sqrt{n_0}\tau_{n_0}^{-1}\alpha_{n_0}^\top \left(\check{\beta}_{\mathcal{B}^c}^{(0)} - \beta_{\mathcal{B}^c}^{(0)}\right) \xrightarrow{d} N(0,1)$, $\mathbb{P}\left(\check{\beta}_{\mathcal{B}}^{(1)} = \beta_{\mathcal{B}}^{(1)}\right) \to 1$,

$$\begin{aligned}
\min_{j\in\mathcal{I}_2} \left|\tilde{\delta}_j\right| = \min_{j\in\mathcal{I}_2} \left|\check{\beta}_j^{(0)} - \check{\beta}_j^{(1)}\right| &= \min_{j\in\mathcal{I}_2} \left|\check{\beta}_j^{(0)} - \beta_j^{(0)} - \left(\check{\beta}_j^{(1)} - \beta_j^{(1)}\right) - \left(\beta_j^{(1)} - \beta_j^{(0)}\right)\right| \\
&\geq \min_{j\in\mathcal{I}_2} \left|\beta_j^{(1)} - \beta_j^{(0)}\right| + \min_{j\in\mathcal{I}_2} \left|\check{\beta}_j^{(1)} - \beta_j^{(1)}\right| - \min_{j\in\mathcal{I}_2} \left|\check{\beta}_j^{(0)} - \beta_j^{(0)}\right| \\
&\geq l_{\mathcal{A}^c} + \min_{j\in\mathcal{I}_2} \left|\check{\beta}_j^{(1)} - \beta_j^{(1)}\right| = l_{\mathcal{A}^c}
\end{aligned}$$

for $n_1 > N$. Therefore, $\exists\ \xi_l > 0$, $\forall n_1 > N$,

$$\mathbb{P}\left(\min_{j\in\mathcal{I}_2} \left|\tilde{\delta}_j\right| > \xi_l l_{\mathcal{A}^c}\right) \geq \mathbb{P}\left(l_{\mathcal{A}^c} > \xi_l l_{\mathcal{A}^c}\right) = 1.$$

Analogous proof could be shown for $j \in \mathcal{I}_3$ as $\mathbb{P}\left(\min_{j\in\mathcal{I}_3} \left|\tilde{\delta}_j\right| > \xi_l l_{\mathcal{A}^c}\right) \to 1$ as $n_0 \to \infty$. For $j \in \mathcal{I}_5$, $\min_{j\in\mathcal{I}_5} |\tilde{\delta}_j| \geq$

$l_{\mathcal{A}^c} + O_p(n_0^{-1/2})$, then $\exists\ \xi_l > 0$, $\forall n_0, n_1 > N$,

$$\mathbb{P}\left(\min_{j\in\mathcal{I}_5}\left|\tilde{\delta}_j\right| > \xi_l l_{\mathcal{A}^c}\right) \geq \mathbb{P}\left(l_{\mathcal{A}^c} + O_p(n_0^{-1/2}) > \xi_l l_{\mathcal{A}^c}\right) = 1$$

as long as $\xi_l = o_p(n_0^{-1/2})$. Combine the cases for $\mathcal{I}_2, \mathcal{I}_3$ and $\mathcal{I}_5$, set $\xi_l = o_p(n_0^{-1/2}) > 0$,

$$\begin{aligned}\mathbb{P}\left(\min_{j\in\mathcal{A}^c}\left|\tilde{\delta}_j\right| > \xi_l\ l_{\mathcal{A}^c}\right) &\geq \mathbb{P}\left[\left(\min_{j\in\mathcal{I}_2}\left|\tilde{\delta}_j\right| > \xi_l\ l_{\mathcal{A}^c}\right)\bigcup\left(\min_{j\in\mathcal{I}_3}\left|\tilde{\delta}_j\right| > \xi_l\ l_{\mathcal{A}^c}\right)\bigcup\left(\min_{j\in\mathcal{I}_5}\left|\tilde{\delta}_j\right| > \xi_l\ l_{\mathcal{A}^c}\right)\right] \\ &\geq \mathbb{P}\left(\min_{j\in\mathcal{I}_m}\left|\tilde{\delta}_j\right| > \xi_l\ l_{\mathcal{A}^c}\right), \quad m \in \{2,3,5\} \\ &\to 1\end{aligned}$$

□

A.2. Non-Asymptotic Analysis for $q = 1$

Define the RSC condition for $p \times p$ covariance matrix $\widehat{\Sigma}^{(0)}$ and $\widehat{\Sigma}^{(1)}$ if for any $v \in \mathbb{R}^p$,

$$v^\top\widehat{\Sigma}^{(0)}v \geq \frac{1}{4}v^\top\Sigma^{(0)}v - c_0\frac{\|v\|_1^2\log p}{n_0}, \quad v^\top\widehat{\Sigma}^{(1)}v \geq \frac{1}{4}v^\top\Sigma^{(1)}v - c_1\frac{\|v\|_1^2\log p}{n_1}.$$

Lemma 2 Under *(A1),(A2)*, and *(A5)*, with probability at least $1 - \exp(-c_1 n_0)$,

$$\hat{u}^\top\widehat{\Sigma}^{(1)}\hat{u} \vee \|\hat{u}\|_2^2 \lesssim s\lambda_{n_1}^2 + \lambda_{n_1}C_\Sigma h, \quad \|\hat{u}\|_1 \lesssim s\lambda_{n_1} + C_\Sigma h.$$

where $\hat{u} \equiv \hat{\beta}^{(1)} - \beta^{(1)}$, $C_\Sigma = 1 + \max_{j\leq p}\|e_j^\top\left(\Sigma^{(1)} - \Sigma^{(0)}\right)\left(\Sigma^{(1)}\right)^{-1}\|_1$, e_j is with j-th element being 1 while keeping other elements zero.

Proof. Related proof could be found in Li et al. (2022) appendix A. □

Theorem 3 *(convergence rate)* Under *(A1),(A2)*, and *(A5)*, take $\lambda_{n_1} = c_1\sqrt{\mathbb{E}[(y_i^{(1)})^2]\log p/n_1}$ and $\lambda_{n_0} = c_0\sqrt{\log p/n_0}$ for large c_1, c_0, with $C_\Sigma h \lesssim s\sqrt{\log p/n_0}$ and $n_0 \lesssim n_1$. If $s\log p/n_1 + C_\Sigma h\,(\log p/n_0)^{1/2} = o(1)$, then

$$\begin{aligned}\inf_{B\in\Theta_1(s,h)}\mathbb{P}\left(\frac{1}{n_0}\left\|X^{(0)}(\hat{\beta}^{(0)} - \beta^{(0)})\right\|_2^2 \vee \|\hat{\beta}^{(0)} - \beta^{(0)}\|_2^2 \lesssim \frac{s\log p}{n_0+n_1} + \frac{s\log p}{n_0} \wedge \lambda_{n_0}\|w\|_\infty C_\Sigma h \wedge \bar{w}^2C_\Sigma^2h^2\right) \\ \geq 1 - \exp(-c_2\log p),\end{aligned}$$

where $\bar{w} \equiv \frac{\|w\|_\infty}{\min_j\{w_j - 1/\kappa\}} > 0$ for κ being a large positive constant.

Proof. Denote $\hat{v} \equiv \hat{\delta} - \delta$, $\hat{u} \equiv \hat{\beta}^{(1)} - \beta^{(1)}$, the following oracle inequality holds as for **Algorithm 1**:

$$\frac{1}{2n_0}\|X^{(0)}\hat{v}\|_2^2 \leq \lambda_{n_0}\langle w, |\delta|\rangle - \lambda_{n_0}\langle w, |\hat{\delta}|\rangle + \frac{1}{n_0}|\langle X^{(0)}\hat{v}, \epsilon^{(0)} - X^{(0)}\hat{u}\rangle|,$$

where $w \equiv |\tilde{\delta}|^{\gamma_0} \in \mathbb{R}^p$. If $\frac{1}{n_0}\|{X^{(0)}}^\top\epsilon^{(0)}\|_\infty \leq \lambda_{n_0}/\kappa$, since $|ab| \leq a^2 + b^2/4$,

$$\frac{1}{n_0}|\langle X^{(0)}\hat{v}, \epsilon^{(0)} - X^{(0)}\hat{u}\rangle| \leq \frac{\lambda_{n_0}}{\kappa}\langle e, |\hat{v}|\rangle + \frac{1}{n_0}\|X^{(0)}\hat{u}\|_2^2 + \frac{1}{4n_0}\|X^{(0)}\hat{v}\|_2^2$$

where $e \equiv (1, 1, ..., 1)^\top \in \mathbb{R}^p$. Then by back-substitution,

$$\begin{aligned}\frac{1}{4n_0}\|X^{(0)}\hat{v}\|_2^2 &\leq \lambda_{n_0}\langle w, |\delta|\rangle - \lambda_{n_0}\langle w, |\hat{\delta}|\rangle + \frac{\lambda_{n_0}}{\kappa}\langle e, |\hat{v}|\rangle + \frac{1}{n_0}\|X^{(0)}\hat{u}\|_2^2 \\ &\leq \lambda_{n_0}\langle w, |\delta|\rangle - \left(\lambda_{n_0}\langle w, |\hat{\delta} - \delta|\rangle - \lambda_{n_0}\langle w, |\delta|\rangle\right) + \frac{\lambda_{n_0}}{\kappa}\langle e, |\hat{v}|\rangle + \frac{1}{n_0}\|X^{(0)}\hat{u}\|_2^2 \\ &= 2\lambda_{n_0}\langle w, |\delta|\rangle - \lambda_{n_0}\left\langle w - \frac{e}{\kappa}, |\hat{v}|\right\rangle + \frac{1}{n_0}\|X^{(0)}\hat{u}\|_2^2 \\ &\leq 2\lambda_{n_0}\|w\|_\infty\|\delta\|_1 - \lambda_{n_0}\min_{j=1,\ldots,p}\left\{w_j - \frac{1}{\kappa}\right\}\|\hat{v}\|_1 + \frac{1}{n_0}\|X^{(0)}\hat{u}\|_2^2.\end{aligned}$$

(i) If $\frac{1}{n_0}\|X^{(0)}\hat{u}\|_2^2 \le 2\lambda_{n_0}\|w\|_\infty\|\delta\|_1$, then

$$\lambda_{n_0}\min_{j=1,\dots,p}\left\{w_j - \frac{1}{\kappa}\right\}\|\hat{v}\|_1 \le 4\lambda_{n_0}\|w\|_\infty\|\delta\|_1.$$

Denote $\bar{w} \equiv \frac{\|w\|_\infty}{\min_j\{w_j - 1/\kappa\}}$, when $\min_j\{w_j - 1/\kappa\} > 0$, we have

$$\|\hat{v}\|_2 \le \|\hat{v}\|_1 \le 4\bar{w}\|\delta\|_1 \le 4\bar{w}C_\Sigma h;$$

Also the bound for

$$\frac{1}{n_0}\|X^{(0)}\hat{v}\|_2^2 \le 16\lambda_{n_0}\|w\|_\infty\|\delta\|_1 \le 16\lambda_{n_0}\|w\|_\infty C_\Sigma h.$$

Thus,

$$\frac{1}{n_0}\|X^{(0)}\hat{v}\|_2^2 \vee \|\hat{v}\|_2^2 \lesssim \bar{w}^2C_\Sigma^2h^2 \wedge \lambda_{n_0}\|w\|_\infty C_\Sigma h.$$

(ii) If $2\lambda_{n_0}\|w\|_\infty\|\delta\|_1 \le \frac{1}{n_0}\|X^{(0)}\hat{u}\|_2^2$, then

$$\lambda_{n_0}\min_{j=1,\dots,p}\left\{w_j - \frac{1}{\kappa}\right\}\|\hat{v}\|_1 \le \frac{2}{n_0}\|X^{(0)}\hat{u}\|_2^2.$$

When $\min_j\{w_j - 1/\kappa\} > 0$,

$$\|\hat{v}\|_1 \le \frac{2\bar{w}}{n_0\lambda_{n_0}\|w\|_\infty}\|X^{(0)}\hat{u}\|_2^2, \quad \frac{1}{n_0}\|X^{(0)}\hat{v}\|_2^2 \le \frac{8}{n_0}\|X^{(0)}\hat{u}\|_2^2.$$

If $\frac{1}{n_0}\|X^{(0)}\hat{u}\|_2^2 \le 2\Lambda_{\max}(\Sigma^{(0)})\|\hat{u}\|_2^2$, by RSC condition and **Lemma 2**,

$$\|\hat{v}\|_1 \lesssim \frac{2\bar{w}}{\lambda_{n_0}\|w\|_\infty}\left(s\lambda_{n_1}^2 + \lambda_{n_1}C_\Sigma h\right), \quad \frac{1}{n_0}\|X^{(0)}\hat{v}\|_2^2 \lesssim s\lambda_{n_1}^2 + \lambda_{n_1}C_\Sigma h.$$

Choose $\lambda_{n_1} = c_1\sqrt{\mathbb{E}[(y_i^{(1)})^2]\log p/n_1}$ and $\lambda_{n_0} = c_0\sqrt{\log p/n_0}$ for large c_1, c_0, then

$$\frac{1}{n_0}\|X^{(0)}\hat{v}\|_2^2 \vee \|\hat{v}\|_2^2 \lesssim s\frac{\log p}{n_0+n_1} + C_\Sigma h\sqrt{\frac{\log p}{n_0}}.$$

With $C_\Sigma h \lesssim s\sqrt{\log p/n_0}$, combine case (i) and (ii), we have

$$\frac{1}{n_0}\|X^{(0)}\hat{v}\|_2^2 \vee \|\hat{v}\|_2^2 \lesssim s\frac{\log p}{n_0+n_1} + s\sqrt{\frac{\log p}{n_0}} \wedge \lambda_{n_0}\|w\|_\infty C_\Sigma h \wedge \bar{w}^2C_\Sigma^2h^2$$

under probability

$$\mathbb{P}\left\{\frac{1}{n_0}\|X^{(0)^\top}\epsilon^{(0)}\|_\infty \le \frac{\lambda_\delta}{\kappa},\ \frac{1}{n_0}\|X^{(0)}\hat{u}\|_2^2 \le 2\Lambda_{\max}(\Sigma^{(0)})\|\hat{u}\|_2^2,\ \text{RSC holds for } \Sigma^{(0)} \text{ and } \Sigma^{(1)}\right\} \tag{2}$$

for $s\log p/n_1 + C_\Sigma h\left(\log p/n_0\right)^{1/2} = o(1)$.

Next, to show that (2) goes to 1, the sub-Gaussian $x_i^{(0)}\epsilon_i^{(0)}$ tells that

$$\mathbb{P}\left(\frac{1}{n_0}\|X^{(0)^\top}\epsilon^{(0)}\|_\infty \ge c_1\sqrt{\frac{\log p}{n_0}}\right) \le \exp\left\{-c_2\log p\right\}$$

for large c_2. On the other hand, by independence between $X^{(0)}$ and $\hat{u}$, conditioning on $\hat{u}$ to have

$$\mathbb{P}\left(\frac{1}{n_0}\|X^{(0)}\hat{u}\|_2^2 \ge 2\Lambda_{\max}(\Sigma^{(0)})\|\hat{u}\|_2^2\right) \le \exp\left\{-c_1n_0\right\}.$$

Finally,

$$\mathbb{P}\left(\text{RSC holds for } \Sigma^{(0)} \text{ and } \Sigma^{(1)}\right) \ge 1 - \exp\left\{-c_1n_0\right\}.$$

is promised by Theorem 1 in Raskutti et al. (2010). □

Theorem 4 *(minimax lower bound)* Under *(A1),(A2),(A5)*, and *(A6)*, for constant $0 < C_{01} < 1$, if

$\max\{s\log p/\left(n_{\mathcal{A}}+n_0\right), h\sqrt{\log p/n_0}\} = o(1)$, $1 \le s \le \frac{2}{3}p$ for large p, then

$$\inf_{\hat{\beta}^{(0)}} \sup_{\Theta_1(s,h)} \mathbb{P}\left(\|\hat{\beta}^{(0)} - \beta^{(0)}\|_2^2 \gtrsim \frac{s\log p}{n_0+n_1} + \frac{s\log p}{n_0} \wedge \sqrt{\log p/n_0}\|w\|_\infty C_\Sigma h \wedge \bar{w}^2 C_\Sigma^2 h^2\right) \ge C_{01}.$$

Proof. Through discretization of the target sparse parameter space $\mathcal{B}(s)$ of $\beta^{(0)}$ into $\mathcal{B}(\varepsilon, s, \|\cdot\|_2)$, the ε-packing with l_2 metric, the minimax problem is bounded by

$$\inf_{\hat{\beta}^{(0)}} \sup_{\beta^{(0)}\in\mathcal{B}(s)} \mathbb{P}\left(\|\hat{\beta}^{(0)} - \beta^{(0)}\|_2 \ge \varepsilon\right) \ge \inf_{\hat{\beta}^{(0)}} \sup_{\beta^{(0)}\in\mathcal{B}(\varepsilon,s,\|\cdot\|_2)} \mathbb{P}\left(\|\hat{\beta}^{(0)} - \beta^{(0)}\|_2 \ge \varepsilon\right).$$

Consider the testing of $\hat{\phi} = j$ if $\hat{\beta}^{(0)}$ is closest to unique $\beta^{(0)}(j)$ within the packing set under l_2 metric, then the transformation into a multiple testing issue provided the bound of

$$\inf_{\hat{\beta}^{(0)}} \sup_{\beta^{(0)}\in\mathcal{B}(\varepsilon,s,\|\cdot\|_2)} \mathbb{P}\left(\|\hat{\beta}^{(0)} - \beta^{(0)}\|_2 \ge \varepsilon\right) \ge \max_{1\le j\le|\mathcal{B}(\varepsilon,s,\|\cdot\|_2)|} \mathbb{P}_j(\hat{\phi}\ne j).$$

And employing Fano's method, the minimax lower bound is controlled by both the metric entropy of ε-packing as well as the maximum KL-divergence among all $|\mathcal{B}(\varepsilon, s, \|\cdot\|_2)|$ distributions

$$\inf_{\hat{\beta}^{(0)}} \sup_{\beta^{(0)}\in\mathcal{B}(s)} \mathbb{P}\left(\|\hat{\beta}^{(0)} - \beta^{(0)}\|_2 \ge \varepsilon\right) \ge 1 - \frac{\rho + \log 2}{\log|\mathcal{B}(\varepsilon, s, \|\cdot\|_2)|},$$

where $\rho \equiv \max_{1\le i,j\le|\mathcal{B}(\varepsilon,s,\|\cdot\|_2)|} KL(f_i, f_j)$. What's following analyzes different cases of dominating terms as in **Theorem 3**.

(i) When $\frac{s\log p}{n_0+n_1} \ge \sqrt{\log p/n_0}\|w\|_\infty C_\Sigma h \wedge \bar{w}^2 C_\Sigma^2 h^2$, let $h = 0$, here I define a packing of $\mathcal{B}(s)$ similar as Raskutti et al. (2011):

$$\mathcal{Z}(s) = \{z \in \{-1, 0, 1\}^p : \|z\|_0 = s\}.$$

Thus existing $\widetilde{\mathcal{Z}}(s) \subset \mathcal{Z}(s)$ such that $\forall z \ne z' \in \widetilde{\mathcal{Z}}(s)$, $\|z - z'\|_0 \ge s/2$ and $|\widetilde{\mathcal{Z}}(s)| \ge \exp\left(\frac{s}{2}\log\frac{p-s}{s/2}\right)$, for well-defined cardinality, here requires $1 \le s \le \frac{2}{3}p$. Then under l_2 metric, $\varepsilon\sqrt{2/s}\widetilde{\mathcal{Z}}(s)$ is a ε-packing of the target sparse parameter space $\mathcal{B}(s)$ of $\beta^{(0)}$. Also by *(A2)*, $\Lambda_{\max}(\Sigma^{(k)}_{\mathcal{B}^c}) < M_2 < \infty$, assume $\sigma^{(1)} = \sigma^{(0)} = \sigma$ and let $\varepsilon = c_0\sqrt{s\log p/(n_0+n_1)}$ for sufficiently small number c_0, then

$$KL\left(\beta^{(0)}(i), \beta^{(0)}(j)\right) = \frac{\mathbb{E}\left[\sum_{k\in\{0,1\}}\left\|X^{(k)}\left(\beta^{(0)}(i) - \beta^{(0)}(j)\right)\right\|_2^2\right]}{2\sigma^2} \le \frac{2(n_0+n_1)M_2 s\varepsilon^2}{\sigma^2} = \frac{2c_0^2 M_2 s^2\log p}{\sigma^2}.$$

Finally,

$$\begin{aligned}\inf_{\hat{\beta}^{(0)}} \sup_{\Theta_1(s,h)} \mathbb{P}\left(\|\hat{\beta}^{(0)} - \beta^{(0)}\|_2 \ge \varepsilon\right) &\ge \inf_{\hat{\beta}^{(0)}} \sup_{\beta^{(0)}\in\varepsilon\sqrt{\frac{2}{s}}\widetilde{\mathcal{Z}}(s), h=0} \mathbb{P}\left(\|\hat{\beta}^{(0)} - \beta^{(0)}\|_2 \ge c_0\sqrt{\frac{s\log p}{n_0+n_1}}\right)\\ &\ge 1 - \frac{2c_0^2 M_2 s^2\log p/\sigma^2 + \log 2}{s/2\log(2(p-s)/s)} \ge C_{01}\end{aligned}$$

for sufficiently small c_0 and large p, where C_{01} is a positive constant smaller than 1.

(ii) When $\frac{s\log p}{n_0+n_1} \le s\log p/n_0 \wedge \sqrt{\log p/n_0}\|w\|_\infty C_\Sigma h \wedge \bar{w}^2 C_\Sigma^2 h^2$, consider three possible dominating terms separately:

(ii-1) If $s\log p/n_0 \le \sqrt{\log p/n_0}\|w\|_\infty C_\Sigma h$ and $s\log p/n_0 \le \bar{w}^2 C_\Sigma^2 h^2$, i.e.,

$$h \ge \left(s \vee \sqrt{s}\min_j\{w_j - 1/\kappa\}\right)\frac{\sqrt{\log p/n_0}}{C_\Sigma\|w\|_\infty}$$

by expanding $\bar{w} \equiv \frac{\|w\|_\infty}{\min_j\{w_j - 1/\kappa\}}$. Under least dissimilar assumption $s \ge \min_j\{w_j - 1/\kappa\}^2$, $s\sqrt{\log p/n_0} \lesssim h$. Thus, consider the packing of $\left\{\beta^{(0)} \in \varepsilon\sqrt{\frac{2}{s}}\widetilde{Z}(s), \beta^{(1)} = 0\right\}$ with $\varepsilon = c_0\sqrt{s\log p/n_0}$, the inclusion condition holds:

$$\|\delta\|_1 = \|\beta^{(0)}\|_1 \le \sqrt{2}c_0 s\sqrt{\log p/n_0} \lesssim h.$$

Finally,

$$\inf_{\hat{\beta}^{(0)}} \sup_{\Theta_1(s,h)} \mathbb{P}\left(\|\hat{\beta}^{(0)} - \beta^{(0)}\|_2 \geq \varepsilon\right) \geq \inf_{\hat{\beta}^{(0)}} \sup_{\beta^{(0)} \in \varepsilon\sqrt{\frac{2}{s}}\widetilde{Z}(s), \beta^{(1)}=0} \mathbb{P}\left(\|\hat{\beta}^{(0)} - \beta^{(0)}\|_2 \geq c_0\sqrt{\frac{s\log p}{n_0}}\right)$$
$$\geq 1 - \frac{2c_0^2(1+n_1/n_0)M_2 s^2 \log p/\sigma^2 + \log 2}{s/2\log(2(p-s)/s)} \geq C_{01}$$

for sufficiently small c_0 and large p.

(ii-2) If $\sqrt{\log p/n_0}\|w\|_\infty C_\Sigma h \leq s\log p/n_0$ and $\sqrt{\log p/n_0}\|w\|_\infty C_\Sigma h \leq \bar{w}^2 C_\Sigma^2 h^2$, i.e.

$$\frac{\sqrt{\log p/n_0}\min_j\{w_j - 1/\kappa\}^2}{C_\Sigma\|w\|_\infty} \leq h \leq \frac{s\sqrt{\log p/n_0}}{C_\Sigma\|w\|_\infty}$$

under least dissimilar assumption. Consider the packing of $\left\{\beta^{(0)} \in \varepsilon\sqrt{\frac{2}{m}}\widetilde{Z}(m), \beta^{(1)} = 0\right\}$ with $\varepsilon = c_0(\log p/n_0)^{1/4}\sqrt{\|w\|_\infty C_\Sigma h}$, where c_0 is small enough. Then the inclusion condition holds:

$$\|\delta\|_1 = \|\beta^{(0)}\|_1 \leq \varepsilon\sqrt{2m} \lesssim \frac{\sqrt{\log p/n_0}\min_j\{w_j - 1/\kappa\}^2}{C_\Sigma\|w\|_\infty} \leq h$$

when choosing the sparsity of the subset $\widetilde{Z}(m)$ to be $m = \sqrt{\log p/n_0}/h$. Finally,

$$\inf_{\hat{\beta}^{(0)}} \sup_{\Theta_1(s,h)} \mathbb{P}\left(\|\hat{\beta}^{(0)} - \beta^{(0)}\|_2 \geq \varepsilon\right) \geq \inf_{\hat{\beta}^{(0)}} \sup_{\beta^{(0)} \in \varepsilon\sqrt{\frac{2}{m}}\widetilde{Z}(m), \beta^{(1)}=0} \mathbb{P}\left(\|\hat{\beta}^{(0)} - \beta^{(0)}\|_2 \geq c_0\left(\frac{\log p}{n_0}\right)^{\frac{1}{4}}\sqrt{\|w\|_\infty C_\Sigma h}\right)$$
$$\geq 1 - \frac{2c_0^2\sqrt{\log p/n_0}\|w\|_\infty C_\Sigma h(n_0+n_1)M_2 s/\sigma^2 + \log 2}{s/2\log(2(p-s)/s)} \geq C_{01}$$

for sufficiently small c_0 and large p.

(ii-3) If $\bar{w}^2 C_\Sigma^2 h^2 \leq s\log p/n_0$ and $\bar{w}^2 C_\Sigma^2 h^2 \leq \sqrt{\log p/n_0}\|w\|_\infty C_\Sigma h$. Consider the packing of $\left\{\beta^{(0)} \in \varepsilon\sqrt{2}\widetilde{Z}(1), \beta^{(1)} = 0\right\}$ with $\varepsilon = c_0\bar{w}C_\Sigma h$, where c_0 is small. Then the inclusion condition holds:

$$\|\delta\|_1 = \|\beta^{(0)}\|_1 \leq \sqrt{2}\varepsilon \lesssim h$$

Finally,

$$\inf_{\hat{\beta}^{(0)}} \sup_{\Theta_1(s,h)} \mathbb{P}\left(\|\hat{\beta}^{(0)} - \beta^{(0)}\|_2 \geq \varepsilon\right) \geq \inf_{\hat{\beta}^{(0)}} \sup_{\beta^{(0)} \in \varepsilon\sqrt{2}\widetilde{Z}(1), \beta^{(1)}=0} \mathbb{P}\left(\|\hat{\beta}^{(0)} - \beta^{(0)}\|_2 \geq c_0\bar{w}C_\Sigma h\right)$$
$$\geq 1 - \frac{2c_0^2\bar{w}^2 C_\Sigma^2 h^2(n_0+n_1)M_2 s^2/\sigma^2 + \log 2}{s/2\log(2(p-s)/s)} \geq C_{01}$$

for sufficiently small c_0 and large p. □

Corollary 2 Based on oracle $\check{\beta}^{(0)}$ and $\check{\beta}^{(1)}$, $\tilde{\delta}_j - \delta_j = O_p(1/\sqrt{n_0})$.

Proof. Remind $\delta \equiv \beta^{(0)} - \beta^{(1)}$, by oracle property of $\check{\beta}^{(0)}$ and $\check{\beta}^{(1)}$ based on the target and source sample respectively,

$$\mathbb{P}\left(\check{\beta}^{(0)}_{\mathcal{B}} = \beta^{(0)}_{\mathcal{B}}\right) \to 1, \quad \sqrt{n_0}\tau_{n_0}^{-1}\alpha_{n_0}^\top\left(\check{\beta}^{(0)}_{\mathcal{B}^c} - \beta^{(0)}_{\mathcal{B}^c}\right) \xrightarrow{d} N(0,1),$$

$$\mathbb{P}\left(\check{\beta}^{(1)}_{\mathcal{B}} = \beta^{(1)}_{\mathcal{B}}\right) \to 1, \quad \sqrt{n_1}\tau_{n_1}^{-1}\alpha_{n_1}^\top\left(\check{\beta}^{(1)}_{\mathcal{B}^c} - \beta^{(1)}_{\mathcal{B}^c}\right) \xrightarrow{d} N(0,1).$$

Then follow the similar argument as in proof of **Corollary 1** under partition of $\{1,...,p\} = \mathcal{K}_1 \cup \mathcal{K}_2 \cup \mathcal{K}_3 \cup \mathcal{K}_4 \equiv \{\mathcal{B}^{(0)} \cap \mathcal{B}^{(1)}\} \cup \{\mathcal{B}^{(0)} \cap \mathcal{B}^{(1)^c}\} \cup \{\mathcal{B}^{(0)^c} \cap \mathcal{B}^{(1)}\} \cup \{\mathcal{B}^{(0)^c} \cap \mathcal{B}^{(1)^c}\}$, for each j, we have

$$\tilde{\delta}_j - \delta_j = O_p(1/\sqrt{n_0}).$$

□

B More Simulation Results

$s/p = 0.01, h = 0.01, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.2246	0.2014	0.1572	0.1461	0.1330
Oracle Trans-Lasso	0.1144	0.0814	0.0760	0.0665	0.0549
Trans-Lasso	0.1216	0.1058	0.1162	0.1032	0.0989
H-Ada-Trans	0.0831	0.0683	0.0425	0.0404	0.0340
S-Ada-Trans	0.0718	0.0510	0.0294	0.0149	0.0072
$s/p = 0.01, h = 0.1, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.2332	0.1884	0.1735	0.1447	0.1393
Oracle Trans-Lasso	0.1089	0.0869	0.0867	0.0663	0.0710
Trans-Lasso	0.1358	0.1083	0.1269	0.1016	0.1035
H-Ada-Trans	0.0819	0.0626	0.0491	0.0442	0.0420
S-Ada-Trans	0.0724	0.0514	0.0317	0.0242	0.0212
$s/p = 0.01, h = 1, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.2499	0.1988	0.1817	0.1495	0.1230
Oracle Trans-Lasso	0.1340	0.1246	0.0933	0.0813	0.0772
Trans-Lasso	0.1777	0.1624	0.1574	0.1320	0.1241
H-Ada-Trans	0.1882	0.1224	0.0548	0.0304	0.0252
S-Ada-Trans	0.1387	0.1305	0.1088	0.1061	0.1001
$s/p = 0.01, h = 10, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.2591	0.1965	0.1803	0.1476	0.1378
Oracle Trans-Lasso	0.5046	0.3193	0.1579	0.0949	0.0749
Trans-Lasso	0.5147	0.2953	0.1647	0.1356	0.1303
H-Ada-Trans	0.1153	0.0573	0.0356	0.0307	0.0213
S-Ada-Trans	0.1481	0.1209	0.1002	0.0949	0.0872
$s/p = 0.01, h = 100, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.2346	0.2006	0.1673	0.1501	0.1354
Oracle Trans-Lasso	3.5167	1.6343	0.8952	0.4238	0.1709
Trans-Lasso	1.0057	0.7696	0.3513	0.2322	0.1524
H-Ada-Trans	0.1654	0.1407	0.1407	0.1383	0.1356
S-Ada-Trans	0.2353	0.2097	0.2063	0.2052	0.2048

$s/p = 0.025, h = 0.01, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.2722	0.1675	0.1436	0.1145	0.1117
Oracle Trans-Lasso	0.0668	0.0496	0.0326	0.0260	0.0218
Trans-Lasso	0.0880	0.0626	0.0496	0.0408	0.0500
H-Ada-Trans	0.0636	0.0432	0.0303	0.0212	0.0163
S-Ada-Trans	0.0552	0.0369	0.0206	0.0121	0.0060
$s/p = 0.025, h = 0.1, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.2913	0.1673	0.1428	0.1229	0.1058
Oracle Trans-Lasso	0.0663	0.0450	0.0398	0.0296	0.0297
Trans-Lasso	0.0830	0.0618	0.0527	0.0524	0.0493
H-Ada-Trans	0.0643	0.0476	0.0333	0.0264	0.0229
S-Ada-Trans	0.0554	0.0371	0.0245	0.0171	0.0147
$s/p = 0.025, h = 1, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.2432	0.1627	0.1302	0.1197	0.1035
Oracle Trans-Lasso	0.1020	0.0864	0.0622	0.0572	0.0454
Trans-Lasso	0.1245	0.1158	0.1101	0.1029	0.0827
H-Ada-Trans	0.1442	0.1369	0.1243	0.1183	0.1145
S-Ada-Trans	0.1070	0.0966	0.0852	0.0829	0.0729

$s/p = 0.025, h = 10, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.2518	0.1787	0.1334	0.1233	0.1008
Oracle Trans-Lasso	0.4821	0.2816	0.1301	0.0666	0.0505
Trans-Lasso	0.5401	0.2934	0.1609	0.0990	0.0914
H-Ada-Trans	0.1541	0.1362	0.1358	0.1310	0.1231
S-Ada-Trans	0.1321	0.1004	0.0937	0.0882	0.0841
$s/p = 0.025, h = 100, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.2474	0.1909	0.1435	0.1233	0.1100
Oracle Trans-Lasso	7.5572	3.8182	1.2964	0.4932	0.1591
Trans-Lasso	0.7780	0.6873	0.5393	0.2795	0.1339
H-Ada-Trans	0.1044	0.0886	0.0878	0.0868	0.0862
S-Ada-Trans	0.1506	0.1424	0.1418	0.1395	0.1387

$s/p = 0.05, h = 0.01, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.2952	0.1775	0.1264	0.1041	0.0968
Oracle Trans-Lasso	0.0446	0.0322	0.0199	0.0142	0.0144
Trans-Lasso	0.0625	0.0455	0.0336	0.0248	0.0354
H-Ada-Trans	0.0496	0.0356	0.0234	0.0155	0.0134
S-Ada-Trans	0.0442	0.0281	0.0173	0.0092	0.0046
$s/p = 0.05, h = 0.1, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.3626	0.1576	0.1463	0.1116	0.0830
Oracle Trans-Lasso	0.0486	0.0302	0.0216	0.0158	0.0138
Trans-Lasso	0.0664	0.0475	0.0320	0.0291	0.0261
H-Ada-Trans	0.0533	0.0348	0.0246	0.0169	0.0164
S-Ada-Trans	0.0484	0.0306	0.0178	0.0115	0.0098
$s/p = 0.05, h = 1, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.2862	0.2075	0.1357	0.1039	0.0864
Oracle Trans-Lasso	0.0796	0.0712	0.0491	0.0396	0.0362
Trans-Lasso	0.0927	0.0887	0.0828	0.0745	0.0680
H-Ada-Trans	0.1015	0.0980	0.0916	0.0876	0.0844
S-Ada-Trans	0.0841	0.0745	0.0643	0.0563	0.0545
$s/p = 0.05, h = 10, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.3449	0.1967	0.1308	0.1091	0.0986
Oracle Trans-Lasso	0.4529	0.2755	0.1019	0.0583	0.0392
Trans-Lasso	0.5060	0.3186	0.1443	0.0907	0.0660
H-Ada-Trans	0.1048	0.1021	0.0959	0.0936	0.0891
S-Ada-Trans	0.0947	0.0750	0.0701	0.0630	0.0608
$s/p = 0.05, h = 100, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.3289	0.1797	0.1351	0.1129	0.0923
Oracle Trans-Lasso	10.4190	5.3750	1.9320	0.5903	0.1661
Trans-Lasso	0.6716	0.5570	0.5137	0.3718	0.1501
H-Ada-Trans	0.0805	0.0669	0.0632	0.0612	0.0606
S-Ada-Trans	0.1105	0.1020	0.0991	0.0989	0.0982

$s/p = 0.1, h = 0.01, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.3169	0.2152	0.1328	0.0951	0.0770
Oracle Trans-Lasso	0.0368	0.0243	0.0145	0.0082	0.0051
Trans-Lasso	0.0499	0.0353	0.0244	0.0187	0.0174
H-Ada-Trans	0.0440	0.0287	0.0181	0.0109	0.0091
S-Ada-Trans	0.0376	0.0246	0.0133	0.0074	0.0038
$s/p = 0.1, h = 0.1, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.3238	0.2341	0.1446	0.0988	0.0866
Oracle Trans-Lasso	0.0353	0.0240	0.0137	0.0079	0.0076
Trans-Lasso	0.0488	0.0349	0.0233	0.0180	0.0181
H-Ada-Trans	0.0405	0.0283	0.0168	0.0130	0.0130
S-Ada-Trans	0.0371	0.0253	0.0128	0.0086	0.0073

$s/p = 0.1, h = 1, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.3565	0.1993	0.1123	0.0997	0.0865
Oracle Trans-Lasso	0.0573	0.0547	0.0424	0.0310	0.0294
Trans-Lasso	0.0659	0.0650	0.0634	0.0592	0.0553
H-Ada-Trans	0.0693	0.0678	0.0661	0.0625	0.0602
S-Ada-Trans	0.0624	0.0552	0.0506	0.0442	0.0407
$s/p = 0.1, h = 10, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.3213	0.1983	0.1400	0.0888	0.0774
Oracle Trans-Lasso	0.3907	0.2420	0.0823	0.0435	0.0335
Trans-Lasso	0.4377	0.2700	0.1375	0.0720	0.0598
H-Ada-Trans	0.0680	0.0640	0.0638	0.0620	0.0620
S-Ada-Trans	0.0656	0.0601	0.0497	0.0468	0.0451
$s/p = 0.1, h = 100, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.3026	0.2043	0.1376	0.0914	0.0771
Oracle Trans-Lasso	7.2099	6.7800	2.4940	0.7445	0.1807
Trans-Lasso	0.4904	0.4502	0.4112	0.3702	0.1580
H-Ada-Trans	0.0647	0.0490	0.0444	0.0434	0.0423
S-Ada-Trans	0.0881	0.0730	0.0715	0.0690	0.0689

$s/p = 0.25, h = 0.01, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.2657	0.1935	0.1318	0.1005	0.0657
Oracle Trans-Lasso	0.0258	0.0171	0.0092	0.0048	0.0024
Trans-Lasso	0.0348	0.0244	0.0157	0.0109	0.0088
H-Ada-Trans	0.0318	0.0212	0.0129	0.0073	0.0061
S-Ada-Trans	0.0282	0.0182	0.0096	0.0052	0.0028
$s/p = 0.25, h = 0.1, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.2741	0.1933	0.1431	0.1077	0.0729
Oracle Trans-Lasso	0.0234	0.0154	0.0091	0.0055	0.0049
Trans-Lasso	0.0318	0.0232	0.0162	0.0124	0.0099
H-Ada-Trans	0.0295	0.0188	0.0126	0.0090	0.0069
S-Ada-Trans	0.0257	0.0170	0.0098	0.0059	0.0043
$s/p = 0.25, h = 1, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.2849	0.2060	0.1226	0.0976	0.0685
Oracle Trans-Lasso	0.0373	0.0344	0.0317	0.0249	0.0210
Trans-Lasso	0.0424	0.0416	0.0412	0.0411	0.0380
H-Ada-Trans	0.0468	0.0421	0.0407	0.0404	0.0393
S-Ada-Trans	0.0418	0.0370	0.0328	0.0294	0.0258
$s/p = 0.25, h = 10, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.2715	0.2193	0.1378	0.0940	0.0732
Oracle Trans-Lasso	0.2657	0.1889	0.0657	0.0316	0.0232
Trans-Lasso	0.3241	0.2216	0.1265	0.0699	0.0433
H-Ada-Trans	0.0513	0.0442	0.0413	0.0412	0.0394
S-Ada-Trans	0.0494	0.0351	0.0337	0.0333	0.0296
$s/p = 0.25, h = 100, (n_0, n_1)$	(20, 100)	(25, 200)	(30, 600)	(35, 2000)	(40, 8000)
Non-Trans	0.2777	0.2204	0.1471	0.0983	0.0714
Oracle Trans-Lasso	3.8271	4.1686	3.3208	1.0792	0.2358
Trans-Lasso	0.3331	0.3131	0.2962	0.2579	0.1692
H-Ada-Trans	0.0456	0.0323	0.0270	0.0268	0.0265
S-Ada-Trans	0.0639	0.0495	0.0442	0.0440	0.0439