On Hypothesis Transfer Learning of Functional Linear Models

Haotian Lin [1] **Matthew Reimherr** [1]

Abstract

We study the *transfer learning* (TL) for the *functional linear regression* (FLR) under the Reproducing Kernel Hilbert Space (RKHS) framework, observing the TL techniques in existing high-dimensional linear regression is not compatible with the truncation-based FLR methods as functional data are intrinsically infinite-dimensional and generated by smooth underlying processes. We measure the similarity across tasks using RKHS distance, allowing the type of information being transferred to be tied to the properties of the imposed RKHS. Building on the hypothesis offset transfer learning paradigm, two algorithms are proposed: one conducts the transfer when positive sources are known, while the other leverages aggregation techniques to achieve robust transfer without prior information about the sources. We establish asymptotic lower bounds for this learning problem and show the proposed algorithms enjoy a matching upper bound. These analyses provide statistical insights into factors that contribute to the dynamics of the transfer. We also extend the results to functional generalized linear models. The effectiveness of the proposed algorithms is demonstrated via extensive synthetic data as well as real-world data applications.

1. Introduction

Advances in technologies enable us to collect and process densely observed data over some temporal or spatial domains, which are coined functional data (Ramsay et al., 2005; Kokoszka & Reimherr, 2017). While functional data analysis (FDA) has been proven useful in various fields like finance, genetics and, etc., and has been researched widely in the statistical community, its effectiveness relies on having sufficient training samples drawn from the same distribution. However, this may not hold for functional data under some applications due to collection expenses or other constraints. Transfer learning (TL)(Torrey & Shavlik, 2010) leverages additional information from some similar (source) tasks to enhance the learning procedure on the original (target) task and is an appealing mechanism when there is a lack of training samples. The goal of this paper is to develop TL algorithms for functional linear regression (FLR), one of the most prevalent models in the FDA. The FLR concerned in this paper is Scalar-on-Function regression, which takes the form:

$$Y = \alpha + \langle \beta, X \rangle_{L^2} + \epsilon = \alpha + \int_{\mathcal{T}} X(s)\beta(s)\, ds + \epsilon,$$

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 function respectively over a compact domain $\mathcal{T} \subset \mathbb{R}$, and ϵ is a random noise with zero mean.

A classical approach to estimating β is to reduce the problem to classical multivariate linear regression by expanding the X and β under the same finite basis, like deterministic basis functions, e.g. Fourier basis, or the eigenbasis of the covariance function of X (Cardot et al., 1999; Yao et al., 2005; Hall & Hosseini-Nasab, 2006; Hall & Horowitz, 2007), which we refer to truncation-based FLR methods in this paper. Conceptually, the offset transfer learning techniques developed in the existing multivariate/high-dimensional linear regression framework (Kuzborskij & Orabona, 2013; 2017; Li et al., 2022; Bastani, 2021) can be applied to truncation-based FLR methods to conduct transfer learning in FLR, though they lack a theoretical foundation in this context due to the truncation error inherent in a basis expansion of β. In particular, a key property distinguishing functional data from multivariate data is that they are inherently infinite-dimensional and generated through smooth underlying processes. Omitting this fact, using finite-dimensional approximations to β, and leveraging existing multivariate OTL techniques on the finite coefficients lose the benefit that data are generated from smooth processes and are less interpretable for the transfer process; see detailed discussion in Section 2. Observing these limitations we develop the first TL algorithms for FLR with statistical convergence rate guarantees under the supervised learning setting.

We summarize our main contributions as follows.

[1] Department of Statistics, The Pennsylvania State University, University Park, PA, USA. Correspondence to: Haotian Lin <hzl435@psu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. We propose using the Reproducing Kernel Hilbert Space (RKHS) distance between tasks' coefficients as a measure of task similarity. The transferred information is thus tied to the RKHS's properties and makes the transfer more interpretable. One can tailor the employed RKHS to the task's nature, offering flexibility to embed diverse structural elements, like smoothness or periodicity, into TL processes.

2. Built upon the offset transfer learning (OTL) paradigm, we propose TL-FLR, a variant of OTL for multiple positive transfer sources. We establish the minimax optimality for TL-FLR. Intriguingly, the result reveals that the faster statistical rate of TL-FLR, compared to non-transfer learning, not only depends on source sample size and the magnitude of discrepancy across tasks like most existing works, but also the signal ratio between offset and source model.

3. To deal with the practical scenario in which there is no available prior task similarity information, we propose Aggregation-based TL-FLR (ATL-FLR), utilizing sparse aggregation to mitigate negative transfer effects. We establish the upper bound for ATL-FLR and show the aggregation cost decreases faster than the transfer learning risk, demonstrating an ability to identify optimal sources without too much extra cost compared to TL-FLR. We further extend this framework to Functional Generalized Linear Models (FGLM) with theoretical guarantees, broadening its applicability.

4. In developing statistical guarantees, we uncovered unique requirements for making OTL theoretically feasible in the functional data context. These include the necessity for covariate functions across tasks to exhibit similar structural properties to ensure statistical convergence, and the coefficient functions of negative sources can be separated from positive ones within a finite-dimensional space to ensure optimal source identification.

Literature review. Apart from truncation-based FLR approaches mentioned before, another line of research proposed that one can obtain a smooth estimator via smoothness regularization (Yuan & Cai, 2010; Cai & Yuan, 2012), and has been widely used in other functional models like the FGLM, functional Cox-model, etc. (Cheng & Shang, 2015; Qu et al., 2016; Reimherr et al., 2018; Sun et al., 2018).

Turning to the TL regime in supervised learning, the hypothesis transfer learning (HTL) framework has become popular (Li & Bilmes, 2007; Orabona et al., 2009; Kuzborskij & Orabona, 2013; Perrot & Habrard, 2015; Du et al., 2017). Offset transfer learning (OTL) (a.k.a. biases regularization transfer learning) has been widely analyzed and applied as one of the most popular HTL paradigms. It assumes the target's function/parameter is a summation of the source's and the offset's function/parameter. A series of works have derived theoretical analysis under different settings. For example, in Kuzborskij & Orabona (2013; 2017), the authors provide the first theoretical study of OTL in the context of linear regression with stability analysis and generalization bounds. Later, in Wang & Schneider (2015); Wang et al. (2016), the authors derive similar theoretical guarantees for non-parametric regression via Kernel Ridge Regression. A unified framework that generalizes many previous works is proposed in Du et al. (2017), and the authors also present an excess risk analysis for their framework. Apart from the regression setting, generalization bounds for classification with surrogate losses have been studied in Aghbalou & Staerman (2023). Other results that study HTL outside OTL can be found in Li & Bilmes (2007); Cheng & Shang (2015). Besides, OTL can also be viewed as a case of representation learning (Du et al., 2020; Tripuraneni et al., 2020; Xu & Tewari, 2021) by viewing the estimated source model as a representation for target tasks. Finally, the bias regularization technique on which OTL relies is also widely used in other learning settings, e.g., Meta, Multi-task, and unsupervised learning, see Denevi et al. (2018; 2019); Balcan et al. (2019); Tian et al. (2022; 2023).

The statistics community has recently adopted OTL for various high-dimensional models with statistical risk guarantees. For example, Bastani (2021) proposed using OTL for high-dimensional (generalized) linear regression but only includes one positive transfer source. Later, Li et al. (2022) extended this idea to multiple sources scenario and leveraged aggregation to alleviate negative transfer effects. In Tian & Feng (2022), the learning procedure gets extended to the high-dimensional generalized linear model, and the authors also proposed a positive sources detection algorithm via a validation approach. In these works, the similarity among tasks is quantified via ℓ^1-norm, which captures the sparsity structure in high-dimensional parameters. There is no TL for FDA that we are aware of, but the closest work would be in the area of domain adaptation. Zhu et al. (2021) studied the domain adaptation problem between two separable Hilbert spaces by proposing algorithms to estimate the optimal transport mapping between two spaces.

Notation. For two sequence $\{a_k\}_{k\geq 1}$ and $\{b_k\}_{k\geq 1}$, we denotes $a_n \asymp b_n$ and $a_n \lesssim b_n$ if $|a_n/b_n| \to c$ and $|a_n/b_n| \leq c$ for some universal constant c when $n \to \infty$. For two random variable sequence $\{A_k\}_{k\geq 1}$ and $\{B_k\}_{k\geq 1}$, if for any $\delta > 0$, there exists $M_\delta > 0$ and $N_\delta > 0$ such that $\mathbb{P}(A_k < M_\delta B_k) \geq 1-\delta, \forall k \geq N_\delta$, we say $A_k = O_{\mathbb{P}}(B_k)$. For a set A, let $|A|$ denote its cardinality, A^c denote its complement. For an integer n, denote $[n] := \{1, \cdots, n\}$.

We denote the covariance function of X as $C(s,t) = \mathrm{E}[X(s) - \mathrm{E}\,X(s)][X(t) - \mathrm{E}\,X(t)]$ for $s,t \in \mathcal{T}$. For a real, symmetric, square-integrable, and nonnegative kernel, $K : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$, we denote its associated RKHS on $\mathcal{T}$ as $\mathcal{H}_K$ and corresponding norm as $\|\cdot\|_K$. We also denote its integral operator as $L_K(f) = \int_{\mathcal{T}} K(\cdot, t)f(t)dt$ for $f \in L^2$. For two kernels, K_1 and K_2, their composition is $(K_1K_2)(s,t) = \int_{\mathcal{T}} K_1(s,u)K_2(u,t)du$. For a given kernel K and covariance kernel C, define bivariate function Γ and its integral operator as $\Gamma := K^{1/2}CK^{1/2}$ and $L_\Gamma(f) = L_{K^{\frac{1}{2}}}(L_C(L_{K^{\frac{1}{2}}}(f)))$.

2. Preliminaries and Backgrounds

Problem Set-up. We now formally set the stage for the transfer learning problem in the context of FLR. Consider the following series of FLRs,

$$Y_i^{(t)} = \alpha^{(t)} + \left\langle X_i^{(t)}, \beta^{(t)} \right\rangle_{L^2} + \epsilon_i^{(t)} \tag{1}$$

for $i \in [n_t]$, $t = 0 \cup [T]$, where $t = 0$ denotes the target model and $t \in [T]$ denotes source models. Denote the sample space $\mathcal{Z}$ as the Cartesian product of the covariate space $\mathcal{X}$ and response space $\mathcal{Y}$. For each $t \in 0 \cup [T]$, we denote $\mathcal{D}^{(t)} = \{(X_i^{(t)}, Y_i^{(t)})\}_{i=1}^{n_t} = \{Z_i^{(t)}\}_{i=1}^{n_t}$. Throughout the paper we assume $\epsilon_i^{(t)}$ are i.i.d. across both i and t with zero mean and finite variance σ^2.

As estimating $\beta^{(0)}$ is our primary interest, we assume for simplicity that $\alpha^{(t)} = 0$ for all t. We assume $n_0 \ll \sum_{t=1}^T n_t$, a condition commonly validated in most TL literature and numerous practical applications. While our framework is designed primarily for the posterior drift setting, i.e., the marginal distributions of $X^{(t)}$ remain the same, but $\beta^{(t)}$ vary, the excess risk bounds we establish are based on a comparatively more relaxed condition, see Section 4.

In the absence of source data, estimating $\beta^{(0)}$ is termed as target-only learning, and one can obtain a smooth estimator of β through the regularized empirical risk minimization (RERM) (Yuan & Cai, 2010; Cai & Yuan, 2012), i.e.

$$\hat{\beta} = \underset{\beta \in \mathcal{H}_K}{\operatorname{argmin}} \left\{ \frac{1}{n_0} \sum_{i=1}^{n_0} \ell(\beta, Z_i^{(0)}) + \lambda \|\beta\|_K^2 \right\},$$

where K is an employed kernel and $\ell : \mathcal{H}_K \times \mathcal{Z} \to \mathbb{R}^+$ is the loss function. This approach has been proven to achieve the optimal rate in terms of excess risk, and we refer to it as *Optimal Functional Linear Regression* (OFLR) in this paper, which serves as a non-transfer learning baseline.

Similarity Measure. We first state the limitations of using ℓ^1/ℓ^2-norm as a similarity measure in the truncation-based FLR method, which converts the problem into a classic multivariate one. For a given series of basis functions $\{\phi_j\}_{j\geq 1}$ and truncated number M, one can model the t-th FLR as

$$Y_i^{(t)} \approx \sum_{j=1}^{M} X_{ij}^{(t)} \beta_j^{(t)} + \epsilon_i^{(t)} \tag{2}$$

where $X_{ij}^{(t)} = \langle X_i^{(t)}, \phi_j \rangle_{L^2}$ and $\beta_j^{(t)} = \langle \beta^{(t)}, \phi_j \rangle_{L^2}$. Denote $\beta_{\text{trunc}}^{(t)} \in \mathbb{R}^M$ as the coefficient vector in (2), one can then measure the similarity between the target and the t-th FLR model by the ℓ^1 or ℓ^2 norm of $\beta_{\text{trunc}}^{(t)} - \beta_{\text{trunc}}^{(0)}$ like the previous works did for multivariate linear regression. However, from the functional data analysis literature, since the functional data are generated from some structural underlying process, it is well known that one has to have the same kind of structure in the estimator, like smoothness, for theoretical reliability. For example, when the coefficient functions are smooth, the above approach cannot measure the similarity since $\{\beta_{\text{trunc}}^{(t)}\}_{t=0}^T$ are not necessarily sparse or might require regularization via an ℓ^2-norm, but the employed basis functions might not reflect the desired smoothness. Besides, the basis functions and M should be consistent across tasks, which reduces the flexibility of the learning procedure.

To explore the similarity tied to the structure of coefficient functions, one should quantify the similarity between tasks within certain functional spaces that possess the same structures. These structural properties, e.g., continuity/smoothness/periodicity, can be naturally encapsulated via kernels and their corresponding RKHS. Consequently, quantifying the similarity within a certain RKHS provides interpretability since the type of information transferred is tied to the structural properties of the used RKHS. We also note that this method is broadly applicable since the reproducing kernel can be tailored to the application problem accordingly. For example, one can transfer the information about continuity or smoothness by picking K to be a Sobolev kernel, and about periodicity by picking periodic kernels like $K(x_1, x_2) = exp\left(-2/l^2 \sin\left(\pi|x_1 - x_2|/p\right)\right)$ where l is the lengthscale and p is the period.

Given the reasoning above, for $t = 0 \cup [T]$, we assume $\beta^{(t)} \in \mathcal{H}_K$, and define the t-th contrast function $\delta^{(t)} := \beta^{(0)} - \beta^{(t)}$. Given a constant $h \geq 0$, we say the t-th source model is "h-transferable" if $\|\delta^{(t)}\|_K \leq h$. The magnitude of h characterizes the similarity between the target model and source models. We also define $\mathcal{S}_h = \{t \in [T] : \|\delta^{(t)}\|_K \leq h\}$ as a subset of $[T]$, which consists of the indexes of all "h-transferable" source models. It is worth mentioning that the quantity h is introduced for theoretical purposes to establish optimality, which is prevalent in recent studies such as Bastani (2021); Li et al. (2022); Tian & Feng (2022); He et al. (2024). However, for the implementation of the algorithm, it is not necessary to know the actual value of h. We abbreviate $\mathcal{S}_h$ as $\mathcal{S}$ to generally represent the h-transferable sources index.

Learning Framework. This paper leverages the widely used OTL paradigm, see reviews in Section 1. Formally, in the FLR and single source $\beta^{(1)}$ context, the OTL obtains the target function via $\hat{\beta}^{(0)} = \hat{\beta}^{(1)} + \hat{\delta}$ where $\hat{\beta}^{(1)}$ is the estimator trained on source dataset and $\hat{\delta}$ is obtained from target dataset via following minimization problem:

$$\hat{\delta} = \underset{\delta \in \mathcal{H}_K}{\operatorname{argmin}} \frac{1}{n_0} \sum_{i=1}^{n_0} \ell(\delta + \hat{\beta}^{(0)}, Z_i^{(0)}) + \lambda \|\delta\|_K^2,$$

where the loss function can be square loss (Orabona et al., 2009; Kuzborskij & Orabona, 2013) or surrogate losses (Aghbalou & Staerman, 2023). The main idea is that the estimator $\hat{\beta}^{(1)}$ can be learned well given sufficiently large source samples and the simple offset estimator $\hat{\delta}^{(0)}$ can be learned with much fewer target samples.

3. Methodology

3.1. Transfer Learning with $\mathcal{S}$ Known

For multiple sources, the idea of data fusion inspires us to obtain a centered source estimator $\beta_{\mathcal{S}}$ via all source datasets in place of $\beta^{(1)}$. Therefore, we can generalize single source OTL to the multiple sources scenario as follows.

Algorithm 1 TL-FLR

1: **Input:** Target/Source datasets $\{\mathcal{D}^{(t)}\}_{t=0}^{T}$; index set of source datasets $\mathcal{S}$; Loss function ℓ as square loss.

2: Transfer Step: Obtain $\hat{\beta}_{\mathcal{S}}$ via

$$\hat{\beta}_{\mathcal{S}} = \underset{\beta \in \mathcal{H}_K}{\operatorname{argmin}} \sum_{t \in \mathcal{S}} \frac{1}{n_t} \sum_{i=1}^{n_t} \ell(\beta, Z_i^{(t)}) + \lambda_1 \|\beta\|_K^2. \quad (3)$$

3: Calibration Step: Obtain offset $\hat{\delta}$ via

$$\hat{\delta} = \underset{\delta \in \mathcal{H}_K}{\operatorname{argmin}} \frac{1}{n_0} \sum_{i=1}^{n_0} \ell(\delta + \hat{\beta}_{\mathcal{S}}, Z_i^{(0)}) + \lambda_2 \|\delta\|_K^2. \quad (4)$$

4: Return $\hat{\beta}_{\mathcal{S}} + \hat{\delta}$.

Since the probabilistic limit of $\hat{\beta}_{\mathcal{S}}$ is not consistent with $\beta^{(0)}$, calibration of $\hat{\beta}_{\mathcal{S}}$ is performed in (4). The regularization term in (4) is consistent with our similarity measure, i.e. it restricts $\hat{\beta}^{(0)}$ to lie in a $\mathcal{H}_K$ ball centered at $\hat{\beta}_{\mathcal{S}}$. Therefore, this term pushes the $\hat{\beta}^{(0)}$ close to $\hat{\beta}_{\mathcal{S}}$ while the mean square error loss over the target dataset allows calibration for the bias. Intuitively, if $\hat{\beta}_{\mathcal{S}}$ is close to $\beta^{(0)}$, then TL-FLR can boost the learning on the target model.

3.2. Transfer Learning with Unknown $\mathcal{S}$

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 tasks might have little or even a negative contribution to the target one, it could be practically harmful to directly apply Algorithm 1 by assuming all sources belong to $\mathcal{S}$. Inspired by the idea of aggregating multiple estimators in Li et al. (2022), we develop ATL-FLR, which can be directly applied without knowing $\mathcal{S}$ while being robust to negative transfer sources.

The general idea of ATL-FLR is that one can 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 estimators $\mathcal{F} = \{\hat{\beta}(\hat{\mathcal{S}}_1), \cdots, \hat{\beta}(\hat{\mathcal{S}}_J)\}$ via TL-FLR. Then, one aggregates the candidate estimators in $\mathcal{F}$ such that the aggregated estimator $\hat{\beta}_a$ satisfies the following oracle inequality in high probability,

$$R(\hat{\beta}_a) \le \min_{\beta \in \mathcal{F}} R(\beta) + r(\mathcal{F}, n), \quad (5)$$

where $R(f) = \mathrm{E}_{(X,Y)}[\ell(Y, f(X)) | \{\mathcal{D}^{(t)} : t \in 0 \cup [T]\}]$, and $r(\mathcal{F}, n)$ is the aggregation cost. Thus, the $\hat{\beta}_a$ can achieve similar performance as TL-FLR up to some aggregation cost. The proposed aggregation-based TL-FLR is as follows:

Algorithm 2 Aggregation-based TL-FLR (ATL-FLR)

Input: Target/Source datasets $\{\mathcal{D}^{(t))}\}_{t=0}^{T}$; index set of source datasets $\mathcal{S}$; Loss function ℓ as square loss; A given integer M.

Step 1: Split the target dataset $\mathcal{D}^{(0)}$ into $\mathcal{D}_{\mathcal{I}}^{(0)}$ and $\mathcal{D}_{\mathcal{I}^c}^{(0)}$ with $\mathcal{I}$ be a random subset of $[n_0]$ such that $|\mathcal{I}| = \lfloor \frac{n_0}{2} \rfloor$.

Step 2: Built candidate sets of $\mathcal{S}$, $\{\hat{\mathcal{S}}_0, \hat{\mathcal{S}}_1, \cdots, \hat{\mathcal{S}}_T\}$ as:

1. Obtain $\hat{\beta}_0$ by OFLR using $\mathcal{D}_{\mathcal{I}}^{(0)}$ and let $\hat{\mathcal{S}}_0 = \emptyset$.
2. For each $t \in [T]$, obtain $\hat{\beta}_t$ by OFLR using $\mathcal{D}^{(t)}$ and obtain truncated RKHS norm $\hat{\Delta}_t = \|\hat{\beta}_0 - \hat{\beta}_t\|_{K^M} := \sum_{j=1}^{M} \langle \hat{\beta}_0 - \hat{\beta}_t, v_j \rangle^2 / \tau_j$.
3. Set $\hat{\mathcal{S}}_t = \left\{ k : \hat{\Delta}_k \text{ is among the first } t \text{ smallest.} \right\}$

Step 3: For $t \in [T]$, fit TL-FLR by setting $\mathcal{S} = \hat{\mathcal{S}}_t$ with dataset $\mathcal{D}_{\mathcal{I}}^{(0)}$. Let $\mathcal{F} = \{\hat{\beta}(\hat{\mathcal{S}}_0), \hat{\beta}(\hat{\mathcal{S}}_1), \cdots \hat{\beta}(\hat{\mathcal{S}}_T)\}$.

Step 4: Implement the sparse aggregation procedure in Algoritm 3 with $\mathcal{F}$ as the dictionary and training dataset as $\mathcal{D}_{\mathcal{I}^c}^{(0)}$. Obtain the sparse aggregated estimator $\hat{\beta}_a$.

Remark 3.1. While exploring the estimated similarity across sources to the target in Step 2, we use a truncated RKHS norm, which is the distance between $\hat{\beta}_0$ and $\hat{\beta}_t$ after projecting them onto the space spanned by the first M eigenfunctions of K. Here, $\{\tau_j\}_{j \ge 1}$ and $\{v_j\}_{j \ge 1}$ are the eigenvalues and eigenfunctions of K. Such a truncated norm guarantees the identifiability of $\mathcal{S}$, see Section 4.2 for detail.

Step 2 ensures the target-only baseline $\hat{\beta}_0$ lies in $\mathcal{F}$ while the construction of $\hat{\mathcal{S}}_t$ ensures thorough exploration of $\mathcal{S}$. If $\mathcal{S}$ can be identified by one of the $\hat{\mathcal{S}}_t$, then inequality (5) indicates that even without knowing $\mathcal{S}$, the $\hat{\beta}_a$ can mimic the performance of the TL-FLR estimator, while not being worse than the target-only $\hat{\beta}_0$, up to an aggregation cost.

The sparse aggregation is adopted from Gaîffas & Lecué (2011), see Appendix C. Although we note that other aggregation methods like aggregate with cumulated exponential weights (ACEW) (Juditsky et al., 2008; Audibert, 2009), aggregate with exponential weights (AEW) (Leung & Barron, 2006; Dalalyan & Tsybakov, 2007), and Q-aggregation (Dai et al., 2012) can replace sparse aggregation in Step 4, sparse aggregation is often preferred due to its computational efficiency and ability to eliminate negative transfer effects. Specifically, the final aggregated estimator is usually represented as a convex combination of elements in $\mathcal{F}$ i.e., $\hat{\beta}_a = \sum_{j=1}^{J} c_j \hat{\beta}(\hat{\mathcal{S}}_j)$. The sparse aggregation sets most of the c_j to zero, which effectively excludes the negative transfer sources. On the other hand, none of the ACEW, AEW, and Q-aggregation will set the c_j to 0 most of the time, meaning that negative transfer sources can still affect $\hat{\beta}_a$. Although one can manually tune temperature parameters in these approaches to shrink the c_j close to zero, they are less computationally efficient given the fact that sparse aggregation does not require such a tuning process. In Section 6, we verify that sparse aggregation outperforms other aggregation methods under various settings.

4. Theoretical Analysis

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

$$\mathcal{E}(\hat{\beta}^{(0)}) := \mathrm{E}_{Z^{(0)}}\left[\ell(\hat{\beta}^{(0)}, Z^{(0)}) - \ell(\beta^{(0)}, Z^{(0)})\right]$$

where the expectation is taken over an independent test data point $Z^{(0)}$ from the target distribution. To study the excess risk of TL-FLR and ATL-FLR, we denote $\beta_{\mathcal{S}}$ the population version of $\hat{\beta}_{\mathcal{S}}$ which also lies in $\mathcal{H}_K$ and define the parameter space as

$$\Theta(h, R) = \left\{\{\beta^{(t)}\}_{t\in\{0\}\cup\mathcal{S}} : \|\beta_{\mathcal{S}}\|_K \le R, \|\delta^{(t)}\|_K \le h\right\}.$$

To establish the theoretical analysis of the proposed algorithms, we first state some assumptions. For $t \in 0 \cup [T]$, denote $\{s_j^{(t)}\}_{j\ge1}$ and $\{\phi_j^{(t)}\}_{j\ge1}$ as the eigenvalues and eigenfunctions of $\Gamma^{(t)} := K^{\frac{1}{2}} C^{(t)} K^{\frac{1}{2}}$ respectively.

Assumption 4.1 (Eigenvalue Decay Rate (EDR))**.** Suppose that the eigenvalue decay rate (EDR) of $L_{\Gamma^{(0)}}$ is $2r$, i.e.

$$s_j^{(0)} \asymp j^{-2r}, \quad \forall j \ge 1.$$

The polynomial EDR assumption is standard in FLR literature like Cai & Yuan (2012); Reimherr et al. (2018). RKHSs that satisfy this assumption, like Sobolev spaces, are natural choices when considering smoothness as the structural properties in the TL processes. For target-only learning with RERM (Cai & Yuan, 2012), the minimax convergence rate of the excess risk is $n_0^{-2r/(2r+1)}$.

Assumption 4.2. We assume either one of the following conditions holds.

1. $L_{\Gamma^{(t)}}$ commutes with $L_{\Gamma^{(0)}}$, $\forall t \in \mathcal{S}$, i.e. $L_{\Gamma^{(0)}} L_{\Gamma^{(t)}} = L_{\Gamma^{(t)}} L_{\Gamma^{(0)}}$, and
$$a_j^{(t)} := \langle L_{\Gamma^{(t)}}(\phi_j^{(0)}), \phi_j^{(0)}\rangle \asymp s_j^{(0)} \quad \forall j \ge 1.$$
2. Or the following linear operator is Hilbert–Schmidt.
$$\mathbf{I} - (L_{\Gamma^{(0)}})^{-1/2} L_{\Gamma^{(t)}} (L_{\Gamma^{(0)}})^{-1/2}, \quad \forall t \in \mathcal{S}.$$

We note that under the posterior drift setting, both conditions in Assumption 4.2 are satisfied automatically, and thus, our theoretical results are built on assumptions that are more relaxed than posterior drift. Although neither condition implies the other, both conditions primarily focus on how the smoothness of the source kernel $\Gamma^{(t)}$ relates to that of the target kernel $\Gamma^{(0)}$. Specifically, Condition 1 implies $L_{\Gamma^{(0)}}$ and $L_{\Gamma^{(t)}}$ not only share the same eigenspace but also have similar magnitudes of the projection onto the j-th dimension of the eigenspace, which commonly appears in FDA literature (Yuan & Cai, 2010; Balasubramanian et al., 2022). Condition 2 implies the probability measures of $X^{(0)}$ and $X^{(t)}$ are equivalent, see Baker (1973). Collectively, both conditions indicate the feasibility of OTL for FLR relies on the fact that the regularity of source operator $L_{\Gamma^{(t)}}$ should behave similarly to the target's. Either a too "smooth" or a too "rough" source can degrade the optimality. Besides, these conditions help to prevent the excess risk of $\hat{\beta}_S$ over the target domain from diverging, and we refer readers to Appendix A for a technical discussion.

4.1. Minimax Excess Risk of TL-FLR

We first provide the upper bound of excess risk on TL-FLR.

Theorem 4.3 (Upper Bound)**.** *Suppose Assumption 4.2 and 4.1 hold. If $n_0/n_{\mathcal{S}} \nrightarrow 0$, let $\xi(h,R) = \frac{h^2}{R^2}$, then for the output $\hat{\beta}$ of Algorithm 1,*

$$\sup_{\Theta(h,R)} \mathcal{E}\left(\hat{\beta}\right) = O_{\mathbb{P}}\left(n_{\mathcal{S}}^{-\frac{2r}{2r+1}} + n_0^{-\frac{2r}{2r+1}}\xi(h,R)\right), \quad (6)$$

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

Theorem 4.3 provides the excess risk upper bound of $\hat{\beta}$, which bounds the excess risk of two terms. The first term comes from the transfer step and depends on the sample size of sources in $\mathcal{S}$, while the second term is due to only using the target dataset to learn the offset. In the trivial case when $\mathcal{S} = \emptyset$, the upper bound becomes $O_{\mathbb{P}}(n_0^{-2r/(2r+1)})$, which coincides with the upper bound of target-only baseline OFLR (Cai & Yuan, 2012). When $\mathcal{S} \neq \emptyset$, compared with the excess risk of the target-only baseline, we can see the sample size $n_{\mathcal{S}}$ in source models and the factor $\xi(h, R)$ are jointly affecting the transfer learning. The factor $\xi(h, R)$ represents the relative task signal strength between the source and target tasks. Geometrically, one can interpret $\xi(h, R)$ as the factor that roughly controls the angle between the source and target models within the RKHS.

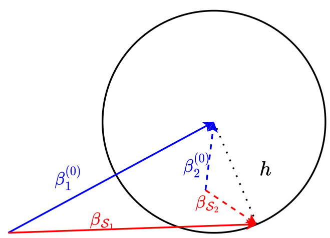


Figure 1. Geometric illustration for how $\xi(h, R)$ will affect the transfer dynamic. The circle represents an RKHS ball centered around $\beta^{(0)}$ with radius h. With the same h, larger signal strength of $\beta_{\mathcal{S}}$, i.e. $\|\beta_{\mathcal{S}_1}\|_K$ leads to smaller $\xi(h, R)$, while smaller signal strength of $\beta_{\mathcal{S}}$, i.e. $\|\beta_{\mathcal{S}_2}\|_K$ leads to larger $\xi(h, R)$.

Figure 1 shows how $n_{\mathcal{S}}$ and $\xi(h, R)$ impact the learning rate. When the $\beta_{\mathcal{S}}$ and $\beta^{(0)}$ are more concordant ($\beta_{\mathcal{S}_1}$ and $\beta_1^{(0)}$), the angle between them are small and thus so the $\xi(h, R)$, making the second term in the upper bound negligible in the excess risk, and thus the risk converges faster compared to baseline $n_0^{-2r/(2r+1)}$ given sufficiently large $n_{\mathcal{S}}$. If $\beta_{\mathcal{S}}$ and $\beta^{(0)}$ are less concordant ($\beta_{\mathcal{S}_2}$ and $\beta_2^{(0)}$), leveraging $\beta_{\mathcal{S}}$ will be less effective since a large $\xi(h, R)$ will make the second term the dominate term.

It is worth noting that most of the existing literature fails to identify how $\xi(h, R)$ affects the effectiveness of OTL. For example, in Wang et al. (2016); Du et al. (2017), this factor does not appear in the upper bound, and they claim $n_{\mathcal{S}} \gg n_0$ provide successful transfer from source to target. In high-dimensional linear regression (Li et al., 2022; Tian & Feng, 2022), the authors only identify $\xi(h, R)$ is proportional to h and claims a small h can provide a faster convergence rate excess risk. However, our analysis (Figure 1) shows even with the same h, the similarity of the two tasks can be different given different signal strengths of $\beta_{\mathcal{S}}$, which will also affect the effectiveness of OTL. To this end, this reveals that one cannot obtain a faster excess risk in OTL by simply including more source datasets (larger $n_{\mathcal{S}}$), but should also carefully select or construct the $\mathcal{S}$, i.e., the more source data are available, the more strict one should be with what sources one uses to build $\hat{\beta}_{\mathcal{S}}$ in the OTL framework.

Theorem 4.4 (Lower Bound)**.** *Under the same condition of Theorem 4.3, for any possible estimator $\tilde{\beta}$ based on $\{\mathcal{D}^{(t)} : t \in \{0\} \cup \mathcal{S}\}$, the excess risk of $\tilde{\beta}$ satisfies*

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

Combining Theorems 4.3 and 4.4 implies that the estimator from TL-FLR is rate-optimal in excess risk. The proof of the lower bound is based on considering the lower bound of two cases: (1) the ideal case where $\beta^{(t)} = \beta^{(0)}$ for all $t \in \mathcal{S}$ and (2) the worst case where $\beta^{(t)} \equiv 0$, meaning no knowledge should be transferred at all.

4.2. Excess Risk of ATL-FLR

In this subsection, we study the excess risk for ATL-FLR. As we discussed before, making ATL-FLR achieve similar performance to TL-FLR relies on the fact that there exists a $\hat{\mathcal{S}}_t$ such that it equals to the true $\mathcal{S}$ (so $\hat{\beta}(\hat{\mathcal{S}}_t) = \hat{\beta}(\mathcal{S})$) with high probability. Therefore, to ensure the $\mathcal{F}$ constructed in Step 2 of Algorithm 2 satisfies such a property, we impose the following assumption to guarantee the identifiability of $\mathcal{S}$ and thus ensure the existence of such $\hat{\mathcal{S}}_t$.

Assumption 4.5 (Identifiability of $\mathcal{S}$)**.** Suppose for any h, there is an integer M such that

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

where $\|\cdot\|_{K^M}$ is the truncated version of $\|\cdot\|_K$ defined in Algorithm 2.

Assumption 4.5 ensures that $\forall t \in \mathcal{S}^c$, there exists a finite-dimensional subspace of $\mathcal{H}_K$, such that the norm of the projection of the contrast function, $\delta^{(t)}$, on this subspace is already greater than h. This assumption indeed eliminates the existence of $\beta^{(t)}$, for $t \in \mathcal{S}^c$, that live on the boundary of the RKHS-ball centered at $\beta^{(0)}$ with radius h in $\mathcal{H}_K$. Under Assumption 4.5, we now show the $\mathcal{F}$ constructed in Algorithm 2 guarantees the existence of $\hat{\mathcal{S}}_t$.

Theorem 4.6. *Suppose Assumption 4.5 holds, then*

$$\max_{t\in\mathcal{S}} \Delta_t < \min_{t\in\mathcal{S}^c} \Delta_t \quad \text{and} \quad \mathbb{P}\left(\max_{t\in\mathcal{S}} \hat{\Delta}_t < \min_{t\in\mathcal{S}^c} \hat{\Delta}_t\right) \to 1,$$

and hence there exists a t s.t. $\hat{\mathcal{S}}_t \in \mathcal{F}$ and

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

Remark 4.7. Assumption 4.5 ensures a sufficient gap between those Δ_t that belong to $\mathcal{S}$ and those that don't, which ensures their estimated counterpart will also possess this gap with high probability, making one of the $\hat{\mathcal{S}}$ consistent with $\mathcal{S}$.

With Proposition D.1, which states the cost of sparse aggregation in Appendix D.5, and the excess risk of TL-FLR in Theorem 4.6, we can establish the excess risk for ATL-FLR.

Theorem 4.8 (Upper Bound of ATL-FLR)**.** *Let $\hat{\beta}_a$ be the output of Algorithm 2, then under the same settings of Theorem 4.3 and Assumption 4.5,*

$$\sup_{\Theta(h,R)} \mathcal{E}\left(\hat{\beta}_a\right) = O_{\mathbb{P}}\Bigg(\underbrace{n_{\mathcal{S}}^{-\frac{2r}{2r+1}} + n_0^{-\frac{2r}{2r+1}}\xi(h,R)}_{\text{transfer learning risk}} + \underbrace{\frac{log(T)log(n_0)}{n_0}}_{\text{aggregation cost}}\Bigg).$$

One interesting note is that the transfer learning risk is the classical nonparametric rate while the aggregation cost is parametric (or nearly parametric). Therefore, the aggregation cost usually decays substantially faster than the transfer learning risk. However, in the high-dimensional linear regression TL, such a faster-decayed aggregation cost is diminished since the transfer learning risk is also parametric, see (Li et al., 2022).

5. Extension to Functional Generalized Linear Models

In this section, we show our approaches in the FLR model can be naturally extended to the functional generalized linear model (FGLM) settings, which includes wider application scenarios like classification. To start, consider the following series of FGLM models similar to the FLR setting (1),

$$\mathbb{P}(Y_i^{(t)}|X_i^{(t)}) = \rho(Y_i^{(t)}) \exp\left\{\frac{Y_i^{(t)}\eta(\theta_i^{(t)}) - \psi(\theta_i^{(t)})}{d(\tau)}\right\},$$

where $i \in [n_t]$ and $t \in [T]$, $\theta_i^{(t)} = \langle X_i^{(t)}, \beta^{(t)}\rangle_{L^2}$ is the canonical parameter. The functions ρ, η, ψ, d are known, and τ is either known or out-of-interest parameter that is independent of $X^{(t)}$. In this paper, we consider η to take the canonical form, i.e., $\eta(x) = x$. The GLMs are characterized by the different ψ. For example, in linear regression with Gaussian response, $\psi(x) = x^2/2$; in the logistic regression with binary response, $\psi(x) = log(1+e^x)$; and in Poisson regression with non-negative integer response, $\psi(x) = e^x$.

A standard method for addressing GLM involves minimizing the loss function defined as the negative log-likelihood. Therefore, to implement the transfer learning for FGLM, one can naturally substitute the square loss in TL-FLR (Algorithm 1) and ATL-FLR (Algorithm 2) with the negative log-likelihood loss, i.e.

$$\ell(\beta, Z_i^{(t)}) = -Y_i^{(t)}\eta(\theta_i^{(t)}) + \psi(\theta_i^{(t)}).$$

We refer to these transfer learning algorithms for FGLM as TL-FGLM and ATL-FGLM. To establish the optimality of TL-FGLM and ATL-FGLM, the following technical assumptions are required.

Assumption 5.1. Assume ψ is Lipschitz continuous on its domain, and $\psi' < \infty$.

Assumption 5.2. Assume there exist constants $0 < \mathcal{A}_1 \leq \mathcal{A}_2 < \infty$ such that the function ψ'' satisfies

$$\mathcal{A}_1 \leq \inf_{s\in\mathcal{T}} \psi''(s) \leq \psi''(s) \leq \sup_{s\in\mathcal{T}} \psi''(s) \leq \mathcal{A}_2.$$

Assumption 5.1 is natural in most GLM literature and is satisfied by many popular exponential families. Assumption 5.2 restricts the ψ'' in the bounded region and thus restricts the variance of y.

Since the conditional mean for FGLM is $\mathrm{E}[Y_i|X_i] = \eta'(\langle\beta, X^{(0)}\rangle_{L^2})$, we evaluate the accuracy by excess risk, i.e. $\mathcal{E}(\hat{\beta}) := \mathrm{E}_{X^{(0)}}[\eta'(\langle\hat{\beta}, X^{(0)}\rangle_{L^2}) - \eta'(\langle\beta^{(0)}, X^{(0)}\rangle_{L^2})]^2$.

Theorem 5.3. *Under the same assumption of Theorem 4.3 and suppose Assumptions 5.1, 5.2 holds.*

1. *(Lower Bound) For any possible estimator $\tilde{\beta}$ based on target and source datasets, the excess risk of $\tilde{\beta}$ satisfies*

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

2. *(Upper Bound) If $n_0/n_{\mathcal{S}} \nrightarrow 0$ and $\lambda_1 \asymp n_{\mathcal{S}}^{-\frac{2r}{2r+1}}$ and $\lambda_2 \asymp n_0^{-\frac{2r}{2r+1}}$, then for the output $\hat{\beta}$ of TL-FGLM,*

$$\sup_{\Theta(h,R)} \mathcal{E}\left(\hat{\beta}\right) = O_{\mathbb{P}}\left(n_{\mathcal{S}}^{-\frac{2r}{2r+1}} + n_0^{-\frac{2r}{2r+1}}\xi(h,R)\right).$$

Remark 5.4. The error bound of TL-FLR and TL-FGLM are the same, which is consistent with the case in the target-only learning between FLR and FGLM, see Cai & Yuan (2012); Du & Wang (2014). However, we note the proof is not a trivial extension of FLR since minimizing the regularized negative likelihood usually will not provide an analytical solution.

Remark 5.5. Due to the same upper bound for TL-FLR and TL-FGLM, the upper bound of ATL-FGLM is the same as ATL-TLR, i.e., with the same aggregation cost.

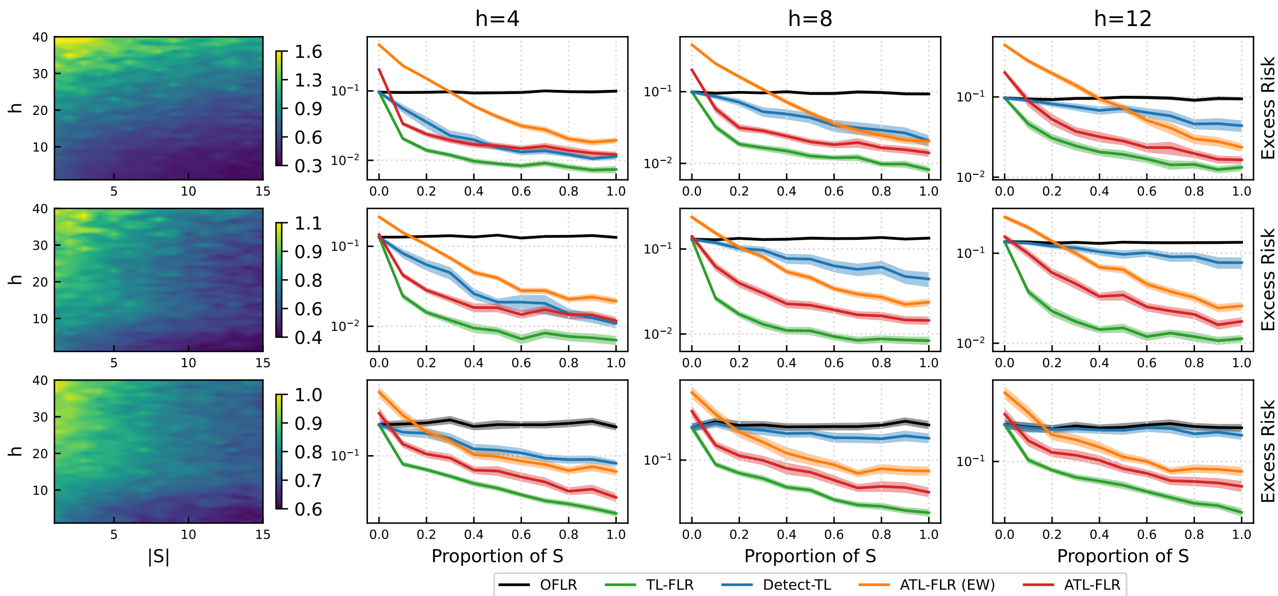


Figure 2. Left panel (Heatmap): log relative excess risk of TL-FLR to OFLR. Right panels (Line Chart): Excess Risk of different transfer learning algorithms. Each row corresponds to a $\beta^{(0)}$, and the y-axes for each row are under the same scale. The result for each sample size is an average of 100 replicate experiments with the shaded area indicating $\pm$ 2 standard error.

6. Experiments

We illustrate our algorithms for FLR by providing results using simulated data and defer two real-world applications, including financial market data application for FLR and wearable device detection application for FGLM, to Appendix F [1]. We consider the following algorithms: OFLR, TL-FLR, ATL-FLR, Detection Transfer Learning (Detect-TL) (Tian & Feng, 2022) and Exponential Weighted ATL-FLR (ATL-FLR (EW)) Li et al. (2022). To set up the RKHS, we consider the setting in Cai & Yuan (2012). Let $\psi_k(s) = \sqrt{2}\cos(\pi k s)$ for $j \geq 1$ and define the reproducing kernel K of $\mathcal{H}_K$ as $K(\cdot,\cdot) = \sum_{k=1}^{\infty} k^{-2}\psi_k(\cdot)\psi_k(\cdot)$.

For the target model, $\beta^{(0)}(s)$ is set to be (1) $\beta_1^{(0)}(s) = \sum_{k=1}^{\infty} 4\sqrt{2}(-1)^{k-1}k^{-2}\psi_k(s)$; (2) $\beta_2^{(0)}(s) = 4\cos(3\pi s)$; (3) $\beta_3^{(0)}(s) = 4\cos(3\pi s) + 4\sin(3\pi s)$. For a specific h, let $\mathcal{S} = \{l : \|\beta_0 - \beta_t\|_K \leq h\}$, then we generate source models as follows. We scale each target model such that their RKHS norm is 20. If $t \in \mathcal{S}$, then $\beta_t(t)$ is set to be $\beta_t(s) = \beta_0(s) + \sum_{k=1}^{\infty} \left(\mathcal{U}_k(\sqrt{12}h/\pi k^2)\right)\psi_k(s)$ with $\mathcal{U}_k$'s i.i.d. uniform random variable on $[-1,1]$. If $t \in \mathcal{S}^c$, then $\beta_t(s)$ is generated from a Gaussian process with mean function $\cos(2\pi s)$ with kernel $\exp(-15|s-t|)$ as covariance kernel. The predictors $X^{(t)}$ are i.i.d. generated from a Gaussian process with the mean function $\sin(\pi s)$ and the covariance kernels are set to be Matérn kernel $C_{\nu,\rho}$ (Cressie & Huang, 1999) where the parameter ν controls the smoothness of $X^{(t)}$. We set the covariance kernel of $X^{(t)}$ as $C_{1/2,1}$ for the target tasks and $C_{3/2,1}$ for source tasks, to fulfill Assumptions 4.2. We note that such a setting is more challenging than assuming that the target and source tasks have the same covariance kernel. All functions are generated on $[0,1]$ with 50 evenly spaced points and we set $n_0 = 150$ and $n_t = 100$. For each algorithm, we set the regularization parameters as λ_1 and λ_2 as the optimal values in Theorem 4.3 and select the constants using cross-validation. The excess risk for the target tasks is calculated via the Monte-Carlo method by using 1000 newly generated predictors $X^{(0)}$.

In the left panel of Figure 2 we compare TL-FLR with OFLR by considering *relative excess risk*, i.e. the ratio of TL-FLR's excess risk to OFLR's. We note that since RKHS of $\beta^{(0)}$ is fixed, the magnitude of h is proportional to $\xi(h, \beta_{\mathcal{S}})$ and a large h indicates less similarity between sources and target tasks. Overall, the effectiveness of TL-FLR for different $\beta^{(0)}$ presents a consistent pattern, i.e, with more transferable sources involved and smaller h (bottom right), TL-FLR has a more significant improvement, while with fewer sources and larger h (top left), the transfer may be worse than OFLR.

In the right panel of Figure 2, we evaluate ATL-FLR under unknown $\mathcal{S}$ cases. We set $\mathcal{S}$ as a random subset of $\{1, 2, \cdots, 20\}$ such that $|\mathcal{S}|$ is equal to $0, 2, \cdots, 20$. We also implement TL-FLR by using true $\mathcal{S}$ and OFLR as

[1]The R code and the application datasets are available in https://github.com/haotianlin/HTL-FLM.

baseline. In all scenarios, ATL-FLR outperforms all its competitors. Comparing ATL-FLR with ATL-FLR(EW), even though ATL-FLR(EW) has similar patterns as ATL-FLR, we can see the gaps between the two curves are larger when the proportion of $\mathcal{S}$ is small, showing that ATL-FLR(EW) is more sensitive to source tasks in $\mathcal{S}^c$, while ATL-FLR is less affected. Detect-TL only has a considerable reduction on the excess risk with relatively small h, but provides limited improvement when h is large, indicating its limited performance when limited knowledge is available in sources.

7. Discussion

We conclude by summarizing our results and discussing potential future research directions.

Summary of Results. This paper studies transfer learning under the functional linear model framework, including FLR and FGLM. We derive the asymptotic rates for the excess risk over the target domain and show a faster statistical rate depending on both source sample size and the magnitude of similarity across tasks. Our theoretical analysis helps researchers better understand the transfer dynamic of OTL. Moreover, we leverage the sparse aggregation to implement the transfer practically, alleviate the negative transfer effect, and achieve nearly optimal statistical rates.

Future Directions. We discuss two potential future directions of this work.

(1) In the current analysis, we assume that the offset slope function resides in $\mathcal{H}_K$ and shares the same smoothness as the target and source functions. This assumption is reasonable given the unknown true smoothness of the offset slope function. However, the success of OTL hinges on the offset slope function possessing a simpler structure (well-regularized) that can be effectively learned from a small sample size. Consequently, a critical open question emerges: if the offset slope function exhibits higher smoothness, how do we identify the different smoothness for the source and offset slope functions and subsequently apply the appropriate kernel to achieve optimal statistical rates? Recently, Lin & Reimherr (2024) explored this issue within the nonparametric regression setting, identifying the Gaussian kernel as a universal solution to achieve adaptive OTL under varying smoothness scenarios. Although their findings are specific to Sobolev spaces, it is worth investigating whether a similar solution exists for FLR and FGLM since the kernel in these contexts is a composition of the covariance kernel and the RKHS kernel.

(2) When constructing the source hypothesis $\hat{\beta}_{\mathcal{S}}$, Algorithm 1 employs a data fusion technique in the transfer step. Specifically, it consolidates all the source datasets in $\mathcal{S}$ and simultaneously trains a unified $\hat{\beta}\mathcal{S}$. The time and space complexities for this step are $O(n_{\mathcal{S}}^3)$ and $O(n_{\mathcal{S}}^2)$, respectively. In real-world transfer learning scenarios, the involvement of numerous large sample-size source datasets can lead to significant algorithmic scaling issues, thereby limiting the practicality of the transfer step and the proposed algorithm. A potential solution is to perform the learning in a distributed manner (Tong, 2021; Liu & Shi, 2022). Additionally, if the learned source models for each source domain are available, one might consider fusing these learned hypotheses instead of the raw data.

Impact Statement

This paper aims to study the minimax asymptotic rates for hypothesis transfer learning under the framework of (generalized) functional linear regression, and the goal is to advance the theoretical understanding of transfer learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Aghbalou, A. and Staerman, G. Hypothesis transfer learning with surrogate classification losses: Generalization bounds through algorithmic stability. In *International Conference on Machine Learning*, pp. 280–303. PMLR, 2023.

Anguita, D., Ghio, A., Oneto, L., Parra Perez, X., and Reyes Ortiz, J. L. A public domain dataset for human activity recognition using smartphones. In *Proceedings of the 21th international European symposium on artificial neural networks, computational intelligence and machine learning*, pp. 437–442, 2013.

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

Baker, C. R. On equivalence of probability measures. *The Annals of Probability*, pp. 690–698, 1973.

Balasubramanian, K., Müller, H.-G., and Sriperumbudur, B. K. Unified rkhs methodology and analysis for functional linear and single-index models. *arXiv preprint arXiv:2206.03975*, 2022.

Balcan, M.-F., Khodak, M., and Talwalkar, A. Provable guarantees for gradient-based meta-learning. In *International Conference on Machine Learning*, pp. 424–433. PMLR, 2019.

Bastani, H. Predicting with proxies: Transfer learning in high dimension. *Management Science*, 67(5):2964–2984, 2021.

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

Cardot, H., Ferraty, F., and Sarda, P. Functional linear model. *Statistics & Probability Letters*, 45(1):11–22, 1999.

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

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

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

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

Denevi, G., Ciliberto, C., Stamos, D., and Pontil, M. Learning to learn around a common mean. *Advances in neural information processing systems*, 31, 2018.

Denevi, G., Ciliberto, C., Grazzi, R., and Pontil, M. Learning-to-learn stochastic gradient descent with biased regularization. In *International Conference on Machine Learning*, pp. 1566–1575. PMLR, 2019.

Du, P. and Wang, X. Penalized likelihood functional regression. *Statistica Sinica*, pp. 1017–1041, 2014.

Du, S. S., Koushik, J., Singh, A., and Póczos, B. Hypothesis transfer learning via transformation functions. *Advances in neural information processing systems*, 30, 2017.

Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., and Lei, Q. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

Gaîffas, S. and Lecué, G. Hyper-sparse optimal aggregation. *The Journal of Machine Learning Research*, 12:1813–1833, 2011.

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

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

He, Z., Sun, Y., and Li, R. Transfusion: Covariate-shift robust transfer learning for high-dimensional regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 703–711. PMLR, 2024.

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

Kokoszka, P. and Reimherr, M. *Introduction to functional data analysis*. Chapman and Hall/CRC, 2017.

Kuzborskij, I. and Orabona, F. Stability and hypothesis transfer learning. In *International Conference on Machine Learning*, pp. 942–950. PMLR, 2013.

Kuzborskij, I. and Orabona, F. Fast rates by transferring from auxiliary hypotheses. *Machine Learning*, 106:171–195, 2017.

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

Li, S., Cai, T. T., and Li, H. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):149–173, 2022.

Li, X. and Bilmes, J. A bayesian divergence prior for classiffier adaptation. In *Artificial Intelligence and Statistics*, pp. 275–282. PMLR, 2007.

Lin, H. and Reimherr, M. Smoothness adaptive hypothesis transfer learning. *arXiv preprint arXiv:2402.14966*, 2024.

Liu, J. and Shi, L. Statistical optimality of divide and conquer kernel-based functional linear regression. *arXiv preprint arXiv:2211.10968*, 2022.

Orabona, F., Castellini, C., Caputo, B., Fiorilla, A. E., and Sandini, G. Model adaptation with least-squares svm for adaptive hand prosthetics. In *2009 IEEE international conference on robotics and automation*, pp. 2897–2903. IEEE, 2009.

Perrot, M. and Habrard, A. A theoretical analysis of metric hypothesis transfer learning. In *International Conference on Machine Learning*, pp. 1708–1717. PMLR, 2015.

Qu, S., Wang, J.-L., and Wang, X. Optimal estimation for the functional cox model. *The Annals of Statistics*, 44(4): 1708–1738, 2016.

Ramsay, J., Ramsay, J., Silverman, B., et al. *Functional Data Analysis*. Springer Science & Business Media, 2005.

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

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

Tian, Y. and Feng, Y. Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, pp. 1–14, 2022.

Tian, Y., Weng, H., and Feng, Y. Unsupervised multi-task and transfer learning on gaussian mixture models. *arXiv preprint arXiv:2209.15224*, 2022.

Tian, Y., Gu, Y., and Feng, Y. Learning from similar linear representations: adaptivity, minimaxity, and robustness. *arXiv preprint arXiv:2303.17765*, 2023.

Tong, H. Distributed least squares prediction for functional linear regression. *Inverse Problems*, 38(2):025002, 2021.

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

Tripuraneni, N., Jordan, M., and Jin, C. On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, 33: 7852–7862, 2020.

Vaart, A. W. and Wellner, J. A. Weak convergence. In *Weak convergence and empirical processes*, pp. 16–28. Springer, 1996.

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

Wang, X. and Schneider, J. G. Generalization bounds for transfer learning under model shift. In *UAI*, pp. 922–931, 2015.

Wang, X., Oliva, J. B., Schneider, J. G., and Póczos, B. Nonparametric risk and stability analysis for multi-task learning problems. In *IJCAI*, pp. 2146–2152, 2016.

Wendland, H. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

Xu, Z. and Tewari, A. Representation learning beyond linear prediction functions. *Advances in Neural Information Processing Systems*, 34:4792–4804, 2021.

Yao, F., Müller, H.-G., and Wang, J.-L. Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6):2873–2903, 2005.

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

Zhu, J., Guha, A., Do, D., Xu, M., Nguyen, X., and Zhao, D. Functional optimal transport: map estimation and domain adaptation for functional data. *arXiv preprint arXiv:2102.03895*, 2021.

Contents of Appendix

A. Appendix: Additional Remarks

Eigenspace Assumption Discussion. While we have discussed Assumption 4.2 intuitively in Section 4, we would like to discuss why the assumption is taking such a complex form and how it contributes to our theoretical analysis technically. A technical reason behind such a complex form (an asymptotic behavior between the eigenvalue of $\Gamma^{(0)}$, $s_j^{(0)}$, and the projection of $\Gamma^{(0)}$ onto the eigenspace of $\Gamma^{(0)}$, $a_j^{(t)}$) is due to the natural difficulty of the functional data problem as one needs to handle quantities in infinite dimensions. Specifically, we are dealing with vectors in $\mathbb{R}^\infty$ when bounding the approximation error in the excess risk of the learned source model on the target domain, i.e., $\mathrm{E}\langle X^{(0)}, \hat{\beta}_S - \beta_S\rangle_{L^2}^2$. The constant term for the approximation error will be affected by the maximum ratio of $s_j^{(0)}$ and $a_j^{(t)}$ overall dimensions, i.e., $\forall j \geq 1$, and Assumption 4.2 indeed gives an appropriate control about this maximum ratio to avoid an exploded approximation error. In the multivariate case, this is not a problem as these quantities are more easily manipulated. It's crucial to highlight that this unique challenge arises when dealing with functional objects due to their infinite-dimensional nature.

The Role of Smoothness of Offset. As we have discussed in 7, in this paper, we assume all the slope functions reside in $\mathcal{H}_K$, including source, target, and offset. This means they all possess the same smoothness, i.e., the eigenvalue decay rate. In the theoretical analysis, the smoothness of δ affects the excess risk in the exponential part of the offset error, i.e., $n_0^{-2r/2r+1}$. While assuming $\delta \in \mathcal{H}_K$ is a safe assumption without knowing its true smoothness, one can provide a fine-grained analysis when the regularity of δ is known. Consider δ lies in a subspace of $\mathcal{H}_K$, namely $\mathcal{H}_{K_1}$, with higher smoothness (thus it is even easier to learn). Suppose the $L_{\Gamma_1^{(0)}} = L_{K_1^{1/2} C^{(0)} K_1^{1/2}}$ and the EDR of $L_{\Gamma_1^{(0)}}$ is $2r_1$ with $r_1 > r$. Then the excess risk in Theorem 4.3 becomes $n_S^{-2r/2r+1} + n_0^{-2r_1/2r_1+1}\xi(h, \mathcal{S})$. It provides a slightly faster rate than the one in Theorem 4.3. However, the challenge for achieving such a rate is identifying the subspace where δ lies and selecting the correct kernel in the calibration step. Recently, in the context of nonparametric regression, Lin & Reimherr (2024) considered the source function resides in Sobolev space H^{m_0} while offset function resides in H^m with $m \geq m_0$. They proposed using Gaussian kernels and a training-validation approach to adaptively learn both functions with unknown Sobolev smoothness m_0 and m. However, we note that this is not a trivial extension as in the nonparametric regression context, one only needs to be concerned with the employed RKHS kernel, while in the context of functional data, the kernel we are concerned with is the composition of covariance kernel of X and the RKHS kernel.

B. Appendix: Background of RKHS and Integral Operators

In this section, we will present some facts about the RKHS and also the integral operator of a kernel that are useful in our proof and refer readers to Wendland (2004) for a more detailed discussion.

Let $\mathcal{T}$ be a compact set of $\mathbb{R}$. For a real, symmetric, square-integrable, and semi-positive definite kernel $K : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$, we denote its associated RKHS as $\mathcal{H}_K$. For the reproducing kernel K, we can define its integral operator $L_K : L^2 \to L^2$ as

$$L_K(f)(\cdot) = \int_{\mathcal{T}} K(s, \cdot) f(s) ds.$$

L_K is self-adjoint, positive-definite, and trace class (thus Hilbert-Schmidt and compact). By the spectral theorem for self-adjoint compact operators, there exists an at most countable index set N, a non-increasing summable positive sequence $\{\tau_j\}_{j\geq 1}$ and an orthonormal basis of L^2, $\{e_j\}_{j\geq 1}$ such that the integrable operator can be expressed as

$$L_K(\cdot) = \sum_{j\in N} \tau_j \langle \cdot, e_j\rangle_{L^2} e_j.$$

The sequence $\{\tau_j\}_{j\geq 1}$ and the basis $\{e_j\}_{j\geq 1}$ are referred as the eigenvalues and eigenfunctions. The Mercer's theorem shows that the kernel K itself can be expressed as

$$K(x, x') = \sum_{j\in N} \tau_j e_j(x) e_j(x'), \quad \forall x, x' \in \mathcal{T},$$

where the convergence is absolute and uniform.

We now introduce the fractional power integral operator and the composite integral operator of two kernels. For any $s \geq 0$, the fractional power integral operator $L_{K^s} : L^2 \to L^2$ is defined as

$$L_{K^s}(\cdot) = \sum_{j\in N} \tau_j^s \langle \cdot, e_j\rangle_{L^2} e_j.$$

For two kernels K_1 and K_2, we define their composite kernel as

$$(K_1K_2)(x, x') = \int_{\mathcal{T}} K_1(x, s) K_2(s, x') ds,$$

and thus $L_{K_2K_2} = L_{K_1} \circ L_{K_2}$. Given these definitions, for a given reproducing kernel K and covariance function C, the definition of Γ in the main paper is

$$L_\Gamma = L_{K^{\frac{1}{2}}} \circ L_C \circ L_{K^{\frac{1}{2}}} \quad \text{and} \quad \Gamma := K^{\frac{1}{2}} C K^{\frac{1}{2}}.$$

If both $L_{K^{\frac{1}{2}}}$ and L_C are bounded linear operators, the spectral algorithm guarantees the existence of eigenvalues $\{s_j\}_{j\geq 1}$ and eigenfunctions $\{\psi_j\}_{j\geq 1}$.

C. Appendix: Sparse Aggregation Process

We provide the procedure of sparse aggregation in Step 4 of ATL-FLR (Algorithm 2) for completeness and refer readers to Gaîffas & Lecué (2011) for more detail.

Algorithm 3 Sparse Aggregation (Gaîffas & Lecué, 2011)

1: **Input:** The candidate set $\mathcal{F}$; target dataset $\mathcal{D}^{(0)}_{\mathcal{I}^c}$; pre-specified parameter c, ϕ.
2: Split $\mathcal{D}^{(0)}_{\mathcal{I}^c}$ into equal size set, with index set $\mathcal{I}^c_1$ and $\mathcal{I}^c_2$
3: Use $\mathcal{D}^{(0)}_{\mathcal{I}^c_1}$ to define a random subset of $\mathcal{F}$ as

$$\mathcal{F}_1 = \left\{\beta \in \mathcal{F} : R_{n,\mathcal{I}^c_1}(\beta) \le R_{n,\mathcal{I}^c_1}(\hat{\beta}_{n1}) + c\max\left(\phi\left\|\hat{\beta}_{n1} - \beta\right\|_{n,\mathcal{I}^c_1}, \phi^2\right)\right\}$$

where

$$\|\beta\|^2_{n,\mathcal{I}^c_1} = \frac{1}{|\mathcal{I}^c_1|}\sum_{i\in\mathcal{I}^c_1}\langle X^{(0)}_i, \beta\rangle^2_{L^2}, \quad R_{n,\mathcal{I}}(\beta) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}(Y^{(0)}_i - \langle X^{(0)}_i, \beta\rangle_{L^2})^2, \quad \hat{\beta}_{n1} = \underset{\beta\in\mathcal{F}}{argmin} R_{n,\mathcal{I}^c_1}(\beta)$$

4: Set $\mathcal{F}_2$ as following:

$$\mathcal{F}_2 = \{c_1\beta_1 + c_2\beta_2 : \beta_1, \beta_2 \in \mathcal{F}_1 \text{ and } c_1 + c_2 = 1\}$$

then, return

$$\hat{\beta}_a = \underset{\beta\in\mathcal{F}_2}{argmin} R_{n,\mathcal{I}^c_2}(\beta).$$

The sparse aggregation algorithm is stated in Algorithm 3. For the Oracle inequality and the pre-specified parameter c and ϕ, we refer the reader to Appendix D.5 for model detail. In general, the final aggregated estimator $\hat{\beta}_a$ will only select two of the best-performed candidates from the candidates set $\mathcal{F}$. This guarantees some of the incorrectly constructed $\hat{\mathcal{S}}$ are not involved in building $\hat{\beta}_a$ and thus alleviate the negative transfer sources.

Remark C.1. In (Gaîffas & Lecué, 2011), the authors indicated $\hat{\beta}_a$ has a explicit solution with the form

$$\hat{\beta}_a = \hat{t}\hat{\beta}_1 + (1-\hat{t})\hat{\beta}_2$$

with $\hat{\beta}_1$ and $\hat{\beta}_2$ in belongs to $\mathcal{F}$ and $\hat{t}$ has an analytical form.

D. Appendix: Proof of Section 4

D.1. Proof of Upper Bound for TL-FLR (Theorem 4.3)

Proof. We first prove the upper bound under Assumption 4.2 condition 1 and defer the proof under condition 2 at the end. WLOG, we assume the eigenfunction of $L_{\Gamma^{(0)}}$ and $L_{\Gamma^{(t)}}$ are perfectly aligned, i.e. $\phi^{(0)}_j = \phi^{(t)}_j$ for all $j \in \mathbb{N}$. We also We also recall that we set all the intercept $\alpha^{(t)} = 0$ since $\alpha^{(t)}$ will not affect the convergence rate of estimating $\beta^{(t)}$ (Du & Wang, 2014).

Let $L^2 = \{f : \mathcal{T} \to \mathbb{R} : \|f\|_{L^2} < \infty\}$ represent the he set of all square integrable functions over $\mathcal{T}$. 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 $f^{(t)}$ as $\beta^{(t)}$'s corresponding element in L^2. Therefore, we can rewrite the minimization problem in the transfer step and the calibrate step as

$$\hat{f}_{\mathcal{S}\lambda_1} = \underset{f\in L^2}{\operatorname{argmin}}\left\{\frac{1}{n_{\mathcal{S}}}\sum_{t\in\mathcal{S}}\sum_{i=1}^{n_t}\left(Y^{(t)}_i - \langle X^{(t)}_i, L_{K^{\frac{1}{2}}}(f)\rangle\right)^2 + \lambda_1\|f\|^2_{L^2}\right\},$$

where $n_{\mathcal{S}} = \sum_{t\in\mathcal{S}} n_t$ and

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

Thus the excess risk of $\hat{\beta}$ can be rewritten as

$$\mathcal{E}(\hat{\beta}) = \left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}(\hat{f} - f_0)\right\|^2_{L^2} \quad \text{where} \quad \hat{f} = \hat{f}_{\mathcal{S}\lambda_1} + \hat{f}_{\delta\lambda_2}$$

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

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

and let

$$L_{\Gamma_n^{(t)}} = L_{K^{\frac{1}{2}}} L_{C_n^{(t)}} L_{K^{\frac{1}{2}}}.$$

To bound the excess risk $\mathcal{E}(\hat{\beta})$, by triangle inequality,

$$\left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}(\hat{f} - f_0)\right\|_{L^2} \le \left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}(\hat{f}_{\mathcal{S}} - f_{\mathcal{S}})\right\|_{L^2} + \left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}(\hat{f}_{\delta} - f_{\delta})\right\|_{L^2}$$

where each term at the r.h.s. corresponds to the excess risk from the transfer and calibrate steps, respectively.

Transfer Step. For the transfer step, the solution of minimization is

$$\hat{f}_{\mathcal{S}\lambda_1} = \left(\sum_{t\in\mathcal{S}} \alpha_t L_{\Gamma_n^{(t)}} + \lambda_1 \mathbf{I}\right)^{-1} \left(\sum_{t\in\mathcal{S}} \alpha_t L_{\Gamma_n^{(t)}}(f^{(t)}) + \sum_{t\in\mathcal{S}} g_n^{(t)}\right),$$

where $\mathbf{I}$ is identity operator, $\alpha_t = \frac{n_t}{n_{\mathcal{S}}}$ and

$$g_n^{(t)} = \frac{1}{n_{\mathcal{S}}}\sum_{i=1}^{n_t} \epsilon_i^{(t)} L_{K^{\frac{1}{2}}}(X_i^{(t)}).$$

Besides, the solution of the transfer learning step, $\hat{f}_{\mathcal{S}\lambda_1}$, converges to its population version, which is defined by the following moment condition

$$\sum_{t\in\mathcal{S}} \alpha_t \,\mathrm{E}\left\{L_{K^{\frac{1}{2}}}(X^{(t)})\left(Y^{(t)} - \langle L_{K^{\frac{1}{2}}}(X^{(t)}), f_{\mathcal{S}}\rangle_{L^2}\right)\right\} = 0,$$

and therefore leads to the explicit form of $f_{\mathcal{S}}$ as

$$\left(\sum_{t\in\mathcal{S}} \alpha_t L_{\Gamma^{(t)}}\right) f_{\mathcal{S}} = \sum_{t\in\mathcal{S}} \alpha_t L_{\Gamma^{(t)}}\left(f^{(t)}\right).$$

Define

$$f_{\mathcal{S}\lambda_1} = \left(\sum_{t\in\mathcal{S}} \alpha_t L_{\Gamma^{(t)}} + \lambda_1 \mathbf{I}\right)^{-1} \left(\sum_{t\in\mathcal{S}} \alpha_t L_{\Gamma^{(t)}}(f^{(t)})\right).$$

By triangle inequality

$$\left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}(\hat{f}_{\mathcal{S}\lambda_1} - f_{\mathcal{S}})\right\|_{L^2} \le \underbrace{\left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}(\hat{f}_{\mathcal{S}\lambda_1} - f_{\mathcal{S}\lambda_1})\right\|_{L^2}}_{\text{estimation error}} + \underbrace{\left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}(f_{\mathcal{S}\lambda_1} - f_{\mathcal{S}})\right\|_{L^2}}_{\text{approximation error}}.$$

For approximation error, by Lemma D.4 and taking $v = \frac{1}{2}$, the second term on r.h.s. can be bounded by

$$\left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}(f_{\mathcal{S}\lambda_1} - f_{\mathcal{S}})\right\|_{L^2}^2 = O_{\mathbb{P}}\left(\lambda_1 \left\|f_{\mathcal{S}}\right\|_{L^2}^2\right).$$

Now, we turn to the estimation error. We further introduce an intermedia term

$$\tilde{f}_{\mathcal{S}\lambda_1} = f_{\mathcal{S}\lambda_1} + \left(\sum_{t\in\mathcal{S}} \alpha_t L_{\Gamma^{(t)}} + \lambda_1 \mathbf{I}\right)^{-1} \left(\sum_{t\in\mathcal{S}} \alpha_t L_{\Gamma_n^{(t)}}(f_{\mathcal{S}} - f_{\mathcal{S}\lambda_1}) + \sum_{t\in\mathcal{S}} g_n^{(t)} - \lambda_1 f_{\mathcal{S}\lambda_1}\right).$$

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

$$\lambda_1 f_{\mathcal{S}\lambda_1} = \sum_{t\in\mathcal{S}} \alpha_t L_{\Gamma^{(t)}}\left(f^{(t)} - f_{\mathcal{S}\lambda_1}\right) = \sum_{t\in\mathcal{S}} \alpha_t L_{\Gamma^{(t)}}\left(f_{\mathcal{S}} - f_{\mathcal{S}\lambda_1}\right),$$

we have

$$\begin{aligned}
&\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}(f_{\mathcal{S}\lambda_1}-\tilde{f}_{\mathcal{S}\lambda_1})\|_{L^2}\\
&\leq \left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}(\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}}+\lambda_1\mathbf{I})^{-1}(\sum_{t\in\mathcal{S}}g_n^{(t)})\right\|_{L^2}+\\
&\quad\left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}(\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}}+\lambda_1\mathbf{I})^{-1}(\sum_{t\in\mathcal{S}}\alpha_t(L_{\Gamma_n^{(t)}}-L_{\Gamma^{(t)}})(f_{\mathcal{S}}-f_{\mathcal{S}\lambda_1}))\right\|_{L^2}\\
&=\left\{\sum_{j=1}^{\infty}\left(\left\langle(L_{\Gamma^{(0)}})^{\frac{1}{2}}(\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}}+\lambda_1\mathbf{I})^{-1}\left(\sum_{t\in\mathcal{S}}g_n^{(t)}\right),\phi_j^{(0)})\right\rangle_{L^2}\right)^2\right\}^{\frac{1}{2}}+\\
&\quad\left\{\sum_{j=1}^{\infty}\left(\left\langle(L_{\Gamma^{(0)}})^{\frac{1}{2}}(\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}}+\lambda_1\mathbf{I})^{-1}\left(\sum_{t\in\mathcal{S}}\alpha_t(L_{\Gamma_n^{(t)}}-L_{\Gamma^{(t)}})(f_{\mathcal{S}}-f_{\mathcal{S}\lambda_1})\right),\phi_j^{(0)})\right\rangle_{L^2}\right)^2\right\}^{\frac{1}{2}}
\end{aligned}$$

For the first term in the above inequality, by Lemma D.6,

$$\left\{\sum_{j=1}^{\infty}\left(\left\langle(L_{\Gamma^{(0)}})^{\frac{1}{2}}(\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}}+\lambda_1\mathbf{I})^{-1}\left(\sum_{t\in\mathcal{S}}g_n^{(t)}\right),\phi_j^{(0)})\right\rangle_{L^2}\right)^2\right\}=O_{\mathbb{P}}\left(\sigma^2\left(n_{\mathcal{S}}\right)^{-1}\lambda_1^{\frac{1}{2r}}\right).$$

For second one, by Lemma D.5 and D.7,

$$\sum_{j=1}^{\infty}\left(\left\langle(L_{\Gamma^{(0)}})^{\frac{1}{2}}(\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}}+\lambda_1\mathbf{I})^{-1}\left(\sum_{t\in\mathcal{S}}\alpha_t(L_{\Gamma_n^{(t)}}-L_{\Gamma^{(t)}})(f_{\mathcal{S}}-f_{\mathcal{S}\lambda_1})\right),\phi_j^{(0)})\right\rangle_{L^2}\right)^2=O_{\mathbb{P}}\left(\left(n_{\mathcal{S}}\right)^{-1}\lambda_1^{\frac{1}{2r}}\right).$$

Therefore,

$$\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}(f_{\mathcal{S}\lambda_1}-\tilde{f}_{\mathcal{S}\lambda_1})\|_{L^2}^2=O_{\mathbb{P}}\left(\|f_{\mathcal{S}}\|_{L^2}^2\left(n_{\mathcal{S}}\right)^{-1}\lambda_1^{\frac{1}{2r}}\right).$$

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

$$\hat{f}_{\mathcal{S}\lambda_1}-\tilde{f}_{\mathcal{S}\lambda_1}=\left(\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}}+\lambda_1\mathbf{I}\right)^{-1}\left(\sum_{t\in\mathcal{S}}\alpha_t(L_{\Gamma^{(t)}}-L_{\Gamma_n^{(t)}})(\hat{f}_{\mathcal{S}\lambda_1}-f_{\mathcal{S}\lambda_1})\right).$$

Thus, by Lemma D.5

$$\begin{aligned}
&\left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}(\hat{f}_{\mathcal{S}\lambda_1}-\tilde{f}_{\mathcal{S}\lambda_1})\right\|_{L^2}^2\\
&\leq\left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}(\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}}+\lambda_1\mathbf{I})^{-1}(\sum_{t\in\mathcal{S}}\alpha_t\left(L_{\Gamma^{(t)}}-L_{\Gamma_n^{(t)}}\right))(L_{\Gamma^{(0)}})^{-\frac{1}{2}}\right\|_{op}^2\left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}(\hat{f}_{\mathcal{S}\lambda_1}-f_{\mathcal{S}\lambda_1})\right\|_{L^2}^2\\
&=O_{\mathbb{P}}\left(n_{\mathcal{S}}^{-1}\lambda_1^{\frac{1}{2r}}\left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}(\hat{f}_{\mathcal{S}\lambda_1}-f_{\mathcal{S}\lambda_1})\right\|_{L^2}^2\right)\\
&=o_{\mathbb{P}}\left(\left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}(\hat{f}_{\mathcal{S}\lambda_1}-f_{\mathcal{S}\lambda_1})\right\|_{L^2}^2\right).
\end{aligned}$$

Combine three parts, we get

$$\left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}\left(\hat{f}_{\mathcal{S}\lambda_1}-f_{\mathcal{S}}\right)\right\|_{L^2}^2=O_{\mathbb{P}}\left(\|f_{\mathcal{S}}\|_{L^2}^2\lambda_1+\left(\sigma^2+\|f_{\mathcal{S}}\|_{L^2}^2\right)n_{\mathcal{S}}^{-1}\lambda_1^{-\frac{1}{2r}}\right),$$

by taking $\lambda_1\asymp(n_{\mathcal{S}})^{-\frac{2r}{2r+1}}$ and notice the fact that $\frac{\sigma^2}{\|\beta_{\mathcal{S}}\|_K^2}$ is bounded above (This is a reasonable condition since the signal-to-noise ratio can't be 0, otherwise one can hardly learn anything from the data), we have

$$\left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}\left(\hat{f}_{\mathcal{S}\lambda_1}-f_{\mathcal{S}}\right)\right\|_{L^2}^2=O_{\mathbb{P}}\left(\|f_{\mathcal{S}}\|_{L^2}^2 n_{\mathcal{S}}^{-\frac{2r}{2r+1}}\right)=O_{\mathbb{P}}\left(R^2 n_{\mathcal{S}}^{-\frac{2r}{2r+1}}\right).$$

Calibrate Step. The estimation scheme in the calibrate step is in the same form as the transfer step, and thus, its proof follows the same logic as the transfer step. The solution to the minimization problem in the calibration step is

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

where

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

Similarly, define

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

where f_δ is the population version of the estimator, i.e. $\hat{f}_\delta$.

$$\left(\sum_{t\in\mathcal{S}} \alpha_t L_{\Gamma^{(t)}}\right) f_\delta = \left(\sum_{t\in\mathcal{S}} \alpha_t L_{\Gamma^{(t)}}\left(f_\delta^{(t)}\right)\right)$$

By triangle inequality,

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

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

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

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

By Lemma D.4 with $\mathcal{S} = \emptyset$,

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

where the second inequality holds with the fact the $\mathcal{S} = \left\{1 \leq l \leq L : \|f_0 - f^{(t)}\|_{L^2} \leq h\right\}$. Therefore,

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

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

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

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

$$\tilde{f}_{\delta\lambda_2} - f_{\delta\lambda_2} = (L_{\Gamma^{(0)}} + \lambda_2\mathbf{I})^{-1}\left((L_{\Gamma_n^{(0)}} - L_{\Gamma^{(0)}})(f_{\mathcal{S}} - \hat{f}_{\mathcal{S}} + f_\delta - f_{\delta\lambda_2}) + g_n^{(0)}\right).$$

Therefore,

$$\begin{aligned}\left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}\left(\tilde{f}_{\delta\lambda_2} - f_{\delta\lambda_2}\right)\right\|_{L^2} \leq &\left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}(L_{\Gamma^{(0)}} + \lambda_2\mathbf{I})^{-1} g_n^{(0)}\right\|_{L^2} + \\ &\left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}\left(L_{\Gamma^{(0)}} + \lambda_2\mathbf{I}\right)^{-1}\left(L_{\Gamma_n^{(0)}} - L_{\Gamma^{(0)}}\right)(L_{\Gamma^{(0)}})^{-\frac{1}{2}}\right\|_{op} \cdot \\ &\left\{\left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}\left(f_{\mathcal{S}} - \hat{f}_{\mathcal{S}} + f_\delta - f_{\delta\lambda_2}\right)\right\|_{L^2}\right\}\end{aligned}$$

leading to

$$\left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}\left(\tilde{f}_{\delta\lambda_2}-f_{\delta\lambda_2}\right)\right\|_{L^2}=O_{\mathbb{P}}\left(\left(n_0\lambda_2^{\frac{1}{2r}}\right)^{-1}+\left(n_0\lambda_2^{\frac{1}{2r}}\right)^{-1}\left[n_{\mathcal{S}}^{-\frac{2r}{2r+1}}+\lambda_2h^2\right]\right),$$

where the first term and operator norm comes from Lemma D.5 and D.6 with $\mathcal{S}=\emptyset$, and bounds on $\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}(f_{\mathcal{S}}-\hat{f}_{\mathcal{S}}+f_\delta-f_{\delta\lambda_2})\|_{L^2}^2$ comes from **transfer step** and bias term of **calibrate step**.

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

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

thus by Lemma D.5,

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

Combine three parts, we get

$$\left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}\left(\hat{f}_{\delta\lambda_2}-f_\delta\right)\right\|_{L^2}^2=O_{\mathbb{P}}\left(\lambda_2h^2+\left(h^2+\sigma^2\right)(n_0\lambda_2^{\frac{1}{2r}})^{-1}\right),$$

taking $\lambda_2\asymp n_0^{-\frac{2r}{2r+1}}$ and notice the fact that $\frac{\sigma^2}{h^2}$ is bounded above (similar reasoning as the transfer step), we have

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

Combining the results from **transfer step** and **calibrate step**, and reorganizing the constants for each term, we have

$$\mathcal{E}(\hat{\beta})=\left\|(L_{\Gamma^{(0)}})^{\frac{1}{2}}(\hat{f}-f_0)\right\|_{L^2}=O_{\mathbb{P}}\left(n_{\mathcal{S}}^{-\frac{2r}{2r+1}}+n_0^{-\frac{2r}{2r+1}}\xi(h,R)\right).$$

To prove the same upper bound under Assumption 2, we only need to show Lemma D.4 to Lemma D.8 still hold under Assumption 2. Let $\{(s_j^{(0)},\phi_j^{(0)})\}_{j\geq 1}$ be the eigen-pairs of $L_{\Gamma^{(0)}}$. We show that

$$\left\langle L_{\Gamma^{(t)}}(\phi_j^{(0)}),\phi_j^{(0)}\right\rangle_{L^2}=s_j^{(0)}(1+o(1)). \tag{8}$$

Consider

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

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

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

which leads to

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

Therefore, Equation (8) holds. One can now replace the common eigenfunctions ϕ_j by $\phi_j^{(0)}$ in the proofs of Lemma D.4 to Lemma D.8, and it is not hard to check the results still hold. □

D.2. Proof of Lower Bound for TL-FLR (Theorem 4.4)

In this part, we proof the alternative version for lower bound, i.e.

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

This alternative form is also proved in other contexts like high-dimensional linear regression or GLM to show optimality. However, the upper bound we derive for TL-FLR can still be sharp since in the TL regime, it is always assumed $n_{\mathcal{S}}\gg n_0$, and leads to $(n_{\mathcal{S}}+n_0)^{-\frac{2r}{2r+1}}\asymp n_{\mathcal{S}}^{-\frac{2r}{2r+1}}$.

On the other hand, one can modify the transfer step in TL-FLR by including the target dataset $\mathcal{D}^{(0)}$ to estimate $\beta_{\mathcal{S}}$, which produces an alternative upper bound $(n_{\mathcal{S}}+n_0)^{-\frac{2r}{2r+1}}+n_0^{-\frac{2r}{2r+1}}\xi(h,R)$, and mathematically aligns with the alternative lower bound we mention above. However, we would like to note that such a modified TL-FLR is not computationally efficient for transfer learning, since for each new upcoming target task, TL-FLR needs to recalculate a new $\hat{\beta}_{\mathcal{S}}$ with the huge datasets $\{\mathcal{D}^{(t)}: t\in 0\cup\mathcal{S}\}$.

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.

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

In this case, all the source model shares the same coefficient function as the target model, i.e., $\beta^{(t)}=\beta^{(0)}$ for all $t\in\mathcal{S}$, and therefore the estimation process is equivalent to estimate β under target model with sample size equal to $n_{\mathcal{S}}$. The Proposition D.3 implies

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

where the constant is proportional to R^2.

(2) Consider $\beta^{(0)}\in\mathcal{B}_{\mathcal{H}}(h)$ where $\mathcal{B}_{\mathcal{H}}(h)$ is a ball in RKHS centered at 0 with radius h, and $\beta^{(t)}=0$ for all $t\in\{0\}\cup\mathcal{S}$ and $\sigma\geq h$. That is, all the source datasets contain no information about $\beta^{(0)}$. Applying Proposition D.3 again leads to

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

where the constant is proportional to h^2.

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

D.3. Proof of Consistency (Theorem 4.6)

Proof. Under Assumption 3,

$$\max_{t\in\mathcal{S}}\Delta_t<\min_{t\in\mathcal{S}^c}\Delta_t$$

holds automatically. To prove

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

we only need to show the following fact holds

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

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 & Cai (2010)

$$\left\|(\hat{\beta}_0 - \hat{\beta}_t) - (\beta_0 - \beta_t)\right\|_{K^M} \lesssim \left\|(\hat{\beta}_0 - \hat{\beta}_t) - (\beta_0 - \beta_t)\right\|_{L^2} = o_{\mathbb{P}}(1).$$

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

$$\left\|\hat{\beta}_0 - \hat{\beta}_t\right\|_{K^M} \geq (1 - o_{\mathbb{P}}(1)) \left\|\beta_0 - \beta_t\right\|_{K^M}$$

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

$$\left\|\hat{\beta}_0 - \hat{\beta}_t\right\|_{K^M} \leq (1 + o_{\mathbb{P}}(1)) \left\|\beta_0 - \beta_t\right\|_{K^M} \leq (1 + o_{\mathbb{P}}(1)) \left\|\beta_0 - \beta_t\right\|_{K}$$

with high probability. Finally,

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

where the convergence in probability is guaranteed by Assumption 4.5. □

D.4. Proof of Upper Bound for ATL-FLR (Theorem 4.8)

The Theorem directly holds by combining Theorem 4.3, Proposition D.1 with setting (2), and Theorem 4.6.

D.5. Proposition

Proposition D.1 (Gaîffas & Lecué (2011))**.** *Given a confidence level δ, assume either setting (1) or (2) holds for a constant b,*

1. $\max\{|Y^{(0)}|, \max_{\beta\in\mathcal{H}_K} |\langle X^{(0)}, \beta\rangle_{L^2}|\} \leq b$
2. $\max\{\|\epsilon^{(0)}\|_\psi, \sup_{\beta\in\mathcal{H}_K} \|\langle X^{(0)}, \beta - \beta^{(0)}\rangle\|\} \leq b$

where $\|\epsilon^{(0)}\|_\psi := inf\{c > 0 : \mathrm{E}[\exp\{|\epsilon^{(0)}|/c\}] \leq 2\}$. Let $\sigma^2_{\epsilon^{(0)}}$ be the upper bound for $\mathrm{E}\left[(\epsilon^{(0)})^2|X^{(0)}\right]$. The pre-specified parameter ϕ in Algorithm 3 is defined as

$$\phi = \begin{cases} b\sqrt{\dfrac{(log(|\mathcal{F}| + \delta))}{|\mathcal{I}^c|}}, & \textit{if setting (1) holds} \\ (\sigma_{\epsilon^{(0)}} + b)\sqrt{\dfrac{(log(|\mathcal{F}| + \delta)log(|\mathcal{I}^c|))}{|\mathcal{I}^c|}}. & \textit{if setting (2) holds} \end{cases}$$

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

$$\mathcal{E}\left(\hat{\beta}_a\right) \leq \min_{t=0,1,\cdots,T} \mathcal{E}\left(\hat{\beta}(\hat{\mathcal{S}}_t)\right) + r_\delta(\mathcal{F}, n_0) \tag{9}$$

holds with probability at least $1 - 4e^{-\delta}$ where

$$r_\delta(\mathcal{F}, n_0) = \begin{cases} C_{b1}\dfrac{(1+\delta)log(T)}{n_0}, & \textit{if setting (1) holds} \\ C_{b2}\dfrac{(1+\log(4\delta^{-1})log(T)log(n_0)}{n_0}, & \textit{if setting (2) holds} \end{cases}$$

and C_{b1}, C_{b2} are some constants depend on b.

Remark D.2. We call the setting (1) bounded setting and (2) sub-exponential setting. The latter one is milder but leads to a suboptimal cost. We refer readers to Gaîffas & Lecué (2011) for more detailed discussions about the optimal cost in sparse aggregation.

Proposition D.3 (Lower bound for target-only FLR). *Suppose the observed data $\{(Y_i, X_i)\}_{i=1}^{n_0}$ are generated from FLR model, with the true slope function resides in $\mathcal{B}_{\mathcal{H}}(R) = \{\beta \in \mathcal{H} : \|\beta\|_K \leq R\}$, then*

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

Proof. Consider slope functions $\beta_1, \cdots, \beta_M \in \mathcal{B}_{\mathcal{H}}(R)$ 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

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

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 D.9

$$\begin{aligned}\left\|L_{(C^{(0)})^{\frac{1}{2}}}\left(\tilde{\beta} - \beta_i\right)\right\|_K^2 &\geq P_i\left(\tilde{\beta} \neq \beta_i\right)\min_{i,j}\left\|L_{(C^{(0)})^{\frac{1}{2}}}\left(\beta_i - \beta_j\right)\right\|_K^2 \\ &\geq \left(1 - \frac{\frac{n_0}{2\sigma^2}\max_{i,j}\left\|L_{(C^{(0)})^{\frac{1}{2}}}\left(\beta_i - \beta_j\right)\right\|_K^2 + log(2)}{log(M-1)}\right)\min_{i,j}\left\|L_{(C^{(0)})^{\frac{1}{2}}}\left(\beta_i - \beta_j\right)\right\|_K^2.\end{aligned} \tag{10}$$

Our target is to construct a sequence of $\beta_1, \cdots, \beta_M \in \mathcal{B}_{\mathcal{H}}(R)$ such that the above lower bound matches with the upper bound. We consider Varshamov-Gilbert bound in Varshamov (1957), which we state as Lemma D.10. Now we define,

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

where $(b_{i,1}, b_{i,2}, \cdots, b_{i,N}) \in 0, 1^N$. Then,

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

hence $\beta_\theta \in \mathcal{B}_{\mathcal{H}}(R)$. Besides,

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

where the last inequality is by Lemma D.10, and

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

Therefore, one can bound the KL divergence by

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

Using the above results, the r.h.s. of Equation 10 becomes

$$\left(1 - \frac{\frac{4n_0R^2}{\sigma^2}s_N^{(0)} + 8log(2)}{N}\right)\frac{s_{2N}^{(0)}R^2}{4}.$$

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

$$\begin{aligned}\left(1 - \frac{\frac{4n_0h^2}{\sigma^2}s_N^{(0)} + 8log(2)}{N}\right)\frac{s_{2N}^{(0)}R^2}{4} &\asymp \left(\frac{1}{2} - \frac{8log(2)}{N}\right)R^2N^{-2r} \\ &\asymp n_0^{-\frac{2r}{2r+1}}R^2\end{aligned}$$

□

D.6. Lemmas

In this part, we prove the lemmas that are used in the proof of Theorem 4.3. We prove them under the Assumption 4.2 condition 1 and let ϕ_j denote perfectly aligned eigenfunctions of $\Gamma^{(t)}$ with $t \in \{0\} \cup \mathcal{S}$.

Lemma D.4.

$$\|(L_{\Gamma^{(0)}})^v (f_{\mathcal{S}\lambda_1} - f_{\mathcal{S}})\|_{L^2}^2 \le (1-v)^{2(1-v)}v^{2v}\lambda_1^{2v}\|f_{\mathcal{S}}\|_{L^2}^2 \max_j\left\{\left(\frac{s_j^{(0)}}{\sum_{t\in\mathcal{S}}\alpha_t s_j^{(t)}}\right)^{2v}\right\}.$$

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

$$\left(\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}} + \lambda_1 I\right)f_{\mathcal{S}\lambda_1} = \sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}}\left(f^{(t)}\right) \quad \textit{and} \quad \left(\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}}\right)f_{\mathcal{S}} = \sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}}\left(f^{(t)}\right)$$

then

$$f_{\mathcal{S}\lambda_1} - f_{\mathcal{S}} = -\left(\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}} + \lambda_1 I\right)^{-1}\lambda_1 f_{\mathcal{S}}.$$

Hence,

$$\begin{aligned}\|(L_{\Gamma^{(0)}})^v (f_{\mathcal{S}\lambda_1} - f_{\mathcal{S}})\|_{L^2}^2 &\le \lambda_1^2\left\|(L_{\Gamma^{(0)}})^v(\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}} + \lambda_1 I)^{-1}\right\|_{op}^2\|f_{\mathcal{S}}\|_{L^2}^2 \\ &\le \lambda_1^2\max_j\left\{\frac{(s_j^{(0)})^{2v}}{(\sum_{t\in\mathcal{S}}\alpha_t s_j^{(t)} + \lambda_1)^2}\right\}\|f_{\mathcal{S}}\|_{L^2}^2\end{aligned}$$

By Young's inequality, $\lambda_1 + \sum_{t\in\mathcal{S}}\alpha_t s_j^{(t)} \ge (1-v)^{-(1-v)}v^{-v}\lambda_1^{1-v}(\sum_{t\in\mathcal{S}}\alpha_t s_j^{(t)})^v$

$$\|(L_{\Gamma^{(0)}})^v (f_{\mathcal{S}\lambda_1} - f_{\mathcal{S}})\|_{L^2}^2 \le (1-v)^{2(1-v)}v^{2v}\lambda_1^{2v}\|f_{\mathcal{S}}\|_{L^2}^2 \max_j\left\{\left(\frac{s_j^{(0)}}{\sum_{t\in\mathcal{S}}\alpha_t s_j^{(t)}}\right)^{2v}\right\}.$$

□

Lemma D.5.

$$\left\|(L_{\Gamma^{(0)}})^v\left(\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}} + \lambda_1\mathbf{I}\right)^{-1}\left(\sum_{t\in\mathcal{S}}\alpha_t\left(L_{\Gamma^{(t)}} - L_{\Gamma_n^{(t)}}\right)\right)(L_{\Gamma^{(0)}})^{-v}\right\|_{op} = O_{\mathbb{P}}\left(\left(n_{\mathcal{S}}\lambda_1^{1-2v+\frac{1}{2r}}\right)^{-\frac{1}{2}}\right)$$

Proof.

$$\begin{aligned}
&\left\|(L_{\Gamma^{(0)}})^v\left(\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}}+\lambda_1\mathbf{I}\right)^{-1}\left(\sum_{t\in\mathcal{S}}\alpha_t\left(L_{\Gamma^{(t)}}-L_{\Gamma_n^{(t)}}\right)\right)(L_{\Gamma^{(0)}})^{-v}\right\|_{op}\\
&=\sup_{h:\|h\|_{L^2}=1}\left|\left\langle h,(L_{\Gamma^{(0)}})^v\left(\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}}+\lambda_1\mathbf{I}\right)^{-1}\left(\sum_{t\in\mathcal{S}}\alpha_t\left(L_{\Gamma^{(t)}}-L_{\Gamma_n^{(t)}}\right)\right)(L_{\Gamma^{(0)}})^{-v}h\right\rangle_{L^2}\right|.
\end{aligned}$$

Let

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

then

$$\begin{aligned}
&\left\langle h,(L_{\Gamma^{(0)}})^v\left(\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}}+\lambda_1\mathbf{I}\right)^{-1}\left(\sum_{t\in\mathcal{S}}\alpha_t\left(L_{\Gamma^{(t)}}-L_{\Gamma_n^{(t)}}\right)\right)(L_{\Gamma^{(0)}})^{-v}h\right\rangle_{L^2}\\
&=\sum_{j,k}\frac{(s_j^{(0)})^v(s_k^{(0)})^{-v}h_jh_k}{\sum_{t\in\mathcal{S}}\alpha_t s_j^{(t)}+\lambda_1}\left\langle\phi_j,\sum_{t\in\mathcal{S}}\left(L_{\Gamma^{(t)}}-L_{\Gamma_n^{(t)}}\right)\phi_k\right\rangle_{L^2}.
\end{aligned}$$

By Cauchy-Schwarz inequality,

$$\begin{aligned}
&\left\|(L_{\Gamma^{(0)}})^v\left(\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}}+\lambda_1\mathbf{I}\right)^{-1}\left(\sum_{t\in\mathcal{S}}\alpha_t\left(L_{\Gamma^{(t)}}-L_{\Gamma_n^{(t)}}\right)\right)(L_{\Gamma^{(0)}})^{-v}\right\|_{op}\\
&\leq\left(\sum_{j,k}\frac{(s_j^{(0)})^{2v}(s_k^{(0)})^{-2v}}{(\sum_{t\in\mathcal{S}}\alpha_t s_j^{(t)}+\lambda_1)^2}\left\langle\phi_j,\sum_{t\in\mathcal{S}}\alpha_t\left(L_{\Gamma^{(t)}}-L_{\Gamma_n^{(t)}}\right)\phi_k\right\rangle_{L^2}^2\right)^{\frac{1}{2}}.
\end{aligned}$$

Consider $\mathrm{E}\langle\phi_j,\sum_{t\in\mathcal{S}}(L_{\Gamma^{(t)}}-L_{\Gamma_n^{(t)}})\phi_k\rangle_{L^2}^2$, note that

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

By Jensen's inequality

$$\begin{aligned}
&\mathrm{E}\left(\sum_{j,k}\frac{(s_j^{(0)})^{2v}(s_k^{(0)})^{-2v}}{(\sum_{t\in\mathcal{S}}\alpha_t s_j^{(t)}+\lambda_1)^2}\left\langle\phi_j,\sum_{t\in\mathcal{S}}\alpha_t\left(L_{\Gamma^{(t)}}-L_{\Gamma_n^{(t)}}\right)\phi_k\right\rangle_{L^2}^2\right)^{\frac{1}{2}}\\
&\leq\left(\sum_{j,k}\frac{(s_j^{(0)})^{2v}(s_k^{(0)})^{-2v}}{(\sum_{t\in\mathcal{S}}\alpha_t s_j^{(t)}+\lambda_1)^2}\,\mathrm{E}\left\langle\phi_j,\sum_{t\in\mathcal{S}}\alpha_t\left(L_{\Gamma^{(t)}}-L_{\Gamma_n^{(t)}}\right)\phi_k\right\rangle_{L^2}^2\right)^{\frac{1}{2}},
\end{aligned}$$

thus,

$$\begin{aligned}
&\mathrm{E}\left(\sum_{j,k}\frac{(s_j^{(0)})^{2v}(s_k^{(0)})^{-2v}}{(\sum_{t\in\mathcal{S}}\alpha_t s_j^{(t)}+\lambda_1)^2}\left\langle\phi_j,\sum_{t\in\mathcal{S}}\alpha_t\left(L_{\Gamma^{(t)}}-L_{\Gamma_n^{(t)}}\right)\phi_k\right\rangle_{L^2}^2\right)^{\frac{1}{2}}\\
&\leq\left(\sum_{j,k}\frac{(s_j^{(0)})^{2v}(s_k^{(0)})^{-2v}}{(\sum_{t\in\mathcal{S}}\alpha_t s_j^{(t)}+\lambda_1)^2}\left(\sum_{t\in\mathcal{S}}\alpha_t s_j^{(t)}s_k^{(t)}\right)\frac{|\mathcal{S}|}{(n_\mathcal{S})}\right)^2\\
&\leq\max_{j,k}\left(\frac{\sum_{t\in\mathcal{S}}\alpha_t s_j^{(t)}s_k^{(t)}}{s_j^{(0)}s_k^{(0)}}\right)\left(\sum_{j,k}\frac{(s_j^{(0)})^{1+2v}(s_k^{(0)})^{1-2v}}{(\sum_{t\in\mathcal{S}}\alpha_t s_j^{(t)}+\lambda_1)^2}\frac{|\mathcal{S}|}{(n_\mathcal{S})}\right)^2
\end{aligned}$$

By assumptions of eigenvalues, $\max_{j,k}\left(\frac{\sum_{t\in\mathcal{S}}\alpha_t s_j^{(t)}s_k^{(t)}}{s_j^{(0)}s_k^{(0)}}\right)\leq C_1$ for some constant C_1. Finally, by Lemma D.8

$$\mathrm{E}\left(\sum_{j,k}\frac{(s_j^{(0)})^{2v}(s_k^{(0)})^{-2v}}{(\sum_{t\in\mathcal{S}}\alpha_t s_j^{(t)}+\lambda_1)^2}\left\langle\phi_j,\sum_{t\in\mathcal{S}}\alpha_t\left(L_{\Gamma^{(t)}}-L_{\Gamma_n^{(t)}}\right)\phi_k\right\rangle_{L^2}\right)^{\frac{1}{2}}\lesssim\left((n_\mathcal{S})\lambda_1^{1-2v+\frac{1}{2r}}\right)^{-1}.$$

The rest of the proof can be completed by Markov inequality. ☐

Lemma D.6.

$$\left\|(L_{\Gamma^{(0)}})^v(\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}}+\lambda_1\mathbf{I})^{-1}\sum_{t\in\mathcal{S}}g_n^{(t)}\right\|_{L^2}^2=O_\mathbb{P}\left(\left((n_\mathcal{S})\,\lambda_1^{1-2v+\frac{1}{2r}}\right)^{-1}\right)$$

Proof.

$$\begin{aligned}
\left\|(L_{\Gamma^{(0)}})^v(\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}}+\lambda_1\mathbf{I})^{-1}\sum_{t\in\mathcal{S}}g_n^{(t)}\right\|_{L^2}^2&=\sum_{j\geq1}\left\langle(L_{\Gamma^{(0)}})^v(\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}}+\lambda_1\mathbf{I})^{-1}\sum_{t\in\mathcal{S}}g_n^{(t)},\phi_j\right\rangle_{L^2}^2\\
&=\sum_{j\geq1}\left\langle\sum_{t\in\mathcal{S}}g_n^{(t)},\frac{(s_j^{(0)})^v}{\sum_{t\in\mathcal{S}}\alpha_t s_j^{(t)}+\lambda_1}\phi_j\right\rangle_{L^2}^2\\
&=\sum_{j\geq1}\frac{(s_j^{(0)})^{2v}}{(\sum_{t\in\mathcal{S}}\alpha_t s_j^{(t)}+\lambda_1)^2}\\
&\qquad\left(\frac{1}{n_\mathcal{S}}\sum_{t\in\mathcal{S}}\sum_{i=1}^{n_t}\left\langle\epsilon_i^{(t)}X_i^{(t)},L_{K^{\frac{1}{2}}}(\phi_j)\right\rangle_{L^2}\right)^2.
\end{aligned}$$

Therefore,

$$
\begin{aligned}
\mathrm{E}\left\|(L_{\Gamma^{(0)}})^v(\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}}+\lambda_1\mathbf{I})^{-1}\sum_{t\in\mathcal{S}} g_n^{(t)}\right\|_{L^2}^2 &= \sum_{j\geq 1}\frac{(s_j^{(0)})^{2v}}{(\sum_{t\in\mathcal{S}}\alpha_t s_j^{(t)}+\lambda_1)^2}\cdot \\
&\quad \mathrm{E}\left(\frac{1}{n_{\mathcal{S}}}\sum_{t\in\mathcal{S}}\sum_{i=1}^{n_t}\left\langle \epsilon_i^{(t)}X_i^{(t)}, L_{K^{\frac{1}{2}}}(\phi_j)\right\rangle_{L^2}\right)^2 \\
&= \sum_{j\geq 1}\frac{(s_j^{(0)})^{2v}}{(\sum_{t\in\mathcal{S}}\alpha_t s_j^{(t)}+\lambda_1)^2}\frac{1}{(n_{\mathcal{S}})^2}\cdot \\
&\quad \sum_{t\in\mathcal{S}} n_t\,\mathrm{E}\left(\langle \epsilon_i^{(t)}X_i^{(t)}, L_{K^{\frac{1}{2}}}(\phi_j)\rangle_{L^2}\right)^2 \\
&= \sum_{j\geq 1}\frac{(s_j^{(0)})^{2v}}{(\sum_{t\in\mathcal{S}}\alpha_t s_j^{(t)}+\lambda_1)^2}\frac{\left(\sum_{t\in\mathcal{S}}\sigma^2 n_t s_j^{(t)}\right)}{(n_{\mathcal{S}})^2} \\
&\leq \max_j\left\{\frac{\alpha_0 s_j^{(0)}+\sum_{t\in\mathcal{S}}\alpha_t s_j^{(t)}}{s_j^{(0)}}\right\}\cdot \\
&\quad \left(\frac{C_1}{n_{\mathcal{S}}}\sum_{j\geq 1}\frac{(s_j^{(0)})^{1+2v}}{(\sum_{t\in\mathcal{S}}\alpha_t s_j^{(t)}+\lambda_1)^2}\right),
\end{aligned}
$$

thus by assumption on eigenvalues and Lemma D.8 with $v=\frac{1}{2}$,

$$
\mathrm{E}\left\|(L_{\Gamma^{(0)}})^v(\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma^{(t)}}+\lambda_1\mathbf{I})^{-1}\sum_{t\in\mathcal{S}} g_n^{(t)}\right\|_{L^2}^2 \lesssim \left((n_{\mathcal{S}})\,\lambda_1^{1-2v+\frac{1}{2r}}\right)^{-1},
$$

with the constant proportional to σ^2. The rest of the proof can be completed by Markov inequality. □

Lemma D.7.

$$
\left\|\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma_n^{(t)}}\left(f^{(t)}-f_{\mathcal{S}}\right)\right\|_{L^2}^2 = O_{\mathbb{P}}\left((n_{\mathcal{S}})^{-1}\right)
$$

Proof.

$$
\begin{aligned}
\mathrm{E}\left\|\sum_{t\in\mathcal{S}}\alpha_t L_{\Gamma_n^{(t)}}\left(f^{(t)}-f_{\mathcal{S}}\right)\right\|_{L^2}^2 &= \sum_{j=1}^{\infty}\mathrm{E}\left(\sum_{t\in\mathcal{S}}\alpha_t\left\langle C_n^{(t)}L_{K^{\frac{1}{2}}}(f^{(t)}-f_{\mathcal{S}}), L_{K^{\frac{1}{2}}}(\phi_j)\right\rangle_{L^2}\right)^2 \\
&\lesssim \sum_{j=1}^{\infty}\sum_{t\in\mathcal{S}}\frac{\alpha_t}{n_{\mathcal{S}}}\langle f^{(t)}-f_{\mathcal{S}},\phi_j\rangle_{L^2}^2(s_j^{(t)})^2 \\
&\lesssim (n_{\mathcal{S}})^{-1}\max_{j,l}\left\{\alpha_t(s_j^{(t)})^2\right\}\sum_{t\in\mathcal{S}}\left\|f^{(t)}-f_{\mathcal{S}}\right\|_{L^2}^2 \\
&\lesssim (n_{\mathcal{S}})^{-1},
\end{aligned}
$$

with the universal constant proportional to $\|f_{\mathcal{S}}\|_{L^2}^2$. The rest of the proof can be completed by Markov inequality. □

Lemma D.8.

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

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

Lemma D.9 (Fano's Lemma)**.** *Let $P_1, \cdots, P_M$ be probability measure such that*

$$KL(P_i|P_j) \le \alpha, \quad 1 \le i \ne j \le K$$

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

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

Lemma D.10. *(Varshamov-Gilbert) For any $N \ge 1$, there exists at least $M = exp(N/8)$ N-dimenional vectors, $b_1, \cdots, b_M \subset \{0,1\}^N$ such that*

$$\sum_{l=1}^{N} \mathbf{1}\{b_{ik} \ne b_{jk}\} \ge N/4.$$

E. Appendix: Proof of Section 5

We prove the upper bound and the lower bound of TL-FGLM. We first note that under Assumption 5.2, the excess risk $\mathcal{E}(\hat{\beta})$ for FGLM is equivalent to $\mathrm{E}_{X^{(0)}}\langle\hat{\beta} - \beta^{(0)}, X^{(0)}\rangle_{L^2}^2$ up to universal constants. Thus we focus on bounding the $\mathrm{E}_{X^{(0)}}\langle\hat{\beta} - \beta^{(0)}, X^{(0)}\rangle_{L^2}^2$ in following proofs.

Although we are focusing on $\mathrm{E}_{X^{(0)}}\langle\hat{\beta} - \beta^{(0)}, X^{(0)}\rangle_{L^2}^2$, which is exactly the same as FLR. However, minimizing the regularized negative log-likelihood will not provide an analytical solution of $\hat{\beta}$ as those in FLR, meaning that the proof techniques we used in proving TL-FLR and ATL-FLR are not applicable. Therefore, we use the empirical process to prove the upper bound.

We abbreviate $\langle\cdot,\cdot\rangle_{L^2}$ as $\langle\cdot,\cdot\rangle$ in following proofs. We first introduce some notations. Let

$$\mathcal{L}^{\mathcal{S}}(\beta) = \sum_{t\in\mathcal{S}} \alpha_t \,\mathrm{E}\left[Y^{(t)}\langle X^{(t)}, \beta\rangle - \psi(\langle X^{(t)}, \beta\rangle)\right],$$
$$\mathcal{L}(\beta) = \mathrm{E}\left[Y^{(0)}\langle X^{(0)}, \beta + \hat{\beta}_{\mathcal{S}}\rangle - \psi(\langle X^{(0)}, \beta + \hat{\beta}_{\mathcal{S}}\rangle)\right]$$

and their empirical version are denoted as

$$\mathcal{L}^{\mathcal{S}}_{n_{\mathcal{S}}}(\beta) = \frac{1}{n_0 + n_{\mathcal{S}}} \sum_{t\in\mathcal{S}} \sum_{i=1}^{n_t} \left[Y_i^{(t)}\langle X_i^{(t)}, \beta\rangle - \psi(\langle X_i^{(t)}, \beta\rangle)\right],$$
$$\mathcal{L}_n(\beta) = \frac{1}{n_0} \sum_{i=1}^{n_0} \left[Y_i^{(0)}\langle X_i^{(0)}, \beta + \hat{\beta}_{\mathcal{S}}\rangle - \psi(\langle X_i^{(0)}, \beta + \hat{\beta}_{\mathcal{S}}\rangle)\right]$$

Let $P^{\mathcal{S}}$ and P be the conditional distribution of $\cup_{t\in\mathcal{S}} Y^{(t)}|X^{(t)}$ and $Y^{(0)}|X^{(0)}$ respectively, and $P^{\mathcal{S}}_{n_{\mathcal{S}}}$ and P_n as their empirical version, by define

$$\ell^{\mathcal{S}}(\beta) = \sum_{t\in\mathcal{S}} \alpha_t \left[Y^{(t)}\langle X^{(t)}, \beta\rangle - \psi(\langle X^{(t)}, \beta\rangle)\right],$$
$$\ell(\beta) = \left[Y^{(0)}\langle X^{(0)}, \beta + \hat{\beta}_{\mathcal{S}}\rangle - \psi(\langle X^{(0)}, \beta + \hat{\beta}_{\mathcal{S}}\rangle)\right]$$

we get

$$P^{\mathcal{S}}_{n_{\mathcal{S}}}\ell^{\mathcal{S}}(\beta) = \mathcal{L}^{\mathcal{S}}_{n_{\mathcal{S}}}(\beta), \quad P^{\mathcal{S}}\ell^{\mathcal{S}}(\beta) = \mathcal{L}^{\mathcal{S}}(\beta), \quad P_n\ell(\beta) = \mathcal{L}_n(\beta), \quad P\ell(\beta) = \mathcal{L}(\beta).$$

E.1. Proof of Upper bound for TL-FGLM (Theorem 5.3)

Proof. As mentioned before, we are focusing on $\|\hat{\beta}-\beta^{(0)}\|^2_{C^{(0)}}$, i.e.

$$\begin{aligned}\|\hat{\beta}-\beta^{(0)}\|^2_{C^{(0)}} &= \int_{\mathcal{T}}\int_{\mathcal{T}}(\hat{\beta}(s)-\beta^{(0)}(s))C^{(0)}(s,t)(\hat{\beta}(t)-\beta^{(0)}(t))dsdt \\ &= \mathrm{E}\langle X^{(0)},\hat{\beta}-\beta^{(0)}\rangle^2.\end{aligned}$$

Therefore, we only need to show $\|\hat{\beta}-\beta^{(0)}\|^2_{C^{(0)}}$ is bounded by the error terms in Theorem 5.3. Notice that

$$\left\|\hat{\beta}-\beta^{(0)}\right\|_{C^{(0)}} \le \left\|\hat{\beta}_{\mathcal{S}}-\beta_{\mathcal{S}}\right\|_{C^{(0)}} + \left\|\hat{\delta}_{\mathcal{S}}-\delta_{\mathcal{S}}\right\|_{C^{(0)}}, \tag{11}$$

we then bound the two terms in r.h.s. separately. We denote $\|a-b\|^2_{C^{(t)}} := d_t^2(a,b)$ for all $t\in 0\cup[T]$ and $a,b\in\mathcal{H}_K$.

We first focus on the transfer learning error. Based on the Theorem 3.4.1 in (Vaart & Wellner, 1996), if the following three conditions hold,

1. $\mathrm{E}\sup_{\rho/2\le d_0(\beta,\beta_{\mathcal{S}})\le\rho}\sqrt{n_{\mathcal{S}}}|(\mathcal{L}^{\mathcal{S}}_{n_{\mathcal{S}}}-\mathcal{L}^{\mathcal{S}})(\beta-\beta_{\mathcal{S}})| \lesssim \rho^{\frac{2r-1}{2r}}$;
2. $\sup_{\rho/2\le d_0(\beta,\beta_{\mathcal{S}})\le\rho} P^{\mathcal{S}}\ell^{\mathcal{S}}(\beta)-P^{\mathcal{S}}\ell^{\mathcal{S}}(\beta_{\mathcal{S}}) \lesssim -\rho^2$;
3. $\mathcal{L}^{\mathcal{S}}_{n_{\mathcal{S}}}(\hat{\beta}_{\mathcal{S}}) \ge \mathcal{L}^{\mathcal{S}}(\beta_{\mathcal{S}}) - O_{\mathbb{P}}\left(r^{-2}_{n_{\mathcal{S}}}\|\beta_{\mathcal{S}}\|^2_K\right)$.

then

$$d_0^2(\hat{\beta}_{\mathcal{S}},\beta_{\mathcal{S}}) = O_{\mathbb{P}}(r^{-2}_{n_{\mathcal{S}}}\|\beta_{\mathcal{S}}\|^2_K) = O_{\mathbb{P}}(r^{-2}_{n_{\mathcal{S}}}R^2).$$

For part (1), define

$$\Pi^{\mathcal{S}}_{\rho} = \{\ell^{\mathcal{S}}(\beta)-\ell^{\mathcal{S}}(\beta_{\mathcal{S}}) : \beta\in\mathcal{B}_\rho\} \quad \text{where} \quad \mathcal{B}_\rho = \{\beta\in\mathcal{H}_K : d_0^2(\beta,\beta_{\mathcal{S}})\in[\frac{\rho}{2},\rho]\}.$$

Then $\sup_{\beta\in\mathcal{B}_\rho}|(\mathcal{L}^{\mathcal{S}}_{n_{\mathcal{S}}}-\mathcal{L}^{\mathcal{S}})(\beta-\beta_{\mathcal{S}})| = \sup_{f\in\Pi^{\mathcal{S}}_\rho}|(P^{\mathcal{S}}_{n_{\mathcal{S}}}-P^{\mathcal{S}})f|$ and by Cauchy-Schwarz inequality,

$$\mathrm{E}\sup_{f\in\Pi^{\mathcal{S}}_\rho}|(P^{\mathcal{S}}_{n_{\mathcal{S}}}-P^{\mathcal{S}})f| \le \left\{\mathrm{E}[\sup_{f\in\Pi^{\mathcal{S}}_\rho}|(P^{\mathcal{S}}_{n_{\mathcal{S}}}-P^{\mathcal{S}})f|]^2\right\}^{1/2} := \left\|\sup_{f\in\Pi^{\mathcal{S}}_\rho}|(P^{\mathcal{S}}_{n_{\mathcal{S}}}-P^{\mathcal{S}})f|\right\|_{P^{\mathcal{S}},2}.$$

To bound the right hand side, by Theorem 2.14.1 in (Vaart & Wellner, 1996), we need to find the covering number of $\Pi^{\mathcal{S}}_\rho$, i.e. $\mathcal{N}(\epsilon,\Pi^{\mathcal{S}}_\rho,\|\cdot\|_{P^{\mathcal{S}},2})$. We first show that

$$log\left(\mathcal{N}(\epsilon,\Pi^{\mathcal{S}}_\rho,\|\cdot\|_{P^{\mathcal{S}},2})\right) \le O\left(\epsilon^{-\frac{1}{r}}log(\frac{\rho}{\epsilon})\right).$$

Suppose there exist functions $\beta_1,\cdots,\beta_M\in\mathcal{B}_\rho$ such that

$$\min_{1\le m\le M}\|\ell^{\mathcal{S}}(\beta)-\ell^{\mathcal{S}}(\beta_m)\|_{P^{\mathcal{S}},2} < \epsilon, \quad \forall\beta\in\mathcal{B}_\rho.$$

Since

$$\begin{aligned}\left(\ell^{\mathcal{S}}(\beta)-\ell^{\mathcal{S}}(\beta_i)\right)^2 &= \left[\sum_{t\in\mathcal{S}}\alpha_t Y^{(t)}\langle X^{(t)},\beta-\beta_i\rangle - \left(\psi(\langle X^{(t)},\beta\rangle)-\psi(\langle X^{(t)},\beta_i\rangle)\right)\right]^2 \\ &\le |\mathcal{S}|\sum_{t\in\mathcal{S}}\alpha_t^2((Y^{(t)})^2+(C^{(t)})^2)\langle X^{(t)},\beta-\beta_i\rangle^2_{\mathcal{L}^2} \\ &\le |\mathcal{S}|\max_{t\in\mathcal{S}}\left\{(Y^{(t)})^2+(C^{(t)})^2\right\}\sum_{t\in\mathcal{S}}\alpha_t^2\langle X^{(t)},\beta-\beta_i\rangle^2_{\mathcal{L}^2}\end{aligned}$$

thus

$$\|l^{\mathcal{S}}(\beta) - l^{\mathcal{S}}(\beta_i)\|^2_{P^{\mathcal{S}},2} \le |\mathcal{S}| \max_{t\in\mathcal{S}} \left\{ \mathrm{E}\, \psi'(\langle X^{(t)}, \beta^{(t)}\rangle) + (C^{(t)})^2 \right\} d_0^2(\beta, \beta_i) := C_1 d_0^2(\beta, \beta_i),$$

where the inequality follows the fact for all $t \in \mathcal{S}$, and $d_t(\beta, \beta_i) \asymp d_0(\beta, \beta_i)$ under Assumption 4.2. Hence, the covering number of $\Pi_\rho^{\mathcal{S}}$ under norm $\|\cdot\|_{P^{\mathcal{S}},2}$ is bounded by covering number of $\mathcal{B}_\rho$ under norm d_0, i.e.

$$\mathcal{N}(\epsilon, \Pi_\rho^{\mathcal{S}}, \|\cdot\|_{P^{\mathcal{S}},2}) \le \mathcal{N}(\frac{\epsilon}{C_1}, \mathcal{B}_\rho, d_0).$$

Define $\tilde{\mathcal{B}}_\rho = \{\beta \in \mathcal{H}_K : d_0(\beta, \beta_{\mathcal{S}}) \in [0, \rho]\}$, then

$$\mathcal{N}(\frac{\epsilon}{C_1}, \mathcal{B}_\rho, d_0) \le \mathcal{N}(\frac{\epsilon}{C_1}, \tilde{\mathcal{B}}_\rho, d_0).$$

Next, we will show $\mathcal{N}(\frac{\epsilon}{C_1}, \tilde{\mathcal{B}}_\rho, d_0)$ can be bounded by covering number for a ball in $\mathbb{R}^J$ for some finite integer J. Notice that $\mathcal{H}_K = L_{K^{1/2}}(L^2) = \{\sum_{j\ge1} b_j L_{K^{1/2}}(\phi_j) : (b_j)_{j\ge1} \in \ell^2\}$, hence for any $\beta = \sum_{j\ge1} b_j L_{K^{1/2}}(\phi_j) \in \mathcal{H}_K$,

$$\begin{aligned} d_0^2(\beta, \beta_{\mathcal{S}}) &= \langle \beta - \beta_{\mathcal{S}}, L_{C^{(0)}}(\beta - \beta_{\mathcal{S}})\rangle^2 \\ &= \sum_{j=1}^{\infty} \langle b_j - b_j^{\mathcal{S}}, L_{K^{1/2}CK^{1/2}}(b_j - b_j^{\mathcal{S}})\rangle \\ &= \sum_{j=1}^{\infty} s_j^0 (b_j - b_j^{\mathcal{S}})^2 \end{aligned}$$

which allows one to rewrite $\tilde{\mathcal{B}}_\rho$ as

$$\tilde{\mathcal{B}}_\rho = \left\{ \sum_{j\ge1} b_j L_{K^{1/2}}(\phi_j) : \sum_{j=1}^{\infty} s_j^0 (b_j - b_j^{\mathcal{S}})^2 \le \rho^2 \right\}.$$

Let $J = \lfloor (\frac{\epsilon}{2C_1})^{-\frac{1}{r}} \rfloor$ be a truncation number, and define

$$\tilde{\mathcal{B}}_\rho^* = \left\{ \sum_{j=1}^{J} b_j L_{K^{1/2}}(\phi_j) : \sum_{j=1}^{J} s_j^0 (b_j - b_j^{\mathcal{S}})^2 \le \rho^2 \right\}.$$

For any $\beta \in \tilde{\mathcal{B}}_\rho$, let $\beta^* \in \tilde{\mathcal{B}}_\rho^*$ be its counterpart, then

$$d_0^2(\beta, \beta^*) = \sum_{j=J+1}^{\infty} s_j^0 b_j^2 \le s_J^0 \sum_{j=J+1}^{\infty} b_j^2 \asymp J^{-2r} = (\frac{\epsilon}{2C_1})^2.$$

Suppose there exist function $\beta_1^*, \cdots, \beta_M^* \in \tilde{\mathcal{B}}_\rho^*$ such that

$$\min_{1\le m\le M} d_0(\beta^*, \beta_m^*) < \frac{\epsilon}{2C_1} \quad \forall \beta \in \tilde{\mathcal{B}}_\rho^*,$$

then by triangle inequality

$$\min_{1\le m\le M} d_0(\beta, \beta_i^*) < \frac{\epsilon}{C_1} \quad \forall \beta \in \mathcal{B}_\rho.$$

The above inequality indeed shows that the covering number of $\tilde{\mathcal{B}}_\rho$ with radius $\frac{\epsilon}{C_1}$ can be bounded by the covering of $\tilde{\mathcal{B}}_\rho^*$ with radius $\frac{\epsilon}{2C_1}$, i.e.

$$\mathcal{N}(\frac{\epsilon}{C_1}, \tilde{\mathcal{B}}_\rho, d_0) \le \mathcal{N}(\frac{\epsilon}{2C_1}, \tilde{\mathcal{B}}_\rho^*, d_0).$$

It is known that the covering number for a unit ball in $\mathbb{R}^N$, then the covering number is less than $(\frac{2}{\epsilon} + 1)^N$. Therefore,

$$\mathcal{N}(\frac{\epsilon}{2C_1}, \tilde{\mathcal{B}}_\rho^*, d_0) \le \left(\frac{2\rho + \frac{\epsilon}{2C_1}}{\frac{\epsilon}{2C_1}} \right)^J$$

which leads to

$$log\mathcal{N}(\frac{\epsilon}{2C_1}, \tilde{\mathcal{B}}^*_\rho, d_0) \leq O_{\mathbb{P}}\left(\epsilon^{-\frac{1}{r}} log(\frac{\rho}{\epsilon})\right).$$

By Dudley entropy integral, we know

$$\begin{aligned}
\sup_{f\in L^{\mathcal{S}}_\rho} |(P^{\mathcal{S}}_{n_{\mathcal{S}}} - P^{\mathcal{S}})f| &\lesssim \int_0^\rho \sqrt{\frac{log\mathcal{N}(\frac{\epsilon}{2C_1}, \tilde{\mathcal{B}}^*_\rho, d_0(\cdot,\cdot))}{n_{\mathcal{S}}}} d\epsilon \\
&= \rho^{\frac{2r-1}{2r}} n_{\mathcal{S}}^{-\frac{1}{2}} \int_1^\infty exp\{(1-\frac{1}{2h})u^2\}u^2 du \\
&= O(\rho^{\frac{2r-1}{2r}} n_{\mathcal{S}}^{-\frac{1}{2}})
\end{aligned}$$

Hence, by Theorem 2.14.1 in (Vaart & Wellner, 1996), we finish the proof of (1).

For part (2), let $G(t) = \ell^{\mathcal{S}}(\beta_{\mathcal{S}} + t\tilde{\beta})$ where $\tilde{\beta} = \beta - \beta_{\mathcal{S}}$, then we notice $G(1) = \ell^{\mathcal{S}}(\beta)$ and $G(0) = \ell^{\mathcal{S}}(\beta_{\mathcal{S}})$. We further notice

$$G^{'}(t) = -\sum_{t\in\mathcal{S}} \alpha_t \left\{ Y^{(t)}\langle X^{(t)}, \tilde{\beta}\rangle - \psi'(\langle X^{(t)}, \beta_{\mathcal{S}} + t\tilde{\beta}\rangle)\langle X^{(t)}, \tilde{\beta}\rangle \right\}$$

and thus

$$\begin{aligned}
\mathrm{E}\, G^{'}(0) &= \sum_{t\in\mathcal{S}} \alpha_t \,\mathrm{E}\left\{ Y^{(t)}\langle X^{(t)}, \tilde{\beta}\rangle - \psi'(\langle X^{(t)}, \beta_{\mathcal{S}}\rangle)\langle X^{(t)}, \tilde{\beta}\rangle \right\} \\
&= \sum_{t\in\mathcal{S}} \alpha_t \,\mathrm{E}\left\{ \mathrm{E}\left\{ Y^{(t)} - \psi'(\langle X^{(t)}, \beta_{\mathcal{S}}\rangle) | X^{(t)} \right\} \langle X^{(t)}, \tilde{\beta}\rangle \right\} \\
&= 0
\end{aligned}$$

Besides, by direct calculation,

$$G^{''}(t) = -\sum_{t\in\mathcal{S}} \alpha_t \left\{ \psi''(\langle X^{(t)}, \beta_{\mathcal{S}} + t\tilde{\beta}\rangle)\langle X^{(t)}, \tilde{\beta}\rangle^2 \right\}.$$

By Taylor expansion, there exists a $\gamma \in [0,1]$ such that

$$\begin{aligned}
G(1) - G(0) &= G^{'}(0) + \frac{1}{2}G^{''}(\gamma) \\
&= G^{'}(0) - \frac{1}{2}\sum_{t\in\mathcal{S}} \alpha_t \left\{ \psi''(\langle X^{(t)}, \beta_{\mathcal{S}} + \gamma\tilde{\beta}\rangle)\langle X^{(t)}, \tilde{\beta}\rangle^2 \right\}.
\end{aligned}$$

Notice that $P^{\mathcal{S}}\ell^{\mathcal{S}}(\beta) - P^{\mathcal{S}}\ell^{\mathcal{S}}(\beta_{\mathcal{S}}) = \mathrm{E}[G(1) - G(0)]$, and then

$$\begin{aligned}
P^{\mathcal{S}}\ell^{\mathcal{S}}(\beta) - P^{\mathcal{S}}\ell^{\mathcal{S}}(\beta_{\mathcal{S}}) &= \mathrm{E}[G(1) - G(0)] \\
&= -\frac{1}{2}\sum_{t\in\mathcal{S}} \alpha_t \,\mathrm{E}\left\{ \psi''(\langle X^{(t)}, \beta_{\mathcal{S}} + \gamma\tilde{\beta}\rangle)\langle X^{(t)}, \tilde{\beta}\rangle^2 \right\} \\
&\leq -\frac{\min_{t\in\mathcal{S}}\{\mathcal{A}_1\}}{2} \sum_{t\in\mathcal{S}} \alpha_t \langle X^{(t)}, \tilde{\beta}\rangle^2,
\end{aligned}$$

and

$$\begin{aligned}
P^{\mathcal{S}}\ell^{\mathcal{S}}(\beta) - P^{\mathcal{S}}\ell^{\mathcal{S}}(\beta_{\mathcal{S}}) &= \mathrm{E}[G(1) - G(0)] \\
&= -\frac{1}{2}\sum_{t\in\mathcal{S}} \alpha_t \,\mathrm{E}\left\{ \psi''(\langle X^{(t)}, \beta_{\mathcal{S}} + \gamma\tilde{\beta}\rangle)\langle X^{(t)}, \tilde{\beta}\rangle^2 \right\} \\
&\geq -\frac{\max_{t\in\mathcal{S}}\{\mathcal{A}_2\}}{2} \sum_{t\in\mathcal{S}} \alpha_t \langle X^{(t)}, \tilde{\beta}\rangle^2,
\end{aligned}$$

which leads to

$$P^{\mathcal{S}}\ell^{\mathcal{S}}(\beta) - P^{\mathcal{S}}\ell^{\mathcal{S}}(\beta_{\mathcal{S}}) \asymp -\sum_{t\in\mathcal{S}} \alpha_t d_t^2(\beta, \beta_{\mathcal{S}})$$

Hence, we get

$$\sup_{\rho/2 \leq d_0(\beta,\beta_0) \leq \rho} \left\{ P^{\mathcal{S}} \ell^{\mathcal{S}}(\beta) - P^{\mathcal{S}} \ell^{\mathcal{S}}(\beta_{\mathcal{S}}) \right\} \asymp - \left(\rho^2 + \sum_{t \in \mathcal{S}} \alpha_t d_t^2(\beta, \beta_{\mathcal{S}}) \right) \\ \lesssim -\rho^2,$$

which proves part (2).

Finally for part (3), we pick $r_{n_{\mathcal{S}}} = n_{\mathcal{S}}^{\frac{r}{2r+1}} \|\beta_{\mathcal{S}}\|_K^{-\frac{2r}{2r+1}}$ which satisfies $r_{n_{\mathcal{S}}}^2 \phi_n(r_{n_{\mathcal{S}}}^{-1}) \leq \sqrt{n_{\mathcal{S}}}$ where $\phi_n(x) = \|\beta_{\mathcal{S}}\|_K x^{\frac{2r-1}{2r}}$. Let $\lambda_1 = O(r_{n_{\mathcal{S}}}^{-2})$, since

$$-\mathcal{L}_n^{\mathcal{S}}(\hat{\beta}_{\mathcal{S}}) + \lambda_1 \|\hat{\beta}_{\mathcal{S}}\|_K^2 \leq -\mathcal{L}_n^{\mathcal{S}}(\beta_{\mathcal{S}}) + \lambda_1 \|\beta_{\mathcal{S}}\|_K^2,$$

hence

$$\begin{aligned} \mathcal{L}_n^{\mathcal{S}}(\hat{\beta}_{\mathcal{S}}) &\geq \mathcal{L}_n^{\mathcal{S}}(\beta_{\mathcal{S}}) + \lambda_1 \left(\|\hat{\beta}_{\mathcal{S}}\|_K^2 - \|\beta_{\mathcal{S}}\|_K^2 \right) \\ &\geq \mathcal{L}_n^{\mathcal{S}}(\beta_{\mathcal{S}}) - \lambda_1 \|\beta_{\mathcal{S}}\|_K^2 \\ &\geq \mathcal{L}_n^{\mathcal{S}}(\beta_{\mathcal{S}}) - O(r_{n_{\mathcal{S}}}^{-2} \|\beta_{\mathcal{S}}\|_K^2). \end{aligned}$$

Combining part (1)-(3), based on the Theorem 3.4.1 in (Vaart & Wellner, 1996), we know

$$d_0^2(\hat{\beta}_{\mathcal{S}}, \beta_{\mathcal{S}}) = O_p\left(r_{n_{\mathcal{S}}}^{-2} \|\beta_{\mathcal{S}}\|_K^2) \right).$$

To bound the second term in the r.h.s. of (11), we follow the same proof procedure as the proof of bounding the first term. Specifically, we need to show

1. $\mathrm{E} \sup_{\rho/2 \leq d_0(\delta,\delta_{\mathcal{S}}) \leq \rho} \sqrt{n_0} |(\mathcal{L}_{n_0} - \mathcal{L})(\delta - \delta_{\mathcal{S}})| \lesssim \rho^{\frac{2r-1}{2r}}$;
2. $\sup_{\rho/2 \leq d_0(\delta,\delta_{\mathcal{S}}) \leq \rho} P\ell(\delta) - P\ell(\delta_{\mathcal{S}}) \lesssim -\rho^2$;
3. $\mathcal{L}_{n_0}(\hat{\delta}_{\mathcal{S}}) \geq \mathcal{L}(\delta_{\mathcal{S}}) - O_{\mathbb{P}}\left(r_{n_0}^{-2} \|\delta_{\mathcal{S}}\|_K^2 \right)$.

It is not hard to check, including the estimator from transfer step $\hat{\beta}_{\mathcal{S}}$ into the loss function for the debias step defined at the beginning of the proof will not affect the statements (1)-(3). For example, in part (1), the $\hat{\beta}_{\mathcal{S}}$ will vanish when calculating $(\ell(\delta) - \ell(\delta_i))^2$; in part (2), its effect will vanish since our assumption of the second order derivatives of ψs is bounded from infinity and zero; in part (3), the inequality holds as $\hat{\delta_S}$ is the minimizer of the regularized loss function. Therefore, in the end, we have

$$d_0^2(\hat{\delta}_{\mathcal{S}}, \delta_{\mathcal{S}}) = O_{\mathbb{P}}(r_{n_0}^{-2} \|\delta_{\mathcal{S}}\|_K^2) = O_{\mathbb{P}}(r_{n_0}^{-2} h^2).$$

Combining the bounds of $d_0(\hat{\beta}_{\mathcal{S}}, \beta_{\mathcal{S}})$ and $d_0(\hat{\delta}_{\mathcal{S}}, \delta_{\mathcal{S}})$, we reach to

$$\mathcal{E}(\hat{\beta}) = O_p\left(n_{\mathcal{S}}^{-\frac{2r}{2r+1}} + \left(\frac{h^2}{R^2} \right)^a n_0^{-\frac{2r}{2r+1}} \right),$$

for some $a > 0$.

□

E.2. Proof of Lower Bound for TL-FGLM (Theorem 5.3)

Proof. We calculate the Kullback–Leibler divergence between P_i and P_j under the exponential family. By the definition of KL divergence and density function of the exponential family, we have

$$\begin{aligned} KL(P_i || P_j) &= (n_0 + n_{\mathcal{S}}) \mathrm{E} \Big\{ \langle X^{(0)}, \beta_i - \beta_j \rangle \psi'(\langle X^{(0)}, \beta_i \rangle) \\ &\qquad - \left(\psi(\langle X^{(0)}, \beta_i \rangle) - \psi(\langle X^{(0)}, \beta_j \rangle) \right) \Big\} \\ &= (n_0 + n_{\mathcal{S}}) \mathrm{E} \left\{ \frac{1}{2} \psi''(\langle X^{(0)}, \tilde{\beta} \rangle) < X^{(0)}, \beta_i - \beta_j >^2 \right\} \\ &\lesssim (n_0 + n_{\mathcal{S}}) d_0^2(\beta_i, \beta_j), \end{aligned}$$

for some $\tilde{\beta}$ between β_i and β_j. Therefore, the proof of the lower bound for the target-only FGLM follows exactly the same as FLR, i.e., Proposition D.3.

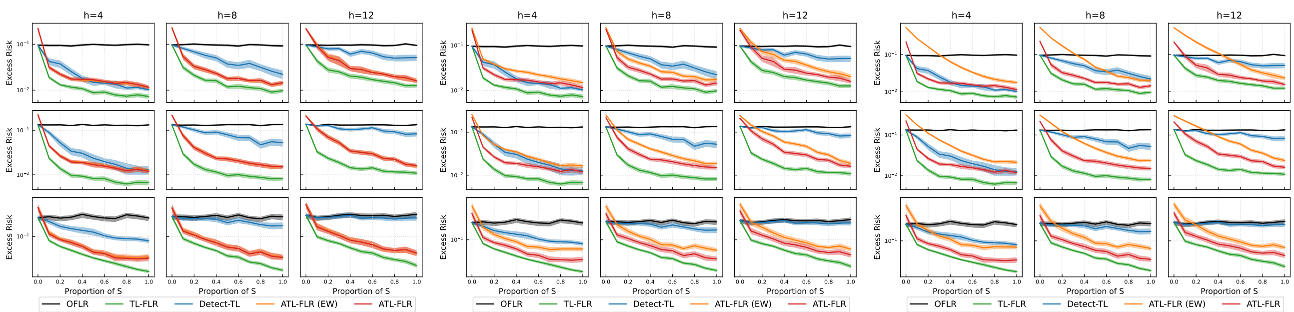
Then, the proof of the lower bound for TL-FGLM follows exactly the same as TL-FLR, i.e., Theorem 4.4, by considering two specific cases. □

F. Appendix: Additional Experiments for TL-FLR/ATL-FLR

In this section, we explore how the smoothness of the coefficient functions, $\beta^{(t)}$, for $t \in \mathcal{S}^c$, will affect the performance of ATL-FLR. We also explore how different temperatures will affect the performance of Exponential Weighted ATL-FLR (ATL-FLR (EW)).

We consider the setting that $\beta^{(t)}$ with $t \in \mathcal{S}^c$ are generated from a much rougher Gaussian process, i.e. β_t are generated from a Gaussian process with mean function $\cos(2\pi t)$ with covariance kernel $\min(s,t)$, which is exactly Wiener process, and thus the β_ts are less smooth than β_ts that are generated from Ornstein–Uhlenbeck process (the one we used in main paper). For ATL-FLR (EW), we consider three different temperatures, i.e., $T = 0.2, 2, 10$, where a lower temperature will usually produce small aggregation coefficients. All the other settings are the same as the simulation section.

The results are presented in Figure 3. In general, the patterns of using the Wiener process are consistent with using the Ornstein–Uhlenbeck process, which demonstrates the robustness of the proposed algorithms to negative transfer source models. We also note that while the temperature is low ($T = 0.2$), the small convex combination coefficients $\{c_j\}$ will make ATL-FLR(EW) have almost the same performance as ATL-FLR, but it still cannot beat ATL-FLR. While we increase the temperature ($T = 2, T = 10$), the gap between ATL-FLR(EW) and ATL-FLR increases, especially when the proportion of $|\mathcal{S}|$ is small. Therefore, selecting the wrong T can hugely degrade the performance of ATL-FLR(EW). This demonstrates the superiority of sparse aggregation in practice since its performance does not depend on the selection of any hyperparameters.



(a) Low Temperature ($T = 0.2$) (b) Middle Temperature ($T = 2$) (c) High Temperature ($T = 10$)

Figure 3. Excess Risk of different transfer learning algorithms. Each row corresponds to different $\beta^{(0)}$, 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 indicating $\pm$ 2 standard error.

G. Appendix: Application

G.1. Application for Functional Linear Regression

In this section, we demonstrate an application of the proposed algorithms in the financial market. The goal of portfolio management is to balance future stock returns and risk, and thus, investors can rebalance their portfolios according to their goals. Some investors may be interested in predicting the future stock returns in a specific sector, and transfer learning can borrow market information from other sectors to improve the prediction of the interest.

In this stock data application, for two given adjacent months, we focus on utilizing the Monthly Cumulative Return (MCR) of the first month to predict the Monthly Return (MR) of the subsequent month and improving the prediction accuracy on a certain sector by transferring market information from other sectors. Specifically, 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 responses are expressed as

$$X(t) = \frac{s^1(t) - s^1(t_0)}{s^1(t_0)} \quad \text{and} \quad Y = \frac{s^2(t_m) - s^2(t_0)}{s^2(t_0)}. \tag{12}$$

The stock price data are collected from Yahoo Finance (https://finance.yahoo.com/), and we focus on stocks whose corresponding companies have 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 raw data obtained from websites are processed to match the format in (12) and both the raw data and processed data are available at https://github.com/haotianlin/HTL-FLM.

After pre-processing, the dataset consists of total 11 sectors: Basic Industries (BI), Capital Goods (CG), Consumer Durable (CD), Consumer Non-Durable (CND), Consumer Services (CS), Energy (E), Finance (Fin), Health Care (HC), Public Utility (PU), Technology (Tech), and Transportation (Trans), with the number of stocks in each sector as $60, 58, 31, 30, 104, 55, 70, 68, 46, 103, 41$. The period of the 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*, *ATL-FLR(EW)* and *ATL-FLR*. Naive-TL implements TL-FLR by setting all source sectors belonging to $\mathcal{S}$, while the Pooled-TL one omits the calibrate step in Naive-TL, and the other three are the same as the former simulation section. The learning of each sector is treated as the target task 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 as the reproducing kernel K again. Specifically, we set $\rho = 1$ and $\nu = 1/2, 3/2, \infty$ (where $\nu = 1/2$ is equivalent to the 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 replicate the experiment 100 times and report the average prediction error with standard error in Figure 4.

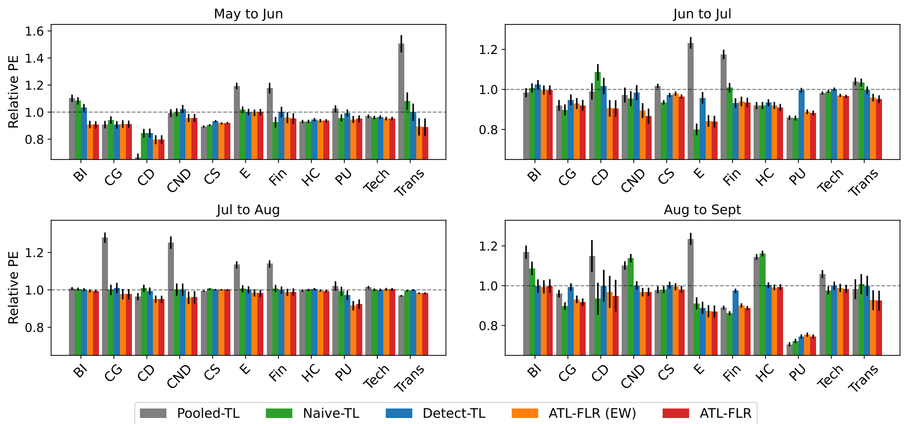
First, we note that the Pooled-TL and Naive-TL only reduce the prediction error in a few sectors but make no improvement or even downgrade the predictions in most sectors. This implies the effect of direct transfer learning can be quite random, as it can benefit the prediction of the target sector when it shares high similarities with other sectors while having worse performance when similarities are low. Besides, Naive-TL shows an overall better performance compared to the Pooled-TL, demonstrating the importance of the calibrate step. For Detect-TL, all the ratios are close to 1, showing its limited improvement, which is as expected as it can miss positive transfer sources easily. Finally, both ATL-FLR(EW) and ATL-FLR provide more robust improvements on average. We can see both of them have improvements 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 consistent across all the sectors and adjacent months, showing the proposed algorithms are also robust to different reproducing kernels.

G.2. Application for Functional Generalized Linear Models

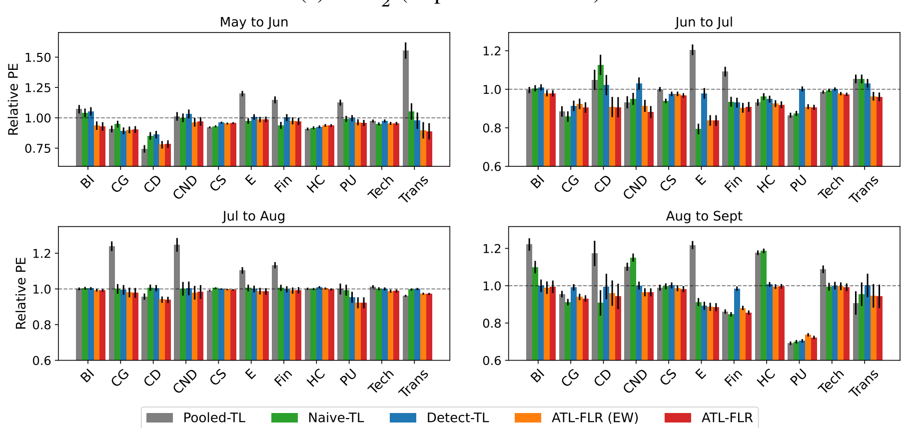
Personal wearable devices have become increasingly popular as they can detect users' movements and provide feedback/records. However, the limited data can make the device's detection inaccurate for a newly registered user. The technology can make detecting new users' actions more accurate by leveraging learned hypotheses from other users with similar features. Under this context, we consider the Human Activity Recognition (HAR) dataset (Anguita et al., 2013), which contains the recordings of volunteers performing daily living activities, including walking, walking upstairs, walking downstairs, sitting, standing, and laying. Each volunteer carried a waist-mounted smartphone with an embedded accelerometer and gyroscope sensors to capture the body acceleration and gravity signal.

In this section, we evaluate the efficacy of our proposed transfer learning method for FGLM and specifically aim at functional logistic regression. The goal is to distinguish the actions of walking upstairs (marked as 1) and walking downstairs (marked as 0) by applying functional logistic regression with the covariate X as the body acceleration signal in the vertical direction over time. We treat the classification for each volunteer as a separate task. After preprocessing, there are a total of 30 volunteer datasets. Each dataset is balanced in terms of label proportion and the sample size of each is between 78 to 127. The covariate of body acceleration signals is measured in equal spacing 128 consecutive time points, and we pick the first 32 points to reduce motion cycles. For each dataset, we randomly split the samples into a training set (80%) and test set (20%). Each volunteer is treated as a target each time and all the other volunteers as sources.

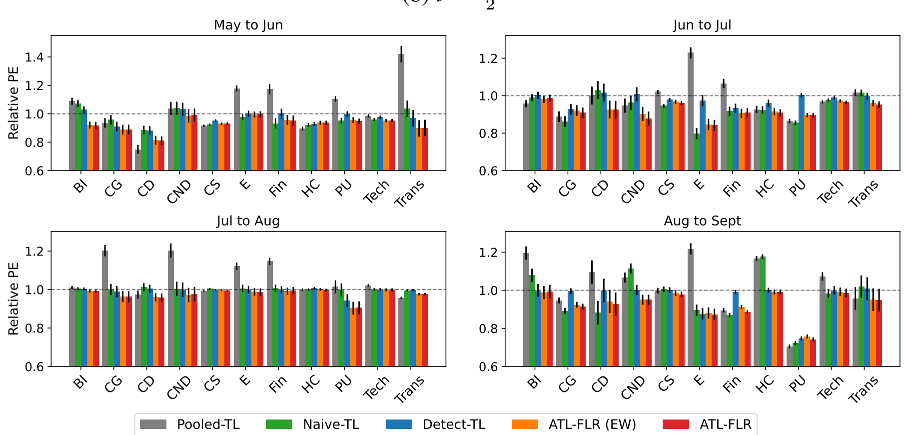
We compare the performance of our proposed *ATL-FLR(EW)* and *ATL-FLR* with the non-transfer baseline *OFLR* and some competitors, including *Pooled Transfer (Pooled-TL)*, *Naive Transfer (Naive-TL)*, *Detect-TL*. We consider the Matérn kernel $K_{\nu,\rho}$ (Cressie & Huang, 1999) as the reproducing kernel and set $\rho = 1$ and $\nu = 1/2$. The regularization parameters are



(a) $\nu = \frac{1}{2}$ (Exponential Kernel)



(b) $\nu = \frac{3}{2}$



(c) $\nu = \infty$ (Gaussian Kernel)

Figure 4. Relative prediction error of Pooled-TL, Naive-TL, Detect-TL, ATL-FLR(EW), and ATL-FLR to OFLR for each target sector. Each bar is an average of 100 replications, with standard error as the black line.

selected via Generalized Cross-Validation (GCV). We repeat the experiment 100 times to assess the variability in train/test data split and report the misclassification rate (in percentage) on test data with standard error. The threshold for all binary classification is set to be 0.5. The results are placed in Table 1.

Based on Table 1, the results demonstrate that the proposed ATL-FGLM can provide the lowest misclassification rate for most of the volunteers. In cases where ATL-FGLM is not the best, its misclassification rate is also close to the lowest (e.g., Volunteer 9,17,27), which verifies its robustness to non-informative sources.

Table 1. Misclassification rates (%) on the test set for each target volunteer with standard errors in subscript.

Target	Transfer Learning Algorithms					
	Non-TL	Pooled-TL	Naive-TL	Detect-TL	ATL-FGLM (EW)	ATL-FGLM
Volunteer1	$5.548_{0.395}$	$8.290_{0.408}$	$5.129_{0.364}$	$5.194_{0.346}$	$4.774_{0.365}$	$\mathbf{4.387}_{0.365}$
Volunteer2	$15.679_{0.522}$	$17.679_{0.541}$	$16.536_{0.514}$	$15.714_{0.520}$	$15.179_{0.521}$	$\mathbf{15.036}_{0.528}$
Volunteer3	$21.879_{0.747}$	$21.576_{0.600}$	$21.121_{0.628}$	$21.121_{0.650}$	$18.667_{0.716}$	$\mathbf{17.758}_{0.680}$
Volunteer4	$25.000_{0.705}$	$25.767_{0.693}$	$23.467_{0.686}$	$\mathbf{21.133}_{0.669}$	$22.467_{0.701}$	$21.300_{0.672}$
Volunteer5	$29.000_{0.913}$	$43.857_{0.791}$	$28.786_{0.837}$	$29.393_{0.804}$	$24.714_{0.800}$	$\mathbf{24.357}_{0.779}$
Volunteer6	$16.448_{0.600}$	$\mathbf{15.345}_{0.544}$	$16.103_{0.572}$	$16.103_{0.550}$	$16.414_{0.572}$	$16.276_{0.565}$
Volunteer7	$12.759_{0.489}$	$15.448_{0.638}$	$14.207_{0.599}$	$13.379_{0.574}$	$12.483_{0.543}$	$\mathbf{12.483}_{0.530}$
Volunteer8	$27.000_{0.697}$	$24.304_{0.801}$	$23.913_{0.684}$	$25.087_{0.762}$	$25.609_{0.786}$	$\mathbf{23.609}_{0.780}$
Volunteer9	$\mathbf{23.536}_{0.614}$	$25.679_{0.623}$	$25.036_{0.667}$	$25.286_{0.651}$	$24.714_{0.665}$	$24.536_{0.657}$
Volunteer10	$31.120_{0.732}$	$32.600_{0.674}$	$30.160_{0.740}$	$30.600_{0.736}$	$30.080_{0.731}$	$\mathbf{29.480}_{0.764}$
Volunteer11	$16.267_{0.575}$	$14.400_{0.503}$	$15.000_{0.487}$	$14.200_{0.552}$	$13.967_{0.552}$	$\mathbf{13.400}_{0.574}$
Volunteer12	$6.667_{0.402}$	$9.100_{0.452}$	$6.633_{0.403}$	$5.600_{0.367}$	$5.467_{0.359}$	$\mathbf{4.767}_{0.352}$
Volunteer13	$19.903_{0.539}$	$18.903_{0.543}$	$18.581_{0.552}$	$19.065_{0.564}$	$18.581_{0.567}$	$\mathbf{18.129}_{0.546}$
Volunteer14	$37.300_{0.578}$	$\mathbf{33.533}_{0.688}$	$35.833_{0.702}$	$35.667_{0.692}$	$35.733_{0.648}$	$35.867_{0.679}$
Volunteer15	$21.926_{0.618}$	$23.000_{0.554}$	$20.778_{0.595}$	$20.926_{0.587}$	$20.481_{0.628}$	$\mathbf{20.333}_{0.604}$
Volunteer16	$18.379_{0.630}$	$19.310_{0.600}$	$17.931_{0.626}$	$18.241_{0.632}$	$17.897_{0.625}$	$\mathbf{17.655}_{0.620}$
Volunteer17	$\mathbf{17.964}_{0.614}$	$20.250_{0.616}$	$22.393_{0.716}$	$20.857_{0.635}$	$19.893_{0.564}$	$18.821_{0.576}$
Volunteer18	$24.971_{0.580}$	$21.412_{0.591}$	$20.029_{0.579}$	$20.176_{0.545}$	$19.559_{0.577}$	$\mathbf{18.471}_{0.567}$
Volunteer19	$20.833_{0.657}$	$19.875_{0.672}$	$19.375_{0.631}$	$\mathbf{18.958}_{0.668}$	$19.000_{0.668}$	$19.167_{0.686}$
Volunteer20	$15.966_{0.617}$	$18.379_{0.618}$	$15.897_{0.548}$	$15.207_{0.565}$	$15.586_{0.605}$	$\mathbf{15.069}_{0.549}$
Volunteer21	$16.036_{0.550}$	$18.321_{0.585}$	$16.429_{0.585}$	$15.357_{0.582}$	$\mathbf{15.107}_{0.534}$	$15.643_{0.546}$
Volunteer22	$19.500_{0.716}$	$19.708_{0.758}$	$19.750_{0.736}$	$18.750_{0.734}$	$18.500_{0.686}$	$\mathbf{18.458}_{0.689}$
Volunteer23	$22.323_{0.648}$	$20.387_{0.637}$	$20.871_{0.626}$	$21.871_{0.641}$	$20.774_{0.642}$	$\mathbf{19.742}_{0.582}$
Volunteer24	$16.314_{0.580}$	$\mathbf{15.171}_{0.529}$	$15.400_{0.536}$	$15.514_{0.550}$	$15.543_{0.559}$	$15.514_{0.554}$
Volunteer25	$32.139_{0.583}$	$30.194_{0.648}$	$29.417_{0.595}$	$29.444_{0.587}$	$29.306_{0.593}$	$\mathbf{29.056}_{0.619}$
Volunteer26	$6.844_{0.376}$	$8.156_{0.389}$	$7.437_{0.392}$	$6.750_{0.392}$	$6.937_{0.384}$	$\mathbf{6.719}_{0.383}$
Volunteer27	$8.500_{0.473}$	$8.786_{0.414}$	$7.643_{0.425}$	$7.536_{0.476}$	$\mathbf{6.286}_{0.394}$	$6.393_{0.407}$
Volunteer28	$25.103_{0.709}$	$24.586_{0.716}$	$23.448_{0.712}$	$24.172_{0.739}$	$23.517_{0.726}$	$\mathbf{23.069}_{0.694}$
Volunteer29	$39.276_{0.761}$	$38.586_{0.738}$	$37.345_{0.775}$	$37.276_{0.752}$	$37.552_{0.819}$	$\mathbf{37.103}_{0.846}$
Volunteer30	$33.474_{0.749}$	$33.263_{0.610}$	$30.526_{0.672}$	$32.632_{0.712}$	$30.000_{0.777}$	$\mathbf{29.263}_{0.800}$