The Annals of Statistics
2024, Vol. 52, No. 2, 654–678
https://doi.org/10.1214/24-AOS2362

TRANSFER LEARNING FOR FUNCTIONAL MEAN ESTIMATION: PHASE TRANSITION AND ADAPTIVE ALGORITHMS

BY T. TONY CAI[a], DONGWOO KIM[b] AND HONGMING PU[c]

Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, [a]*tcai@wharton.upenn.edu,* [b]*dongwooo@wharton.upenn.edu,* [c]*hpu@wharton.upenn.edu*

This paper studies transfer learning for estimating the mean of random functions based on discretely sampled data, where in addition to observations from the target distribution, auxiliary samples from similar but distinct source distributions are available. The paper considers both common and independent designs and establishes the minimax rates of convergence for both designs. The results reveal an interesting phase transition phenomenon under the two designs and demonstrate the benefits of utilizing the source samples in the low sampling frequency regime.

For practical applications, this paper proposes novel data-driven adaptive algorithms that attain the optimal rates of convergence within a logarithmic factor simultaneously over a large collection of parameter spaces. The theoretical findings are complemented by a simulation study that further supports the effectiveness of the proposed algorithms.

1. Introduction. Functional data is commonly observed in a wide range of fields, and there has been significant research on functional data analysis (FDA) with applications in diverse areas such as biomedical studies, neuroscience, linguistics, psychology, demography, economics and engineering. For a comprehensive review of the FDA and its applications, we recommend referring to Ramsay and Silverman [29] and Wang, Chiou and Müller [37].

Estimating the mean function is a fundamental problem in the FDA that has garnered considerable attention in the existing literature. Notable references include Rice and Silverman [31], Page et al. [25], Park et al. [27], Jiang, Aston and Wang [17] and Cai and Yuan [10]. This problem arises naturally in a broad range of practical applications, such as analyzing a diffusion tensor imaging (DTI) data set to investigate multiple sclerosis (MS) patients [13, 28, 33], analyzing a longitudinal CD4 cell count data set to investigate acquired immunodeficiency syndrome (AIDS) [42] and examining a longitudinal data set of trajectories of patient-reported symptom severity of Parkinson's disease [38].

On the other hand, transfer learning has experienced growing popularity as a machine learning technique aimed at improving performance in a target domain by utilizing information from different yet related source domains. This approach proves particularly valuable in situations where acquiring target observations is infrequent or costly, but data from analogous studies are available. Transfer learning has been applied in various machine learning applications, including computer vision [14, 35], speech recognition [15] and genre classification [12]. We refer readers to Pan and Yang [26] and Weiss, Khoshgoftaar and Wang [39] for a more comprehensive discussion on the practical applications of transfer learning. Recently, transfer learning has also attracted increasing attention in statistics and has been successful in various statistical learning problems such as nonparametric classification [8, 30], high-dimensional linear regression [22], large-scale Gaussian graphical model [23], high-dimensional generalized linear models [4, 34], nonparametric regression [7] and contextual multiarmed bandits [3].

Received May 2023; revised October 2023.
MSC2020 subject classifications. Primary 62J05; secondary 62G20.
Key words and phrases. Adaptivity, common design, functional data analysis, independent design, mean function, minimax rate of convergence, phase transition, transfer learning.

Transfer learning can also be effective in the context of functional mean estimation when subjects are divided into groups, such as medical measurement data sets grouped by disease status [29], Chapter 5, gene expression data sets grouped by particular disease [21, 27, 32, 40] or spatial dispersion data sets of marine mammals grouped by species [24]. The application of transfer learning can improve the performance of functional mean estimation in the target group by utilizing knowledge from other groups.

An interesting and motivating example of functional mean estimation with transfer learning is found in the trend analysis of coronavirus disease-19 (COVID-19). The objective is to gain insights into the epidemic's progression and identify overall trends in the United States. Similar to the work by Kalogridis and Van Aelst [18], one can look at the number of new daily COVID-19 cases across the 50 U.S. states. Each state's curve may be regarded as a functional realization and the problem reduces to estimate their functional mean. However, COVID-19 has had a global impact, affecting populations worldwide [19]. Structural similarities in trends across different countries have been observed, as demonstrated by James and Menzies [16]. This suggests that transfer learning can improve the estimation accuracy of the COVID-19 trend in the United States by utilizing the COVID-19 curves from other American countries or other continents, including Africa, Asia, Europe and Oceania.

1.1. *Problem formulation.* The problem of functional mean estimation in a conventional setting can be described as follows. Let $X^{[t]} : [0,1] \to \mathbb{R}$ be a target random function and each target subject $i \in \{1, \dots, n^{[t]}\}$ has an an independent copy $X_i^{[t]}$ (called curve) of $X^{[t]}$. We have noisy observations of these curves at discrete locations:

$$Y_{ij}^{[t]} = X_i^{[t]}\big(T_{ij}^{[t]}\big) + \varepsilon_{ij}^{[t]}, \quad \big(j = 1, \dots, m_i^{[t]} \text{ and } i = 1, \dots, n^{[t]}\big), \tag{1}$$

where $T_{ij}^{[t]}$ are target design points, and $\varepsilon_{ij}^{[t]}$ are an independent random noises. Based on the target sample $\mathcal{D}^{[t]} := \{(T_{ij}^{[t]}, Y_{ij}^{[t]}) : j = 1, \dots, m_i^{[t]},\ i = 1, \dots, n^{[t]}\}$, our primary objective is to estimate the target mean function $f^{[t]}(\cdot) := \mathbb{E}(X^{[t]}(\cdot))$. Such problems naturally arise in a wide range of applications and are typical in FDA. Extensive examples can be found in the works of Ramsay and Silverman [29].

In the transfer learning setup, we also have K source samples $\mathcal{D}^{[s,1]}, \dots, \mathcal{D}^{[s,K]}$ in addition to the target sample $\mathcal{D}^{[t]}$. These source samples are generated similarly. That is, for each source index $k \in \{1, \dots, K\}$, there is a random function $X^{[s,k]} : [0,1] \to \mathbb{R}$ with its source mean function $f^{[s,k]}(\cdot) := \mathbb{E}(X^{[s,k]}(\cdot))$ and we observe

$$\begin{aligned} Y_{ij}^{[s,k]} &= X_i^{[s,k]}\big(T_{ij}^{[s,k]}\big) + \varepsilon_{ij}^{[s,k]}, \\ &\big(j = 1, \dots, m_i^{[s,k]}, i = 1, \dots, n^{[s,k]}, \text{ and } k = 1, \dots, K\big), \end{aligned} \tag{2}$$

where the curves $X_i^{[s,k]}$ are an independent copies of $X^{[s,k]}$, $T_{ij}^{[s,k]}$ are source design points and $\varepsilon_{ij}^{[s,k]}$ are an independent random noises. The estimator for the target mean function $f^{[t]}$ now can utilize the source samples $\mathcal{D}^{[s,1]}, \dots, \mathcal{D}^{[s,K]}$ as well as the target sample $\mathcal{D}^{[t]}$.

We model the relationship between $\mathcal{D}^{[s,1]}, \dots, \mathcal{D}^{[s,K]}$ and $\mathcal{D}^{[t]}$ through their corresponding mean functions $f^{[s,1]}, \dots, f^{[s,K]}$ and $f^{[t]}$. Let us denote by $\mathcal{H}_\alpha(L, M)$ the class of bounded functions with Hölder smoothness $\alpha > 0$. To be more specific, $f \in \mathcal{H}_\alpha(L, M)$ if and only if $f : [0,1] \to \mathbb{R}$ is bounded as $\|f\|_{\mathcal{L}^\infty} \le M$ and α^*-times continuously differentiable such that

$$\big|f^{(\alpha^*)}(t_1) - f^{(\alpha^*)}(t_2)\big| \le L|t_1 - t_2|^{\alpha - \alpha^*}, \quad \text{for any } t_1, t_2 \in [0,1],$$

where $\alpha^* := \omega(\alpha)$ denotes the largest integer strictly smaller than α. We assume

$$\begin{aligned} &f^{[t]},\ f^{[s,k]} \in \mathcal{H}_{\alpha_m}(L_m, M_m), \\ &\delta^{[s,k]} := f^{[t]} - f^{[s,k]} \in \mathcal{H}_{\alpha_\delta}(L_\delta, M_\delta), \end{aligned} \qquad (k = 1, \dots, K) \tag{3}$$

with two smoothness parameters $\alpha_m, \alpha_\delta > 0$ and constants $L_m, M_m, L_\delta, M_\delta > 0$. Our model captures a broader range of flexibility and provides more informative insights into the problem. We have never made any specific assumptions or constraints regarding the relationships between smoothness parameters, α_m and α_δ. This allows us to adapt to a wider variety of scenarios and better address the complexities of the problem.

In line with the conventional framework for functional mean estimation [10], we consider two distinct sampling designs. The first design is characterized by common design points, where observations are gathered at identical locations across curves. In this setting, we have $T_{ij}^{[t]} = T_j^{[t]}$ and $T_{ij}^{[s,k]} = T_j^{[s]}$. The second design employs an independent approach, where $T_{ij}^{[t]}$ and $T_{ij}^{[s,k]}$ are independently sampled from a distribution defined over $[0, 1]$. For a more comprehensive formalization of these designs, Section 2 is dedicated to exploring a common design setting, while Section 3 delves into an independent design case.

1.2. *Our contribution.* To streamline our presentation, we make an assumption that the source samples share the same design type as the target sample. Additionally, we assume that both the target and source samples exhibit identical characteristics, including the number of subjects and design points per subject. In other words,

$$\begin{aligned} &n^{[t]} = n_t \text{ and } m_i^{[t]} = m_t \quad \text{for any } i = 1, \dots, n^{[t]}, \\ &n^{[s,k]} = n_s \text{ and } m_i^{[s,k]} = m_s \quad \text{for any } i = 1, \dots, n^{[s,k]} \text{ and } k = 1, \dots, K. \end{aligned} \tag{4}$$

We propose novel algorithms and develop an optimality theory for the estimation of the target mean function $f^{[t]}$ under both common and independent designs. Under a common design, the optimal minimax rate of convergence up to logarithmic factors is shown to be

$$\left[m_t^{-2\alpha_m} + \frac{1}{n_t}\right] \wedge \left[m_t^{-2\alpha_\delta} + \frac{1}{n_t} + m_s^{-2\alpha_m} + \frac{1}{Kn_s}\right].$$

On the other hand, under an independent design, the rate is given by

$$\left[(m_t n_t)^{-2\alpha_m/(2\alpha_m+1)} + \frac{1}{n_t}\right] \wedge \left[(m_t n_t)^{-2\alpha_\delta/(2\alpha_\delta+1)} + \frac{1}{n_t} + (Km_s n_s)^{-2\alpha_m/(2\alpha_m+1)} + \frac{1}{Kn_s}\right].$$

The examination of minimax convergence rates uncovers several critical aspects of phase transition that significantly impact the effectiveness of transfer learning. First of all, for any model in which $\alpha_\delta \le \alpha_m$ is satisfied, the utilization of source samples and transfer learning becomes futile. Under such a model, the convergence rates are identical to those achieved when $n_s = 0$. Essentially, any model with $\alpha_\delta \le \alpha_m$ cannot outperform the conventional learning setup. The effectiveness of transfer learning is justifiable only for models such that α_δ is strictly greater than α_m. This observation gives rise to the phase transition phenomenon, highlighting the importance of modeling assumption $\alpha_\delta > \alpha_m$ for meaningful discussions regarding the effectiveness of transfer learning.

Under the extra modeling assumption $\alpha_\delta > \alpha_m$, we can find crucial differences between two regimes: high and low sampling frequencies. In the regime of low sampling frequency where $m_t \ll n_t^{1/2\alpha_m}$, transfer learning can lead to faster convergence rates compared to the case of utilizing the target sample only. However, in the high sampling frequency regime where the target sampling frequency m_t grows at least as rapidly as the rate $n_t^{1/2\alpha_m}$, transfer learning cannot improve the performance beyond the optimal parametric rate of n_t^{-1}, resulting in a phase transition at $m_t \asymp n_t^{1/2\alpha_m}$. The remarkable finding is that this transition boundary remains invariant, regardless of whether the design is common or independent, which underscores the structural equivalence between the two designs.

We turn our attention to the low sampling frequency regime where $m_t \ll n_t^{1/2\alpha_m}$ holds. In this regime, we shall explore distinct conditions under which transfer learning can enhance performance under common and independent designs, respectively. Under a common design, transfer learning proves effective when $m_s^{-2\alpha_m} + (Kn_s)^{-1} \ll m_t^{-2\alpha_m}$, whereas it is effective if $(Km_sn_s)^{-2\alpha_m/(2\alpha_m+1)} + (Kn_s)^{-1} \ll (m_tn_t)^{-2\alpha_m/(2\alpha_m+1)}$ under an independent design. To a certain extent, both designs require a sufficient number of total subjects in the source samples ($Kn_s \gg m_t^{2\alpha_m}$ and $Kn_s \gg (m_tn_t)^{2\alpha_m/(2\alpha_m+1)}$, respectively), but neither implies the other. The second condition, however, exhibits a significant disparity between the two designs. In a common design, a higher sampling frequency in the source samples ($m_s \gg m_t$) is necessary, while in an independent design, a larger total number of observations in the source samples ($Km_sn_s \gg m_tn_t$) is required. Finally, another noteworthy observation is that the size of the source group K plays a more influential role in an independent design than in a common design. This is mainly because increasing K leads to a larger total number of observations in the source samples, but not to a higher sampling frequency in the source samples.

A comparison of the minimax risks between common and independent designs reveals that the rate of convergence for an independent design is consistently faster or equal to that of a common design. Consequently, when confronted with the choice between these two designs, assuming all other factors remain constant, selecting an independent design should be always preferred. This observation aligns with our intuition that estimating the functional mean will be more accurate when approached holistically rather than atomistically. In this regard, an independent design possesses inherent advantages as it explores a wider range of design points than a common design does, even though the accuracy of estimation for each design point may be relatively lower.

Although our primary focus is on transfer learning, these results also provide new insights into functional mean estimation in the conventional setup. By setting both the source sample size n_s and sampling frequency m_s to zero, we obtain rates for functional mean estimation without transfer learning under both common and independent designs. Our rates of convergence, $m_t^{-2\alpha_m} + n_t^{-1}$ and $(m_tn_t)^{-2\alpha_m/(2\alpha_m+1)} + n_t^{-1}$, serve as a generalization of the rates derived by Cai and Yuan [10]. This generalization is particularly significant when the target mean function is rough and even not differentiable, as the current framework still guarantees consistent estimation. Such flexibility is a key advantage of our model.

In this paper, we introduce two novel algorithms, $\mathcal{A}_{\mathrm{CL}}$ and $\mathcal{A}_{\mathrm{TL}}$, where the latter utilizes the source samples to transfer knowledge but the former does not. These algorithms can be applied under both designs and achieve the optimal rate with carefully chosen parameters. However, these optimal parameters differ between the two designs. Under an independent design, the optimal values depend on the unknown modeling parameters α_m and α_δ while in a common design, they do not. Therefore, for constructing an adaptive and optimal procedure under a common design, it suffices to incorporate the two algorithms, $\mathcal{A}_{\mathrm{CL}}$ and $\mathcal{A}_{\mathrm{TL}}$, while under an independent design, one additionally needs to tune parameters adaptively. This highlights the difference between the two designs in constructing adaptive estimators. We propose adaptive procedures $\mathcal{A}_{\mathrm{ALC}}$ and $\mathcal{A}_{\mathrm{ALI}}$ for common and independent designs, respectively, and demonstrate their optimality.

1.3. *Organization and notation.* The rest of the paper is organized as follows. We complete this section with basic notation and then study transfer learning for functional mean estimation under a common design in Section 2. The optimal rate of convergence is established and an adaptive algorithm is proposed. We next shift our focus to an independent design in Section 3. In Section 4, we conduct a simulation study to investigate the numerical performance of the proposed adaptive algorithms for both common and independent designs. The numerical results agree with the theoretical findings presented earlier. Section 5 discusses

further research directions. The proofs of the main results are provided in Section 6, while additional proofs are presented in the Supplementary Material [6].

Throughout the paper, we consider subject size (n_t, n_s), sampling frequency (m_t, m_s) and the number of source samples (K) as the primary asymptotic components while all other quantities are treated as constants. We adhere to the standard for big-Oh(O), big-Omega(Ω) and big-Theta(Θ) notation. On occasion, for brevity, we use the symbols $\lesssim$, $\gtrsim$ and $\asymp$ as replacements for big-Oh(O), big-Omega(Ω) and big-Theta(Θ), respectively. Similarly, we follow the convention for the little-Oh(o) notation and briefly express $a \ll b$ or $b \gg a$ if and only if $a = o(b)$ is true. Although the constant term is only allowed in the standard big-Oh(O), big-Omega(Ω) and big-Theta(Θ) notation, their tilde-variants, big-Oh-tilde($\widetilde{O}$), big-Omega-tilde($\widetilde{\Omega}$) and big-Theta-tilde($\widetilde{\Theta}$) accept logarithmic polynomial terms of n_t, n_s, m_t, m_s and K. For any function $f : [0, 1] \to \mathbb{R}$, it is straightforward to define functional $\mathcal{L}^2$-norm, $\|f\|_{\mathcal{L}^2(I)}$ and $\mathcal{L}^\infty$-norm, $\|f\|_{\mathcal{L}^\infty(I)}$, restricted on some interval $I \subset [0, 1]$. When I stands for the whole domain $[0, 1]$, we follow the convention of leaving out I in the norm notation.

2. Transfer learning for functional mean estimation under a common design. In this section, we will explore transfer learning for functional mean estimation under a common design, where curves are observed at the same design points within the same target or source subject groups. The goal is to introduce a novel algorithm for estimating the functional mean within the transfer learning framework and derive an upper bound for its integrated mean squared error (IMSE). We then establish the minimax rate of convergence and the optimality of the proposed algorithm by obtaining a matching lower bound. Finally, we introduce an adaptive and optimal algorithm that can be utilized in practical applications without requiring knowledge of modeling parameters.

While the results presented in this section are valid for randomly selected common design points, for the sake of clarity in our exposition, we simplify by assuming that these design points are deterministic. Additionally, we adopt the convention, without loss of generality, that the common design points are arranged in ascending order. To summarize, a common design is characterized by two collections of common design points, $\{T_j^{[t]} : j = 1, \ldots, m_t\}$ and $\{T_j^{[s]} : j = 1, \ldots, m_s\}$, such that $T_{j_1}^{[t]} \le T_{j_2}^{[t]}$ and $T_{j_1}^{[s]} \le T_{j_2}^{[s]}$ for any $j_1 \le j_2$ as well as the following holds:

$$T_{ij}^{[t]} = T_j^{[t]} \qquad \text{for all } j = 1, \ldots, m_t \text{ and } i = 1, \ldots, n_t,$$

$$T_{ij}^{[s,k]} = T_j^{[s]} \qquad \text{for all } j = 1, \ldots, m_s,\ i = 1, \ldots, n_s \text{ and } k = 1, \ldots, K.$$

To describe the optimality of functional mean estimation, we consider a statistical model denoted by $\mathcal{P}$, which is a collection of certain probability measures. The mean and difference functions of measures in $\mathcal{P}$ satisfy equation (3) and the measures themselves also satisfy the uniform sub-Gaussian condition (Assumption 1). Finally, the collection of target and source samples, $\{\mathcal{D}^{[t]}, \mathcal{D}^{[s,1]}, \ldots, \mathcal{D}^{[s,K]}\}$, is assumed to be independent. We continue to employ the same assumptions in Section 3 where we argue the optimality under an independent design.

ASSUMPTION 1. The random functions and noises are uniformly sub-Gaussian variables with a positive variance proxy τ^2. As per Wainwright [36], we assume that for every $u > 0$, the following holds: for any given $x \in [0, 1]$ and $k = 1, \ldots, K$,

$$\left.\begin{aligned} &\mathbb{P}\big(|X_1^{[t]}(x) - f^{[t]}(x)| \ge u\big) \vee \mathbb{P}\big(|X_1^{[s,k]}(x) - f^{[s,k]}(x)| \ge u\big) \\ &\mathbb{P}\big(|\varepsilon_{11}^{[t]}| \ge u\big) \vee \mathbb{P}\big(|\varepsilon_{11}^{[s,k]}| \ge u\big) \end{aligned}\right\} \le 2e^{-u^2/2\tau^2}$$

2.1. *Methodology and upper bound.* In this subsection, we present a novel algorithm for estimating the target mean function. Before delving into the transfer learning problem of our main interest, it is useful to revisit the conventional framework for functional mean estimation, which is a crucial component of the transfer learning algorithm. This involves estimating the target mean function $f^{[t]}$ when no source samples are available, that is, $n_s = 0$. Although this problem has been extensively investigated by Cai and Yuan [10], and the focus of this paper is on transfer learning, it is still valuable to discuss it separately.

THEOREM 2.1 (The minimax risk under conventional setup and common design). *Suppose no source samples are available, that is, $n_s = 0$ and the fixed and common design points for the target sample satisfy $\max_{1\le j\le m_t+1}(T_j^{[t]} - T_{j-1}^{[t]}) \le C_t/m_t$ for some constant $C_t > 0$, where $T_0^{[t]} = 0$ and $T_{m_t+1}^{[t]} = 1$. Then*

$$\inf_{\widehat{f}^{[t]}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\big\|\widehat{f}^{[t]} - f^{[t]}\big\|_{\mathcal{L}^2}^2 = \widetilde{\Theta}\bigg(L_m^2 m_t^{-2\alpha_m} + \frac{1}{n_t}\bigg),$$

where the infimum is taken over all estimators $\widehat{f}^{[t]} = \widehat{f}^{[t]}(\mathcal{D}^{[t]})$ based on the target sample.

Theorem 2.1 generalizes the minimax rate shown by Cai and Yuan [10], $\Theta(m_t^{-2r} + n_t^{-1})$, where both the curve $X^{[t]}$ and mean function $f^{[t]}$ are assumed to be r times differentiable with $r \in \mathbb{Z}^+$. Notably, the current framework remains consistent in estimating the mean function even when it is not smooth ($\alpha_m < 1$), whereas the previous one does not. Moreover, our approach provides greater flexibility by not requiring the smoothness of the target curve $X^{[t]}$. Therefore, the proposed model represents a significant improvement over the existing literature on functional mean estimation.

The primary contribution of Theorem 2.1 lies in the estimation method, characterized by the algorithm $\mathcal{A}$. This algorithm is equally pivotal in the transfer learning setup. At the core of the algorithm $\mathcal{A}$, it partitions the domain $[0, 1]$ into subintervals and performs polynomial regression on each subinterval to estimate the mean function. Algorithm 1 outlines the step-by-step instructions for implementing this algorithm.

The purpose of introducing a randomized collection $\mathcal{D}_r$ in Algorithm $\mathcal{A}$ is to ensure the independence of the observations in the collection. Unlike Algorithm $\mathcal{A}$, when we naively take the subcollection $\widetilde{\mathcal{D}}_r := \{(T, Y) \in \mathcal{D} : T \in I_r\}$ of all observations whose design points fall into the interval I_r, they are no longer independent in general. This is because $\widetilde{\mathcal{D}}_r$ may contain more than one observation from the same curve depending on a common sampling scheme. This poses some technical challenges when analyzing the resulting estimator. To address this issue, we use a randomized reduction technique. The idea is to randomly select one observation within each interval I_r per curve and discard the rest. This reduction procedure ensures that the observations used for polynomial regression are independent.

We must ensure that observations are neither underrepresented nor overrepresented when randomly selecting them, as only one discrete observation per subject is chosen. Under an independent design, we assume that the density generating random design points is bounded both above and below by constants (See Theorems 3.1 and 3.2), which solves the under and overrepresentation problem. However, under a common design, we assume fixed and common design points are not too far from each other (See Theorems 2.1 and 2.2), which only handles the underrepresentation issue. To address the issue of overrepresentation of common design, the proposed algorithm employs a (randomized) subcollection $\widetilde{\mathcal{D}}^{(i)} \subset \mathcal{D}^{(i)}$. The design points of $\widetilde{\mathcal{D}}^{(i)}$ are required to be spaced at a minimum separation of $1/m$ while still encompassing those found in $\mathcal{D}^{(i)}$ within a maximum distance of $1/2m$. The introduction of subcollection $\widetilde{\mathcal{D}}^{(i)}$ is thus expected to alleviate the issue of overrepresentation.

Algorithm 1 Randomized local polynomial regression with thresholding $\mathcal{A}(\mathcal{D}, b, d, M)$

Require: A collection $\mathcal{D} = \{(T_{ij}, Y_{ij}) : j = 1, \dots, m,\ i = 1, \dots, n\}$ of observations, a bandwidth $b \in 1/\mathbb{Z}^+$, a degree $d \in \mathbb{Z}^+$ of polynomial, and a threshold $M > 0$.

Partition the domain $[0, 1]$ into b^{-1}-many intervals of length b. We denote those intervals by I_r $(r = 1, \dots, b^{-1})$ from left to right.

Let us denote by $\mathcal{D}^{(i)} := \{(T_{ij}, Y_{ij}) : j = 1, \dots, m\}$ the collection of observations from the same subject $i \in \{1, \dots, n\}$. For index $r = 1, \dots, b^{-1}$, we take randomized collection $\mathcal{D}_r := \{(\mathbb{T}_{i,r}, \mathbb{Y}_{i,r}) : i = 1, \dots, n\}$ of observations where $(\mathbb{T}_{i,r}, \mathbb{Y}_{i,r})$ is randomly chosen from the following process:

if the collection $\mathcal{D}$ comes from common design **then**

    Consider any $(1/m)$-packing and $(1/2m)$-covering subcollection $\widetilde{\mathcal{D}}^{(i)}$ of $\mathcal{D}^{(i)}$ where the distance between observations is computed based on design points. The random $(\mathbb{T}_{i,r}, \mathbb{Y}_{i,r})$ is now uniformly chosen from $\{(T, Y) \in \widetilde{\mathcal{D}}^{(i)} : T \in I_r\}$.

else if the collection $\mathcal{D}$ comes from independent design **then**

    The random $(\mathbb{T}_{i,r}, \mathbb{Y}_{i,r})$ is uniformly chosen from $\{(T, Y) \in \mathcal{D}^{(i)} : T \in I_r\}$.

end if

Implement the polynomial regression of degree d on each collection $\mathcal{D}_r$ $(r = 1, \dots, b^{-1})$ of observations. In other words, we are enough to compute for each $r = 1, \dots, b^{-1}$,

$$(\check{a}_{r,0}, \check{a}_{r,1}, \dots, \check{a}_{r,d}) := \operatorname*{argmin}_{(a_0, a_1, \dots, a_d) \in \mathbb{R}^{d+1}} \sum_{(T,Y) \in \mathcal{D}_r} \left[Y - \sum_{s=0}^{d} a_s \left(\frac{T - (q-1)b}{b} \right)^s \right]^2.$$

If the solution is not available, simply take $(\check{a}_{r,0}, \check{a}_{r,1}, \dots, \check{a}_{r,d}) = 0$.

Compute the local polynomial regression estimator $\check{f} : [0, 1] \to \mathbb{R}$ by

$$\check{f}(x) := \sum_{r=1}^{b^{-1}} \sum_{s=0}^{d} \check{a}_{r,s} \left(\frac{x - (r-1)b}{b} \right)^s \mathbb{1}(x \in I_r) \quad (0 \le x \le 1).$$

Output the final estimator $\widehat{f} : [0, 1] \to \mathbb{R}$ through thresholding:

$$\widehat{f} := \begin{cases} \check{f} & \text{if } \|\check{f}\|_{\mathcal{L}^\infty(I_r)} \le M, \\ 0 & \text{otherwise,} \end{cases} \quad \text{on each } I_r \ (r = 1, \dots, b^{-1}).$$

We will now present the optimal algorithm $\mathcal{A}_{\mathrm{CL}}(b_t, d_t, M_t)$ for estimating the mean function under the conventional setup and common design. This algorithm is essentially the same as algorithm $\mathcal{A}$ with the following four inputs: the target sample $\mathcal{D}^{[t]}$, a bandwidth $b_t \in 1/\mathbb{Z}^+$, a degree $d_t \in \mathbb{Z}^+$ of local polynomial and a threshold $M_t > 0$. The output $\widehat{f}^{[t]}_{\mathrm{CL}}$ of the conventional learning algorithm $\mathcal{A}_{\mathrm{CL}}$ indeed attains the following minimax rate of convergence in Theorem 2.1 when input parameters are carefully selected. The detailed selection scheme will be shortly postponed to Theorem 2.2:

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E} \| \widehat{f}^{[t]}_{\mathrm{CL}} - f^{[t]} \|^2_{\mathcal{L}^2} \le O\left(L_m^2 m_t^{-2\alpha_m} + \frac{\log^2 n_t}{n_t} \right).$$

Let us shift our focus back to the transfer learning setup of the primary interest. In addition to the target sample, we have access to the source samples that could be potentially beneficial for estimating the target mean function more accurately. In this setting, the transfer learning algorithm, represented by $\mathcal{A}_{\mathrm{TL}}$, incorporates both the target and source samples and leverages their structures to achieve a more precise estimation of the target mean function. The basic idea is to decompose the target mean function $f^{[t]}$ into the source part $f^{[s]} := K^{-1} \sum_{k=1}^{K} f_k^{[s]}$

Algorithm 2 Transfer learning for mean function $\mathcal{A}_{\mathrm{TL}}(b_s, b_\delta, d_s, d_\delta, M_s, M_\delta)$

Require: Two bandwidths $b_s, b_\delta \in 1/\mathbb{Z}^+$, two degrees $d_s, d_\delta \in \mathbb{Z}^+$ of polynomial and two thresholds $M_s, M_\delta > 0$.

1: Execute $\mathcal{A}(\mathcal{D}^{[s]}, b_s, d_s, M_s)$ with the combined source sample $\mathcal{D}^{[s]} := \bigcup_{k=1}^{K} \mathcal{D}^{[s,k]}$. The result of this algorithm is denoted as $\widehat{f}^{[s]}$.

2: Compute a new sample $\mathcal{D}^{[\delta]} := \{(T_{ij}^{[t]}, Y_{ij}^{[t]} - \widehat{f}^{[s]}(T_{ij}^{[t]})) : j = 1, \ldots, m_t,\ i = 1, \ldots, n_t\}$.

3: Execute $\mathcal{A}(\mathcal{D}^{[\delta]}, b_\delta, d_\delta, M_\delta)$. The output of this algorithm is denoted by $\widehat{\delta}^{[s]}$.

4: Output our final estimator $\widehat{f}^{[t]} := \widehat{f}^{[s]} + \widehat{\delta}^{[s]}$.

and the difference part $\delta^{[s]} := f^{[t]} - f^{[s]}$ and to estimate them separately. The step-by-step guide for implementing this algorithm is provided in Algorithm 2.

It is important to notice that the transfer learning algorithm $\mathcal{A}_{\mathrm{TL}}$ may not always perform well. For instance, if the source samples have too few observations or are generated adversarially from models satisfying $\alpha_\delta < \alpha_m$, we cannot gain any benefits from using them. In such cases, the conventional learning algorithm $\mathcal{A}_{\mathrm{CL}}$ can still be used in the transfer learning setup by simply ignoring the collected source samples. The optimal algorithm for conventional learning, $\mathcal{A}_{\mathrm{CL}}$, thus provides a baseline for estimation performance, as it has been shown to achieve the convergence rate of $\widetilde{\Theta}(m_t^{-2\alpha_m} + n_t^{-1})$. The transfer learning algorithm $\mathcal{A}_{\mathrm{TL}}$ is only advantageous if it performs better than this baseline rate; otherwise, it makes sense to stick with the conventional learning algorithm $\mathcal{A}_{\mathrm{CL}}$. Reflecting this idea, the methodology suggested in Theorem 2.2 appropriately combines $\mathcal{A}_{\mathrm{TL}}$ with $\mathcal{A}_{\mathrm{CL}}$ to yield an optimal estimate.

THEOREM 2.2 (Upper bound under a common design). *Suppose*

$$\max_{1\le j\le m_t+1}\left(T_j^{[t]} - T_{j-1}^{[t]}\right) \le \frac{C_t}{m_t} \quad \text{and} \quad \max_{1\le j\le m_s+1}\left(T_j^{[s]} - T_{j-1}^{[s]}\right) \le \frac{C_s}{m_s}$$

where $C_t, C_s > 0$ are some constants as well as $T_0^{[t]} = T_0^{[s]} = 0$ and $T_{m_t+1}^{[t]} = T_{m_t+1}^{[s]} = 1$. Consider the conventional learning estimator $\widehat{f}_{\mathrm{CL}}^{[t]}$, which is the output of the algorithm $\mathcal{A}_{\mathrm{CL}}(b_t, d_t, M_t)$ with the following specifications:

- *Any constant degree $d_t \ge \omega(\alpha_m)$,*
- *Any bandwidth $b_t = \lceil m_t/2B_t(d_t+1)\rceil^{-1}$ for constant $B_t \ge C_t$,*
- *The threshold $M_t = \log n_t$.*

Plus, let $\widehat{f}_{\mathrm{TL}}^{[t]}$ be the output of the transfer learning algorithm $\mathcal{A}_{\mathrm{TL}}(b_s, b_\delta, d_s, d_\delta, M_s, M_\delta)$ with the following specifications:

- *Any constant degrees $d_s \ge \omega(\alpha_m)$ and $d_\delta \ge \omega(\alpha_\delta)$.*
- *Any bandwidths $b_s = \lceil m_s/2(d_s+1)B_s\rceil^{-1}$ and $b_\delta = \lceil m_t/2(d_\delta+1)B_\delta\rceil^{-1}$ for any given constants $B_s \ge C_s$ and $B_\delta \ge C_t$.*
- *Two thresholds $M_s = \log n_s$ and $M_\delta = \log n_t n_s$.*

The estimator $\widehat{f}^{[t]}$ for the mean function $f^{[t]}$ is now defined as one of them:

$$\widehat{f}^{[t]} = \begin{cases} \widehat{f}_{\mathrm{CL}}^{[t]} & \text{when } R_{\mathrm{C}}(\widehat{f}_{\mathrm{CL}}^{[t]}) \le R_{\mathrm{C}}(\widehat{f}_{\mathrm{TL}}^{[t]}), \\ \widehat{f}_{\mathrm{TL}}^{[t]} & \text{when } R_{\mathrm{C}}(\widehat{f}_{\mathrm{CL}}^{[t]}) > R_{\mathrm{C}}(\widehat{f}_{\mathrm{TL}}^{[t]}), \end{cases}$$

where the following quantities are additionally defined:

$$R_{\mathrm{C}}(\widehat{f}_{\mathrm{CL}}^{[t]}) := L_m^2 m_t^{-2\alpha_m} + \frac{\log^2 n_t}{n_t},$$

$$R_{\mathrm{C}}(\widehat{f}_{\mathrm{TL}}^{[t]}) := L_\delta^2 m_t^{-2\alpha_\delta} + \frac{\log^2(n_t n_s)}{n_t} + L_m^2 m_s^{-2\alpha_m} + \frac{\log^2(Kn_s)}{Kn_s}.$$

In this situation, we obtain

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\|\widehat{f}^{[t]} - f^{[t]}\|_{\mathcal{L}^2}^2 \lesssim R_{\mathrm{C}}(\widehat{f}_{\mathrm{CL}}^{[t]}) \wedge R_{\mathrm{C}}(\widehat{f}_{\mathrm{TL}}^{[t]}).$$

Theorem 2.2 highlights a crucial aspect of the transfer learning algorithm $\mathcal{A}_{\mathrm{TL}}$, which achieves the convergence rate of $\widetilde{\Theta}(m_t^{-2\alpha_\delta} + n_t^{-1} + m_s^{-2\alpha_m} + (Kn_s)^{-1})$ with appropriately chosen parameters. Corresponding to the decomposition $f^{[t]} = f^{[s]} + \delta^{[s]}$ of the target mean function, this rate can also be decomposed into two parts: the optimal estimation error of the source part $\widetilde{\Theta}(m_s^{-2\alpha_m} + (Kn_s)^{-1})$ and the difference part $\widetilde{\Theta}(m_t^{-2\alpha_\delta} + n_t^{-1})$. The final estimator $\widehat{f}^{[t]}$ is now selected as an output of any algorithm that achieves a better convergence rate. We therefore obtain the best possible rate of convergence utilizing both learning algorithms, $\mathcal{A}_{\mathrm{TL}}$ and $\mathcal{A}_{\mathrm{CL}}$.

It should be noted that both conventional learning estimator $\widehat{f}_{\mathrm{CL}}^{[t]}$ and the transfer learning estimator $\widehat{f}_{\mathrm{TL}}^{[t]}$ never depend on unknown parameters. This means that both algorithms, $\mathcal{A}_{\mathrm{TL}}$ and $\mathcal{A}_{\mathrm{CL}}$, are naturally adaptive under a common design with optimally chosen parameters. However, selecting between two outputs from two algorithms is still challenging, and the strategy for the final estimator $\widehat{f}^{[t]}$ suggested in Theorem 2.2 is not effective. In practice, the smoothness parameters, α_m and α_δ, are typically unknown, which makes it impossible to compare the upper bounds of maximal risk, $R_{\mathrm{C}}(\widehat{f}_{\mathrm{TL}}^{[t]})$ and $R_{\mathrm{C}}(\widehat{f}_{\mathrm{CL}}^{[t]})$. Therefore, another adaptive and optimal procedure is necessary to decide whether to employ source samples and transfer learning. One novel solution to this issue will be presented in Section 2.3.

2.2. *Matching lower bound.* The following theorem demonstrates a lower bound of the maximal risk in estimating the target mean function under a common design.

THEOREM 2.3 (Lower bound under a common design). *Under a common design,*

$$\inf_{\widehat{f}^{[t]}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\|\widehat{f}^{[t]} - f^{[t]}\|_{\mathcal{L}^2}^2 \gtrsim \left(L_m^2 m_t^{-2\alpha_m} + \frac{1}{n_t}\right) \wedge \left(L_\delta^2 m_t^{-2\alpha_\delta} + \frac{1}{n_t} + L_m^2 m_s^{-2\alpha_m} + \frac{1}{Kn_s}\right),$$

where the infimum is taken over all possible estimators $\widehat{f}^{[t]} = \widehat{f}^{[t]}(\mathcal{D}^{[t]}, \mathcal{D}^{[s,1]}, \ldots, \mathcal{D}^{[s,K]})$.

The lower bound presented in Theorem 2.3 matches exactly with the upper bound shown in Theorem 2.2, up to logarithmic terms. This establishes the minimax risk under a common design given by

$$\left(m_t^{-2\alpha_m} + \frac{1}{n_t}\right) \wedge \left(m_t^{-2\alpha_\delta} + \frac{1}{n_t} + m_s^{-2\alpha_m} + \frac{1}{Kn_s}\right).$$

This rate is decomposed into two terms, each quantifying the difficulty of estimating the target mean function $f^{[t]}$ but with different scenarios:

- $\widetilde{\Theta}(m_t^{-2\alpha_m} + n_t^{-1})$ represents the conventional learning setup.
- $\widetilde{\Theta}(m_t^{-2\alpha_\delta} + n_t^{-1} + m_s^{-2\alpha_m} + (Kn_s)^{-1})$ stands for the transfer learning setup.

It is reasonable to measure the entire difficulty of the problem by the minimum of these two terms. Furthermore, the second term can be decomposed into $\widetilde{\Theta}(m_s^{-2\alpha_m} + (Kn_s)^{-1})$ and $\widetilde{\Theta}(m_t^{-2\alpha_\delta} + n_t^{-1})$, which correspond to the minimax risk of estimation on the source and target samples, respectively. This decomposition highlights the crux of transfer learning: by paying the cost of estimation on the source sample, the target sample can be treated as if the

target mean function is given by α_δ-smooth, instead of α_m-smooth. As long as the cost is not exceedingly expensive and substituting smoothness results in increased smoothness, this leads to a more accurate estimation and a faster rate of convergence.

By comparing the minimax risks between the conventional and transfer learning settings, we can determine whether and under what conditions the source samples and transfer learning can enhance performance. The comparison is relatively simple because the minimax rate for the transfer learning setup already includes that for the conventional learning setup. What makes this process particularly interesting is the presence of a phase transition under which the efficacy of transfer learning experiences a significant shift.

To begin with, there is no rationale for incurring the cost of substituting smoothness if it does not result in increased smoothness. In other words, if our model is given with $\alpha_\delta \leq \alpha_m$, transfer learning has an adverse effect because estimating α_δ-smooth functions is, at the very least, as challenging as estimating α_m-smooth functions. The effectiveness of transfer learning is justifiable only for models such that α_δ is strictly greater than α_m. In such cases, we can exploit the benefits of substituting smoothness and anticipate surpassing the minimax risk of the conventional learning algorithm. This gives rise to the phase transition phenomenon and further discussions about the effectiveness of transfer learning become meaningful under the additional assumption that $\alpha_\delta > \alpha_m$ holds.

From this point onwards, suppose that our model satisfies $\alpha_\delta > \alpha_m$. In the high-frequency regime, defined as $m_t \gtrsim n_t^{1/2\alpha_m}$, the minimax risk for the conventional learning algorithm is $\widetilde{\Theta}(n_t^{-1})$ because this rate is always as rapid as $\widetilde{\Theta}(m_t^{-2\alpha_\delta} + n_t^{-1} + m_s^{-2\alpha_m} + (Kn_s)^{-1})$, the additional rate introduced by the transfer learning setup. In other words, the conventional learning algorithm already achieves the parametric rate $\widetilde{\Theta}(n_t^{-1})$ concerning the number of target subjects, regardless of the existence of source samples or the level of smoothness. Unless difference functions $\delta^{[s,k]}$ $(k = 1, \ldots, K)$ are exactly known, the minimax risk for estimating the target mean function must be at least the parametric rate $\widetilde{\Theta}(n_t^{-1})$. As a result, transfer learning cannot be effective as there is no more room for improvement in the high-frequency regime. This kind of phenomenon is expected to be universal across problems beyond the scope of the functional mean estimation.

In the low-frequency regime with $m_t \ll n_t^{1/2\alpha_m}$, the minimax risk in the conventional learning setting is $\widetilde{\Theta}(m_t^{-2\alpha_m})$. In this case, transfer learning improves the estimation performance of the target mean function provided the following condition holds:

$$\widetilde{\Theta}\big(m_s^{-2\alpha_m} + (Kn_s)^{-1}\big) \ll \widetilde{\Theta}\big(m_t^{-2\alpha_m}\big).$$

This condition implies that estimating the average of source mean functions, denoted as $f^{[s]}$, is a less challenging task compared to estimating the target mean function $f^{[t]}$ assuming the conventional learning settings. Moreover, this condition remains unaffected by the smoothness parameter α_δ, indicating that the difference functions $\delta^{[s,k]}$ $(k = 1, \ldots, K)$ only need to be slightly smoother than the target mean function $f^{[t]}$.

In breaking down the condition for the low-frequency regime, we can identify two essential requirements. First, the source subjects in total should be abundant, satisfying $Kn_s \gg m_t^{2\alpha_m}$. Second, the frequency of design points in the source samples, m_s, should significantly surpass that of the target sample, m_t, indicating $m_s \gg m_t$. It is worth noting that the size of the source group denoted as K, plays a restricted role under a common design. In cases where the source samples lack an adequate number of design points ($m_s \lesssim m_t$), simply increasing the size of the source group will not yield substantial benefits in transfer learning. As long as the source samples offer a sufficient pool of design points ($m_s \gg m_t$), increasing K can be an effective strategy to obtain a faster rate of convergence with transfer learning.

Algorithm 3 Adaptive transfer learning for mean function under a common design $\mathcal{A}_{\mathrm{ALC}}$

1: Randomly partition the target sample $\mathcal{D}^{[t]}$ into two subsamples, denoted as $\mathcal{D}^{[t]}_{\mathrm{train}}$ and $\mathcal{D}^{[t]}_{\mathrm{test}}$, based on subjects. Specifically, perform a random split of the index set $\{1,\dots,2n_t\}$ into two partitions, $\mathcal{I}^{[t]}_{\mathrm{train}}$ and $\mathcal{I}^{[t]}_{\mathrm{test}}$, such that $|\mathcal{I}^{[t]}_{\mathrm{train}}| = |\mathcal{I}^{[t]}_{\mathrm{test}}| = n_t$. We define

$$\mathcal{D}^{[t]}_{\mathrm{train}} := \{(T^{[t]}_j, Y^{[t]}_{ij}) : i \in \mathcal{I}^{[t]}_{\mathrm{train}},\ j = 1,\dots,m_t\},$$

$$\mathcal{D}^{[t]}_{\mathrm{test}} := \{(T^{[t]}_j, Y^{[t]}_{ij}) : i \in \mathcal{I}^{[t]}_{\mathrm{test}},\ j = 1,\dots,m_t\}.$$

2: Execute both $\mathcal{A}_{\mathrm{CL}}(b_t, d_t, M_t)$ and $\mathcal{A}_{\mathrm{TL}}(b_s, b_\delta, d_s, d_\delta, M_s, M_\delta)$ following the same specifications as outlined in Theorem 2.2, with the only distinction being that the target sample is provided as $\mathcal{D}^{[t]}_{\mathrm{train}}$, not $\mathcal{D}^{[t]}$. The outputs are denoted as $\widehat{f}^{[t]}_{\mathrm{CL}}$ and $\widehat{f}^{[t]}_{\mathrm{TL}}$, respectively.

3: Output the following estimator $\widehat{g}^{[t]}_*$. If a tie occurs, use any randomization to break it:

$$\widehat{g}^{[t]}_* := \operatorname*{argmin}_{\widehat{g}^{[t]} \in \{\widehat{f}^{[t]}_{\mathrm{CL}}, \widehat{f}^{[t]}_{\mathrm{TL}}\}} \sum_{i \in \mathcal{I}^{[t]}_{\mathrm{test}}} \sum_{j=1}^{m_t} (Y^{[t]}_{ij} - \widehat{g}^{[t]}(T^{[t]}_j))^2 (\Delta T^{[t]}_j),$$

where $T^{[t]}_0 := 0$, $T^{[t]}_{m_t+1} := 1$ and $\Delta T^{[t]}_j := T^{[t]}_j - T^{[t]}_{j-1}$ for each $j = 1,\dots,m_t+1$.

2.3. *Adaptive estimation.* Although the optimal estimators, $\widehat{f}^{[t]}_{\mathrm{CL}}$ and $\widehat{f}^{[t]}_{\mathrm{TL}}$, introduced in Theorem 2.2 are naturally adaptive, selecting between them is not. We propose a new adaptive procedure to address this issue. Algorithm 3 is an outline of the adaptive learning algorithm under a common design, called $\mathcal{A}_{\mathrm{ALC}}$. It is further assumed for brevity that the target sample $\mathcal{D}^{[t]}$ contains $2n_t$ subjects, which does not affect the rate of convergence.

The success of Algorithm 3 mainly hinges on the train-test split. Half of the target sample, denoted as $\mathcal{D}^{[t]}_{\mathrm{train}}$, serves as the training data for two algorithms, while the remaining portion, referred to as $\mathcal{D}^{[t]}_{\mathrm{test}}$, is employed to evaluate their empirical performance. Although the design points of the target sample are fixed, we have assumed $\max_{j=1,\dots,m_t+1}(\Delta T^{[t]}_j) = o(m_t^{-1})$, ensuring that the Riemann sum serves as a robust proxy for the functional $\mathcal{L}^2$-loss. Not only that, the computational cost for the adaptation step remains negligible, as it merely involves the selection between two data-driven estimators, $\widehat{f}^{[t]}_{\mathrm{CL}}$ and $\widehat{f}^{[t]}_{\mathrm{TL}}$.

Notice that the base algorithms, $\mathcal{A}_{\mathrm{TL}}$ and $\mathcal{A}_{\mathrm{CL}}$, generate a randomized estimator. Consequently, the adaptive learning algorithm $\mathcal{A}_{\mathrm{ALC}}$ must be randomized as well. To improve the performance in finite samples, we execute this adaptive procedure $r_{\max}$ times and calculate the average of the resulting estimates. This approach shares a similar idea with bagging estimators suggested by Breiman [2]. The optimal minimax rate of convergence, up to logarithmic terms, is achieved by this algorithm, as demonstrated by the subsequent theorem.

THEOREM 2.4 (Adaptive estimation under a common design). *Under the same assumptions as Theorem 2.2, we consider a maximum number of repetitions, $r_{\max} \in \mathbb{Z}^+$ and let $\widehat{g}^{[t]}_r$ $(r = 1,\dots,r_{\max})$ denote the output of the rth execution of the algorithm $\mathcal{A}_{\mathrm{ALC}}$. By averaging these estimates, we obtain the final estimator*:

$$\widehat{f}^{[t]} = \frac{1}{r_{\max}} \sum_{r=1}^{r_{\max}} \widehat{g}^{[t]}_r. \tag{5}$$

This adaptive estimator $\widehat{f}^{[t]}$ attains the same upper bound of Theorem 2.2:

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E} \| \widehat{f}^{[t]} - f^{[t]} \|^2_{\mathcal{L}^2} \lesssim R_{\mathrm{C}}(\widehat{f}^{[t]}_{\mathrm{CL}}) \wedge R_{\mathrm{C}}(\widehat{f}^{[t]}_{\mathrm{TL}}).$$

In short, the data-driven estimator (5) adaptively achieves the optimal rate of convergence over a comprehensive collection of function classes.

3. Transfer learning for functional mean estimation under an independent design. This section investigates the setting in which the design points of both the target and source samples are randomly drawn from the domain $[0, 1]$. Consider two probability distributions on $[0, 1]$, referred to as η_t and η_s, for target and sample design points, respectively. We make the following assumptions:

$$\big(T_{ij}^{[t]} : j = 1, \ldots, m_t,\ i = 1, \ldots, n_t\big) \overset{i.i.d.}{\sim} \eta_t,$$

$$\big(T_{ij}^{[s,k]} : j = 1, \ldots, m_s,\ i = 1, \ldots, n_s,\ k = 1, \ldots, K\big) \overset{i.i.d.}{\sim} \eta_s.$$

Apart from the assumption regarding design points, we maintain the same set of assumptions outlined in Section 2. In essence, these assumptions encompass the requirements for the mean and difference functions, as indicated in equation (3), and the fulfillment of the uniform sub-Gaussian condition (Assumption 1). Lastly, it is assumed that the collection of target and source samples, $\{\mathcal{D}^{[t]}, \mathcal{D}^{[s,1]}, \ldots, \mathcal{D}^{[s,K]}\}$, are independent. With these assumptions at hand, we can proceed to examine the optimal estimation of the target mean function $f^{[t]}$ under an independent design.

3.1. *Methodology and upper bound.* This subsection introduces an optimal algorithm for estimating the target mean function. In parallel to the discussion for a common design, let us reconsider the conventional setting for functional mean estimation but this time under an independent design. We again aim to estimate the target mean function $f^{[t]}$ when we have no source samples available ($n_s = 0$). The subsequent theorem investigates the minimax rate of convergence in this situation.

THEOREM 3.1 (The minimax risk under conventional setup and independent design). *Suppose the distribution η_t for the target design points is dominated by the Lebesgue measure and its density is bounded from below by a constant $C_t > 0$ and from above by $\Gamma_t > 0$. Then*

$$\inf_{\widehat{f}^{[t]}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E} \big\| \widehat{f}^{[t]} - f^{[t]} \big\|_{\mathcal{L}^2}^2 = \widetilde{\Theta}\bigg(L_m^{2/(2\alpha_m+1)} (m_t n_t)^{-2\alpha_m/(2\alpha_m+1)} + \frac{1}{n_t} \bigg),$$

where the infimum is taken over all estimators $\widehat{f}^{[t]} = \widehat{f}^{[t]}(\mathcal{D}^{[t]})$ based on the target sample.

The derived minimax rate of convergence again generalizes the optimal rate shown by Cai and Yuan [10], $\Theta((m_t n_t)^{-2r/(2r+1)} + n_t^{-1})$, where they assumed that both the curve $X^{[t]}$ and the mean function $f^{[t]}$ are differentiable r times ($r \in \mathbb{Z}^+$). Our framework provides greater flexibility and improvement over previous approaches for functional mean estimation, similar to what we have observed under a common design.

Even though the sampling designs are different, the current model has a similar structure to the corresponding one under a common design. We thus employ the same conventional learning algorithm $\mathcal{A}_{\mathrm{CL}}$ for estimating the target mean function. The output of the algorithm, denoted as $\widehat{f}_{\mathrm{CL}}^{[t]}$, achieves the minimax rate of convergence in Theorem 3.1, but the optimal choice of input parameters is distinct from that of a common design. We will provide more details about this choice in Theorem 3.2:

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E} \big\| \widehat{f}_{\mathrm{CL}}^{[t]} - f^{[t]} \big\|_{\mathcal{L}^2}^2 \leq O\bigg(L_m^{2/(2\alpha_m+1)} (m_t n_t)^{-2\alpha_m/(2\alpha_m+1)} + \frac{\log^2 n_t}{n_t} \bigg).$$

We will next shift our attention back to the transfer learning setup of our primary focus. Our approach for estimating the target mean function involves implementing and combining the two learning algorithms, $\mathcal{A}_{\mathrm{CL}}$ and $\mathcal{A}_{\mathrm{TL}}$ optimally. The underlying idea is that the transfer learning algorithm $\mathcal{A}_{\mathrm{TL}}$ may not always perform well, especially when the source samples are limited in number or generated in an adversarial way. In such cases, we can exploit the conventional learning algorithm $\mathcal{A}_{\mathrm{CL}}$ as a baseline. This approach is again not surprising as an analogous type of estimator has been demonstrated to achieve the matching lower bound under a common design. The optimal way to execute and combine both learning algorithms under an independent design is outlined in Theorem 3.2. Although the overall structure of the estimator is comparable to that of a common design, the optimal selection of parameters is substantially different.

THEOREM 3.2 (Upper bound under an independent design). *Suppose the distributions η_t and η_s for target and source design points are dominated by the Lebesgue measure, and their densities are bounded from below by a constant $C_t, C_s > 0$ and from above by $\Gamma_t, \Gamma_s > 0$, respectively. Consider the conventional learning estimator $\widehat{f}^{[t]}_{\mathrm{CL}}$, which is the output of algorithm $\mathcal{A}_{\mathrm{CL}}(b_t, d_t, M_t)$ with the following specifications*:

- *For any constant $B_t \geq \Gamma_t$, bandwidth*:
$$b_t = \lceil (L_m^2 m_t n_t)^{1/(2\alpha_m+1)} (\log n_t)^{-2/(2\alpha_m+1)} \rceil^{-1} \wedge \lceil 2B_t m_t \rceil^{-1},$$
- *Any constant degree $d_t \geq \omega(\alpha_m)$,*
- *The threshold $M_t = \log n_t$.*

Besides, let $\widehat{f}^{[t]}_{\mathrm{TL}}$ be the output of the transfer learning algorithm $\mathcal{A}_{\mathrm{TL}}(b_s, b_\delta, d_s, d_\delta, M_s, M_\delta)$ with the following specifications:

- *For any constants $B_s \geq \Gamma_s$ and $B_\delta \geq \Gamma_t$, bandwidths*:
$$b_s = \lceil (L_m^2 K m_s n_s)^{1/(2\alpha_m+1)} (\log(K n_s))^{-2/(2\alpha_m+1)} \rceil^{-1} \wedge \lceil 2B_s K m_s \rceil^{-1},$$
$$b_\delta = \lceil (L_\delta^2 m_t n_t)^{1/(2\alpha_\delta+1)} (\log(n_t n_s))^{-2/(2\alpha_\delta+1)} \rceil^{-1} \wedge \lceil 2B_\delta m_t \rceil^{-1}.$$
- *Any constant degrees $d_s \geq \omega(\alpha_m)$ and $d_\delta \geq \omega(\alpha_\delta)$.*
- *Two thresholds $M_s = \log n_s$ and $M_\delta = \log n_t n_s$.*

The estimator $\widehat{f}^{[t]}$ of the mean function $f^{[t]}$ is defined as one of them:
$$\widehat{f}^{[t]} = \begin{cases} \widehat{f}^{[t]}_{\mathrm{CL}} & \text{when } R_{\mathrm{I}}(\widehat{f}^{[t]}_{\mathrm{CL}}) \leq R_{\mathrm{I}}(\widehat{f}^{[t]}_{\mathrm{TL}}), \\ \widehat{f}^{[t]}_{\mathrm{TL}} & \text{when } R_{\mathrm{I}}(\widehat{f}^{[t]}_{\mathrm{CL}}) > R_{\mathrm{I}}(\widehat{f}^{[t]}_{\mathrm{TL}}), \end{cases}$$

where the following quantities are further defined:
$$R_{\mathrm{I}}(\widehat{f}^{[t]}_{\mathrm{CL}}) := L_m^{2/(2\alpha_m+1)} \left(\frac{\log^2 n_t}{m_t n_t} \right)^{2\alpha_m/(2\alpha_m+1)} + \frac{\log^2 n_t}{n_t},$$
$$R_{\mathrm{I}}(\widehat{f}^{[t]}_{\mathrm{TL}}) := L_m^{2/(2\alpha_m+1)} \left(\frac{\log^2(K n_s)}{K m_s n_s} \right)^{2\alpha_m/(2\alpha_m+1)} + \frac{\log^2(K n_s)}{K n_s} + L_\delta^{2/(2\alpha_\delta+1)} \left(\frac{\log^2(n_t n_s)}{m_t n_t} \right)^{2\alpha_\delta/(2\alpha_\delta+1)} + \frac{\log^2(n_t n_s)}{n_t}.$$

In this situation, we have
$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E} \| \widehat{f}^{[t]} - f^{[t]} \|_{\mathcal{L}^2}^2 \lesssim R_{\mathrm{I}}(\widehat{f}^{[t]}_{\mathrm{CL}}) \wedge R_{\mathrm{I}}(\widehat{f}^{[t]}_{\mathrm{TL}}).$$

It is worth noting in Theorem 3.2 that the transfer learning algorithm $\mathcal{A}_{\mathrm{TL}}$ achieves the rate $\widetilde{\Theta}((m_t n_t)^{-2\alpha_\delta/(2\alpha_\delta+1)} + n_t^{-1} + (Km_s n_s)^{-2\alpha_m/(2\alpha_m+1)} + (Kn_s)^{-1})$, provided that the input parameters are selected appropriately. However, it is not always superior to the conventional learning algorithm, which has an optimal rate of $\widetilde{\Theta}((m_t n_t)^{-2\alpha_m/(2\alpha_m+1)} + n_t^{-1})$. Therefore, the final estimator $\widehat{f}^{[t]}$ in Theorem 3.2 is selected as any algorithm that achieves a better rate of convergence, which optimizes the utilization of both learning algorithms, $\mathcal{A}_{\mathrm{TL}}$ and $\mathcal{A}_{\mathrm{CL}}$.

Although the final estimator $\widehat{f}^{[t]}$ suggested in Theorem 3.2 will be sufficient to argue the minimax rate of convergence, it confronts two challenges in real-world situations. Since the smoothness parameters, α_m and α_δ, are typically unknown, it is impossible to obtain the optimal bandwidth for both algorithms in practice. Additionally, comparing the rates of the two algorithms also requires knowledge of those unknown parameters. Adaptation to smoothness parameters was relatively simple under a common design since the optimal parameters for both learning algorithms are at least independent of them. As a consequence, it must be more challenging to decide whether or not to employ the source samples and transfer learning under an independent design.

3.2. *Matching lower bound.* The upcoming theorem shows the lower bound for estimating the target mean function under an independent design.

THEOREM 3.3 (Lower bound under an independent design). *Under an independent design,*

$$\begin{aligned}\inf_{\widehat{f}^{[t]}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\big\|\widehat{f}^{[t]} - f^{[t]}\big\|_{\mathcal{L}^2}^2 \gtrsim &\left[L_m^{2/(2\alpha_m+1)}(m_t n_t)^{-2\alpha_m/(2\alpha_m+1)} + \frac{1}{n_t}\right] \\ &\wedge \left[L_\delta^{2/(2\alpha_\delta+1)}(m_t n_t)^{-2\alpha_\delta/(2\alpha_\delta+1)} + \frac{1}{n_t} \right. \\ &\left. + L_m^{2/(2\alpha_m+1)}(Km_s n_s)^{-2\alpha_m/(2\alpha_m+1)} + \frac{1}{Kn_s}\right],\end{aligned}$$

where the infimum is taken over all possible estimators $\widehat{f}^{[t]} = \widehat{f}^{[t]}(\mathcal{D}^{[t]}, \mathcal{D}^{[s,1]}, \ldots, \mathcal{D}^{[s,\ell]})$.

Theorem 3.3 provides the minimax lower bound, which exactly matches the upper bound given in Theorem 3.2, up to logarithmic factors. This leads to the optimal rate for estimating the target mean function $f^{[t]}$ under an independent design, which is given by

$$\left[(m_t n_t)^{-2\alpha_m/(2\alpha_m+1)} + \frac{1}{n_t}\right] \wedge \left[(m_t n_t)^{-2\alpha_\delta/(2\alpha_\delta+1)} + \frac{1}{n_t} + (Km_s n_s)^{-2\alpha_m/(2\alpha_m+1)} + \frac{1}{Kn_s}\right].$$

This rate comprises two primary components:

- $\widetilde{\Theta}((m_t n_t)^{-2\alpha_m/(2\alpha_m+1)} + n_t^{-1})$, arising from the conventional learning setup.
- $\widetilde{\Theta}((m_t n_t)^{-2\alpha_\delta/(2\alpha_\delta+1)} + n_t^{-1} + (Km_s n_s)^{-2\alpha_m/(2\alpha_m+1)} + (Kn_s)^{-1})$, originating from the transfer learning setup.

It makes sense to evaluate the problem's overall difficulty by taking the minimum of these two rates. The latter rate is further decomposed into two parts: $\widetilde{\Theta}((m_t n_t)^{-2\alpha_\delta/(2\alpha_\delta+1)} + n_t^{-1})$ and $\widetilde{\Theta}((Km_s n_s)^{-2\alpha_m/(2\alpha_m+1)} + (Kn_s)^{-1})$. These rates correspond to the minimax risks of estimation on the target and source samples, respectively. In parallel to the discussion under a common design, transfer learning treats the target mean function as α_δ-smooth, instead of α_m-smooth, while paying the cost of estimating the source mean functions. Whenever this cost is reasonable and the substitution of smoothness results in benefits, it leads to more accurate estimation and a faster convergence rate. This high-level structure of the minimax

risk is similar to that of a common design and reveals that the model's formulation does not depend on the sampling designs.

By analyzing the minimax risks within the contexts of both the conventional and transfer learning settings, we can determine the circumstances under which utilizing source samples or transfer learning can enhance the performance of estimating the target mean function. It is straightforward to compare the two settings as the minimax rate for the conventional learning setup comprises half of that for the transfer learning setup. Similar to a common design case, we can identify a phase transition that has a substantial impact on the effectiveness of transfer learning. Remarkably, the boundary of phase transition is the same as that in a common design setting. This also implies that the same fundamental principles govern the functional mean estimation regardless of the sampling designs.

First and foremost, it suffices to focus our analysis of effective transfer learning exclusively on models that adhere to the condition $\alpha_\delta > \alpha_m$. The rationale behind this criterion aligns with that of the common design setting. Within such models, we may fully exploit the benefits of substituting smoothness; the increase in smoothness potentially leads to a faster rate of convergence than the one associated with the conventional learning algorithm. For any model with $\alpha_\delta \le \alpha_m$, transfer learning indeed yields adverse effects because the estimation of α_δ-smooth functions is as challenging as estimating α_m-smooth functions. Under the absence of clear advantage in smoothness, there is no justification for engaging in estimation on the source samples and incurring the cost of $\widetilde{\Theta}((Km_sn_s)^{-2\alpha_m/(2\alpha_m+1)} + (Kn_s)^{-1})$ in the convergence rate. This introduces the intriguing concept of a phase transition where the effectiveness of transfer learning undergoes a sharp change. Further discussions concerning the effectiveness become meaningful for limited models satisfying $\alpha_\delta > \alpha_m$.

We exclusively focus on models that meet the condition $\alpha_\delta > \alpha_m$. In the high-frequency regime where $m_t \gtrsim n_t^{1/2\alpha_m}$, transfer learning cannot be effective, similar to the case of a common design. The convergence rate for the conventional learning setup, $\widetilde{\Theta}(n_t^{-1})$, is always less than or equal to $\widetilde{\Theta}((m_tn_t)^{-2\alpha_\delta/(2\alpha_\delta+1)} + n_t^{-1} + (Km_sn_s)^{-2\alpha_m/(2\alpha_m+1)} + (Kn_s)^{-1})$, the additional rate introduced by the transfer learning setup. Unless difference functions $\delta^{[s,k]}$ $(k = 1, \ldots, \ell)$ are exactly known, the parametric rate $\widetilde{\Theta}(n_t^{-1})$ always gives the best possible rate and there is no further way for enhancement.

However, in the low-frequency regime where $m_t \ll n_t^{1/2\alpha_m}$, transfer learning has the potential to enhance the estimation performance like the case of a common design. To be more specific, the following condition is necessary and sufficient for the source samples and transfer learning to be beneficial:

$$(m_sn_s)^{-2\alpha_m/(2\alpha_m+1)} + (Kn_s)^{-1} \ll (m_tn_t)^{-2\alpha_m/(2\alpha_m+1)}.$$

This condition is quite different from the corresponding one under a common design, which is $m_s^{-2\alpha_m} + n_s^{-1} \ll m_t^{-2\alpha_m}$, and neither condition implies the other. Comparing these two conditions for effective transfer learning, both require a sufficient number of source subjects, but their thresholds are different. For an independent design, the requirement is related to the target observation size ($Kn_s \gg (m_tn_t)^{2\alpha_m/(2\alpha_m+1)}$), while for a common design, it is to the target sampling frequency ($Kn_s \gg m_t^{2\alpha_m}$). Additionally, an independent design requires the source samples to have more total observations ($Km_sn_s \gg m_tn_t$), whereas a common design requires the source samples to have a higher sampling frequency ($m_s \gg m_t$).

The impact of the source group's size K varies between a common design and an independent design. As previously discussed, in a common design, increasing K may or may not result in a more effective transfer learning in the low-frequency regime. However, under an independent design, augmenting K consistently leads to enhanced transfer learning in the same regime. Given that the target sample remains unchanged, it should be an effective approach for both conditions, $Kn_s \gg (m_tn_t)^{2\alpha_m/(2\alpha_m+1)}$ and $Km_sn_s \gg m_tn_t$ to be satisfied.

Although the model structures are similar across sampling designs, the minimax risks, the particular conditions for effective transfer learning and the impact of the source group's size differ significantly. This suggests that the estimation behavior is entirely distinct between the two designs.

It is also interesting to directly compare the minimax risks between common and independent designs. Assuming that all other specifications are the same except for the sampling design, the independent design's rate is always equal to or lower than the common design's, up to some constants. Hence, if given the option to choose between the two designs, and all other factors remain the same, then an independent design should be preferred in terms of the convergence rate:

$$\left[(m_t n_t)^{-2\alpha_m/(2\alpha_m+1)} + \frac{1}{n_t}\right] \wedge \left[(m_t n_t)^{-2\alpha_\delta/(2\alpha_\delta+1)} + \frac{1}{n_t} + (K m_s n_s)^{-2\alpha_m/(2\alpha_m+1)} + \frac{1}{K n_s}\right]$$
$$\lesssim \left[m_t^{-2\alpha_m} + \frac{1}{n_t}\right] \wedge \left[m_t^{-2\alpha_\delta} + \frac{1}{n_t} + m_s^{-2\alpha_m} + \frac{1}{K n_s}\right].$$

3.3. *Adaptive estimation*. We introduce an adaptive algorithm for an independent design, named $\mathcal{A}_{\mathrm{ALI}}$, which addresses two main challenges: selecting an optimal bandwidth and choosing a learning algorithm between $\mathcal{A}_{\mathrm{CL}}$ and $\mathcal{A}_{\mathrm{TL}}$. The algorithm generates multiple candidate bandwidths for each algorithm and compares their performances using empirical risk and cross-validation. This approach addresses both challenges simultaneously. We will provide a detailed description of $\mathcal{A}_{\mathrm{TL}}$ in Algorithm 4. It is further assumed for brevity that the target sample $\mathcal{D}^{[t]}$ contains $2n_t$ subjects, which does not affect the rate of convergence.

It should be noted that the lists of candidate bandwidths proposed in Algorithm 4 always include the optimal bandwidths, presented in Theorem 3.2, for the conventional and transfer learning algorithms. Therefore, any chosen bandwidth through cross-validation is expected to perform better than the optimal one. The same reasoning applies to the selection of an optimal learning algorithm. Any estimator selected via cross-validation will outperform both learning algorithms $\mathcal{A}_{\mathrm{TL}}$ and $\mathcal{A}_{\mathrm{CL}}$ with their respective optimal parameters. Finally, the computational complexity of the adaptation step is quite minimal, as the number of candidate bandwidths is expected to be at most $\log_2(m_t n_t) + \log_2(K m_s n_s)\log_2(m_t n_t)$.

Similar to the adaptive learning algorithm $\mathcal{A}_{\mathrm{ALC}}$ for a common design, the adaptive algorithm $\mathcal{A}_{\mathrm{ALI}}$ for an independent design also involves multiple steps of random downsampling, resulting in a randomized estimator. To improve the performance of estimation in finite samples, we will execute this adaptive process $r_{\max}$ times and average the resulting estimates. This approach is analogous to bagging estimators proposed by Breiman [2]. As demonstrated by the subsequent theorem, this algorithm achieves the optimal minimax rate of convergence, up to logarithmic factors.

THEOREM 3.4 (Adaptive estimation under an independent design). *Suppose the assumptions in Theorem 3.2 hold. For any given* $r_{\max} \in \mathbb{Z}^+$, *let* $\widehat{g}_r^{[t]}$ $(r = 1, \ldots, r_{\max})$ *be the output of the* r*th execution of the algorithm* $\mathcal{A}_{\mathrm{ALI}}$. *Take an average of them to obtain our final estimator*:

$$\widehat{f}^{[t]} = \frac{1}{r_{\max}} \sum_{r=1}^{r_{\max}} \widehat{g}_r^{[t]}. \tag{6}$$

This adaptive estimator $\widehat{f}^{[t]}$ *attains the same upper bound of Theorem 3.2*:

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\|\widehat{f}^{[t]} - f^{[t]}\|_{\mathcal{L}^2}^2 \lesssim R_{\mathrm{I}}\big(\widehat{f}_{\mathrm{CL}}^{[t]}\big) \wedge R_{\mathrm{I}}\big(\widehat{f}_{\mathrm{TL}}^{[t]}\big).$$

Therefore, the data-driven estimator (6) attains the optimal rate of convergence adaptively over a wide collection of function classes.

Algorithm 4 Adaptive transfer learning for mean function under an independent design $\mathcal{A}_{\mathrm{ALI}}$

Initialize $\widehat{\mathcal{G}}^{[t]} = \varnothing$, the collection of candidate estimators.
Randomly partition the target sample $\mathcal{D}^{[t]}$ into two subsamples, denoted as $\mathcal{D}^{[t]}_{\mathrm{train}}$ and $\mathcal{D}^{[t]}_{\mathrm{test}}$, based on subjects. Specifically, perform a random split of the index set $\{1, \ldots, 2n_t\}$ into two partitions, $\mathcal{I}^{[t]}_{\mathrm{train}}$ and $\mathcal{I}^{[t]}_{\mathrm{test}}$ such that $|\mathcal{I}^{[t]}_{\mathrm{train}}| = |\mathcal{I}^{[t]}_{\mathrm{test}}| = n_t$. We define

$$\mathcal{D}^{[t]}_{\mathrm{train}} := \{(T^{[t]}_{ij}, Y^{[t]}_{ij}) : i \in \mathcal{I}^{[t]}_{\mathrm{train}},\ j = 1, \ldots, m_t\},$$

$$\mathcal{D}^{[t]}_{\mathrm{test}} := \{(T^{[t]}_{ij}, Y^{[t]}_{ij}) : i \in \mathcal{I}^{[t]}_{\mathrm{test}},\ j = 1, \ldots, m_t\}.$$

Pick any constants $B_t \geq C_t$ and $d_t \geq \omega(\alpha_m)$.
Take a threshold $M_t = \log n_t$.
for $b_t \in \{2^r \leq m_t n_t : r \in \mathbb{Z}^+\}$ **do**
    Execute $\mathcal{A}_{\mathrm{CL}}(b_t, d_t, M_t)$ as if target sample is given as $\mathcal{D}^{[t]}_{\mathrm{train}}$, not $\mathcal{D}^{[t]}$.
    Add the algorithm's output to the collection $\widehat{\mathcal{G}}^{[t]}$.
end for
Pick any constants $B_s \geq C_s$, $B_\delta \geq C_t$, $d_s \geq \omega(\alpha_m)$ and $d_\delta \geq \omega(\alpha_\delta)$.
Take thresholds $M_s = \log n_s$ and $M_s = \log n_t n_s$.
for $(b_s, b_\delta) \in \{2^r \leq K m_s n_s : r \in \mathbb{Z}^+\} \times \{2^r \leq m_t n_t : r \in \mathbb{Z}^+\}$ **do**
    Execute $\mathcal{A}_{\mathrm{TL}}(b_s, b_\delta, d_s, d_\delta, M_s, M_\delta)$ as if target sample is given as $\mathcal{D}^{[t]}_{\mathrm{train}}$, not $\mathcal{D}^{[t]}$.
    Add the algorithm's output to the collection $\widehat{\mathcal{G}}^{[t]}$.
end for
Output the following estimator $\widehat{g}^{[t]}_*$. If a tie occurs, use any randomization to break it:

$$\widehat{g}^{[t]}_* := \operatorname*{argmin}_{\widehat{g}^{[t]} \in \widehat{\mathcal{G}}^{[t]}} \sum_{(T,Y) \in \mathcal{D}^{[t]}_{\mathrm{test}}} \big(Y - \widehat{g}^{[t]}(T)\big)^2.$$

4. Numerical experiments. The proposed adaptive transfer learning algorithms, $\mathcal{A}_{\mathrm{ALC}}$ and $\mathcal{A}_{\mathrm{ALI}}$, are computationally efficient and easy to implement. In this section, we investigate their numerical performance and practical implications in a simulation study.

In our simulation, the target mean function is given by

$$f^{[t]}(x) = x\cos(25x) + 4|x - 0.5|, \quad (0 \leq x \leq 1).$$

We also have two source samples and their mean functions are expressed in terms of difference functions:

$$f^{[s,k]} = f^{[t]} - \delta^{[s,k]} \quad (k = 1, 2), \quad \text{where } \begin{cases} \delta^{[s,1]}(x) = -x^2, \\ \delta^{[s,2]}(x) = e^x - 1, \end{cases} \quad (0 \leq x \leq 1).$$

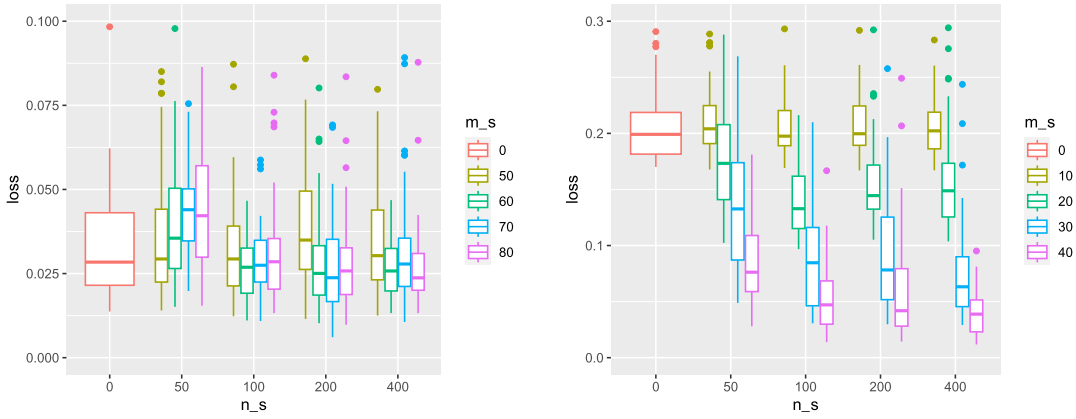
It is worth noticing that the differences $\delta^{[s,1]}$ and $\delta^{[s,2]}$ are smoother than the target $f^{[t]}$. This means that the differences are easier to estimate than the target, and transfer learning from the source samples can improve the estimation performance.

The target and source subjects in each sample have random curves generated as

$$X^{[t]}(x) = f^{[t]}(x) + B^{[t]}_x,$$

$$X^{[s,k]}(x) = f^{[s,1]}(x) + B^{[s,k]}_x \quad (k = 1, 2),$$

where $\{B^{[t]}_x : x \geq 0\}$ and $\{B^{[s,k]}_x : x \geq 0\}$ $(k = 1, 2)$ denote independent standard Brownian motions. It is well known that Brownian motions have wriggly sample paths with smoothness strictly less than $1/2$ almost surely. Traditional frameworks such as [10] do not allow for



(a) High-frequency regime ($n_t = m_t = 50$)

(b) Low-frequency regime ($n_t = 50,\ m_t = 10$)

FIG. 1. *The results for numerical experiments under a common design.*

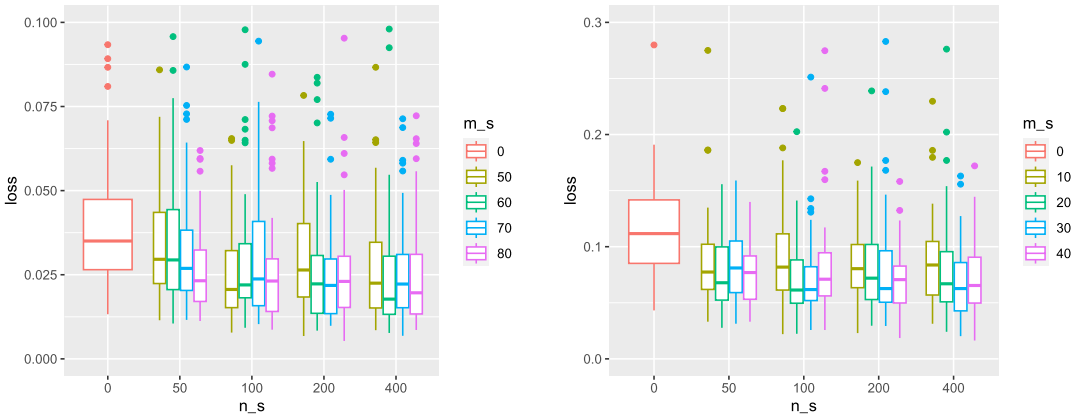
this type of model because the curves are not differentiable. However, the framework in this paper imposes no restrictions on the smoothness of the curves, making it more flexible and general even under conventional learning settings. Finally, each observation at a design point is assumed to be corrupted by additive standard Gaussian noise.

Let us first implement numerical experiments for a common design. The target and source samples consist of n_t and n_s subjects, respectively, and their curves are generated using the previously specified scheme. For each curve, equidistant observations at m_t and m_s design points, respectively, are taken on the domain $[0, 1]$ with standard Gaussian noise. Following the estimation process in Theorem 2.2, the adaptive learning algorithm $\mathcal{A}_{\mathrm{ALC}}$ for a common design is repeated $r_{\max} = 20$ times and the final estimate $\widehat{f}^{[t]}$ is obtained by taking an average of the results. We repeat this simulation 50 times, creating a boxplot of the $\mathcal{L}^2$-integrated mean a squared loss to evaluate the risk for various combinations of (n_t, m_t, n_s, m_s).

Our theory indicates that a phase transition occurs at $m_t \asymp n_t^{1/2\alpha_m}$, which determines the effectiveness of the source samples and transfer learning. To illustrate this phenomenon, we will compare two regimes: the high-frequency regime ($n_t = m_t = 50$) and the low-frequency regime ($n_t = 50,\ m_t = 10$). For the former regime, we shall consider all possible combinations between $n_s \in \{50, 100, 200, 400\}$ and $m_s \in \{50, 60, 70, 80\}$, whereas for the latter regime, we consider $m_s \in \{10, 20, 30, 40\}$ instead. Plus, we will consider the conventional learning setup ($n_s = m_s = 0$), which serves as a baseline for any other setup. Figure 1 shows the aggregated boxplots for all combinations in each regime.

The results from our theory are consistent with Figure 1. Transfer learning does not have a significant impact on the estimation performance in the high-frequency regime, while it is effective in the low-frequency regime. It is noteworthy that when $m_s = 10$ in the low-frequency regime, the performance is similar to the conventional learning ($n_s = m_s = 0$) regardless of the size n_s of the source subject. To enjoy benefits from the transfer learning, the number of design points for the source curves must exceed that of the target curves, that is, $m_t < m_s$.

We next concentrate on numerical experiments under an independent design. The target and source samples are generated using the same method as a common design, except that design points are independently and uniformly drawn from the domain $[0, 1]$. We use the adaptive learning algorithm $\mathcal{A}_{\mathrm{ALI}}$ instead of $\mathcal{A}_{\mathrm{ALC}}$ for estimation while keeping all other simulation details the same. The resulting boxplots, similar to those in Figure 1, are presented in Figure 2.



(a) High-frequency regime ($n_t = m_t = 50$)

(b) Low-frequency regime ($n_t = 50,\ m_t = 10$)

FIG. 2. *The results for numerical experiments under an independent design.*

By comparing the high- and low-frequency regimes depicted in Figure 2, we have come to the same conclusion as we did under a common design: transfer learning is only advantageous under the low-frequency regime. Additionally, we can draw interesting insights from comparing common and independent designs. For instance, under the high-frequency regime, both designs exhibit similar performance, which aligns with our theory as we always achieve the fastest parametric rate n_t^{-1} in this regime. Conversely, under the low-frequency regime, an independent design outperforms a common design when all other factors remain the same. To summarize, the results obtained from the numerical experiments are both explainable and consistent with our theory.

5. Discussion. The present paper thoroughly investigated the transfer learning problem for functional mean estimation. It offers a comprehensive analysis of minimax convergence rates, which not only extends the previous results in conventional learning but also quantifies the advantages of transfer learning in terms of estimation performance. Our findings unveil a compelling phase transition phenomenon under both common and independent designs. Moreover, we developed data-driven algorithms that achieve optimal rates of convergence up to logarithmic factors across a broad spectrum of function spaces.

This paper serves as a critical stepping stone for further theoretical explorations in functional data analysis within the context of transfer learning. One promising avenue for future research is the investigation of transfer learning for the estimation of the covariance function, which plays a pivotal role in the analysis of functional and longitudinal data [9]. In particular, the study of estimating the covariance function,

$$\Sigma^{[t]}(u, v) := \mathbb{E}\big[\big(X^{[t]}(u) - \mathbb{E}X^{[t]}(u)\big)\big(X^{[t]}(v) - \mathbb{E}X^{[t]}(v)\big)\big], \quad (u, v \in [0, 1]),$$

should be of substantial interest, given the availability of both target and source samples, as presented in equations (1) and (2). Establishing the minimax rates of convergence for this problem promises to yield valuable insights.

Another important problem is transfer learning for functional linear regression. In the conventional settings, Yuan and Cai [41] established the minimax rate of convergence for estimating the regression function, while both Cai and Hall [5] and Cai and Yuan [11] investigated the optimal prediction problem. It is of significant interest to explore transfer learning for linear regression and investigate the conditions under which source samples can improve the estimation performance.

While our present work offers valuable insights, it is essential to acknowledge its limitations and explore avenues for further improvement. One notable constraint is our assumption that both the target and source samples share the same sampling design. However, in practice, the designs can differ from each other. For example, the target sample can be generated from a common design while the source samples are obtained through an independent design. Not only that, some source samples are drawn from a common design, while others come from an independent design. Even within source samples sharing the same design, there could be variations in the common design points or the common marginal distributions that generate design points.

An analogous assumption is imposed on the smoothness of mean functions as shown in equation (3), but this could be generalized to cover more diverse data sets. For example, we could assume some source mean functions are smoother while others are rougher than the target: for given $K_1, K_2 \in \mathbb{Z}^+$ with $K_1 + K_2 = K$,

$$\begin{aligned} &f^{[s,1]}, \ldots, f^{[s,K_1]} \in \mathcal{H}_{\alpha_{m_1}}(L_{m_1}, M_{m_1}), \\ &f^{[s,K_1+1]}, \ldots, f^{[s,K]} \in \mathcal{H}_{\alpha_{m_2}}(L_{m_2}, M_{m_2}), \end{aligned} \qquad \text{for some } \alpha_{m_1}, \alpha_{m_2} > 0.$$

Another aspect to consider is the assumption that all source samples share the same number of subjects and design points, as presented in equation (4). However, this assumption can be relaxed to accommodate different numbers of subjects or design points across the source samples. While this relaxation may lead to a more complex minimax rate of convergence, it would better align with the diversity observed in real-world data sets.

Finally, in the independent design setting, we assumed that the marginal distribution of the source samples only trivially differs from that of the target sample. In practice, the marginal density ratio between the target and source samples may not be uniformly bounded away from zero and infinity. This departure from the uniformity of density ratio can significantly impact the minimax rate of convergence. For example, if the marginal distributions of the target and source samples are completely singular, the benefit of transfer learning may be negligible. Incorporating this effect into the analysis could be achieved by introducing a new measure to quantify the singularity between marginal distributions. Similar to the transfer exponent introduced in Kpotufe and Martinet [20], such a measure would be a valuable addition to assess the influence of marginal singularity under the transfer learning setup.

6. Proofs. This section contains the proofs of our main results for common design, especially under transfer learning setup: Theorem 2.2 and Theorem 2.3. We will defer the proofs of the remaining results for common design, as well as the results for independent design and every other technical results, to the Supplementary Material [6].

6.1. *Proof of Theorem 2.2.* Thanks to Theorem 2.1 and the nature of our estimator $\widehat{f}^{[t]}$, it suffices to argue the following rate of the transfer learning estimator $\widehat{f}^{[t]}_{\mathrm{TL}}$:

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\big\|\widehat{f}^{[t]}_{\mathrm{TL}} - f^{[t]}\big\|^2_{\mathcal{L}^2} \lesssim L_\delta^2 m_t^{-2\alpha_\delta} + \frac{\log^2 n_t n_s}{n_t} + L_m^2 m_s^{-2\alpha_m} + \frac{\log^2(Kn_s)}{Kn_s}.$$

Using the notation in the transfer learning algorithm $\mathcal{A}_{\mathrm{TL}}$, we may easily check

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\big\|\widehat{f}^{[t]}_{\mathrm{TL}} - f^{[t]}\big\|^2_{\mathcal{L}^2} \le 2 \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\big\|\widehat{f}^{[s]} - f^{[s]}\big\|^2_{\mathcal{L}^2} + 2 \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\big\|\widehat{\delta}^{[s]} - \delta^{[s]}\big\|^2_{\mathcal{L}^2}$$

We shall present the rates of convergence for those estimators, $\widehat{f}^{[s]}$ and $\widehat{\delta}^{[s]}$, in the following propositions. The proofs are given in the Supplementary Material [6].

PROPOSITION 6.1. *If $m_s b_s \geq 2C_s(d_s+1)$ and $\omega(\alpha_m) \leq d_s \leq O(1)$ are satisfied,*

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\big\|\widehat{f}^{[s]} - f^{[s]}\big\|_{\mathcal{L}^2}^2 \lesssim L_m^2 b_s^{2\alpha_m} + \frac{\log^2(Kn_s)}{Kn_s}.$$

PROPOSITION 6.2. *If $m_t b_\delta \geq 2C_t(d_\delta+1)$ and $\omega(\alpha_\delta) \leq d_\delta \leq O(1)$ are fulfilled,*

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\big\|\widehat{\delta}^{[s]} - \delta^{[s]}\big\|_{\mathcal{L}^2}^2 \lesssim L_\delta^2 b_\delta^{2\alpha_\delta} + \frac{\log^2 n_t n_s}{n_t}.$$

It is therefore optimal to choose the bandwidths $b_s = \Theta(m_s^{-1})$ and $b_\delta = \Theta(m_t^{-1})$ such that they satisfy $m_s b_s \geq 2C_s(d_s+1)$ and $m_t b_\delta \geq 2C_t(d_\delta+1)$, and to select any constant degrees $d_s \geq \omega(\alpha_m)$ and $d_\delta \geq \omega(\alpha_\delta)$. This immediately concludes the proof. □

6.2. *Proof of Theorem 2.3.* It suffices to show that for any estimator $\widehat{f}^{[t]}$,

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\big\|\widehat{f}^{[t]} - f^{[t]}\big\|_{\mathcal{L}^2}^2 \gtrsim \big(L_\delta^2 m_t^{-2\alpha_\delta} \wedge L_m^2 m_t^{-2\alpha_m}\big) + \frac{1}{n_t} + \left(L_m^2 m_t^{-2\alpha_m} + \frac{1}{n_t}\right) \wedge \left(L_m^2 m_s^{-2\alpha_m} + \frac{1}{Kn_s}\right).$$

Consider a submodel $\mathcal{P}_1 \subset \mathcal{P}$ where the source samples $\mathcal{D}^{[s,1]}, \ldots, \mathcal{D}^{[s,K]}$ are completely futile in the sense that $f^{[s,1]} = \cdots = f^{[s,K]} = 0$. Under this submodel $\mathcal{P}_1$, target sample $\mathcal{D}^{[t]}$ forms a sufficient statistic of the target mean function $f^{[t]}$. On top of that, the target mean function should be further restricted by $f^{[t]} \in \mathcal{H}_{\alpha_m}(L_m, M_m) \cap \mathcal{H}_{\alpha_\delta}(L_\delta, M_\delta)$. As a result, Theorem 2.1 immediately implies that

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\big\|\widehat{f}^{[t]} - f^{[t]}\big\|_{\mathcal{L}^2}^2 \geq \sup_{\mathbb{P}\in\mathcal{P}_1} \mathbb{E}\big\|\mathbb{E}\big(\widehat{f}^{[t]} \,|\, \mathcal{D}^{[t]}\big) - f^{[t]}\big\|_{\mathcal{L}^2}^2 \gtrsim \big(L_\delta^2 m_t^{-2\alpha_\delta} \wedge L_m^2 m_t^{-2\alpha_m}\big) + \frac{1}{n_t}.$$

We are now enough to argue the other part of the lower bound:

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\big\|\widehat{f}^{[t]} - f^{[t]}\big\|_{\mathcal{L}^2}^2 \gtrsim \left(L_m^2 m_t^{-2\alpha_m} + \frac{1}{n_t}\right) \wedge \left(L_m^2 m_s^{-2\alpha_m} + \frac{1}{Kn_s}\right).$$

This is equivalent to insisting on the subsequent three propositions:

PROPOSITION 6.3. *For any estimator $\widehat{f}^{[t]}$, we have*

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\big\|\widehat{f}^{[t]} - f^{[t]}\big\|_{\mathcal{L}^2}^2 \gtrsim \frac{1}{n_t} \wedge \frac{1}{Kn_s}.$$

PROPOSITION 6.4. *For any estimator $\widehat{f}^{[t]}$, we have*

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\big\|\widehat{f}^{[t]} - f^{[t]}\big\|_{\mathcal{L}^2}^2 \gtrsim L_m^2\big(m_t^{-2\alpha_m} \wedge m_s^{-2\alpha_m}\big).$$

PROPOSITION 6.5. *For any estimator $\widehat{f}^{[t]}$, we have*

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\big\|\widehat{f}^{[t]} - f^{[t]}\big\|_{\mathcal{L}^2}^2 \gtrsim L_m^2 m_t^{-2\alpha_m} \wedge \frac{1}{Kn_s}.$$

PROPOSITION 6.6. *For any estimator $\widehat{f}^{[t]}$, we have*

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\big\|\widehat{f}^{[t]} - f^{[t]}\big\|_{\mathcal{L}^2}^2 \gtrsim L_m^2 m_s^{-2\alpha_m} \wedge \frac{1}{n_t}.$$

We construct another submodel $\mathcal{P}_2 \subset \mathcal{P}$ as follows. The source samples, $\mathcal{D}^{[s,1]}, \ldots, \mathcal{D}^{[s,K]}$, are the most advantageous in the sense that $f^{[t]} = f^{[s,1]} = \cdots = f^{[s,K]}$. In other words, every sample $\mathcal{D}^{[t]}, \mathcal{D}^{[s,1]}, \ldots, \mathcal{D}^{[s,K]}$ is assumed to share the same data-generating process whose further specifications would be different depending on each case. Let us take any C^∞ function $\phi : \mathbb{R} \to \mathbb{R}$ such that $\phi^{(q)}(x) = 0$ for any $q = 0, 1, 2, \ldots$ and $x \notin (0, 1)$. We will prove Propositions 6.3 and 6.4. The proof of Propositions 6.5 and 6.6 will be presented separately in the Supplementary Material [6].

PROOF OF PROPOSITION 6.3. Under the submodel $\mathcal{P}_2 \subset \mathcal{P}$, let us assume the target random function $X_1^{[t]}$ as well as the source ones $X_1^{[s,1]}, \ldots, X_1^{[s,K]}$ are unknown constant functions. Since their mean functions should coincide, the problem is equivalent to estimating the scalar mean from $n_t + Kn_s$ number of independent and identically distributed observations under mean squared error. It is thus immediate to get the lower bound of the parametric rate:

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\big\|\widehat{f}^{[t]} - f^{[t]}\big\|_{\mathcal{L}^2}^2 \geq \sup_{\mathbb{P}\in\mathcal{P}_2} \mathbb{E}\big\|\widehat{f}^{[t]} - f^{[t]}\big\|_{\mathcal{L}^2}^2 \gtrsim \frac{1}{n_t + Kn_s} \gtrsim \frac{1}{n_t} \wedge \frac{1}{Kn_s}. \qquad \square$$

PROOF OF PROPOSITION 6.4. We begin by letting a quantity $m := m_t + m_s$. For each vector $v = (v_1, \ldots, v_{2m}) \in \{0, 1\}^{2m}$ on hypercube, we define a function $g_v : [0, 1] \to \mathbb{R}$ by

$$g_v(x) := \sum_{r=1}^{2m} C L_m v_r (2m)^{-\alpha_m} \phi\big(2mx - (r-1)\big),$$

where $C > 0$ is a small enough generic constant such that $g_v \in \mathcal{H}_{\alpha_m}(L_m, M_m)$. The submodel $\mathcal{P}_2 \subset \mathcal{P}$ consists of any probability measure $\mathbb{P}_v$ ($v \in \{0, 1\}^{2m}$) whose common mean function $f^{[t]} = f^{[s,1]}$ is given as g_v. This submodel $\mathcal{P}_2$ is Hamming separated in that

$$\|g_v - g_w\|_{\mathcal{L}^2}^2 \geq C^2 L_m^2 (2m)^{-2\alpha_m - 1} \|\phi\|_{\mathcal{L}^2}^2 \sum_{r=1}^{2m} \mathbb{1}(v_r \neq w_r), \quad \text{for any } v, w \in \{0, 1\}^{2m}.$$

For each $r = 1, \ldots, 2m$, let $\overline{\mathbb{P}}_{0,r}$ and $\overline{\mathbb{P}}_{1,r}$ denote the mixture distributions of $\{\mathbb{P}_v : v_r = 0\}$ and $\{\mathbb{P}_v : v_r = 1\}$, respectively. It follows from Assouad's lemma [1] that

$$\begin{aligned} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\big\|\widehat{f}^{[t]} - f^{[t]}\big\|_{\mathcal{L}^2}^2 &\geq \sup_{\mathbb{P}\in\mathcal{P}_2} \mathbb{E}\big\|\widehat{f}^{[t]} - f^{[t]}\big\|_{\mathcal{L}^2}^2 \\ &\geq \frac{1}{2} C^2 L_m^2 (2m)^{-2\alpha_m - 1} \|\phi\|_{\mathcal{L}^2}^2 \sum_{r=1}^{2m} \big(1 - \|\overline{\mathbb{P}}_{0,r} - \overline{\mathbb{P}}_{1,r}\|_{\mathrm{TV}}\big), \end{aligned} \tag{7}$$

where $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation distance. Notice that two mixtures, $\overline{\mathbb{P}}_{0,r}$ and $\overline{\mathbb{P}}_{1,r}$, follow exactly the same distribution when the interval $[(r-1)/2m, r/2m)$ contains nothing among $\{T_j^{[t]} : j = 1, \ldots, m_t\} \cup \{T_j^{[s,1]} : j = 1, \ldots, m_s\}$, the common design points for target and source domains. Since these design points are at most $m = m_t + m_s$, the number of such an $r \in \{1, \ldots, 2m\}$ must be at least m by Pigeonhole principle. In other words, we have

$$\sum_{r=1}^{2m} \big(1 - \|\overline{\mathbb{P}}_{0,r} - \overline{\mathbb{P}}_{1,r}\|_{\mathrm{TV}}\big) \geq m. \tag{8}$$

The desired result is now immediate by combining equations (7) and (8):

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\big\|\widehat{f}^{[t]} - f^{[t]}\big\|_{\mathcal{L}^2}^2 \gtrsim L_m^2 m^{-2\alpha_m} \asymp L_m^2 \big(m_t^{-2\alpha_m} \wedge m_s^{-2\alpha_m}\big). \qquad \square$$

Acknowledgments. The authors extend their gratitude to the Editor, the Associate Editor and the anonymous referees for their constructive and insightful comments and suggestions, which have significantly enhanced the quality of this paper.

Funding. The research of Tony Cai was supported in part by NSF Grant DMS-2015259 and NIH Grant R01-GM129781.

SUPPLEMENTARY MATERIAL

Supplement to "Transfer learning for functional mean estimation: Phase transition and adaptive algorithms." (DOI: 10.1214/24-AOS2362SUPP; .pdf). This supplementary material provides the complete proofs of the main theorems and technical results introduced in the paper, "Transfer Learning for Functional Mean Estimation: Phase Transition and Adaptive Algorithms." The structure of this material is as follows. Supplementary Material A handles the remaining results in common design: Theorems 2.1 and 2.4. Supplementary Material B covers the main results in independent design: Theorems 3.1, 3.2, 3.3 and 3.4. Supplementary Material C includes the proofs of technical results relevant to common design such as Propositions 6.1, 6.2, 6.5 and 6.6 as well as Lemmase A.1 and C.1. Finally, Supplementary Material D contains the proofs of Lemma B.1–B.5, which are technical lemmas for the main results in independent design.

REFERENCES

[1] ASSOUAD, P. (1983). Densité et dimension. *Ann. Inst. Fourier* (*Grenoble*) **33** 233–282. MR0723955
[2] BREIMAN, L. (1996). Bagging predictors. *Mach. Learn.* **24** 123–140. https://doi.org/10.1007/BF00058655
[3] CAI, C., CAI, T. T. and LI, H. (2024). Transfer learning for contextual multi-armed bandits. *Ann. Statist.* **52** 207–232. MR4718413 https://doi.org/10.1214/23-aos2341
[4] LI, S., ZHANG, L., CAI, T. T. and LI, H. (2023). Estimation and inference for high-dimensional generalized linear models with knowledge transfer. *J. Amer. Statist. Assoc.* https://doi.org/10.1080/01621459.2023.2184373
[5] CAI, T. T. and HALL, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34** 2159–2179. MR2291496 https://doi.org/10.1214/009053606000000830
[6] CAI, T. T, KIM, D. and PU, H. (2024). Supplement to "Transfer learning for functional mean estimation: Phase transition and adaptive algorithms." https://doi.org/10.1214/24-AOS2362SUPP
[7] CAI, T. T. and PU, H. (2022). Transfer Learning for Nonparametric Regression: Non-Asymptotic Minimax Analysis and Adaptive Procedure Technical report.
[8] CAI, T. T. and WEI, H. (2021). Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *Ann. Statist.* **49** 100–128. MR4206671 https://doi.org/10.1214/20-AOS1949
[9] CAI, T. T. and YUAN, M. (2010). Nonparametric Covariance Function Estimation for Functional and Longitudinal Data Technical report.
[10] CAI, T. T. and YUAN, M. (2011). Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *Ann. Statist.* **39** 2330–2355. MR2906870 https://doi.org/10.1214/11-AOS898
[11] CAI, T. T. and YUAN, M. (2012). Minimax and adaptive prediction for functional linear regression. *J. Amer. Statist. Assoc.* **107** 1201–1216. MR3010906 https://doi.org/10.1080/01621459.2012.716337
[12] CHOI, K., FAZEKAS, G., SANDLER, M. and CHO, K. (2017). Transfer Learning for Music Classification and Regression Tasks Technical report. https://doi.org/10.48550/arXiv.1703.09179
[13] DEGRAS, D. (2017). Simultaneous confidence bands for the mean of functional data. *Wiley Interdiscip. Rev.: Comput. Stat.* **9** e1397, 15. MR3648600 https://doi.org/10.1002/wics.1397
[14] GONG, B., SHI, Y., SHA, F. and GRAUMAN, K. (2012). Geodesic flow kernel for unsupervised domain adaptation. In 2012 *IEEE Conference on Computer Vision and Pattern*. Recognition 2066–2073. https://doi.org/10.1109/CVPR.2012.6247911
[15] HUANG, J.-T., LI, J., YU, D., DENG, L. and GONG, Y. (2013). Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In 2013 *IEEE International Conference on Acoustics, Speech and Signal Processing* 7304–7308. https://doi.org/10.1109/ICASSP.2013.6639081
[16] JAMES, N. and MENZIES, M. (2021). Trends in COVID-19 prevalence and mortality: A year in review. *Phys. D* **425** Paper No. 132968, 12. MR4275048 https://doi.org/10.1016/j.physd.2021.132968

[17] JIANG, C.-R., ASTON, J. A. D. and WANG, J.-L. (2009). Smoothing dynamic positron emission tomography time courses using functional principal components. *NeuroImage* **47** 184–193. https://doi.org/10.1016/j.neuroimage.2009.03.051

[18] KALOGRIDIS, I. and VAN AELST, S. (2023). Robust optimal estimation of location from discretely sampled functional data. *Scand. J. Stat.* **50** 411–451. MR4599920

[19] KOZLOFF, N., MULSANT, B. H., STERGIOPOULOS, V. and VOINESKOS, A. N. (2020). The Covid-19 global pandemic: Implications for people with schizophrenia and related disorders. *Schizophr. Bull.* **46** 752–757. https://doi.org/10.1093/schbul/sbaa051

[20] KPOTUFE, S. and MARTINET, G. (2018). Marginal singularity, and the benefits of labels in covariate-shift. In *Proceedings of the* 31*st Conference on Learning Theory* 1882–1886. *PMLR*.

[21] LENG, X. and MÜLLER, H.-G. (2006). Classification using functional data analysis for temporal gene expression data. *Bioinformatics* **22** 68–76. https://doi.org/10.1093/bioinformatics/bti742

[22] LI, S., CAI, T. T. and LI, H. (2022). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 149–173. MR4400393

[23] LI, S., CAI, T. T. and LI, H. (2023). Transfer learning in large-scale Gaussian graphical models with false discovery rate control. *J. Amer. Statist. Assoc.* **118** 2171–2183. MR4646634 https://doi.org/10.1080/01621459.2022.2044333

[24] MANTÉ, C., DURBEC, J. P. and DAUVIN, J. C. (2005). A functional data-analytic approach to the classification of species according to their spatial dispersion. Application to a marine macrobenthic community from the Bay of Morlaix (Western English Channel). *J. Appl. Stat.* **32** 831–840. MR2214313 https://doi.org/10.1080/02664760500080124

[25] PAGE, A., AYALA, G., LEÓN, M. T., PEYDRO, M. F. and PRAT, J. M. (2006). Normalizing temporal patterns to analyze sit-to-stand movements by using registration of functional data. *J. Biomech.* **39** 2526–2534. https://doi.org/10.1016/j.jbiomech.2005.07.032

[26] PAN, S. J. and YANG, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22** 1345–1359. https://doi.org/10.1109/TKDE.2009.191

[27] PARK, C., KOO, J.-Y., KIM, S., SOHN, I. and LEE, J. W. (2008). Classification of gene functions using support vector machine for time-course gene expression data. *Comput. Statist. Data Anal.* **52** 2578–2587. MR2411960 https://doi.org/10.1016/j.csda.2007.09.002

[28] POMANN, G.-M., STAICU, A.-M. and GHOSH, S. (2016). A two-sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **65** 395–414. MR3470583 https://doi.org/10.1111/rssc.12130

[29] RAMSAY, J. O. and SILVERMAN, B. W. (2002). *Applied Functional Data Analysis*: *Methods and Case Studies*. *Springer Series in Statistics*. Springer, New York. MR1910407 https://doi.org/10.1007/b98886

[30] REEVE, H. W. J., CANNINGS, T. I. and SAMWORTH, R. J. (2021). Adaptive transfer learning. *Ann. Statist.* **49** 3618–3649. MR4352543 https://doi.org/10.1214/21-aos2102

[31] RICE, J. A. and SILVERMAN, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B* **53** 233–243. MR1094283

[32] SONG, J. J., DENG, W., LEE, H.-J. and KWON, D. (2008). Optimal classification for time-course gene expression data using functional data analysis. *Comput. Biol. Chem.* **32** 426–432. MR2474550 https://doi.org/10.1016/j.compbiolchem.2008.07.007

[33] STAICU, A.-M., CRAINICEANU, C. M., REICH, D. S. and RUPPERT, D. (2012). Modeling functional data with spatially heterogeneous shape characteristics. *Biometrics* **68** 331–343. MR2959599 https://doi.org/10.1111/j.1541-0420.2011.01669.x

[34] TIAN, Y. and FENG, Y. (2023). Transfer learning under high-dimensional generalized linear models. *J. Amer. Statist. Assoc.* **118** 2684–2697. MR4681613 https://doi.org/10.1080/01621459.2022.2071278

[35] TZENG, E., HOFFMAN, J., SAENKO, K. and DARRELL, T. (2017). Adversarial discriminative domain adaptation. In 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) 2962–2971. https://doi.org/10.1109/CVPR.2017.316

[36] WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics*: *A Non-asymptotic Viewpoint*. *Cambridge Series in Statistical and Probabilistic Mathematics* **48**. Cambridge Univ. Press, Cambridge. MR3967104 https://doi.org/10.1017/9781108627771

[37] WANG, J.-L., CHIOU, J.-M. and MÜLLER, H.-G. (2016). Functional data analysis. *Annu. Rev. Stat. Appl.* **3** 257–295. https://doi.org/10.1146/annurev-statistics-041715-033624

[38] WEAVER, C., XIAO, L. and LU, W. (2023). Functional data analysis for longitudinal data with informative observation times. *Biometrics* **79** 722–733. MR4606310

[39] WEISS, K., KHOSHGOFTAAR, T. M. and WANG, D. (2016). A survey of transfer learning. *J. Big Data* **3**. https://doi.org/10.1186/s40537-016-0043-6

[40] WU, P.-S. and MÜLLER, H.-G. (2010). Functional embedding for the classification of gene expression profiles. *Bioinformatics* **26** 509–517. https://doi.org/10.1093/bioinformatics/btp711

[41] YUAN, M. and CAI, T. T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *Ann. Statist.* **38** 3412–3444. MR2766857 https://doi.org/10.1214/09-AOS772

[42] ZHOU, L., LIN, H. and LIANG, H. (2018). Efficient estimation of the nonparametric mean and covariance functions for longitudinal and sparse functional data. *J. Amer. Statist. Assoc.* **113** 1550–1564. MR3902229 https://doi.org/10.1080/01621459.2017.1356317