On Transfer Learning in Functional Linear Regression

Haotian Lin and Matthew Reimherr

Department of Statistics
Pennsylvania State University

Abstract

This work studies the problem of transfer learning under the functional linear model framework, which aims to improve the fit of the target model by leveraging the knowledge from related source models. We measure the relatedness between target and source models using Reproducing Kernel Hilbert Spaces, allowing the type of knowledge being transferred to be interpreted by the structure of the spaces. Two algorithms are proposed: one transfers knowledge when the index of transferable sources is known, while the other one utilizes aggregation to achieve knowledge transfer without prior information about the sources. Furthermore, we establish the optimal convergence rates for excess risk, making the statistical gain via transfer learning mathematically provable. The effectiveness of the proposed algorithms is demonstrated on synthetic data as well as real financial data.

1 Introduction

Machine learning models have been widely and successfully applied in many practical applications. However, their applications to some real-world scenarios might be poorly performing with limited available training data due to collection expenses or other constraints. Under these scenarios, *transfer learning* [Torrey and Shavlik, 2010], which transfers knowledge from related source tasks to enhance the learning of the target task, is an appealing mechanism to solve the above problem. However, even though transfer learning has been gaining increased attention in statistical learning, most works focus on classical scalar statistical models like classification and regression and how transfer learning can make impacts in *functional data* is still unclarified.

Transfer learning usually can be broken down into two subproblems. First, one needs to come up with some criteria to quantify the relatedness among target and source tasks. Intuitively, a high relatedness promises performance improvement, while it can be harmful if the opposite holds, a phenomenon called *negative transfer* [Torrey and Shavlik, 2010]. Second, one needs to design the transfer procedure, i.e. how the knowledge gets transferred. A well designed procedure should be able to identify the positive transfer sources and enlarge their impact while reducing or even avoiding the negative transfer. However, even with so many transfer learning methods available, only a few of them provide interpretability of the transfer process, i.e. what type of knowledge is being transferred, and theoretical guarantees, i.e. a mathematical provable gain in prediction or estimation error.

1.1 Contributions

In this work, we provide the first interpretable transfer learning framework for *functional linear regression* (FLR) with theoretical guarantees on excess risk. Our contributions can be summarized as follows:

1. We propose two novel multi-source transfer learning algorithms for FLR. The first one implements knowledge transfer when those positive transfer sources are given; while in the second one, we aggregate multiple models to achieve knowledge transfer without the prior information about relatedness among target and source tasks. We also establish the excess risk of the algorithms and show they are rate-optimal.

2. We use the *Reproducing Kernel Hilbert Space* (RKHS) norm of the contrast of the target and source coefficient functions as a measure of relatedness between models. By using different RKHS, we can measure the relatedness in different aspects or even multiple aspects simultaneously, which indeed provides generous freedom to incorporate structural information or requirements into the transfer learning process, e.g., smoothness or periodicity.

3. We adopt sparse-aggregation [Gaîffas and Lecué, 2011] to aggregate multiple FLR models which are fitted from different combinations of the target and source datasets. Compared to existing approaches like Q-aggregation [Li et al., 2020] or weighted ensemble [Gao et al., 2008; Yao and Doretto, 2010], sparse-aggregation makes the final model less susceptible to negative transfer sources.

The rest of the paper is is organized as follows. In Section 2, we introduce the background of functional linear regression, formularize the transfer learning problem. In Section 3, two transfer learning algorithms are proposed to deal with the known transferable sources and unknown ones. Section 4 provides the theoretical analysis on the two proposed algorithms and shows the excess risk of the algorithms are rate-optimal. Section **??** provides the numerical illustrate of the effectiveness of the proposed algorithms on extensive synthetic data setting. We also apply the algorithms to stock market real dataset.

1.2 Related Literature

Numerous machine learning approaches have been proposed to achieve knowledge transfer, which can be divided into self-taught learning, multi-task learning, domain adaptation, and covariance shift based on different conditions on target and source distribution, see [Pan and Yang, 2009]. These approaches have been applied to various applications, ranging from medical research, recommendation systems, and natural language processing [Daumé III, 2009; Pan and Yang, 2013; Turki et al., 2017]. Below we review the work that is directly relevant to ours, and refer the interested reader to the [Zhuang et al., 2020] for a comprehensive review.

[Bastani, 2018] first studied using intermediate estimates in one source domain as proxies to enhance estimation and prediction on the target domain in high-dimensional linear models. However, they only consider one positive transferable source domain scenario and the sample size of the source domain is greater than the target domain. Recently, [Li et al., 2020] developed two-step transfer learning algorithms to achieve knowledge transfer high-dimensional linear model under more general scenarios, i.e. multiple source domains with some of them can be negatively transferable, by constructing multiple candidate estimators from source domains and aggregating these estimators to alleviate the negative transfer effect. They proved the proposed algorithms are rate-optimal in l_2-estimation error bound under regularized conditions. In [Tian and Feng, 2021], the algorithms get modified and extended to the high-dimensional generalized linear model. Some succeeding works like [Li et al., 2021, 2022] used a similar two-step procedure to extend transfer learning to large-scale Gaussian graphical model and Federated learning, and [Cai and Wei, 2021; Reeve et al., 2021] studied the transfer learning under the non-parametric classification setting. In functional data context, [Ma et al., 2022] studied the covariance shift problem in non-parametric regression, which is a specific form of transfer learning. [Zhu et al., 2021] studies the problem of domain adaptation between two separable Hilbert spaces by proposing algorithms to estimate the optimal transport mapping between two spaces. However, domain adaptation is limited given responses are available in the source domain.

Apart from transfer learning, some works explored transfer/multi-task learning under parameter sharing regression setting, i.e. the target and source models share common parameters or latent structures. For example, [Chen et al., 2015] studied the "data enrich model" for linear regression. [Tripuraneni et al., 2020; Duan and Wang, 2022] studied multi-task learning problems by applying l_1 penalties to all the contrasts between different tasks.

2 Preliminaries and Notation

In this section, we provide the background on functional Linear regression and formulate the transfer learning problem.

2.1 Functional Linear Regression

Consider the following functional linear model:

$$Y = \alpha + \int_{\mathcal{T}} X(t)\beta(t)dt + \epsilon \tag{1}$$

where Y is a scalar response, $X : \mathcal{T} \to \mathbb{R}$ and $\beta : \mathcal{T} \to \mathbb{R}$ are the square integrable functional predictor and coefficient over a compact domain $\mathcal{T} \subset \mathbb{R}$, and ϵ is a random noise with zero mean.

A classical approach to estimate β is to boil down the problem to classical linear regression by expanding the X and β under the same finite basis, like the Fourier basis or the eigenbasis of the covariance function of X [Cardot et al., 1999; Yao et al., 2005; Hall and Hosseini-Nasab, 2006; Hall and Horowitz, 2007]. Moreover, work from [Yuan and Cai, 2010; Cai and Yuan, 2012] proposed that one can obtain a smooth estimator, $\hat{\beta}$, by restricting it to an RKHS.

This approach has been widely used in other functional models like functional generalized linear model, cox-model etc. [Cheng and Shang, 2015; Qu et al., 2016; Reimherr et al., 2018; Sun et al., 2018].

2.2 Problem setup and Transferability

We now formally set the stage for the transfer learning problem in FLR. Consider the following series of models,

$$M_l : Y_i^{(l)} = \alpha^{(l)} + \int_{\mathcal{T}} X_i^{(l)}(t)\beta^{(l)}(t)dt + \epsilon_i^{(l)} = \eta_l(X_i^{(l)}) + \epsilon_i^{(l)}, \quad i = 1, \cdots, n_l, \; l = 0, 1, \cdots, L, \tag{2}$$

where $l = 0$ denotes the target model and $l \in \{1, 2, \cdots, L\}$ denotes source models. Estimating the slope function, $\beta^{(0)}$, is of primary interest. To measure the relatedness between target and source models, for $l = 0, 1, \cdots, L$, we assume $\beta^{(l)} \in \mathcal{H}(K)$, where $\mathcal{H}(K)$ is an RKHS with reproducing kernel K, and define the l-th contrast function $\delta^{(l)} = \beta^{(0)} - \beta^{(l)}$. Given a constant $h \geq 0$, we say the l-th source model is "h-transferable" if $\|\delta^{(l)}\|_K \leq h$. The magnitude of h characterizes the relatedness between the target model and source models. We also define $\mathcal{S}_h = \{1 \leq l \leq L : \|\delta^{(l)}\|_K \leq h\}$ as a subset of $\{1, 2, \cdots, L\}$, which consists of the indexes of all "h-transferable" source models.

Quantifying the relatedness by using the RKHS norm provides interpretability of the transfer process since the type of knowledge being transferred is tied to the structural information of the used RKHS. Besides, since different kernels endow $\mathcal{H}(K)$ with different structures, one can interpret the relatedness and the type of structural information that is being transferred in different aspects, and thus transfer different types of structural information from sources by picking different K. For example, one is able to transfer the structural information about continuity or smoothness by picking K to be a Sobolev kernel, and about periodicity by picking periodic kernels. Moreover, one is able to transfer structural information in multiple aspects simultaneously by measuring $\delta^{(l)}$ in the intersection of all these RKHSs, which still turns out to be a valid RKHS with its reproducing kernel is determined in quadratic form, see Theorem 2.2.3 in [Okutmuştur, 2005].

2.3 Notation

WLOG, we assume all functional elements' domain is over a compact set $T \subset \mathbb{R}$. For $l = 0, 1, \cdots, L$, we denote the covariance function of $X^{(l)}(\cdot)$ as $C^{(l)}(s,t) = \mathrm{E}[X^{(l)}(s) - \mathrm{E}\, X^{(l)}(s)][X^{(l)}(t) - \mathrm{E}\, X^{(l)}(t)]$ for $s, t \in T$. For a real, symmetric, square-integrable, and nonnegative kernels, $\Gamma : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$, define its integral operator as

$$L_\Gamma(f) = \int_T \Gamma(\cdot, t) f(t) dt \quad \textit{for} \quad f \in L^2.$$

Therefore, for a given reproducing kernel K, we can define a new nonnegative definite kernel $T^{(l)} = K^{\frac{1}{2}} C^{(l)} K^{\frac{1}{2}}$ and its corresponding integral operator,

$$L_{T^{(l)}}(f) := L_{K^{\frac{1}{2}} C^{(l)} K^{\frac{1}{2}}}(f) = L_{K^{\frac{1}{2}}}(L_{C^{(0)}}(L_{K^{\frac{1}{2}}}(f)))$$

Based on the Karhunen–Loève theorem, we denote $\{s_j^{(l)}\}$ and $\{\phi_j^{(l)}\}$ as the eigenvalues and eigenfunctions respectively. Let $a_n \asymp b_n$ and $a_n \lesssim b_n (a_n = O(b_n))$ denote $|a_n/b_n| \to c$ and $|a_n/b_n| \le c$ for some constant c when $n \to \infty$. Let $a_n = O_P(b_n)$ denote $\sup_n P(|a_n/b_n| \le c) \to 1$ as $c \to \infty$. We abbreviate $\mathcal{S}_h$ as $\mathcal{S}$ to generally represent h-transferable sources index without specific h. For a set A, let $|A|$ denote its cardinality and A^c denote its complement.

3 Transfer Methodology

In this section, we first propose the *$\mathcal{S}$-Known Transfer Learning* to transfer knowledge with a known $\mathcal{S}$ and then propose the *Sparse-Aggregation Transfer Learning* to deal with a unknown $\mathcal{S}$.

3.1 $\mathcal{S}$-Known Transfer Learning

Given a known $\mathcal{S}$, the main idea to transfer information from these known source datasets is one can first obtain a "rough" estimate by pooling target and source datasets, and then adapt the "rough" estimate to the target dataset. As the relatedness is quantified in the RKHS, we embrace the estimation scheme from Yuan and Cai [2010]; Cai and Yuan [2012] by using the roughness regularization method Ramsay and Silverman [2008] to restrict the estimator to be within $\mathcal{H}(K)$. This approach has been proved to hit the optimal rate, and we refer it as *Optimal Functional Linear Regression* (OFLR).

Algorithm 1 $\mathcal{S}$-Known Transfer Learning (SKTL)

Input: Target and source datasets $\{\{X_i^{(l)}, Y_i^{(l)}\}_{i=1}^{n_l}\}_{l=0}^{L}$; index set of source datasets $\mathcal{S}$.
Transfer Step: Compute

$$\hat{\beta}_{\mathcal{S}} = \underset{\beta \in \mathcal{H}(K)}{\operatorname{argmin}} \left\{ \frac{1}{(n_{\mathcal{S}} + n_0)} \sum_{l \in \{0\} \cup \mathcal{S}} \sum_{i=1}^{n_l} \left(Y_i^{(l)} - \int_{\mathcal{T}} X_i^{(l)}(t) \beta(t) dt \right)^2 + \lambda_1 \|\beta\|_{\mathcal{H}(K)}^2 \right\} \tag{3}$$

Debias Step: Compute

$$\hat{\delta}_{\mathcal{S}} = \underset{\delta \in \mathcal{H}(K)}{\operatorname{argmin}} \left\{ \frac{1}{n_0} \sum_{i=1}^{n_0} \left(Y_i^{(0)} - \int_{\mathcal{T}} X_i^{(0)}(t) (\hat{\beta}_{\mathcal{S}}(t) + \delta(t)) dt \right)^2 + \lambda_2 \|\delta\|_{\mathcal{H}(K)}^2 \right\} \tag{4}$$

Return $\hat{\beta} = \hat{\beta}_{\mathcal{S}} + \hat{\delta}_{\mathcal{S}}$ and $\hat{\alpha} = \overline{Y}^{(0)} - \int_{\mathcal{T}} \overline{X}^{(0)}(t) \hat{\beta}(t) dt$ where $\overline{Y}^{(0)}$, $\overline{X}^{(0)}$ are the sample mean of target dataset.

The algorithm first fits OFLR with all the target and source datasets, and then fits OFLR for the contrast with the target dataset only. The transfer step is pooling all of the source data and target data to get a single estimate, where all the structural information about $\beta^{(0)}$ is gathered in $\hat{\beta}_{\mathcal{S}}$. The debias step is necessary since the probabilistic limit of the transfer step, $\beta_{\mathcal{S}}$, is not consistent with $\beta^{(0)}$. Therefore, the debias step corrects the bias $\beta^{(0)} - \beta_{\mathcal{S}}$ using the target dataset. The transfer learning boosts the target model in the sense that the excess risk converges fast as the sample size in the transfer step is much larger than n_0. We provide the explicit error bounds for excess risk and their explanation in Section 4.

3.2 Sparse-Aggregation Transfer Learning

Assuming the index set $\mathcal{S}$ is known in Algorithm 1 can be unrealistic in practice without prior information or investigation. Moreover, as some source datasets might have little or even negative contribution to the target one, it could be practically harmful to directly apply Algorithm 1 by assuming all sources belong to $\mathcal{S}$. Thus, one should reduce the negative transfer sources' influence by excluding them from $\mathcal{S}$. From this perspective, work from Tian and Feng [2021] proposed an algorithm to detect those positive transfer sources. However, their algorithm requires data splitting and lacks the debias step during detection, which will lead to potential problems like selection bias, lower convergence and false positive error i.e. including false positive transfer source. Li *et al.*Li et al. [2020], on the other hand, didn't directly estimate $\mathcal{S}$, but obtained multiple weak models from all sources and then aggregated them with different weights, which is more robust.

Our proposed algorithm is also motivated by aggregating multiple models. The main idea to transfer information with the unknown $\mathcal{S}$ is that we first construct a collection of candidates for $\mathcal{S}$, named $\{\hat{\mathcal{S}}_1, \hat{\mathcal{S}}_2, \cdots, \hat{\mathcal{S}}_J\}$, such that there exists at least one $\hat{\mathcal{S}}_j$ satisfying $\hat{\mathcal{S}}_j = \mathcal{S}$ with high probability and then obtain their corresponding candidate estimators $\mathcal{F} = \{\hat{\beta}(\hat{\mathcal{S}}_1), \hat{\beta}(\hat{\mathcal{S}}_2), \cdots, \hat{\beta}(\hat{\mathcal{S}}_J)\}$ via SKTL. Then we aggregate the candidate estimators in $\mathcal{F}$ such that the aggregated estimator $\hat{\beta}_a$ satisfies the following oracle inequality

$$R(\hat{\beta}_a) \leq \min_{\beta \in \mathcal{F}} R(\beta) + r(\mathcal{F}, n), \quad \text{where} \quad R(f) = \mathrm{E}_{(X,Y)}[l(Y, f(X))|D_n] \tag{5}$$

with a high probability or in expectation, where $l(\cdot,\cdot)$ is a loss function and D_n are all the target and source datasets. Under some conditions, the "aggregation cost", $r(\mathcal{F}, n)$, is not a higher order than $\min_{\beta\in\mathcal{F}} R(\beta)$. Therefore, with the condition of $P(\hat{\mathcal{S}}_j = \mathcal{S}) \to 1$, the aggregated estimator $\hat{\beta}_a$ is supposed to have the same upper error bound as assuming we know the ground truth $\mathcal{S}$ and then applying SKTL. Li *et al.* Li et al. [2020] considered high-dimensional linear regression by using Q-aggregation to establish inequality (5) in estimation error with high probabilities and $r(\mathcal{F}, n)$ of order n^{-1}.

Usually, the aggregated model is represented as a convex combination of elements in $\mathcal{F}$, i.e. $\hat{\beta}_a = \sum_{j=1}^{J} c_j \hat{\beta}(\hat{\mathcal{S}}_j)$. Designing an aggregation procedure such that it fulfills inequality (5) has been well studied in the past two decades Tsybakov [2003]; Audibert [2007]; Gaîffas and Lecué [2011]; Dai et al. [2012]; Lecué and Rigollet [2014]. Despite the fact that there are numerous aggregation procedures available, such as aggregation with cumulated exponential weights (ACEW) Juditsky et al. [2008]; Audibert [2009], aggregation with exponential weights (AEW) Leung and Barron [2006]; Dalalyan and Tsybakov [2007] and Q-aggregation Dai et al. [2012], none of them sets the c_j to 0, meaning the negative transfer sources still strongly affect $\hat{\beta}_a$. Besides, it can also be computationally inefficient if total source number is large. To address these issues, we introduce sparsity into $\{c_j\}$ by adopting Sparse-aggregation Gaîffas and Lecué [2011], which will set some c_j to 0. Although in ACEW and AEW, the coefficients can be very small by perfectly tuning the temperature parameters, the tuning process is non-adaptive, whereas sparse-aggregation doesn't require turning, and our empirical results show that sparse aggregation lower bounds ACEW and AEW. The procedure above is summarized in Algorithm 2, though we note that other aggregation procedures can be used in place.

Algorithm 2 Sparse-Aggregation Transfer Learning (SATL)

Inputs: Target and source datasets $\{\{X_i^{(l)}, Y_i^{(l)}\}_{i=1}^{n_l}\}_{l=0}^{L}$; A given integer M.

Construct $\mathcal{F}$:

Step 1: Split the target dataset $\{X_i^{(0)}, Y_i^{(0)}\}_i$ into two equal size sets and let $\mathcal{I}$ be a random subset of $\{1, \cdots, n_0\}$ such that $|\mathcal{I}| = \lfloor \frac{n_0}{2} \rfloor$ and let $\mathcal{I}^c = \{1, \cdots, n_0\} \setminus \mathcal{I}$.

Step 2: Construct $L+1$ candidate set of $\mathcal{S}$, $\{\hat{\mathcal{S}}_0, \hat{\mathcal{S}}_1, \cdots, \hat{\mathcal{S}}_L\}$, by following process:

(1) Calculate $\hat{\beta}_0(\cdot)$ by fitting OFLR on $\{(X_i^{(0)}, Y_i^{(0)})\}_{i\in\mathcal{I}}$ and let $\hat{\mathcal{S}}_0 = \emptyset$.

(2) For each $l = 1, 2, ..., L$, calculate $\hat{\beta}_l(\cdot)$ by fitting OFLR on $\{(X_i^{(l)}, Y_i^{(l)})\}_{i=1}^{n_l}$ and $\hat{\Delta}_l = \|\hat{\beta}_0 - \hat{\beta}_l\|_{K^M}$

(3) For each $l = 1, 2, ..., L$, set
$\hat{\mathcal{S}}_l = \Big\{1 \le k \le L : \hat{\Delta}_k \text{ is among the first } l \text{ smallest of all}\Big\}$.

Step 3: For $l = 1, 2, \cdots, L$, fit Algorithm 1 with $\mathcal{S} = \hat{\mathcal{S}}_l$ and $\{(X_i^{(0)}, Y_i^{(0)})\}_{i\in\mathcal{I}}$. Denote the output as $\hat{\beta}(\hat{\mathcal{S}}_l)$ for $l = 1, 2, \cdots, L$ and set $\mathcal{F} = \{\hat{\beta}(\hat{\mathcal{S}}_0), \hat{\beta}(\hat{\mathcal{S}}_1), \cdots, \hat{\beta}(\hat{\mathcal{S}}_L)\}$.

Sparse-Aggregation:

Step 1: Given a confidence level x, assume either of setting (1) or (2) holds for a constant b.

Setting (1): $\max\{|Y^{(0)}|, \max_{\beta\in\mathcal{H}(K)} |\langle X^{(0)}, \beta\rangle_{L^2}|\} \le b$,

Setting (2): $\max\{\|\epsilon^{(0)}\|_\Psi, \sup_{\beta\in\mathcal{H}(K)} \|\langle X^{(0)}, \beta - \beta^{(0)}\rangle\|\} \le b$,

where $\|\epsilon^{(0)}\|_\Psi := inf\{c > 0 : \mathrm{E}[\exp\{|\epsilon^{(0)}|/c\}] \le 2\}$. Define $\psi = b\sqrt{(log(L)+x)/n_0}$ or $b\sqrt{(log(L)+x)log(n_0))/n_0}$ and $c = 4(1+9b)$ or $(1+b)$ if (1) or (2) holds.

Step 2: Split $\{(X_i^{(0)}, Y_i^{(0)})\}_{i\in\mathcal{I}^c}$ into equal size set, with index set $\mathcal{I}_1^c$ and $\mathcal{I}_2^c$

Step 3: Use $\{(X_i^{(0)}, Y_i^{(0)})\}_{i\in\mathcal{I}_1^c}$ to define a random subset of $\mathcal{F}$ as

$$\mathcal{F}_1 = \left\{\beta \in \mathcal{F} : R_{|\mathcal{I}_1^c|}(\beta) \le R_{|\mathcal{I}_1^c|}(\hat{\beta}_{|\mathcal{I}_1^c|}) + c\max(\psi \left\|\hat{\beta}_{|\mathcal{I}_1^c|} - \beta\right\|_{|\mathcal{I}_1^c|}, \psi^2)\right\}$$

where $\|\beta\|_{|\mathcal{I}|}^2 = 1/|\mathcal{I}| \sum_{i\in\mathcal{I}} \langle X_i^{(0)}, \beta\rangle_{L^2}^2$, $R_{|\mathcal{I}|}(\beta) = 1/|\mathcal{I}| \sum_{i\in\mathcal{I}} (Y_i^{(0)} - \langle X_i^{(0)}, \beta\rangle_{L^2})^2$, $\hat{\beta}_{|\mathcal{I}|} = \underset{\beta\in\mathcal{F}}{argmin} R_{|\mathcal{I}|}(\beta)$.

Step 4: Set $\hat{\mathcal{F}}$ to be one of the following:

- $\mathcal{F}_2 = seg(\mathcal{F}_1) = \{c_1\beta_1 + c_2\beta_2 : \beta_1, \beta_2 \in \mathcal{F}_1 \textit{ and } c_1 + c_2 = 1\}$
- $\mathcal{F}_2 = star(\mathcal{F}_1) = \{c_1\hat{\beta}_{n1} + c_2\beta_2 : \beta_2 \in \mathcal{F}_1 \textit{ and } c_1 + c_2 = 1\}$,

then, return

$$\hat{\beta}_a = \underset{\beta\in\mathcal{F}_2}{argmin} R_{|\mathcal{I}_2^c|}(\beta).$$

Here $\|\cdot\|_{K^M}$ is a truncated version of $\|\cdot\|_K$, which is used to guarantee the identifiability of $\mathcal{S}$, see Theorem 3. Besides, having $\mathcal{F}_2 = star(\mathcal{F}_1)$ is more efficient from a computational perspective, as its time complexity of $O(|\mathcal{F}_1|)$ while for $\mathcal{F}_2 = seg(\mathcal{F}_1)$, its time complexity is $O(|\mathcal{F}_1|^2)$.

4 Theoretical Results

In this section, we study the theoretical properties of the proposed algorithms. We evaluate the prediction accuracy via excess risk, i.e.

$$\mathcal{E}(\hat{\beta}) := \mathrm{E}[Y^{(0)} - \hat{\eta}_0(X^{(0)})]^2 - \mathrm{E}[Y^{(0)} - \eta_0(X^{(0)})]^2.$$

All the proofs of lemmas and theorems are in Appendix.

4.1 Minimax Excess Risk of SKTL

To study the excess risk of SKTL and SATL, we define the parameter space as $\Theta(h) = \{\beta^{(l)} : \max_{l\in\{0\}\cup\mathcal{S}} \|\delta^{(l)}\|_K \leq h\}$. Since the excess risk of FLR heavily depends on the $C^{(l)}$ and K, we now state the assumptions that will be used to establish the convergence rate. We suppose that either of the following two assumptions holds:

Assumption 1. *For all $l \in \mathcal{S}$, $L_{T^{(l)}}$ commutes with $L_{T^{(0)}}$, i.e. $L_{T^{(0)}}L_{T^{(l)}} = L_{T^{(l)}}L_{T^{(0)}}$. Denote $a_j^{(l)} = \langle L_{T^{(l)}}(\phi_j^{(0)}), \phi_j^{(0)}\rangle$ and $a_j^{(l)}$ satisfies $a_j^{(l)} \lesssim s_j^{(0)}$.*

Assumption 2. *For $l \in \mathcal{S}$, the linear operator $\mathbf{I}-(L_{T^{(0)}})^{-1/2}L_{T^{(l)}}(L_{T^{(0)}})^{-1/2}$ is Hilbert–Schmidt.*

In the functional data literature Assumption 1 is more prevalent, whereas in the stochastic processes literature Assumption 2 is more common, though neither one implies each other. For Assumption 1, we note that by assuming $L_{T^{(l)}}$ commutes with $L_{T^{(0)}}$, it immediately shares common eigenspaces spanned by the eigenfunctions of $L_{T^{(0)}}$. While Assumption 2 relaxes the common eigenspace condition by only assuming $L_{T^{(l)}}$ is similar to $L_{T^{(0)}}$ in the eigenspace of $L_{T^{(0)}}$. By further assuming the covariance kernel C and the reproducing kernel K are commutative, Assumption 2 holds if the predictors of target and sources are equivalent Gaussian processes Kuo [1975] which is common in the stochastic processes literature.

Theorem 1 (Upper Bound)**.** *Suppose either Assumption 1 or 2 holds. If $s_j^{(0)} \asymp j^{-2r}$ and $n_0/n_{\mathcal{S}} \nrightarrow 0$, let $\xi(h,\mathcal{S}) = (h/\|\beta_{\mathcal{S}}\|_K)^{\frac{2}{2r+1}}$, then for the output $\hat{\beta}$ of Algorithm 1,*

$$\sup_{\beta^{(0)}\in\Theta(h)} \mathcal{E}\left(\hat{\beta}\right) = O_P\left((n_{\mathcal{S}}+n_0)^{-\frac{2r}{2r+1}} + n_0^{-\frac{2r}{2r+1}}\xi(h,\mathcal{S})\right), \tag{6}$$

if $\lambda_1 \asymp (n_{\mathcal{S}}+n_0)^{-\frac{2r}{2r+1}}$ and $\lambda_2 \asymp n_0^{-\frac{2r}{2r+1}}$ where λ_1 and λ_2 are tuning parameters in Algorithm 1.

Theorem 1 provides the excess risk upper bound of $\hat{\beta}$ when $\beta^{(0)} \in \Theta(h)$. The excess risk consists of the error from the transfer step and the debias step; a faster decay rate of $s_j^{(0)}$ will lead to a faster convergence rate. In the trivial case when $\mathcal{S} = \emptyset$, the upper bound becomes $O_P(n_0^{-2r/(2r+1)})$, which coincides with the upper bound of fitting OFLR with the target dataset only. With $h \ll \|\beta_{\mathcal{S}}\|_K$ and $n_0 \ll n_{\mathcal{S}}$, the excess risk is sharper than the risk of not using any sources, indicating that when the sample size of sources are sufficiently large and the contrast's RKHS norm doesn't exceed $\|\beta_{\mathcal{S}}\|_K$, transfer learning improves learning performance in the target model. As h increases, the second term dominates and will be worse than fitting OFLR with the target dataset only, which is expected as more negative sources get involved.

We note that either assumption implies that the projection coefficient of $L_{T^{(l)}}$ onto j-th eigenspace of $L_{T^{(0)}}$ shouldn't decay slower than $s_j^{(0)}$, meaning that the knowledge transfer only happens when $L_{T^{(l)}}$ is "smoother" than $L_{T^{(0)}}$ in the eigenspaces of $L_{T^{(0)}}$. This can be understood as one can't improve the excess risk by transferring structural information from some "rougher" $\beta^{(l)}$.

Theorem 2 (Lower Bound)**.** *Under the same condition of Theorem 1, for any possible estimator $\tilde{\beta}$ based on $\{(X_i^{(l)}, Y_i^{(l)}) : l \in \{0\} \cup \mathcal{S}, i = 1, \cdots, n_l\}$, the excess risk of $\tilde{\beta}$ satisfies*

$$\lim_{a \to 0} \lim_{n \to \infty} \inf_{\tilde{\beta}} \sup_{\beta^{(0)} \in \Theta(h)} P\left\{\mathcal{E}\left(\tilde{\beta}\right) \geq a\left((n_{\mathcal{S}} + n_0)^{-\frac{2r}{2r+1}} + n_0^{-\frac{2r}{2r+1}} \xi(h, \mathcal{S})\right)\right\} = 1. \tag{7}$$

Combining Theorem 1 and 2, it implies that the estimator from SKTL is rate-optimal in excess risk. The proof of the lower bound is based on considering the optimal convergence rate of two cases: (1) the ideal case where $\beta^{(l)} = \beta^{(0)}$ for all $l \in \mathcal{S}$ and (2) the worst case where $\beta^{(l)} \equiv 0$, meaning no knowledge should be transferred at all.

4.2 Excess Risk of SATL

In this subsection, we study the excess risk for SATL. The following lemma establishes the oracle inequality for sparse-aggregation in Algorithm 2.

Lemma 1. *Let $\hat{\beta}_a$ be the output of Algorithm 2, then*

$$\mathcal{E}\left(\hat{\beta}_a\right) \leq \min_{l=0,1,\cdots,L} \mathcal{E}\left(\hat{\beta}(\hat{\mathcal{S}}_l)\right) + r_x(\mathcal{F}, n) \tag{8}$$

holds with some constants C_{b1}, C_{b2} where

$$r_x(\mathcal{F}, n) = \begin{cases} C_{b1} \dfrac{(1+x) log(L)}{n_0}, & \textit{if setting (1) holds} \\ C_{b2} \dfrac{(1 + \log(4x^{-1}) log(L) log(n_0)}{n_0}. & \textit{if setting (2) holds} \end{cases}$$

Remark 1. *We call the setting (1) bounded setting and (2) sub-exponential setting. The later one is milder but leads to a suboptimal cost. The proof of Lemma 1 can be found in Gaïffas and Lecué [2011].*

The main idea of constructing an aggregate to achieve knowledge transfer relies on the fact that there exists a $\hat{\mathcal{S}}_l$ such that it equals to the groud truth $\mathcal{S}$ (so $\hat{\beta}(\hat{\mathcal{S}}_l) = \hat{\beta}(\mathcal{S})$) with high probability. Thus, to ensure the $\mathcal{F}$ constructed in Algorithm 2 satisfies such a property, we impose an assumption to guarantee the identifiability of $\mathcal{S}$ and thus ensure the existence of such $\hat{\mathcal{S}}_l$.

Assumption 3 (Identifiability of $\mathcal{S}$)**.** *Suppose for any h, there is an integer M such that*

$$\min_{l\in\mathcal{S}^c} \|\beta_0 - \beta_k\|_{K^M} > h,$$

where $\|f\|_{K^M} = \sum_{j=1}^{M} \frac{\langle f, v_j\rangle_{L^2}^2}{\tau_j}$, where τ_j and v_j is the j-th eigenvalue and eigenfunction of K.

We can interpret this assumption as for all $l \in \mathcal{S}^c$, there exists a finite dimensional subspace of $\mathcal{H}(K)$, such that the norm of the projection of the contrast function, $\delta^{(l)}$, on this subspace is already greater than h. This assumption eliminates the existence of $\beta^{(l)}$, for $l \in \mathcal{S}^c$, that live on the boundary of the ball centered at $\beta^{(0)}$ with radius h in $\mathcal{H}(K)$. Under Assumption 3, we now show the $\mathcal{F}$ constructed in Algorithm 2 guarantees the existence of $\hat{\mathcal{S}}_l$.

Theorem 3. *Under Assumption 3, $P(\max_{l\in\mathcal{S}} \hat{\Delta}_l < \min_{l\in\mathcal{S}^c} \hat{\Delta}_l) \to 1$, hence there exists a l s.t.*

$$P\left(\hat{\mathcal{S}}_l = \mathcal{S}\right) \to 1.$$

With Lemma 1 and Theorem 3, we establish the excess risk for SATL.

Theorem 4 (SATL)**.** *Let $\hat{\beta}_a$ be the output estimator function of Algorithm 2, then under the same assumptions of Theorem 1 and Assumption 3,*

$$\sup_{\beta^{(0)}\in\Theta(h)} \mathcal{E}\left(\hat{\beta}_a\right) = O_P\left((n_{\mathcal{S}} + n_0)^{-\frac{2r}{2r+1}} + n_0^{-\frac{2r}{2r+1}}\xi(h,\mathcal{S}) + r(\mathcal{F}, n)\right). \tag{9}$$

where

$$r(\mathcal{F}, n) = \begin{cases} \dfrac{log(L)}{n_0}, & \text{if setting (1) holds} \\ \dfrac{log(L)log(n_0)}{n_0}. & \text{if setting (2) holds} \end{cases}$$

One interesting note is that the excess risk from transfer learning is the classical non-parametric rate and the aggregation cost is parametric (or nearly parametric). Therefore, the aggregation cost is asymptotically negligible. However, in finite dimensional (including high-dimensional) linear regression, the advantage of a negligible aggregation cost disappears as the estimation error or the prediction error is of the order n_0^{-1}, which is also a parametric rate and thus the aggregation cost is not asymptotically negligible. We remind the reader that once the aggregation cost is negligible, SATL is still rate-optimal based on Theorem 2.

Turning back to the excess risk of SATL, we can see that it is sharper than the excess risk of fitting OFLR with the target dataset only for a certain range of $n_{\mathcal{S}}$ and h, which is consistent with Theorem 1. However, once the sources' sample size $n_{\mathcal{S}}$ is sufficiently large such that the first term in (9) no longer dominates the later two, then we can't reduce the excess risk by simply increasing $n_{\mathcal{S}}$ and discerning the aggregation process is more crucial. Under such a scenario, Sparse-aggregation takes advantage of its sparsity property and is less vulnerable than classical aggregation approaches (like ACEW, AEW, Q-aggregation). In Section **??**, we will show that when negative transfer sources account for a larger proportion of all sources, Sparse-aggregation outperforms ACEW and AEW.

5 Simulation

In this section, we conduct the simulations on the proposed transfer learning algorithms and some of their competitors. We specifically consider the following five algorithms, *OFLR*, *SKTL*, *SATL*, *Detection Transfer Learning (Detect-TL)* and *Exponential Weighted Aggregation Transfer Learning (EWATL)*. OFLR fits the model with the target dataset only and serves as the baseline; SKTL and SATL fit models based on Algorithm 1 and 2; Detect-TL implements the detection algorithm proposed in [Tian and Feng, 2021]; EWATL utilizes AEW to aggregated models in Algorithm 2. We pick Star-Aggregation in SATL to boost the computational efficiency.

To set up the RKHS, we consider the setting in [Cai and Yuan, 2012]. Let $\psi_k(t) = \sqrt{2}\cos(\pi k t)$ for $j \geq 1$ and define the reproducing kernel K of $\mathcal{H}(K)$ as

$$K(s,t) = \sum_{k=1}^{\infty} k^{-2}\psi_k(s)\psi_k(t)$$

For the target model, the true coefficient function, $\beta_0(t)$, is set to be (1) $\sum_{k=1}^{\infty} 4\sqrt{2}(-1)^{k-1}k^{-2}$ $\psi_k(t)$; (2) $4\cos(3\pi t)$; (3) $4\cos(3\pi t) + 4\sin(3\pi t)$. For source models,

- if $l \in \mathcal{S}_h$, then $\beta_l(t)$ is set to be $\beta_l(t) = \beta_0(t) + \sum_{k=1}^{\infty}\left(\mathcal{U}_k(\sqrt{12}h/\pi k^2)\right)\psi_k(t)$ with $\mathcal{U}_k$'s i.i.d. uniform random variable on $[-1, 1]$.
- if $l \in \mathcal{S}_h^c$, then $\beta_l(t)$ is generated from a Gaussian process with mean function $\cos(2\pi t)$ with Ornstein–Uhlenbeck kernel $K(s,t) = \exp(-15|s-t|)$ or kernel $K(s,t) = min(s,t)$.

The predictors $X^{(l)}$ are i.i.d. generated from Gaussian process with mean function $\sin(\pi t)$ and covariance kernel $K_{1/2,1}$ for target domain and $K_{3/2,1}$ for source domians, where $K_{\nu,\rho}(s,t)$ is the Matérn kernel. The Matérn kernels with different ν control the smoothness of $X^{(l)}$ since their resulting RKHSs ties to particular Sobolev spaces.

All functions are generated on $[0, 1]$ with 50 evenly spaced points and we set $n_0 = 150$ and $n_l = 100$. For each algorithm, we set the regularization parameters as the optimal values in Theorem 1 by adjusting n accordingly and select c from $\{0.05, 0.1, 0.15, \cdots, 1\}$ using 10-fold cross validation. For each sample size and each algorithm, the excess risk is calculated via Monte-Carlo method by using 1000 new generated predictors $X^{(0)}$.

5.1 Evaluation of Algorithm 1

To evaluate the performance of SKTL, we compare it with OFLR under different $\mathcal{S}_h$. The criteria we use for comparison is the *logarithmic relative excess risk*, i.e.

$$log\left(\frac{\text{the excess risk of SKTL}}{\text{the excess risk of OFLR}}\right)$$

and a ratio smaller than 0 indicates SKTL achieves a smaller excess risk on target domain.

In Figure 1, we report the logarithmic relative excess risk of OFLR to SKTL by varying h from 1 to 40 and $|\mathcal{S}_h|$ from 1 to 15. One can see that with more transferable sources with smaller h involved (right bottom corner), the SKTL has a more significant improvement; while for fewer sources and large h (left top corner), the transfer can be even worse than OFLR. Also, we note that the shape of the yellow area, which indicates a worse peformance on SKTL, is in the shape of triangular, illustrate that the transfer learning effect are determined mutually by the sample size in $\mathcal{S}_h$ and the h.

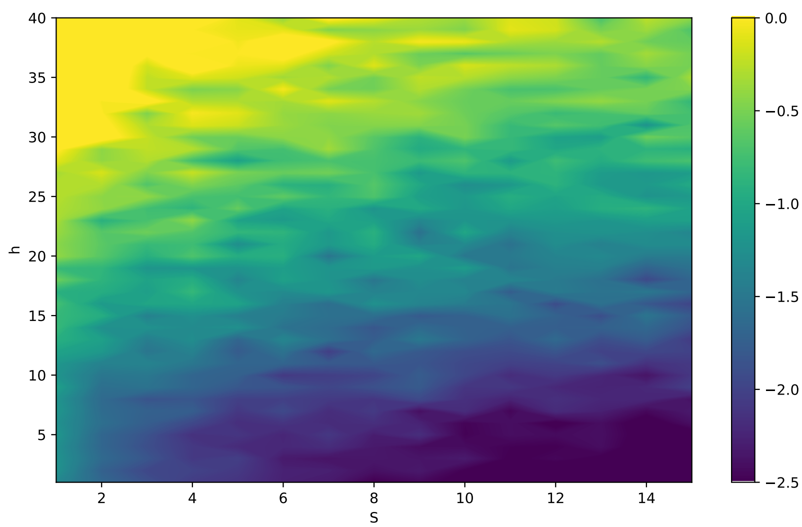


Figure 1: Logarithmic relative excess risk of SKTL to OFLR with different h and $|\mathcal{S}|$.

5.2 Evaluation of Algorithm 2

To evaluate SATL, we compare it with algorithms who deal with unknown $\mathcal{S}$ like Detect-TL, EWATL. We also implement SKTL by setting $\mathcal{S}$ as the true transferable set and OFLR for empirical lower and upper bound comparison. To shuffling the $\mathcal{S}$ with in all source datasets, we let $L = 20$ and $\mathcal{S}$ be a random subset of $\{1, 2, \cdots, L\}$ such that $|\mathcal{S}|$ is equal to $0, 2, \cdots, 20$. We report the excess risks in the right panel of Figure 2. In both panels, one can see the Detect-TL only has considerable reduction on the excess risk with relatively small h, but provides limited improvement with large h, indicating its limitation to transfer knowledge when there is only limited knowledge available in sources. Comparing SATL with EWATL, we can see the gap between the two curves are larger when the proportion of $\mathcal{S}$ is small, showing that EWATL is more sensitive to those $\beta^{(l)}$ with $l \in \mathcal{S}^c$, while SATL

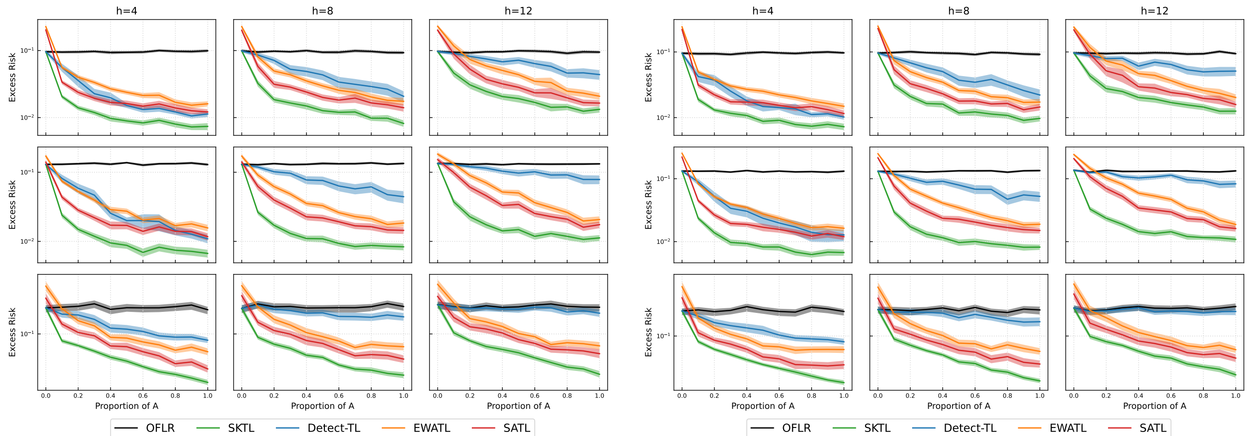


Figure 2: Excess Risk of different transfer learning algorithms. Each row corresponds to a β_0 and the y-axes for each row are under the same scale. The result for each sample size is an average of 100 replicate experiments with the shaded area indicates $\pm$ 2 standard error. Left: $\beta_l(t), l \in \mathcal{S}^c$ are generated from GP with $K(s,t) = exp(-15|s-t|)$. Right: $\beta_l(t), l \in \mathcal{S}^c$ are generated from GP with $K(s,t) = min(s,t)$.

get less affected. Comparing two panels, even non-transferble sources' coefficient function generated from $K(s,t) = min(s,t)$ is rougher than $K(s,t) = exp(-15|s-t|)$, making the non-transferable $\beta^{(l)}$'s trajectories rougher has slight effect on the performance of aggregation based approaches.

We also note that the temperature parameter, T, controls the relative magnitude of the coefficients $\{c_j\}$ in AEW. Previously, we set the temperature to be $T = 1$, and now we set it to be $T = 0.2$ and $T = 10$ to see how it affects the transfer learning performance. The results are reported in Figure 3. While the temperature is low, the convex combination coefficients $\{c_j\}$ will be very small, making EWATL has almost the same performance as SATL (left panel), but it still can't beat SATL. While we set the temperature to be relatively high, the gap between EWATL and SATL increases comparing to $T = 1$ and $T = 0.2$, especially when the proportion of $|\mathcal{S}|$ is small.

6 Real Data Application

In this section, we demonstrate an application of the proposed algorithms in the financial market. The goal of portfolio management is to predict the future stock return, and thus one can rebalance his portfolio. Some people may be interested in predicting the future stock returns in a specific sector, and transfer learning can borrow the market information from other sectors to improve the prediction of the interested one. In this section, for two given adjacent months, we focus on utilizing the Monthly Cumulative Return (MCR) of the early month to predict the Monthly Return (MR) of the later month, and improving the prediction accuracy on a certain sector by transferring market information from other sectors.

In the stock data application, the predictor X is Monthly Cumulative Return (MCR) of the first month and the response Y is the Monthly Return (MR) of the second month

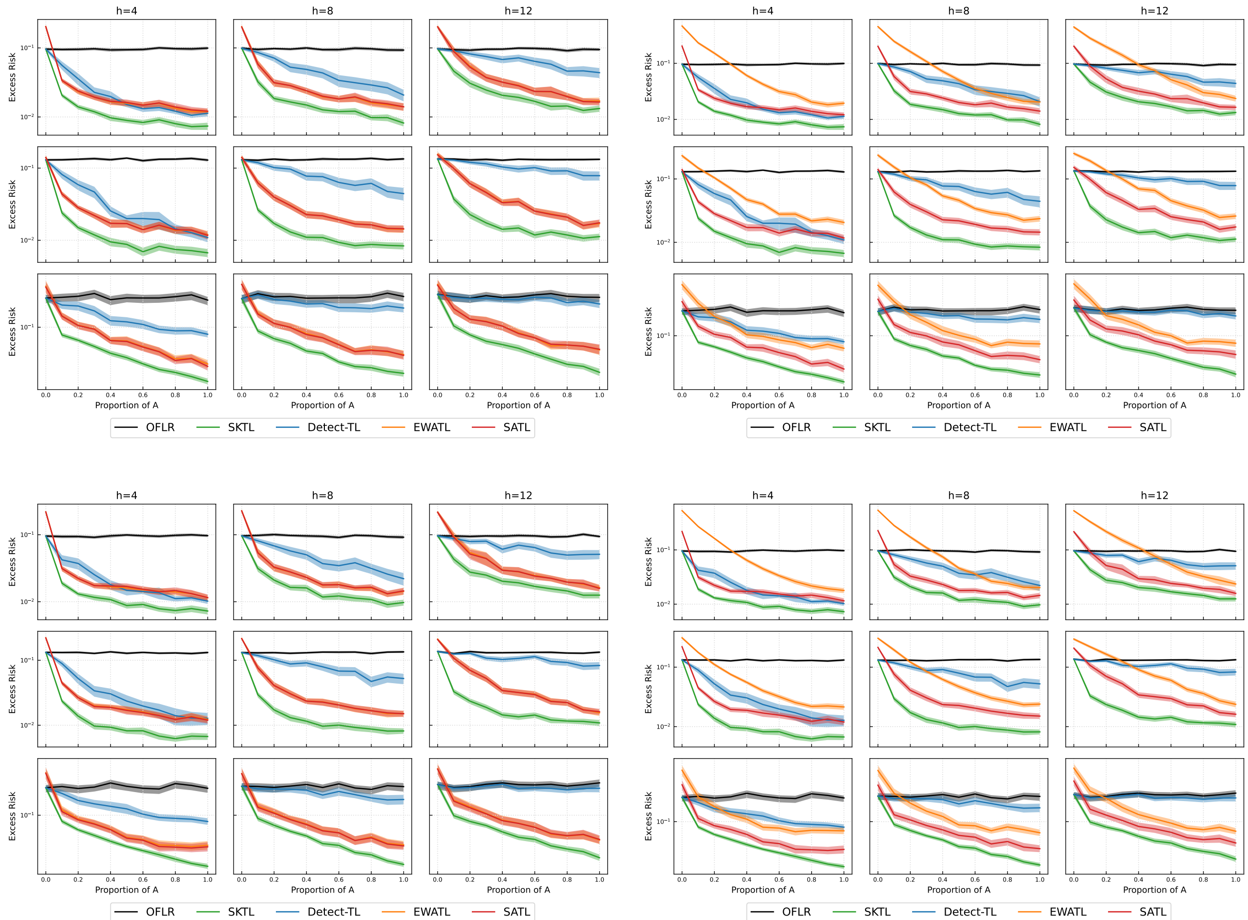


Figure 3: Left: $T = 0.2$ for EWATL. Right: $T = 10$ for EWATL. The upper row corresponds to $\beta^{(l)}$ generate from GP with $K(s,t) = exp(-15|s-t|)$ and the lower row corresponds to $\beta^{(l)}$ generate from GP with $K(s,t) = min(s,t)$.

Kokoszka and Zhang [2012]. Suppose for a specific stock, the daily price for the first month is $\{s^1(t_0), s^1(t_1), \cdots, s^1(t_m)\}$ and for the second month is $\{s^2(t_0), s^2(t_1), \cdots, s^2(t_m)\}$, then the predictors and responds are expressed as

$$X(t) = \frac{s^1(t) - s^1(t_0)}{s^1(t_0)} \quad \text{and} \quad Y = \frac{s^2(m) - s^2(t_0)}{s^2(t_0)}.$$

The stocks price data are collected from Yahoo Finance (https://finance.yahoo.com/) and we force on the stocks whose company has a market cap over 20 Billion. We divide the sectors based on the division criteria on Nasdaq (https://www.nasdaq.com/market-activity/stocks/screener). The dataset consists of total 11 sectors, Basic Industries, Capital Goods, Consumer Durable, Consumer Non-Durable, Consumer Services, Energy, Finance, Health Care, Public Utility, Technology and Transportation with number of stocks as $60, 58, 31, 30, 104, 55, 70, 68, 46, 103, 41$ in each sector. The time period of stocks' price is 05/01/2021 to 09/30/2021.

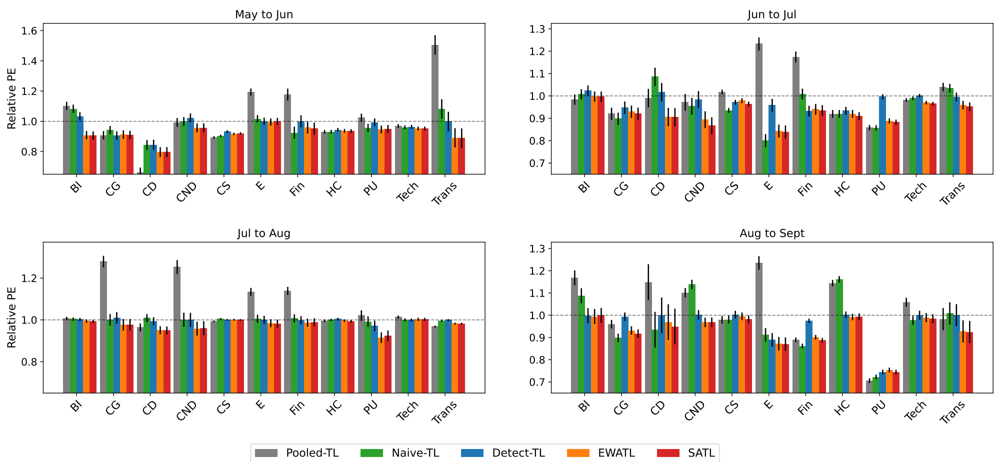
We compare the performance of *Pooled Transfer (Pooled-TL)*, *Naive Transfer (Naive-TL)*, *Detect-TL*, *EWATL* and *SATL*. Naive-TL is the same as SKTL by assuming all source sectors belong to $\mathcal{S}$, while the Pooled-TL one omits the debias step in Naive-TL, and the other three are the same as before. Each sector is treated as the target each time, and all the other sectors are sources. We randomly split the target sector into the train (80%) and test (20%) set and report the ratio of the four approaches' prediction errors to OFLR's on the test set. We consider the Matérn kernel [Cressie and Huang, 1999] as the reproducing kernel K. We set $\rho = 1$ and $\nu = 1/2, 3/2, \infty$ (where $\nu = 1/2$ is equivalent to exponential kernel and $\nu = \infty$ is equivalent to Gaussian kernel) which endows K with different smoothness properties. The tuning parameters are selected via Generalized Cross-Validation(GCV). Again, we repeat the experiment 100 times and report the average prediction error with standard error in Figure 4.

First, we note that the Pooled-TL and Naive-TL only reduce the prediction error in a few sectors, but make no improvement or even downgrade the predictions in most sectors. This implies the target sector still benefits from direct transfer when it shares high similarities with other sectors, while direct transfer regardless of similarity may lead to poor results. Besides, Naive-TL shows an overall better performance compared to the Pooled-TL, demonstrating the importance of the debias step. For Detect-TL, all the ratios are close to 1, showing its limited improvement, which is as expected as it can miss positive transferable sources easily. Finally, both EWATL and SATL provide more robust and significant improvements. We can see both of them have considerable improvement across almost all the sectors, regardless of the similarity between the target sector and source sectors. Comparing the results from different kernels, we can see the improvement patterns are almost the same across all the sectors and adjacent months, showing the proposed algorithms' robustness w.r.t. different reproducing kernel K. Besides, the combination of July and August, there are only a few sectors like CG, CD, CND and PU get improvement.

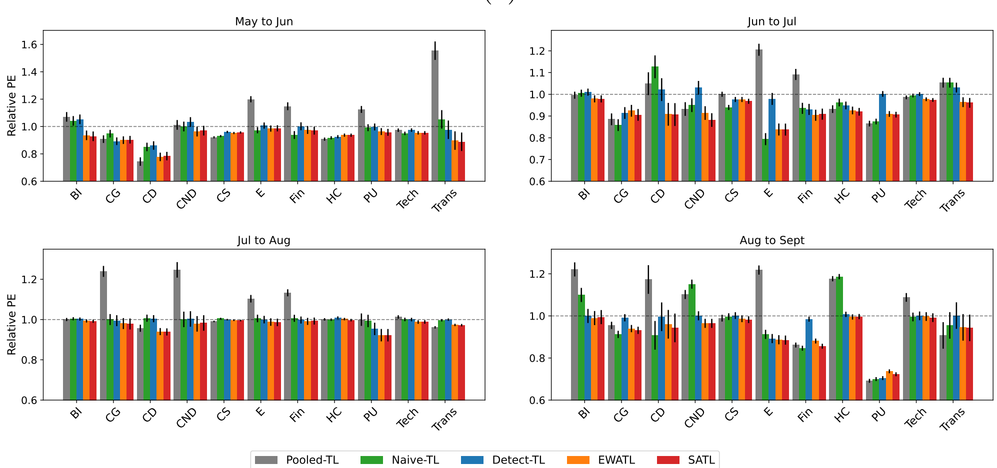
7 Conclusion and Future Work

In this work, we studied transfer learning under functional linear regression. By measuring the relatedness of the different coefficient functions using RKHS, our novel framework provided an interpretable transfer learning process and allows different types of structural information to be transferred. We also proposed two algorithms to achieve knowledge transfer under the functional linear regression to separately deal with the scenario of known transferable sources (SKTL) and unknown ones (SATL), and provided mathematically provable transfer learning gain by establishing the theoretical error bounds for them.

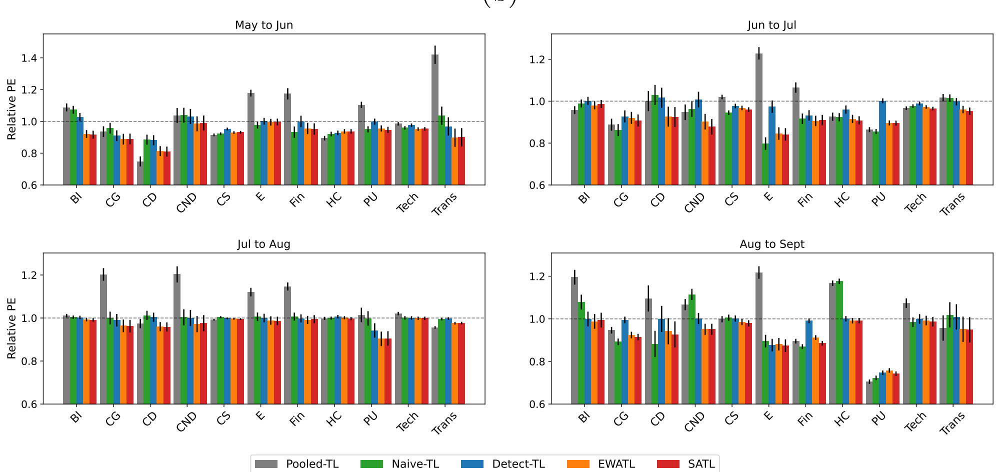
There are several promising directions for future studies. First, even though the same technique in this work can be applied to some more complicated functional linear models like functional GLM, functional Cox-model, etc. by changing L^2-loss to log-likelihood, it is still unclear how to develop transfer learning algorithms in more complex models like functional single index model and even more broad non-parametric regression problems. Another interesting direction is that it will be interesting to investigate how transfer learning will boost the statistical inference for FLR with the existence of source datasets, like increasing the power of hypotheses testing and obtaining a more narrow confidence band for the coefficient function.



(a)



(b)



(c)

Figure 4: Relative prediction error of Pooled-TL, Naive-TL, Detect-Transfer, Q-agg Transfer, and SATL Transfer to OFLR for each target sectors. Each bar is an average of 100 replications, with standard error as black line. (a): $\nu = 1/2$; (b): $\nu = 3/2$; (c): $\nu = \infty$

References

Jean-Yves Audibert. Progressive mixture rules are deviation suboptimal. *Advances in Neural Information Processing Systems*, 20:41–48, 2007.

Jean-Yves Audibert. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591–1646, 2009.

Hamsa Bastani. Predicting with proxies: Transfer learning in high dimension. *arXiv preprint arXiv:1812.11097*, 2018.

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

T Tony Cai and Ming Yuan. Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association*, 107(499):1201–1216, 2012.

Hervé Cardot, Frédéric Ferraty, and Pascal Sarda. Functional linear model. *Statistics & Probability Letters*, 45(1):11–22, 1999.

Aiyou Chen, Art B Owen, and Minghui Shi. Data enriched linear regression. *Electronic journal of statistics*, 9(1):1078–1112, 2015.

Guang Cheng and Zuofeng Shang. Joint asymptotics for semi-nonparametric regression models with partially linear structure. *The Annals of Statistics*, 43(3):1351–1390, 2015.

Noel Cressie and Hsin-Cheng Huang. Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical association*, 94(448):1330–1339, 1999.

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

Arnak S Dalalyan and Alexandre B Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *International Conference on Computational Learning Theory*, pages 97–111. Springer, 2007.

Hal Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.

Yaqi Duan and Kaizheng Wang. Adaptive and robust multi-task learning. *arXiv preprint arXiv:2202.05250*, 2022.

Stéphane Gaïffas and Guillaume Lecué. Hyper-sparse optimal aggregation. *The Journal of Machine Learning Research*, 12:1813–1833, 2011.

Jing Gao, Wei Fan, Jing Jiang, and Jiawei Han. Knowledge transfer via multiple model local structure mapping. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 283–291, 2008.

Peter Hall and Joel L Horowitz. Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1):70–91, 2007.

Peter Hall and Mohammad Hosseini-Nasab. On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68 (1):109–126, 2006.

Anatoli Juditsky, Philippe Rigollet, and Alexandre B Tsybakov. Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206, 2008.

Piotr Kokoszka and Xi Zhang. Functional prediction of intraday cumulative returns. *Statistical Modelling*, 12(4):377, 2012.

Hui-Hsiung Kuo. Gaussian measures in banach spaces. In *Gaussian Measures in Banach Spaces*, pages 1–109. Springer, 1975.

Guillaume Lecué and Philippe Rigollet. Optimal learning with q-aggregation. *The Annals of Statistics*, 42(1):211–224, 2014.

Gilbert Leung and Andrew R Barron. Information theory and mixing least-squares regressions. *IEEE Transactions on information theory*, 52(8):3396–3410, 2006.

Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation, and minimax optimality. *arXiv preprint arXiv:2006.10593*, 2020.

Sai Li, Tianxi Cai, and Rui Duan. Targeting underrepresented populations in precision medicine: A federated transfer learning approach. *arXiv preprint arXiv:2108.12112*, 2021.

Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning in large-scale gaussian graphical models with false discovery rate control. *Journal of the American Statistical Association*, (just-accepted):1–27, 2022.

Cong Ma, Reese Pathak, and Martin J Wainwright. Optimally tackling covariate shift in rkhs-based nonparametric regression. *arXiv preprint arXiv:2205.02986*, 2022.

Baver Okutmuştur. *Reproducing kernel Hilbert spaces*. PhD thesis, bilkent university, 2005.

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

Weike Pan and Qiang Yang. Transfer learning in heterogeneous collaborative filtering domains. *Artificial intelligence*, 197:39–55, 2013.

Simeng Qu, Jane-Ling Wang, and Xiao Wang. Optimal estimation for the functional cox model. *The Annals of Statistics*, 44(4):1708–1738, 2016.

James O Ramsay and Bernhard W Silverman. Functional data analysis. *İnternet Adresi: http*, 2008.

Henry WJ Reeve, Timothy I Cannings, and Richard J Samworth. Adaptive transfer learning. *The Annals of Statistics*, 49(6):3618–3649, 2021.

Matthew Reimherr, Bharath Sriperumbudur, and Bahaeddine Taoufik. Optimal prediction for additive function-on-function regression. *Electronic Journal of Statistics*, 12(2):4571–4601, 2018.

Xiaoxiao Sun, Pang Du, Xiao Wang, and Ping Ma. Optimal penalized function-on-function regression under a reproducing kernel hilbert space framework. *Journal of the American Statistical Association*, 113(524):1601–1611, 2018.

Ye Tian and Yang Feng. Transfer learning under high-dimensional generalized linear models. *arXiv preprint arXiv:2105.14328*, 2021.

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

Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in Neural Information Processing Systems*, 33: 7852–7862, 2020.

Alexandre B Tsybakov. Optimal rates of aggregation. In *Learning theory and kernel machines*, pages 303–313. Springer, 2003.

Turki Turki, Zhi Wei, and Jason TL Wang. Transfer learning approaches to improve drug sensitivity prediction in multiple myeloma patients. *IEEE Access*, 5:7381–7393, 2017.

Rom Rubenovich Varshamov. Estimate of the number of signals in error correcting codes. *Docklady Akad. Nauk, SSSR*, 117:739–741, 1957.

Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6):2873–2903, 2005.

Yi Yao and Gianfranco Doretto. Boosting for transfer learning with multiple sources. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 1855–1862. IEEE, 2010.

Ming Yuan and T Tony Cai. A reproducing kernel hilbert space approach to functional linear regression. *The Annals of Statistics*, 38(6):3412–3444, 2010.

Jiacheng Zhu, Aritra Guha, Mengdi Xu, Yingchen Ma, Rayleigh Lei, Vincenzo Loffredo, XuanLong Nguyen, and Ding Zhao. Functional optimal transport: Mapping estimation and domain adaptation for functional data. 2021.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

A Appendix: Theorems' Proofs

To simplifies the notation, denote $\mathcal{S}^* = \{0\} \cup \mathcal{S}$.

A.1 Proof of Theorem 1

Proof. We first prove the upper bound under Assumption 1. WLOG, we assume the eigenfunction of $L_{T^{(0)}}$ and $L_{T^{(l)}}$ are perfectly aligned, i.e. $\phi_j^{(0)} = \phi_j^{(l)}$ for all $j \in \mathbb{N}$. To simplifies the notation, we denote $\{0\} \cup \mathcal{S}$ as $\mathcal{S}^*$ and keep this notation across all the Proofs.

Since $L_{K^{\frac{1}{2}}}(L^2) = \mathcal{H}(K)$, for any $\beta \in \mathcal{H}(K)$, there exist a $f \in L^2$ such that $\beta = L_{K^{\frac{1}{2}}}(f)$. In following proofs, we denote f_l as $\beta^{(l)}$'s corresponding element in L^2. Therefore, we can rewrite the minimization problem in transfer step and debias step as

$$\hat{f}_{\mathcal{S}} = \underset{f \in L^2}{\operatorname{argmin}} \left\{ \frac{1}{n_{\mathcal{S}} + n_0} \sum_{l \in \mathcal{S}^*} \sum_{i=1}^{n_l} \left(Y_i^{(l)} - \langle X_i^{(l)}, L_{K^{\frac{1}{2}}}(f) \rangle \right)^2 + \lambda_1 \|f\|_{L^2}^2 \right\},$$

and

$$\hat{f}_{\delta} = \underset{f_\delta \in L^2}{\operatorname{argmin}} \left\{ \frac{1}{n_0} \sum_{i=1}^{n_0} \left(Y_i^{(0)} - \langle X_i^{(0)}, L_{K^{\frac{1}{2}}}(\hat{f}_{\mathcal{S}} + f_\delta) \rangle \right)^2 + \lambda_2 \|f_\delta\|_{L^2}^2 \right\}.$$

and thus the excess risk we want to bound can be rewritten as

$$\mathcal{E}(\hat{\beta}) = \left\| (L_{T^{(0)}})^{\frac{1}{2}} (\hat{f} - f_0) \right\|_{L^2}^2 \quad \text{where} \quad \hat{f} = \hat{f}_{\mathcal{S}} + \hat{f}_\delta$$

Define the empirical version of $C^{(l)}$ as

$$C_n^{(l)}(s,t) = \frac{1}{n_l} \sum_{i=1}^{n_l} X_i^{(l)}(s) X_i^{(l)}(t),$$

and $L_{T_n^{(l)}} = L_{K^{\frac{1}{2}}} L_{C_n^{(l)}} L_{K^{\frac{1}{2}}}$.

transfer step:

For transfer step, the solution of minimization is

$$\hat{f}_{\mathcal{S}\lambda_1} = \left(\sum_{l \in \mathcal{S}^*} \alpha_l L_{T_n^{(l)}} + \lambda_1 \mathbf{I} \right)^{-1} \left(\sum_{l \in \mathcal{S}^*} \alpha_l L_{T_n^{(l)}}(f_0^{(l)}) + \sum_{l \in \mathcal{S}^*} g_{ln} \right),$$

where $\mathbf{I}$ is identity operator and

$$g_{ln} = \frac{1}{n_{\mathcal{S}} + n_0} \sum_{i=1}^{n_l} \epsilon_i^{(l)} L_{K^{\frac{1}{2}}}(X_i^{(l)}).$$

Define

$$f_{\mathcal{S}\lambda_1} = \left(\sum_{l \in \mathcal{S}^*} \alpha_l L_{T^{(l)}} + \lambda_1 \mathbf{I} \right)^{-1} \left(\sum_{l \in \mathcal{S}^*} \alpha_l L_{T^{(l)}}(f_0^{(l)}) \right).$$

By triangular inequality

$$\left\|(L_{T^{(0)}})^{\frac{1}{2}}(\hat{f}_{\mathcal{S}\lambda_1}-f_{\mathcal{S}})\right\|_{L^2} \leq \left\|(L_{T^{(0)}})^{\frac{1}{2}}(\hat{f}_{\mathcal{S}\lambda_1}-f_{\mathcal{S}\lambda_1})\right\|_{L^2} + \left\|(L_{T^{(0)}})^{\frac{1}{2}}(f_{\mathcal{S}\lambda_1}-f_{\mathcal{S}})\right\|_{L^2}.$$

By Lemma 2 and taking $v=\frac{1}{2}$, the second term on r.h.s. can be bounded by

$$\left\|(L_{T^{(0)}})^{\frac{1}{2}}(f_{\mathcal{S}\lambda_1}-f_{\mathcal{S}})\right\|_{L^2}^2 = O_P\left((\lambda_1 D_1)\|f_{\mathcal{S}}\|_{L^2}^2\right).$$

Now we turn to the first term. We further introduce an intermedia term

$$\tilde{f}_{\mathcal{S}\lambda_1} = f_{\mathcal{S}\lambda_1} + \left(\sum_{l\in\mathcal{S}^*}\alpha_l L_{T^{(l)}} + \lambda_1\mathbf{I}\right)^{-1}\left(\sum_{l\in\mathcal{S}^*}\alpha_l L_{T_n^{(l)}}(f_{\mathcal{S}}-f_{\mathcal{S}\lambda_1}) + \sum_{l\in\mathcal{S}^*} g_{ln} - \lambda_1 f_{\mathcal{S}\lambda_1}\right).$$

We first bound $\|(L_{T^{(0)}})^{\frac{1}{2}}(f_{\mathcal{S}\lambda_1}-\tilde{f}_{\mathcal{S}\lambda_1})\|_{L^2}^2$. By the fact that

$$\lambda_1 f_{\mathcal{S}\lambda_1} = \sum_{l\in\mathcal{S}^*}\alpha_l L_{T^{(l)}}\left(f_0^{(l)} - f_{\mathcal{S}\lambda_1}\right) = \sum_{l\in\mathcal{S}^*}\alpha_l L_{T^{(l)}}\left(f_{\mathcal{S}}-f_{\mathcal{S}\lambda_1}\right)$$

Therefore,

$$\begin{aligned}
&\|(L_{T^{(0)}})^{\frac{1}{2}}(f_{\mathcal{S}\lambda_1}-\tilde{f}_{\mathcal{S}\lambda_1})\|_{L^2}\\
&\quad\leq \left\|(L_{T^{(0)}})^{\frac{1}{2}}(\sum_{l\in\mathcal{S}^*}\alpha_l L_{T^{(l)}}+\lambda_1\mathbf{I})^{-1}(\sum_{l\in\mathcal{S}^*} g_{ln})\right\|_{L^2} + \\
&\qquad\left\|(L_{T^{(0)}})^{\frac{1}{2}}(\sum_{l\in\mathcal{S}^*}\alpha_l L_{T^{(l)}}+\lambda_1\mathbf{I})^{-1}(\sum_{l\in\mathcal{S}^*}\alpha_l(L_{T_n^{(l)}}-L_{T^{(l)}})(f_{\mathcal{S}}-f_{\mathcal{S}\lambda_1}))\right\|_{L^2}\\
&\quad= \left\{\sum_{j=1}^{\infty}\left(\left\langle (L_{T^{(0)}})^{\frac{1}{2}}(\sum_{l\in\mathcal{S}^*}\alpha_l L_{T^{(l)}}+\lambda_1\mathbf{I})^{-1}\left(\sum_{l\in\mathcal{S}^*} g_{ln}\right), \phi_j\right\rangle_{L^2}\right)^2\right\}^{\frac{1}{2}} + \\
&\qquad\left\{\sum_{j=1}^{\infty}\left(\left\langle (L_{T^{(0)}})^{\frac{1}{2}}(\sum_{l\in\mathcal{S}^*}\alpha_l L_{T^{(l)}}+\lambda_1\mathbf{I})^{-1}\left(\sum_{l\in\mathcal{S}^*}\alpha_l(L_{T_n^{(l)}}-L_{T^{(l)}})(f_{\mathcal{S}}-f_{\mathcal{S}\lambda_1})\right), \phi_j\right\rangle_{L^2}\right)^2\right\}^{\frac{1}{2}}
\end{aligned}$$

For first term in above inequality, by Lemma 4,

$$\left\{\sum_{j=1}^{\infty}\left(\left\langle (L_{T^{(0)}})^{\frac{1}{2}}(\sum_{l\in\mathcal{S}^*}\alpha_l L_{T^{(l)}}+\lambda_1\mathbf{I})^{-1}\left(\sum_{l\in\mathcal{S}^*} g_{ln}\right), \phi_j\right\rangle_{L^2}\right)^2\right\} \lesssim_P (n_{\mathcal{S}}+n_0)^{-1}\lambda_1^{\frac{1}{2r}}.$$

For second one, by Lemma 3 and 5,

$$\sum_{j=1}^{\infty}\left(\left\langle (L_{T^{(0)}})^{\frac{1}{2}}(\sum_{l\in\mathcal{S}^*}\alpha_l L_{T^{(l)}}+\lambda_1\mathbf{I})^{-1}\left(\sum_{l\in\mathcal{S}^*}\alpha_l(L_{T_n^{(l)}}-L_{T^{(l)}})(f_{\mathcal{S}}-f_{\mathcal{S}\lambda_1})\right), \phi_j\right\rangle_{L^2}\right)^2 \lesssim_P (n_{\mathcal{S}}+n_0)^{-1}\lambda_1^{\frac{1}{2r}}.$$

Therefore,

$$\|(L_{T^{(0)}})^{\frac{1}{2}}(f_{\mathcal{S}\lambda_1} - \tilde{f}_{\mathcal{S}\lambda_1})\|_{L^2}^2 \lesssim_P (n_{\mathcal{S}} + n_0)^{-1} \lambda_1^{\frac{1}{2r}}.$$

Finally, we bound $\|(L_{T^{(0)}})^{\frac{1}{2}}(\hat{f}_{\mathcal{S}\lambda_1} - \tilde{f}_{\mathcal{S}\lambda_1})\|_{L^2}^2$. Once again, by the definition of $\tilde{f}_{\mathcal{S}\lambda_1}$

$$\hat{f}_{\mathcal{S}\lambda_1} - \tilde{f}_{\mathcal{S}\lambda_1} = \left(\sum_{l \in \mathcal{S}^*} \alpha_l L_{T^{(l)}} + \lambda_1 \mathbf{I}\right)^{-1} \left(\sum_{l \in \mathcal{S}^*} \alpha_l (L_{T^{(l)}} - L_{T_n^{(l)}})(\hat{f}_{\mathcal{S}\lambda_1} - f_{\mathcal{S}\lambda_1})\right).$$

Thus, by Lemma 3

$$\begin{aligned}
&\left\|(L_{T^{(0)}})^{\frac{1}{2}}(\hat{f}_{\mathcal{S}\lambda_1} - \tilde{f}_{\mathcal{S}\lambda_1})\right\|_{L^2}^2 \\
&\quad \leq \left\|(L_{T^{(0)}})^{\frac{1}{2}}(\sum_{l \in \mathcal{S}^*} \alpha_l L_{T^{(l)}} + \lambda_1 \mathbf{I})^{-1}(\sum_{l \in \mathcal{S}^*} \alpha_l \left(L_{T^{(l)}} - L_{T_n^{(l)}}\right))(L_{T^{(0)}})^{-\frac{1}{2}}\right\|_{op}^2 \|(L_{T^{(0)}})^{\frac{1}{2}}(\hat{f}_{\mathcal{S}\lambda_1} - f_{\mathcal{S}\lambda_1})\|_{L^2}^2 \\
&\quad = O_P\left((n_{\mathcal{S}} + n_0)^{-1} \lambda_1^{\frac{1}{2r}} \|(L_{T^{(0)}})^{\frac{1}{2}}(\hat{f}_{\mathcal{S}\lambda_1} - f_{\mathcal{S}\lambda_1})\|_{L^2}^2\right) \\
&\quad = o_P\left(\|(L_{T^{(0)}})^{\frac{1}{2}}(\hat{f}_{\mathcal{S}\lambda_1} - f_{\mathcal{S}\lambda_1})\|_{L^2}^2\right).
\end{aligned}$$

Combine three parts, we get

$$\left\|(L_{T^{(0)}})^{\frac{1}{2}}\left(\hat{f}_{\mathcal{S}\lambda_1} - f_{\mathcal{S}}\right)\right\|_{L^2}^2 \lesssim_P \lambda_1 + (n_{\mathcal{S}} + n_0)^{-1} \lambda_1^{-\frac{1}{2r}},$$

by taling $\lambda_1 \asymp (n_{\mathcal{S}} + n_0)^{-\frac{2r}{2r+1}}$

$$\left\|(L_{T^{(0)}})^{\frac{1}{2}}\left(\hat{f}_{\mathcal{S}\lambda_1} - f_{\mathcal{S}}\right)\right\|_{L^2}^2 \lesssim_P (n_{\mathcal{S}} + n_0)^{-\frac{2r}{2r+1}}.$$

debias step: The estimation scheme in debias step is in the same form as transfer step and thus its proof follows the same spirit of transfer step. The solution of minimization problem in debias step is

$$\hat{f}_{\delta\lambda_2} = \left(T_n^{(0)} + \lambda_2 \mathbf{I}\right)^{-1}\left(T_n^{(0)}(f_{\mathcal{S}} - \hat{f}_{\mathcal{S}\lambda_1} + f_\delta) + g_{0n}\right),$$

where

$$g_{0n} = \frac{1}{n_0}\sum_{i=1}^{n_0} \epsilon_i^{(0)} L_{K^{\frac{1}{2}}}(X_i^{(0)}).$$

Similarly, define

$$f_{\delta\lambda_2} = (L_{T^{(0)}} + \lambda_2 \mathbf{I})^{-1}\left(T_0(f_{\mathcal{S}} - \hat{f}_{\mathcal{S}} + f_\delta)\right),$$

where

$$f_\delta = \left(\sum_{l \in \mathcal{S}^*} \alpha_l L_{T^{(l)}}\right)^{-1}\left(\sum_{l \in \mathcal{S}} \alpha_l L_{T^{(l)}}\left(f_{0\delta}^{(l)}\right)\right).$$

By triangular inequality,

$$\left\|(L_{T^{(0)}})^{\frac{1}{2}}(\hat{f}_\delta - f_\delta)\right\|_{L^2} \leq \left\|(L_{T^{(0)}})^{\frac{1}{2}}(\hat{f}_\delta - f_{\delta\lambda_2})\right\|_{L^2} + \left\|(L_{T^{(0)}})^{\frac{1}{2}}(f_{\delta\lambda_2} - f_\delta)\right\|_{L^2}.$$

For the second term in r.h.s.,

$$\begin{aligned}\left\|(L_{T^{(0)}})^{\frac{1}{2}}(f_{\delta\lambda_2}-f_\delta)\right\|_{L^2} &\leq \left\|(L_{T^{(0)}})^{\frac{1}{2}}\left(T_0+\lambda_2\mathbf{I}\right)^{-1}T_0\left(f_{\mathcal{S}}-\hat{f}_{\mathcal{S}\lambda_1}\right)\right\|_{L^2}+\left\|(L_{T^{(0)}})^{\frac{1}{2}}\left(f^*_{\delta\lambda_2}-f_\delta\right)\right\|_{L^2}\\ &\leq \left\|(L_{T^{(0)}})^{\frac{1}{2}}\left(T_0+\lambda_2\mathbf{I}\right)^{-1}(L_{T^{(0)}})^{\frac{1}{2}}\right\|_{op}\left\|T_0^{\frac{1}{2}}\left(f_{\mathcal{S}}-\hat{f}_{\mathcal{S}}\right)\right\|_{L^2}\\ &\quad+\left\|(L_{T^{(0)}})^{\frac{1}{2}}\left(f^*_{\delta\lambda_2}-f_\delta\right)\right\|_{L^2},\end{aligned}$$

where $f^*_{\delta\lambda_2}=\left(L_{T^{(0)}}+\lambda_2\mathbf{I}\right)^{-1}L_{T^{(0)}}(f_\delta)$.

By Lemma 2 with $\mathcal{S}=\emptyset$,

$$\begin{aligned}\left\|(L_{T^{(0)}})^{\frac{1}{2}}\left(f^*_{\delta\lambda_2}-f_\delta\right)\right\|^2_{L^2} &\leq \frac{\lambda_2}{4}\left\|f_\delta\right\|^2_{L^2}\\ &\lesssim \lambda_2 h^2.\end{aligned}$$

The second inequality holds with the fact the $\mathcal{S}=\{1\leq l\leq L:\|f_0-f_l\|_{L^2}\leq h\}$.

Therefore,

$$\left\|(L_{T^{(0)}})^{\frac{1}{2}}(f_{\delta\lambda_2}-f_\delta)\right\|^2_{L^2}=O_P\left((n_{\mathcal{S}}+n_0)^{-\frac{2r}{2r+1}}+\lambda_2 h^2\right).$$

For first term, we play the same game as **transfer step**. Define

$$\tilde{f}_{\delta\lambda_2}=f_{\delta\lambda_2}+\left(L_{T^{(0)}}+\lambda_2\mathbf{I}\right)^{-1}\left(T_n^{(0)}(f_{\mathcal{S}\lambda_1}-\hat{f}_{\mathcal{S}\lambda_1}+f_\delta)+g_{0n}-T_n^{(0)}(f_{\delta\lambda_2})-\lambda_1 f_{\delta\lambda_2}\right),$$

and the definition of $f_{\delta\lambda_2}$ leads to

$$\tilde{f}_{\delta\lambda_2}-f_{\delta\lambda_2}=\left(L_{T^{(0)}}+\lambda_2\mathbf{I}\right)^{-1}\left((T_n^{(0)}-L_{T^{(0)}})(f_{\mathcal{S}\lambda_1}-\hat{f}_{\mathcal{S}\lambda_1}+f_\delta-f_{\delta\lambda_2}+g_{0n})\right).$$

Therefore,

$$\begin{aligned}\left\|(L_{T^{(0)}})^{\frac{1}{2}}\left(\tilde{f}_{\delta\lambda_2}-f_{\delta\lambda_2}\right)\right\|_{L^2}\leq &\left\|(L_{T^{(0)}})^{\frac{1}{2}}(L_{T^{(0)}}+\lambda_1\mathbf{I})^{-1}g_{0n}\right\|_{L^2}+\\ &\left\|(L_{T^{(0)}})^{\frac{1}{2}}\left(L_{T^{(0)}}+\lambda_2\mathbf{I}\right)^{-1}(T_n^{(0)}-L_{T^{(0)}})(L_{T^{(0)}})^{-\frac{1}{2}}\right\|_{op}\cdot\\ &\left\{\left\|(L_{T^{(0)}})^{\frac{1}{2}}\left(f_{\mathcal{S}\lambda_1}-\tilde{f}_{\mathcal{S}\lambda_1}+f_\delta-f_{\delta\lambda_2}\right)\right\|_{L^2}\right\}\end{aligned}$$

leading to

$$\left\|(L_{T^{(0)}})^{\frac{1}{2}}\left(\tilde{f}_{\delta\lambda_2}-f_{\delta\lambda_2}\right)\right\|_{L^2}\lesssim_P\left(n_0\lambda_2^{\frac{1}{2r}}\right)^{-1}+\left(n_0\lambda_2^{\frac{1}{2r}}\right)^{-1}\left[(n_{\mathcal{S}}+n_0)^{-\frac{2r}{2r+1}}+\lambda_2 h^2\right],$$

where the first term and operator norm comes from Lemma 3 and 4 with $\mathcal{S}=\emptyset$, and bounds on $\|(L_{T^{(0)}})^{\frac{1}{2}}(f_{\mathcal{S}\lambda_1}-\tilde{f}_{\mathcal{S}\lambda_1}+f_\delta-f_{\delta\lambda_2})\|^2_{L^2}$ comes from **transfer step** and bias term of **debias step**.

Finally, for $\|(L_{T^{(0)}})^{\frac{1}{2}}(\hat{f}_{\delta\lambda_2}-\tilde{f}_{\delta\lambda_2})\|^2_{L^2}$, notice that

$$\hat{f}_{\delta\lambda_2}-\tilde{f}_{\delta\lambda_2}=\left(L_{T^{(0)}}+\lambda_2\mathbf{I}\right)^{-1}\left((L_{T^{(0)}}-T_n^{(0)})(\tilde{f}_{\delta\lambda_2}-f_{\delta\lambda_2})\right),$$

thus by Lemma 3,

$$\left\|(L_{T^{(0)}})^{\frac{1}{2}}(\hat{f}_{\delta\lambda_2}-\tilde{f}_{\delta\lambda_2})\right\|_{L^2}^2 = o_P\left(\|(L_{T^{(0)}})^{\frac{1}{2}}(\tilde{f}_{\delta\lambda_2}-f_{\delta\lambda_2})\|_{L^2}^2\right).$$

Combine three parts, we get

$$\left\|(L_{T^{(0)}})^{\frac{1}{2}}\left(\hat{f}_{\delta\lambda_2}-f_\delta\right)\right\|_{L^2}^2 \lesssim_P \lambda_2 h^2 + (n_0\lambda_2^{\frac{1}{2r}})^{-1},$$

taking $\lambda_2 \asymp n_0^{-\frac{2r}{2r+1}}$ leads to

$$\left\|(L_{T^{(0)}})^{\frac{1}{2}}\left(\hat{f}_{\delta\lambda_2}-f_\delta\right)\right\|_{L^2}^2 \lesssim_P n_0^{-\frac{2r}{2r+1}}h^2.$$

Combining **transfer step** and **debias step**, and notice the constant term for **transfer step**

$$\mathcal{E}(\hat{\beta}) \lesssim_P \left\{(n_{\mathcal{S}}+n_0)^{-\frac{2r}{2r+1}} + n_0^{-\frac{2r}{2r+1}}\xi(h,\mathcal{S})\right\}.$$

To prove the same upper bound under Assumption 2, we only need to show Lemma 2 to Lemma 6 still hold under Assumption 2.

Let $\{(s_j^{(0)},\phi_j^{(0)})\}_j$ be the eigen-pairs of $L_{T^{(0)}}$. We show that

$$\left\langle L_{T^{(l)}}\phi_j^{(0)},\phi_j^{(0)}\right\rangle_{L^2} = \lambda_j^{(0)}(1+o(1)). \tag{10}$$

Consider

$$\begin{aligned}\left|\left\langle (L_{T^{(l)}}-L_{T^{(0)}})\phi_j^{(0)},\phi_j^{(0)}\right\rangle_{L^2}\right| &= \left|\left\langle (L_{T^{(0)}})^{\frac{1}{2}}\left((L_{T^{(0)}})^{-\frac{1}{2}}L_{T^{(l)}}(L_{T^{(0)}})^{-\frac{1}{2}}-\mathbf{I}\right)(L_{T^{(0)}})^{\frac{1}{2}}\phi_j^{(0)},\phi_j^{(0)}\right\rangle_{L^2}\right| \\ &= \lambda_j^{(0)}\left|\left\langle \left((L_{T^{(0)}})^{-\frac{1}{2}}L_{T^{(l)}}(L_{T^{(0)}})^{-\frac{1}{2}}-\mathbf{I}\right)\phi_j^{(0)},\phi_j^{(0)}\right\rangle_{L^2}\right|.\end{aligned}$$

Since $(L_{T^{(0)}})^{-\frac{1}{2}}L_{T^{(l)}}(L_{T^{(0)}})^{-\frac{1}{2}}-\mathbf{I}$ is Hilbert–Schmidt, then

$$\left\|(L_{T^{(0)}})^{-\frac{1}{2}}L_{T^{(l)}}(L_{T^{(0)}})^{-\frac{1}{2}}-\mathbf{I}\right\|_{HS} = \sum_{i,j}\left|\langle\phi_i^{(0)},\left((L_{T^{(0)}})^{-\frac{1}{2}}L_{T^{(l)}}(L_{T^{(0)}})^{-\frac{1}{2}}-\mathbf{I}\right)\phi_j^{(0)}\rangle_{L^2}\right|^2 < \infty$$

which leads to

$$\left|\left\langle \left((L_{T^{(0)}})^{-\frac{1}{2}}L_{T^{(l)}}(L_{T^{(0)}})^{-\frac{1}{2}}-\mathbf{I}\right)\phi_j^{(0)},\phi_j^{(0)}\right\rangle_{L^2}\right| = o(1) \quad as \quad j\to\infty.$$

Therefore, Equation (10) holds. One can now replace the common eigenfunctions ϕ_j in the proofs of Lemma 2 to Lemma 6 by $\phi_j^{(0)}$, and it is not hard to check the results still hold. □

A.2 Proof of Theorem 2

Proof. Note that any lower bound for a specific case will immediately yield a lower bound for the general case. Therefore, we consider the following two cases.

(1) Consider $h = 0$, i.e.

$$y_i^{(l)} = \left\langle X_i^{(l)}, \beta \right\rangle + \epsilon_i^{(l)}, \quad \textit{for all } l \in \{0\} \cup \mathcal{S}.$$

In this case, all the source model shares the same coefficient function as target model and therefore the estimation process is equivalent to estimate β under target model with sample size equal to $n_{\mathcal{S}} + n_0$. The lower bound in Cai and Yuan [2012] can be applied here and leading to

$$\lim_{a \to 0} \lim_{n \to \infty} \inf_{\tilde{\beta}} \sup_{\beta^{(0)} \in \Theta(h)} P\left\{ \mathcal{E}\left(\tilde{\beta}\right) \geq a \left(n_{\mathcal{S}} + n_0\right)^{-\frac{2r}{2r+1}} \right\} = 1.$$

(2) Consider $\beta^{(0)} \in \mathcal{B}_{\mathcal{H}}(h)$ and $\beta^{(l)} = 0$ for all $l \in \{0\} \cup \mathcal{S}$ and $\sigma_0 \geq h$. That is all the source datasets contain no information about $\beta^{(0)}$. Consider slope functions $\beta_1, \cdots, \beta_M \in \mathcal{B}_{\mathcal{H}}(h)$ and $P_1, \cdots, P_M$ as the probability distribution of $\{(X_i^{(0)}, Y_i^{(0)}) : i = 1, \cdots, n_0\}$ under $\beta_1, \cdots, \beta_M$. Then the KL divergence between P_i and P_j is

$$KL\left(P_i | P_j\right) = \frac{n_0}{2\sigma_0^2} \left\| L_{(C^{(0)})^{\frac{1}{2}}} \left(\beta_i - \beta_j\right) \right\|_{\mathcal{H}(K)}^2 \quad \textit{for} \quad i, j \in \{1, \cdots, K\}.$$

Let $\tilde{\beta}$ be any estimator based on $\{(X_i^{(0)}, Y_i^{(0)}) : i = 1, \cdots, n_0\}$ and consider testing multiple hypotheses, by Markov inequality and Lemma 7

$$\begin{aligned} \left\| L_{(C^{(0)})^{\frac{1}{2}}} \left(\tilde{\beta} - \beta_i\right) \right\|_{\mathcal{H}(K)}^2 &\geq P_i\left(\tilde{\beta} \neq \beta_i\right) \min_{i,j} \left\| L_{(C^{(0)})^{\frac{1}{2}}} \left(\beta_i - \beta_j\right) \right\|_{\mathcal{H}(K)}^2 \\ &\geq \left(1 - \frac{\frac{n_0}{2\sigma_0^2} \max_{i,j} \left\| L_{(C^{(0)})^{\frac{1}{2}}} \left(\beta_i - \beta_j\right) \right\|_{\mathcal{H}(K)}^2 + log(2)}{log(M-1)} \right) \min_{i,j} \left\| L_{(C^{(0)})^{\frac{1}{2}}} \left(\beta_i - \beta_j\right) \right\|_{\mathcal{H}(K)}^2 . \end{aligned} \tag{11}$$

Our target is to construct a sequence of $\beta_1, \cdots, \beta_M \in \mathcal{B}_{\mathcal{H}}(h)$ such the above lower bound match with upper bound. We consider Varshamov- Gilbert bound in Varshamov [1957], which we state it as Lemma 8. Now we define,

$$\beta_i = \sum_{k=N+1}^{2N} \frac{b_{i,k-N} h}{\sqrt{N}} L_{K^{\frac{1}{2}}}(\phi_k) \quad \textit{for} \quad i = 1, 2, \cdots, M.$$

Then,

$$\|\beta_i\|_K^2 = \sum_{k=N+1}^{2N} \frac{b_{i,k-N}^2 h^2}{N} \left\| L_{K^{\frac{1}{2}}}(\phi_k) \right\|_K^2 \leq h^2,$$

hence $\beta_\theta \in \mathcal{B}_{\mathcal{H}(K)}(h)$.

Besides,

$$\begin{aligned}\left\|L_{(C^{(0)})^{\frac{1}{2}}}\left(\beta_i-\beta_j\right)\right\|_K^2 &= \frac{h^2}{N}\sum_{k=N+1}^{2N}\left(b_{i,k-N}-b_{j,k-N}\right)^2 s_l^{(0)} \\ &\geq \frac{h^2 s_{2N}^{(0)}}{N}\sum_{k=N+1}^{2N}\left(b_{i,k-N}-b_{j,k-N}\right)^2 \\ &\geq \frac{h^2 s_{2N}^{(0)}}{4},\end{aligned}$$

where the last inequality is by Lemma 8, and

$$\begin{aligned}\left\|L_{(C^{(0)})^{\frac{1}{2}}}\left(\beta_i-\beta_j\right)\right\|_K^2 &= \frac{h^2}{N}\sum_{k=N+1}^{2N}\left(b_{i,k-N}-b_{j,k-N}\right)^2 s_k^{(0)} \\ &\leq \frac{h^2 s_{N}^{(0)}}{N}\sum_{k=M+1}^{M}\left(b_{i,k-N}-b_{j,k-N}\right)^2 \\ &\leq h^2 s_N^{(0)}.\end{aligned}$$

Therefore, one can bound the KL divergence by

$$KL\left(P_i|P_j\right) \leq \max_{i,j}\left\{\frac{n_0}{2\sigma_0^2}\left\|L_{(C^{(0)})^{\frac{1}{2}}}\left(\beta_i-\beta_j\right)\right\|_{\mathcal{H}(K)}^2\right\}.$$

Using the above results, the r.h.s. of Equation 11 becomes

$$\left(1-\frac{\frac{4n_0h^2}{\sigma_0^2}s_N^{(0)}+8log(2)}{N}\right)\frac{s_{2N}^{(0)}h^2}{4}.$$

Taking $N=\frac{8h^2}{\sigma_0^2}n_0^{\frac{1}{2r+1}}$, which implies $N\to\infty$, would produce

$$\left(1-\frac{\frac{4n_0h^2}{\sigma_0^2}s_N^{(0)}+8log(2)}{N}\right)\frac{s_{2N}^{(0)}h^2}{4}\asymp\left(\frac{1}{2}-\frac{8log(2)}{N}\right)h^2N^{-2r}\asymp n_0^{-\frac{2r}{2r+1}}h^2$$

Combining the lower bound in case (1) and case (2), we obtain the desired lower bound. □

A.3 Proof of Theorem 3

Proof. Under Assumption 3,

$$\max_{l\in\mathcal{S}}\Delta_l<\min_{l\in\mathcal{S}^c}\Delta_l$$

holds automatically. To prove

$$\mathbb{P}\left(\hat{\mathcal{S}}_j=\mathcal{S}\right)\to 1,$$

we only need to show

$$\mathbb{P}\left(\max_{l\in\mathcal{S}}\hat{\Delta}_l < \min_{l\in\mathcal{S}^c}\hat{\Delta}_l\right) \to 1$$

holds. Observe that

$$\|f\|_{K^M} = \sum_{j=1}^{M}\frac{f_j^2}{v_j} \leq \frac{1}{v_M}\sum_{j=1}^{M} f_j^2 \leq \frac{1}{v_M}\|f\|_{L^2} \lesssim \|f\|_{L^2}$$

for any finite M, then by Corollary 10 in Yuan and Cai [2010]

$$\left\|(\hat{\beta}_0-\hat{\beta}_l)-(\beta_0-\beta_l)\right\|_{K^M} \lesssim \left\|(\hat{\beta}_0-\hat{\beta}_l)-(\beta_0-\beta_l)\right\|_{L^2} = o(1).$$

Therefore, for $l\in\mathcal{S}^c$

$$\left\|\hat{\beta}_0-\hat{\beta}_l\right\|_{K^M} \geq (1-o(1))\left\|\beta_0-\beta_l\right\|_{K^M}$$

and also for $l\in\mathcal{S}$

$$\left\|\hat{\beta}_0-\hat{\beta}_l\right\|_{K^M} \leq (1+o(1))\left\|\beta_0-\beta_l\right\|_{K^M} \leq (1+o(1))\left\|\beta_0-\beta_l\right\|_{K}$$

with high probability. Finally,

$$\mathbb{P}\left(\max_{l\in\mathcal{S}}\hat{\Delta}_l < \min_{l\in\mathcal{S}^c}\hat{\Delta}_k\right) \geq \mathbb{P}\left((1+o(1))\max_{l\in\mathcal{S}}\|\beta_0-\beta_l\|_K < (1-o(1))\min_{l\in\mathcal{S}^c}\|\beta_0-\beta_l\|_{K^M}\right) \to 1.$$

□

A.4 Proof of Theorem 4

The Theorem directly holds by combining Theorem 1, Lemma 1 and Theorem 3.

A.5 Lemmas

Lemma 2.

$$\left\|\left(L_{T^{(0)}}\right)^v\left(f_{\mathcal{S}\lambda_1}-f_{\mathcal{S}}\right)\right\|_{L^2}^2 \leq (1-v)^{2(1-v)}v^{2v}\lambda_1^{2v}\|f_{\mathcal{S}}\|_{L^2}^2\max_j\left\{\left(\frac{s_j^{(0)}}{\sum_{l\in\mathcal{S}^*}\alpha_l s_j^{(l)}}\right)^{2v}\right\}.$$

Proof. By the definition of $f_{\mathcal{S}}$ and $f_{\mathcal{S}\lambda_1}$,

$$\left(\sum_{l\in\mathcal{S}^*}\alpha_l L_{T^{(l)}}+\lambda_1 I\right)f_{\mathcal{S}\lambda_1} = \sum_{l\in\mathcal{S}^*}\alpha_l L_{T^{(l)}}\left(f_0^{(l)}\right) \quad \textit{and} \quad \left(\sum_{l\in\mathcal{S}^*}\alpha_l L_{T^{(l)}}\right)f_{\mathcal{S}} = \sum_{l\in\mathcal{S}^*}\alpha_l L_{T^{(l)}}\left(f_0^{(l)}\right)$$

then

$$f_{\mathcal{S}\lambda_1}-f_{\mathcal{S}} = -\left(\sum_{l\in\mathcal{S}^*}\alpha_l L_{T^{(l)}}+\lambda_1 I\right)^{-1}\lambda_1 f_{\mathcal{S}}.$$

Hence,

$$\|(L_{T^{(0)}})^v (f_{\mathcal{S}\lambda_1} - f_{\mathcal{S}})\|_{L^2}^2 \leq \lambda_1^2 \left\| (L_{T^{(0)}})^v (\sum_{l\in\mathcal{S}^*} \alpha_l L_{T^{(l)}} + \lambda_1 I)^{-1} \right\|_{op}^2 \|f_{\mathcal{S}}\|_{L^2}^2$$
$$\leq \lambda_1^2 \max_j \left\{ \frac{(s_j^{(0)})^{2v}}{(\sum_{l\in\mathcal{S}^*} \alpha_l s_j^{(l)} + \lambda_1)^2} \right\} \|f_{\mathcal{S}}\|_{L^2}^2$$

By Young's inequality, $\lambda_1 + \sum_{l\in\mathcal{S}^*} \alpha_l s_j^{(l)} \geq (1-v)^{-(1-v)} v^{-v} \lambda_1^{1-v} (\sum_{l\in\mathcal{S}^*} \alpha_l s_j^{(l)})^v$

$$\|(L_{T^{(0)}})^v (f_{\mathcal{S}\lambda_1} - f_{\mathcal{S}})\|_{L^2}^2 \leq (1-v)^{2(1-v)} v^{2v} \lambda_1^{2v} \|f_{\mathcal{S}}\|_{L^2}^2 \max_j \left\{ \left(\frac{s_j^{(0)}}{\sum_{l\in\mathcal{S}^*} \alpha_l s_j^{(l)}} \right)^{2v} \right\}.$$

□

Lemma 3.

$$\left\| (L_{T^{(0)}})^v \left(\sum_{l\in\mathcal{S}^*} \alpha_l L_{T^{(l)}} + \lambda_1 \mathbf{I} \right)^{-1} \left(\sum_{l\in\mathcal{S}^*} \alpha_l \left(L_{T^{(l)}} - L_{T_n^{(l)}} \right) \right) (L_{T^{(0)}})^{-v} \right\|_{op} = O_P \left(\left((n_{\mathcal{S}} + n_0) \lambda_1^{1-2v+\frac{1}{2r}} \right)^{-\frac{1}{2}} \right)$$

Proof.

$$\left\| (L_{T^{(0)}})^v \left(\sum_{l\in\mathcal{S}^*} \alpha_l L_{T^{(l)}} + \lambda_1 \mathbf{I} \right)^{-1} \left(\sum_{l\in\mathcal{S}^*} \alpha_l \left(L_{T^{(l)}} - L_{T_n^{(l)}} \right) \right) (L_{T^{(0)}})^{-v} \right\|_{op}$$
$$= \sup_{h:\|h\|_{L^2}=1} \left| \left\langle h, (L_{T^{(0)}})^v \left(\sum_{l\in\mathcal{S}^*} \alpha_l L_{T^{(l)}} + \lambda_1 \mathbf{I} \right)^{-1} \left(\sum_{l\in\mathcal{S}^*} \alpha_l \left(L_{T^{(l)}} - L_{T_n^{(l)}} \right) \right) (L_{T^{(0)}})^{-v} h \right\rangle_{L^2} \right|.$$

Let

$$h = \sum_{j\geq 1} h_j \phi_j,$$

then

$$\left\langle h, (L_{T^{(0)}})^v \left(\sum_{l\in\mathcal{S}^*} \alpha_l L_{T^{(l)}} + \lambda_1 \mathbf{I} \right)^{-1} \left(\sum_{l\in\mathcal{S}^*} \alpha_l \left(L_{T^{(l)}} - L_{T_n^{(l)}} \right) \right) (L_{T^{(0)}})^{-v} h \right\rangle_{L^2}$$
$$= \sum_{j,k} \frac{(s_j^{(0)})^v (s_k^{(0)})^{-v} h_j h_k}{\sum_{l\in\mathcal{S}^*} \alpha_l s_j^{(l)} + \lambda_1} \left\langle \phi_j, \sum_{l\in\mathcal{S}^*} \left(L_{T^{(l)}} - L_{T_n^{(l)}} \right) \phi_k \right\rangle_{L^2}.$$

By Cauchy-Schwarz inequality,

$$\left\| (L_{T^{(0)}})^v \left(\sum_{l\in\mathcal{S}^*} \alpha_l L_{T^{(l)}} + \lambda_1 \mathbf{I} \right)^{-1} \left(\sum_{l\in\mathcal{S}^*} \alpha_l \left(L_{T^{(l)}} - L_{T_n^{(l)}} \right) \right) (L_{T^{(0)}})^{-v} \right\|_{op}$$
$$\leq \left(\sum_{j,k} \frac{(s_j^{(0)})^{2v} (s_k^{(0)})^{-2v}}{(\sum_{l\in\mathcal{S}^*} \alpha_l s_j^{(l)} + \lambda_1)^2} \left\langle \phi_j, \sum_{l\in\mathcal{S}^*} \alpha_l \left(L_{T^{(l)}} - L_{T_n^{(l)}} \right) \phi_k \right\rangle_{L^2}^2 \right)^{\frac{1}{2}}.$$

Consider $\mathrm{E}\langle\phi_j, \sum_{l\in\mathcal{S}^*}(L_{T^{(l)}} - L_{T_n^{(l)}})\phi_k\rangle_{L^2}^2$, note that

$$
\begin{aligned}
& \mathrm{E}\left\langle \phi_j, \sum_{l\in\mathcal{S}^*} \alpha_l \left(L_{T^{(l)}} - L_{T_n^{(l)}}\right)\phi_k \right\rangle_{L^2}^2 \\
&= \mathrm{E}\left(\sum_{l\in\mathcal{S}^*} \alpha_l \left\langle L_{K^{\frac{1}{2}}}(\phi_k), (C^{(l)} - L_{C_n^{(l)}})L_{K^{\frac{1}{2}}}(\phi_j)\right\rangle_{L^2}\right)^2 \\
&= \mathrm{E}\left(\sum_{l\in\mathcal{S}^*} \alpha_l \frac{1}{n_l}\sum_{i=1}^{n_l}\int_{\mathcal{T}^2} L_{K^{\frac{1}{2}}}(\phi_k)(s)\left(X_i^{(l)}(s)X_i^{(l)}(t) - \mathrm{E}\,X_i^{(l)}(s)X_i^{(l)}(t)\right) L_{K^{\frac{1}{2}}}(\phi_j)(t)dtds\right)^2 \\
&\le |\mathcal{S}^*|\sum_{l\in\mathcal{S}^*}\frac{\alpha_l^2}{n_l}s_j^{(l)}s_k^{(l)}
\end{aligned}
$$

By Jensen's inequality

$$
\begin{aligned}
& \mathrm{E}\left(\sum_{j,k}\frac{(s_j^{(0)})^{2v}(s_k^{(0)})^{-2v}}{(\sum_{l\in\mathcal{S}^*}\alpha_l s_j^{(l)} + \lambda_1)^2}\left\langle \phi_j, \sum_{l\in\mathcal{S}^*} \alpha_l \left(L_{T^{(l)}} - L_{T_n^{(l)}}\right)\phi_k \right\rangle_{L^2}^2\right)^{\frac{1}{2}} \\
&\le \left(\sum_{j,k}\frac{(s_j^{(0)})^{2v}(s_k^{(0)})^{-2v}}{(\sum_{l\in\mathcal{S}^*}\alpha_l s_j^{(l)} + \lambda_1)^2}\,\mathrm{E}\left\langle \phi_j, \sum_{l\in\mathcal{S}^*} \alpha_l \left(L_{T^{(l)}} - L_{T_n^{(l)}}\right)\phi_k \right\rangle_{L^2}^2\right)^{\frac{1}{2}},
\end{aligned}
$$

thus,

$$
\begin{aligned}
& \mathrm{E}\left(\sum_{j,k}\frac{(s_j^{(0)})^{2v}(s_k^{(0)})^{-2v}}{(\sum_{l\in\mathcal{S}^*}\alpha_l s_j^{(l)} + \lambda_1)^2}\left\langle \phi_j, \sum_{l\in\mathcal{S}^*} \alpha_l \left(L_{T^{(l)}} - L_{T_n^{(l)}}\right)\phi_k \right\rangle_{L^2}^2\right)^{\frac{1}{2}} \\
&\le \left(\sum_{j,k}\frac{(s_j^{(0)})^{2v}(s_k^{(0)})^{-2v}}{(\sum_{l\in\mathcal{S}^*}\alpha_l s_j^{(l)} + \lambda_1)^2}\left(\sum_{l\in\mathcal{S}^*}\alpha_l s_j^{(l)}s_k^{(l)}\right)\frac{|\mathcal{S}^*|}{(n_{\mathcal{S}} + n_0)}\right)^2 \\
&\le \max_{j,k}\left(\frac{\sum_{l\in\mathcal{S}^*}\alpha_l s_j^{(l)}s_k^{(l)}}{s_j^{(0)}s_k^{(0)}}\right)\left(\sum_{j,k}\frac{(s_j^{(0)})^{1+2v}(s_k^{(0)})^{1-2v}}{(\sum_{l\in\mathcal{S}^*}\alpha_l s_j^{(l)} + \lambda_1)^2}\frac{|\mathcal{S}^*|}{(n_{\mathcal{S}} + n_0)}\right)^2
\end{aligned}
$$

By assumptions of eigenvalues, $\max_{j,k}\left(\frac{\sum_{l\in\mathcal{S}^*}\alpha_l s_j^{(l)}s_k^{(l)}}{s_j^{(0)}s_k^{(0)}}\right) \le C_1$ for some constant C_1. Finally, by Lemma 6

$$
\mathrm{E}\left(\sum_{j,k}\frac{(s_j^{(0)})^{2v}(s_k^{(0)})^{-2v}}{(\sum_{l\in\mathcal{S}^*}\alpha_l s_j^{(l)} + \lambda_1)^2}\left\langle \phi_j, \sum_{l\in\mathcal{S}^*} \alpha_l \left(L_{T^{(l)}} - L_{T_n^{(l)}}\right)\phi_k \right\rangle_{L^2}\right)^{\frac{1}{2}} \lesssim \left((n_{\mathcal{S}} + n_0)\lambda_1^{1-2v+\frac{1}{2r}}\right)^{-1}.
$$

The rest of the proof can be complete by Markov inequality. □

Lemma 4.

$$
\left\|(L_{T^{(0)}})^v(\sum_{l\in\mathcal{S}^*}\alpha_l L_{T^{(l)}} + \lambda_1\mathbf{I})^{-1}\sum_{l\in\mathcal{S}^*}g_{ln}\right\|_{L^2}^2 = O_P\left(\left((n_{\mathcal{S}} + n_0)\,\lambda_1^{1-2v+\frac{1}{2r}}\right)^{-1}\right)
$$

Proof.

$$\begin{aligned}
\left\|(L_{T^{(0)}})^v(\sum_{l\in\mathcal{S}^*}\alpha_l L_{T^{(l)}}+\lambda_1\mathbf{I})^{-1}\sum_{l\in\mathcal{S}^*}g_{ln}\right\|_{L^2}^2 &= \sum_{j\geq 1}\left\langle(L_{T^{(0)}})^v(\sum_{l\in\mathcal{S}^*}\alpha_l L_{T^{(l)}}+\lambda_1\mathbf{I})^{-1}\sum_{l\in\mathcal{S}^*}g_{ln},\phi_j\right\rangle_{L^2}^2\\
&= \sum_{j\geq 1}\left\langle\sum_{l\in\mathcal{S}^*}g_{ln},\frac{(s_j^{(0)})^v}{\sum_{l\in\mathcal{S}^*}\alpha_l s_j^{(l)}+\lambda_1}\phi_j\right\rangle_{L^2}^2\\
&= \sum_{j\geq 1}\frac{(s_j^{(0)})^{2v}}{(\sum_{l\in\mathcal{S}^*}\alpha_l s_j^{(l)}+\lambda_1)^2}\\
&\qquad\left(\frac{1}{n_{\mathcal{S}}+n_0}\sum_{l\in\mathcal{S}^*}\sum_{i=1}^{n_l}\left\langle\epsilon_i^{(l)}X_i^{(l)},L_{K^{\frac12}}(\phi_j)\right\rangle_{L^2}\right)^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{E}\left\|(L_{T^{(0)}})^v(\sum_{l\in\mathcal{S}^*}\alpha_l L_{T^{(l)}}+\lambda_1\mathbf{I})^{-1}\sum_{l\in\mathcal{S}^*}g_{ln}\right\|_{L^2}^2 &= \sum_{j\geq 1}\frac{(s_j^{(0)})^{2v}}{(\sum_{l\in\mathcal{S}^*}\alpha_l s_j^{(l)}+\lambda_1)^2}\cdot\\
&\qquad\mathrm{E}\left(\frac{1}{n_{\mathcal{S}}+n_0}\sum_{l\in\mathcal{S}^*}\sum_{i=1}^{n_l}\left\langle\epsilon_i^{(l)}X_i^{(l)},L_{K^{\frac12}}(\phi_j)\right\rangle_{L^2}\right)^2\\
&= \sum_{j\geq 1}\frac{(s_j^{(0)})^{2v}}{(\sum_{l\in\mathcal{S}^*}\alpha_l s_j^{(l)}+\lambda_1)^2}\frac{1}{(n_{\mathcal{S}}+n_0)^2}\cdot\\
&\qquad\sum_{l\in\mathcal{S}^*}n_l\,\mathrm{E}\left(\langle\epsilon_i^{(l)}X_i^{(l)},L_{K^{\frac12}}(\phi_j)\rangle_{L^2}\right)^2\\
&= \sum_{j\geq 1}\frac{(s_j^{(0)})^{2v}}{(\sum_{l\in\mathcal{S}^*}\alpha_l s_j^{(l)}+\lambda_1)^2}\frac{\left(\sum_{l\in\mathcal{S}^*}\sigma_l^2 n_l s_j^{(l)}\right)}{(n_{\mathcal{S}}+n_0)^2}\\
&\leq \max_j\left\{\frac{\alpha_0 s_j^{(0)}+\sum_{l\in\mathcal{S}}\alpha_l s_j^{(l)}}{s_j^{(0)}}\right\}\cdot\\
&\qquad\left(\frac{C_1}{n_{\mathcal{S}}+n_0}\sum_{j\geq 1}\frac{(s_j^{(0)})^{1+2v}}{(\sum_{l\in\mathcal{S}^*}\alpha_l s_j^{(l)}+\lambda_1)^2}\right),
\end{aligned}$$

thus by assumption on eigenvalus and Lemma 6 with $v=\frac{1}{2}$,

$$\mathrm{E}\left\|(L_{T^{(0)}})^v(\sum_{l\in\mathcal{S}^*}\alpha_l L_{T^{(l)}}+\lambda_1\mathbf{I})^{-1}\sum_{l\in\mathcal{S}^*}g_{ln}\right\|_{L^2}^2\lesssim\left((n_{\mathcal{S}}+n_0)\,\lambda_1^{1-2v+\frac{1}{2r}}\right)^{-1}.$$

The rest of the proof can be complete by Markov inequality. □

Lemma 5.

$$\left\|\sum_{l\in\mathcal{S}^*}\alpha_l L_{T_n^{(l)}}\left(f_0^{(l)}-f_{\mathcal{S}}\right)\right\|_{L^2}^2=O_P\left((n_{\mathcal{S}}+n_0)^{-1}\right)$$

Proof.

$$\mathrm{E}\left\|\sum_{l\in\mathcal{S}^*}\alpha_l L_{T_n^{(l)}}\left(f_0^{(l)}-f_{\mathcal{S}}\right)\right\|_{L^2}^2 = \sum_{j=1}^{\infty}\mathrm{E}\left(\sum_{l\in\mathcal{S}^*}\alpha_l\left\langle C_n^{(l)}L_{K^{\frac{1}{2}}}(f_0^{(l)}-f_{\mathcal{S}}),L_{K^{\frac{1}{2}}}(\phi_j)\right\rangle_{L^2}\right)^2$$
$$\lesssim \sum_{j=1}^{\infty}\sum_{l\in\mathcal{S}^*}\frac{\alpha_l}{n_{\mathcal{S}}+n_0}\langle f_0^{(l)}-f_{\mathcal{S}},\phi_j\rangle_{L^2}^2(s_j^{(l)})^2$$
$$\lesssim (n_{\mathcal{S}}+n_0)^{-1}\max_{j,l}\left\{\alpha_l(s_j^{(l)})^2\right\}\sum_{l\in\mathcal{S}^*}\left\|f_0^{(l)}-f_{\mathcal{S}}\right\|_{L^2}^2$$
$$\lesssim (n_{\mathcal{S}}+n_0)^{-1}.$$

The rest of the proof can be complete by Markov inequality. □

Lemma 6.

$$\lambda_1^{-\frac{1}{2r}} \lesssim \sum_{j\geq 1}\frac{(s_j^{(0)})^{1+2v}}{(\sum_{l\in\mathcal{S}^*}\alpha_l s_j^{(l)}+\lambda_1)^{1+2v}} \lesssim 1+\lambda_1^{-\frac{1}{2r}}.$$

Proof. The proof is exactly the same as Lemma 6 in Cai and Yuan [2012] once we know that $\max_j\left(\frac{s_j^{(0)}}{\sum_{l\in\mathcal{S}^*}\alpha_l s_j^{(l)}}\right)\leq C$, which got satisfied under the assumptions of eigenvalues. □

Lemma 7 (Fano's Lemma)**.** *Let $P_1,\cdots,P_M$ be probability measure such that*

$$KL(P_i|P_j)\leq\alpha,\quad 1\leq i\neq j\leq K$$

then for any test function ψ taking value in $\{1,\cdots,M\}$, we have

$$P_i(\psi\neq i)\geq 1-\frac{\alpha+log(2)}{log(M-1)}.$$

Lemma 8. *(Varshamov-Gilbert) For any $N\geq 1$, there exists at least $M=exp(N/8)$ N-dimenional vectors, $b_1,\cdots,b_M\subset\{0,1\}^N$ such that*

$$\sum_{l=1}^{N}\mathbf{1}\left\{b_{ik}\neq b_{jk}\right\}\geq N/4.$$