Minimax And Adaptive Transfer Learning for Nonparametric Classification under Distributed Differential Privacy Constraints

Arnab Auddy

Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104.

T. Tony Cai

Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104.

Abhinav Chakraborty

Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104.

Summary. This paper considers minimax and adaptive transfer learning for nonparametric classification under the posterior drift model with distributed differential privacy constraints. Our study is conducted within a heterogeneous framework, encompassing diverse sample sizes, varying privacy parameters, and data heterogeneity across different servers.

We first establish the minimax misclassification rate, precisely characterizing the effects of privacy constraints, source samples, and target samples on classification accuracy. The results reveal interesting phase transition phenomena and highlight the intricate trade-offs between preserving privacy and achieving classification accuracy. We then develop a data-driven adaptive classifier that achieves the optimal rate within a logarithmic factor across a large collection of parameter spaces while satisfying the same set of differential privacy constraints. Simulation studies and real-world data applications further elucidate the theoretical analysis with numerical results.

1. Introduction

Massive and diverse datasets are now routinely collected across a wide range of scientific fields, including genomics, neuroimaging, astrophysics, climate studies, and signal processing. In many applications, alongside the primary data from the target study, additional datasets from different populations or environments with similar structures have also been collected. Transfer learning, which aims to improve learning performance in a target domain by transferring knowledge from different but related source domains, has become a vibrant and promising area of research in machine learning. This concept has found applications in areas such as computer vision, speech recognition, and genre classification.

Alongside the transfer learning framework, another crucial consideration in modern data science is the preservation of privacy. Sensitive data are often spread across various sources, each presenting unique challenges in privacy preservation (see, e.g., Guo et al. [2024]). Differential privacy (DP) has emerged as a leading framework for ensuring that statistical analysis results do not compromise the confidentiality of individual data points.

Originally introduced by Dwork et al. [2006], DP has garnered significant academic attention and has been embraced by industry leaders like Google, Microsoft, and Apple, as well as governmental entities such as the US Census Bureau (Abowd [2016]).

Addressing the distributed nature of data collection and analysis is crucial due to its implications for privacy preservation and collaboration. This raises a number of natural questions: firstly, when and how can inference under the target distribution be improved by leveraging data points distributed across several sources with disparate privacy constraints? Secondly, what are the fundamental limits of inference for the target with privacy-constrained learning from the sources?

The two questions outlined above frequently arise in modern data analysis, spurring a flurry of recent research. The answers to both questions fundamentally depend on the specific inference task at hand. Only recently has rigorous research explored the theoretical performance of transfer learning under differential privacy constraints, addressing both various parametric problems (Li et al. [2024]) and nonparametric problems (Ma and Yang [2023]).

In this paper, we focus on the task of binary nonparametric classification, also referred to as domain adaptation in the literature. Our objective is to achieve statistically optimal transfer learning under distributed DP constraints. We tackle the challenge of heterogeneous data sources with distinct distributions, examining how privacy parameters, sample sizes, and data heterogeneity affect classification performance. Our proposed classifiers are designed to adapt to unknown data heterogeneity and parameters, while maintaining statistical optimality. This adaptability enhances performance even under strict privacy constraints, balancing privacy preservation with classification accuracy. Before delving into further details, we outline some fundamental concepts of classification with transfer learning.

In a classification problem from a single source, we observe independent copies of a tuple (X, Y) from a distribution P where $X \in \mathbb{R}^d$ are the covariates, and $Y \in \{0, 1\}$ denotes the binary class labels. To measure the efficacy of transferring information from a source distribution P to a different target distribution Q, several methods have been proposed to quantify the similarity of P and Q. Building upon the intuition that transfer is easier if both P and Q have high masses in a common region, divergence measures have been used on either the covariate space or as discrepancy between labels. See, e.g., Ben-David et al. [2010], Germain et al. [2013], Cortes et al. [2019], Sugiyama et al. [2007] and references therein. While these are general metrics of measuring the difference between P and Q, such approaches often tend to be pessimistic, as shown by Kpotufe and Martinet [2021]. Incorporating the structure of the classification problem leads to describing more specific transfer models such as covariate shift: where one posits a difference in the marginal distribution of X from P to Q, but assumes the posterior probabilities of $\mathbb{P}(Y = 1|X)$ to be the same for both P and Q. On the other hand, label shift assumes that the class probability $\mathbb{P}(Y = 1)$ differs from P to Q, but the class-conditional covariate distributions $\mathbb{P}(X|Y = 1)$ remains the same. For more details we refer the reader to Kpotufe and Martinet [2021], Sugiyama et al. [2007] for the covariate shift, and to Garg et al. [2020], Lipton et al. [2018], Maity et al. [2022] for the label shift paradigms. More flexible transfer mechanisms have also been considered, see, e.g., Reeve et al. [2021], Fan et al. [2023].

In this paper we focus on the posterior drift model, where the covariates have the same marginal distribution under both the source and target distributions, but the posterior probabilities $\mathbb{P}(Y = 1|X)$ undergoes a shift from the source to the target. Posterior drift has often been studied in the literature: see Cai and Wei [2021], Liu et al. [2020a], Maity et al. [2024], Scott [2019] and references therein. The posterior drift model naturally arises in a range of applications where the data is distributed and there are privacy concerns. Here are a few examples:

- **Healthcare Monitoring Across Hospitals:** In applications where multiple hospitals contribute data for healthcare monitoring (Yeung [2019]) (e.g., patient vital signs, medical histories), each hospital's data is sensitive and subject to privacy regulations like HIPAA. As medical practices evolve and patient demographics change, the underlying distribution of health data in each hospital's domain may drift over time. However, due to privacy concerns, the data cannot be aggregated into a central repository for analysis (see for e.g., Ju et al. [2020] which proposes a privacy-preserving federated transfer learning architecture for EEG classification, achieving higher accuracy without data sharing.). Consequently, the transfer learning model trained on data from one hospital may experience posterior drift when applied to another hospital's data, leading to degradation in performance over time.
- **Financial Fraud Detection in Banking Networks:** Banks collaborate to detect financial fraud by sharing transaction data (Phua et al. [2010], Chan et al. [1999]), but due to privacy regulations and competitive concerns, they cannot directly share sensitive customer information. Over time, patterns of financial fraud may evolve due to changes in customer behavior, economic conditions, or fraud tactics. However, because of customer privacy concerns, each bank must maintain control over its own data, making it challenging to aggregate data for analysis. As a result, transfer learning models used for fraud detection (Lebichot et al. [2020]) may experience posterior drift as the data distributions in different banks' domains shift over time.
- **Social Media Analysis Across Platforms:** Social media platforms collect user-generated content and engagement metrics for analyzing trends, sentiment analysis, and targeted advertising (Saura et al. [2019]). However, due to privacy regulations and platform policies, individual user data cannot be shared openly between social media platforms. As online communities evolve and user behaviors shift with trending topics, viral content, and platform updates, the underlying distribution of social media data in each platform's domain may vary. Therefore, learning methods for social media analysis (Wang et al. [2020]) may encounter posterior drift when applied across platforms, while simultaneously requiring to protect sensitive user data.

Nonparametric classification in the posterior drift model has been considered previously by Cai and Wei [2021], where the minimax rate of misclassification risk without the privacy constraints is established. The present work builds upon their framework, and establishes the optimal rates for classification using data that are distributed across servers with varying quality and different privacy constraints. Our results reveal interesting phase transition phenomena and highlight the intricate trade-offs between preserving privacy and achieving classification accuracy. We propose statistically optimal adaptive procedures that effectively balances this trade-off between privacy and accuracy.

Suppose the source and target distributions P and Q have similar covariate distributions, and consider the functions $\eta_P(x) := \mathbb{P}(Y = 1|X = x)$ under P, and $\eta_Q(x) := \mathbb{P}(Y = 1|X = x)$ under Q. Further, suppose that both P and Q have the same decision boundary, that is, $(\eta_Q(x) - 1/2)(\eta_P(x) - 1/2) \geq 0$ for all $x \in \mathbb{R}^d$. In this framework, Cai and Wei [2021] quantified the efficiency of transfer learning at Q using data from P through the so-called relative signal exponent γ, where $\gamma > 0$ is a number such that

$$\left|\eta_P(x) - \frac{1}{2}\right| \geq \left|\eta_Q(x) - \frac{1}{2}\right|^{\gamma}.$$

When $\gamma < 1$, the covariates are better separated into the two classes in the source distribution P than in Q. In such a case, the P-data are especially useful for transfer learning. The situation reverses when $\gamma > 1$. In this paper, we describe a kernel based classifier that leverages the information from sources, while preserving data privacy in the DP framework. For both the source and the target, we compute:

$$T_P(x) = \frac{1}{n_P h^d} \sum_{i=1}^{n_P} \left(Y_i - \frac{1}{2}\right) K\left(\frac{X_i - x}{h}\right) \text{ and } T_Q(x) = \frac{1}{n_Q h^d} \sum_{i=1}^{n_Q} \left(Y_i - \frac{1}{2}\right) K\left(\frac{X_i - x}{h}\right), \tag{1}$$

for a suitable kernel $K(\cdot)$ and bandwidth h. It can be checked that $T_P(x)$ and $T_Q(x)$ are pointwise consistent estimators of $\left(\eta_P(x) - \frac{1}{2}\right) g_P(x)$ and $\left(\eta_Q(x) - \frac{1}{2}\right) g_Q(x)$ respectively, where $g_P(\cdot)$ and $g_Q(\cdot)$ are the joint densities of the covariates under P and Q respectively. Thus the sign of a suitable linear combination of $T_P(x)$ and $T_Q(x)$ defines a reasonable transfer learning classifier.

The above classification procedure is however, not differentially private. Nonparametric classifiers with local privacy have previously been considered by Berrett and Butucea [2019], Ma and Yang [2023]. In this paper, we focus on "server-level" privacy, where each server can access its own set of unperturbed data, but imposes privacy constraints when sharing aggregated information with other servers. Due to its immense applicability in practice, such a distributed privacy setting has recently received significant attention in the literature: see e.g., Cai et al. [2023], Acharya et al. [2023], Liu et al. [2020b], Levy et al. [2021].

To emphasize the impact of distributed data privacy, we consider a scenario with m source servers from distribution P, each holding n_P samples. Similar to (1) we define the estimator for each j-th source server, denoted as $T_P^{(j)}(\cdot)$. For attaining differential privacy, it is standard to add external noise, with an important tradeoff in mind. The noise level must be enough to ensure the (ε, δ)-differential privacy guarantees (see Definition 2.1) for both the source and the target, but it has to be added prudently so as to not worsen the classification accuracy too much. In our case, we output a noisy test function as follows. We use the Gaussian process based mechanism used by Hall et al. [2013] to privatize kernel estimators. With $T_Q(\cdot)$ and $T_P^{(j)}(\cdot)$ as defined above, we compute:

$$\widetilde{T}_Q(x) = T_Q(x) + \sigma_Q \xi_Q(x) \text{ and } \widetilde{T}_P^{(j)}(x) = T_P^{(j)}(x) + \sigma_P \xi_P^{(j)}(x) \text{ for } j = 1, \ldots, m,$$

where $\xi_P(\cdot)$ and $\xi_Q^{(j)}(\cdot)$ are Gaussian processes with covariance function $K(\cdot/h)$, while σ_P^2 and σ_Q^2 are suitably chosen noise variances that depend on the privacy parameters (ε, δ),

sample sizes and the bandwidth h. Next, we choose an appropriate weight $w \in [0,1]$, and define the privatized transfer weighted estimator

$$\widetilde{T}(x) = w\widetilde{T}_Q(x) + (1-w)\left(\frac{1}{m}\sum_{j=1}^{m}\widetilde{T}_P^{(j)}(x)\right).$$

Finally our classifier becomes:

$$\widehat{f}(x) := \mathbb{1}(\widetilde{T}(x) \geq 0). \tag{2}$$

We use the standard notion of excess risk (see, e.g., Audibert and Tsybakov [2007]) to evaluate the performance of our classifier. Let $f_Q^*(x) = \mathbb{1}(\eta_Q(x) \geq 1/2)$ be the Bayes classifier for the target distribution. Then the excess risk of a classifier under the target distribution is defined as:

$$\mathcal{E}_Q(\widehat{f}) = \mathbb{E}\left[\mathbb{P}_Q(Y \neq \widehat{f})\right] - \mathbb{P}_Q(Y \neq f_Q^*).$$

We assume that the margin assumption (see Definition 2.5) holds with parameter α for Q, and $\eta_Q(x)$ is (β, L_β)-Hölder for some $0 < \beta < 1$. Then if the target and the source are both constrained to be (ε, δ) differentially private, we prove the following minimax rate:

$$\begin{aligned}&\inf_{\widetilde{f}\in\mathcal{M}(\varepsilon,\delta)}\sup_{P,Q}\mathcal{E}_Q(\widehat{f})\\ \asymp\ &\left[L_N\left\{\left(n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}} \vee (n_Q^2\varepsilon^2)^{-\frac{\beta(1+\alpha)}{2\beta+2d}}\right)\bigwedge\left((mn_P)^{-\frac{\beta(1+\alpha)}{2\beta\gamma+d}} \vee (mn_P^2\varepsilon^2)^{-\frac{\beta(1+\alpha)}{2\beta\gamma+2d}}\right)\right\}\wedge 1\right]\end{aligned} \tag{3}$$

where L_N is of order at most $\mathrm{polylog}(mn_P + n_Q)$, the infimum is over all possible transfer learning classifiers satisfying (ε, δ)-DP, and the supremum is over all distributions in the posterior drift framework. The minimax rate is attained by the classifier $\widehat{f}$ in (2).

The minimax rate is determined by a trade-off between four quantities: the non-private rates for the source and the target, as well as their privatized counterparts. An interesting phenomenon emerges through the relative effect of the efficacy of transfer and the cost of privacy, characterized by the parameters γ and ε along with server size m and sample sizes n_P, n_Q. This discrepancy essentially means that a higher privacy cost is incurred when data is more distributed, see Figure 1. For a detailed discussion, refer to the remarks following Theorem 3.1.

Figure 1 illustrates the minimax rate as described in Equation (3), showcasing the different regimes of differentially private transfer learning excess risk. Several intriguing phenomena are observed. For instance, across varying privacy budgets ε, the regimes where target and source servers dominate are not contiguous. Additionally, the phase transitions and the order in which they appear depend delicately on how distributed the data is and the quality of the source data. For a detailed account of all the phase transitions and their intricate dependence on the problem parameters, see the discussion following Corollary 3.2.

In practical applications, we strive to borrow information from multiple sources which are very heterogeneous with respect to both the classification transfer quality, in terms of γ, and the privacy constraints. Take for example, the heart disease dataset of Detrano

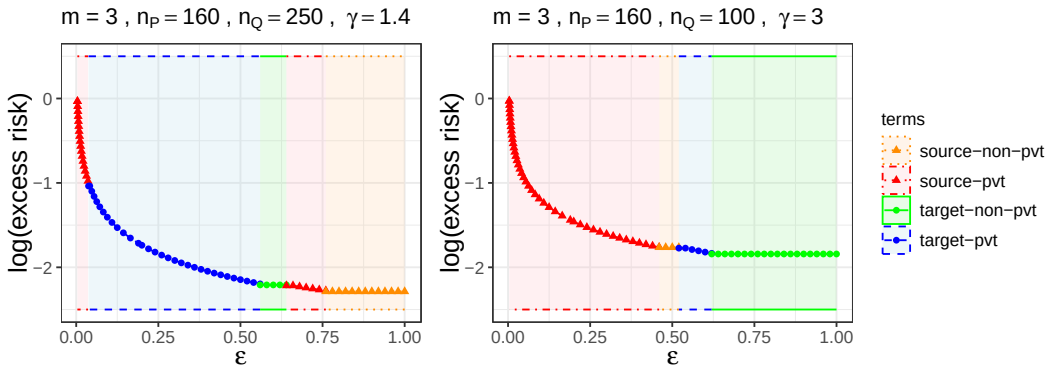

Fig. 1. Relationship of logarithm of excess risk with ε as given by (4), the smoothness level $\beta = 0.25$ and dimension $d = 2$.

et al. [1989], we want to predict the propensity of heart disease in a patient, based on their demographic information and clinical measurements. The data is collected from four different hospitals in Cleveland, Hungary, Long Beach, and Switzerland. The number of patients and disease prevalence are widely varied at each of these four hospitals. Moreover, a model fitted solely on one server shows drastically different performance on other servers, thus pointing to significant heterogeneity in the data. Details can be found in Section 6.2. At the same time, it is reasonable to assume each hospital having its own privacy tolerance level, determined by local guidelines. To tackle these challenges, in Section 4, we develop minimax optimal methods in the heterogeneous setting, where the trade-offs between transfer learning and privacy are more nuanced.

In particular, we show in Section 4 that even when the source distributions are heterogeneous, with their own transfer parameters γ_j and privacy constraints ε_j, δ_j our weighted kernel based classifier continues to be minimax optimal. However the best choices of weights w and bandwidth h depends heavily on the knowledge of γ_j, which is typically unknown. The additional noise due to privacy compounds this issue even further, since choosing a worse bandwidth potentially implies adding a higher amount of noise than is required, thus unnecessarily worsening the performance. To alleviate this issue, we take a data-adaptive approach to choose the best bandwidth from a grid of possible choices. This is based on the popular Lepski method fine-tuned to our setting to accommodate the additional noise for privacy. Given the privacy parameters (ε, δ), our adaptation algorithm outputs a classifier that attains the minimax rate, modulo a $O(\text{polylog}(n_*))$ factor, where n_* is the sample size of the data pooled across all servers. Details can be found in Section 5.

The rest of the paper is organized as follows. In Section 2, we provide the background and formulate our problem in detail. Section 3 presents the minimax rate of excess risk for our problem across various specific cases, as well as the most general case. Section 4 introduces our kernel based classifier, derives its excess risk bounds, and states the minimax lower bound. Section 5 describes the data-driven adaptive procedure for bandwidth and weight selection. We evaluate our proposed method and compare it to existing work via several numerical experiments on simulated and real datasets in Section 6. The paper

concludes with a discussion on possible future work in Section 7. All proofs can be found in the supplementary material Auddy et al. [2024].

2. Problem Formulation

In this section, we outline the general framework for transfer learning under distributed privacy constraints. Our dataset is distributed across $m+1$ servers, indexed by the set $\{0, 1, \ldots, m\}$. The dataset is categorized as target and source. On server 0 (also called the target server), we have n_0 i.i.d. samples from the distribution P_0, while on server j (the source servers) for $j \in 1, \ldots, m$, we have n_j i.i.d. samples from distribution P_j. All of the probability measures $\{P_j\}_{j=0}^m$ are defined on the measurable space $(\mathcal{Z}, \mathscr{Z})$. Let $Z^{(0)} = \{Z_i^{(0)}\}_{i=1}^{n_0}$ denote the n_0 realizations from P_0 on the target server. Let us denote by $Z^{(j)} = \{Z_i^{(j)}\}_{i=1}^{n_j}$ the n_j realizations from P_j on the jth source server for $j = 1, \ldots, m$. These servers serve as the source data, and our goal is to learn the model for our target distribution P_0.

For each source server i.e $j = 1, \ldots, m$, we send a (randomized) transcript $\widetilde{T}^{(j)}$ based on $Z^{(j)}$ to the target server 0, where the law of the transcript is given by a distribution conditional on $Z^{(j)}$, $\mathbb{P}(\cdot|Z^{(j)})$, on a measurable space $(\mathcal{T}, \mathscr{T})$. For $j = 1, \ldots, m$ the transcript $\widetilde{T}^{(j)}$ has to satisfy a $(\varepsilon_j, \delta_j)$-differential privacy constraint.

Definition 2.1. *The transcript $\widetilde{T}^{(j)}$ is $(\varepsilon_j, \delta_j)$-differentially private if for all $A \in \mathcal{A}$ and z, z' differing in one individual datum, it holds that*

$$\mathbb{P}\left(\widetilde{T}^{(j)} \in A \,|\, Z^{(j)} = z\right) \leq e^{\varepsilon_j} \mathbb{P}\left(\widetilde{T}^{(j)} \in A \,|\, Z^{(j)} = z'\right) + \delta_j.$$

The target server can look at the private transcripts $\{\widetilde{T}^{(j)}\}_{j=1}^m$ and the target data $Z^{(0)}$ while constructing the final private transcript $\widetilde{T}$. Hence $\widetilde{T}$ satisfies $(\varepsilon_0, \delta_0)$-interactive differential privacy constraint, which is defined as follows:

Definition 2.2. *The transcript $\widetilde{T}$ is $(\varepsilon_0, \delta_0)$-differentially private if for all $A \in \mathcal{A}$ and z, z' differing in one individual datum and for all $t_j \in \mathcal{T}$ for $j = 1, \ldots, m$, it holds that*

$$\begin{aligned} &\mathbb{P}\left(\widetilde{T} \in A \,|\, Z^{(0)} = z, \widetilde{T}^{(j)} = t_j \text{ for } 1 \leq j \leq m\right) \\ &\leq e^{\varepsilon_0} \mathbb{P}\left(\widetilde{T} \in A \,|\, Z^{(0)} = z', \widetilde{T}^{(j)} = t_j \text{ for } 1 \leq j \leq m\right) + \delta_0. \end{aligned}$$

This privacy constraint can be understood as follows: if we condition on the outcome of all other servers then the distribution of the final private transcript $\widetilde{T}$ does not change much if one of the datum on the target server is changed.

In transfer learning, the focus is on scenarios where multiple parties, such as hospitals, possess heterogeneous data with differing underlying distributions. Employing distributed protocols in such contexts ensures differential privacy while yielding outputs from each participating party. Within this framework, transcripts generated by each source server rely solely on its local data, with no exchange of information occurring between source servers. Communication is solely between the source and target servers. Each of the source

server transmits its transcripts to the target server. The target server utilizing all the transcripts $(\widetilde{T}^{(1)}, \ldots, \widetilde{T}^{(m)})$ from the other servers and target data $Z^{(0)}$, computes the final private transcript $\widetilde{T}$. This scenario often arises when multiple trials involving a population similar to that of the target server are conducted, yet individual locations, such as hospitals, opt against consolidating their original data due to privacy apprehensions.

In the context of transfer learning for nonparametric classification our data looks like a couple $Z_i^{(j)} := (X_i^{(j)}, Y_i^{(j)})$, for $i = 1, \ldots, n_j$; $j = 1, \ldots, m$ for the source servers, and $Z_i^{(0)} := (X_i^{(0)}, Y_i^{(0)})$, for $i = 1, \ldots, n_0$ for the target server. We assume that $Z_i^{(j)}$ takes values in $\mathcal{Z} := [0,1]^d \times \{0,1\}$. We regard $X \in [0,1]^d$ as a vector of features corresponding to an object and $Y \in \{0,1\}$ as a label indicating that the object belongs to one of two classes. Our goal is to propose distributed DP protocols $\widetilde{T}^{(j)}$ for each server and construct classifier $\widehat{f} : [0,1]^d \to \{0,1\}$ based on the final private transcript $\{\widetilde{T}\}$. Unlike the traditional federated learning framework, there's no central server; alternatively, we can consider the target server as acting in a central capacity. We denote the vector of privacy budgets as $(\boldsymbol{\varepsilon}, \boldsymbol{\delta}) = \{(\varepsilon_j, \delta_j)\}_{j=0}^m$ and the class of distributed DP classifiers $\widehat{f}$ by $\mathcal{M}_{\boldsymbol{\varepsilon},\boldsymbol{\delta}}$. Next we denote

$$\eta_j(X^{(j)}) := \mathbb{P}(Y^{(j)} = 1 | X^{(j)}) \text{ for the source servers } j = 1, \ldots, m; \text{ and}$$
$$\eta_0(X^{(0)}) := \mathbb{P}(Y^{(0)} = 1 | X^{(0)}) \text{ for the target server,}$$

as the (source and target) regression functions of Y on X. We denote the marginal distribution of X for the jth server, $j = 0, \ldots, m$ as P_j^X. Define the classification error of a classifier f under the target distribution P_0 as

$$R_0(f) := P_0(Y \neq f(X))$$

The Bayes decision rule is a minimizer of the of the risk $R_0(f)$ which has the form $f_0^*(X) = \mathbb{1}\{\eta_0(X) \geq 1/2\}$. The goal of transfer learning is to transfer the knowledge gained from the source data together with the information in the target data to construct a classifier which minimizes the excess risk on the target data

$$\mathcal{E}_0(\widehat{f}) = \mathbb{E}[R_0(\widehat{f})] - R_0(f_0^*)$$

Under the posterior drift model we quantify the similarity between the regression functions $\{\eta_j\}_{j=1}^m$ and η_0 as follows:

Definition 2.3 (**Relative Signal Exponent (RSE)**). *The class $\Gamma(\boldsymbol{\gamma}, C_{\boldsymbol{\gamma}})$ with relative signal exponent $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m) \in \mathbb{R}_+^m$ and constants $C_{\boldsymbol{\gamma}} = (C_1, \ldots, C_m) \in \mathbb{R}_+^m$, is the set of distribution tuples $(P_0, P_1, \ldots, P_m)$ that satisfy for $1 \leq j \leq m$*

(a) $\operatorname{sign}\left(\eta_j(x) - \frac{1}{2}\right) = \operatorname{sign}\left(\eta_0(x) - \frac{1}{2}\right)$ *for all $1 \leq j \leq m$ and all $x \in [0,1]^d$.*

(b) $\left|\eta_j(x) - \frac{1}{2}\right| \geq C_j \left|\eta_0(x) - \frac{1}{2}\right|^{\gamma_j}$ *for some $\gamma_j > 0$, for all $1 \leq j \leq m$ and all $x \in [0,1]^d$.*

Remark 2.1. The first part follows from the assumption that the Bayes classifier f^* is the same for both source and target populations. The second part introduces a parameter γ which controls the signal strength of the source data from P, with respect to the target data from Q. See Definition 1 and Remark 1 of Cai and Wei [2021].

In addition to the RSE assumption we also need to assume smoothness of η_0 and characterize it behavior near 1/2. These assumptions are standard in nonparametric classification and was first introduced in Audibert and Tsybakov [2007].

DEFINITION 2.4 (**Hölder Smoothness**). *The regression function η_0 belongs to the Hölder class of functions denoted by $\Sigma(\beta, L)$ $(0 < \beta \leq 1)$ which is defined as the set of functions satisfying:*

$$|\eta_0(x) - \eta_0(x')| \leq L\|x - x'\|^\beta \quad \text{for } x, x' \in [0,1]^d.$$

DEFINITION 2.5 (**Margin Assumption (MA)**). *The margin class $\mathcal{M}(\alpha, C_\alpha)$ with $\alpha \geq 0$ and $C_\alpha > 0$ is defined as the set of distributions P_0 such that*

$$P_0^X(0 \leq |\eta_0(X) - 1/2| \leq t) \leq C_\alpha t^\alpha \text{ for all } t > 0.$$

Another definition is about marginal density of X, P_j^X for $j = 1, \ldots, m$.

DEFINITION 2.6 (**Common Support and Strong Density Assumption (SD)**). *We assume that P_j^X for $j = 0, \ldots, m$ have the identical support on a compact (c_μ, r_μ) regular set $A \subset [0,1]^d$ and has a density g_j w.r.t. the Lebesgue measure bounded away from zero and infinity on A:*

$$g_{\min} \leq g_j(x) \leq g_{\max} \text{ for } x \in A \text{ and } g_j(x) = 0 \text{ otherwise},$$

where $c_0, r_0 > 0$ and $0 < g_{\min} < g_{\max} < \infty$ are fixed constants. We denote the set of marginal distributions $(P_0^X, \ldots, P_m^X)$ which satisfy the above constraints as $\mathcal{S}(\mu, c_\mu, r_\mu)$ where $\mu = (g_{\min}, g_{\max})$.

REMARK 2.2. In this paper we focus our attention to the case when the marginal densities have regular support and are bounded from below and above on their support. Moreover we assume that $\alpha\beta \leq d$ throughout the paper. This is because in the other regime ($\alpha\beta > d$), there is no distribution P_0^X such that the regression function η_0 crosses 1/2 in the interior of the support of P_0^X (Audibert and Tsybakov [2007]) and hence this case only contains the trivial cases for classification.

We put all the definitions together to define the class of distributions we consider in the posterior drift model as

$$\begin{aligned}
&\Pi(\boldsymbol{\gamma}, C_{\boldsymbol{\gamma}}, \beta, L, \alpha, C_\alpha, \mu, c_\mu, r_\mu) \\
&:= \{(P_0, P_1, \ldots, P_m) : (P_0, P_1, \ldots, P_m) \in \Gamma(\boldsymbol{\gamma}, C_{\boldsymbol{\gamma}}), \eta_0 \in \Sigma(\beta, L), \\
&\qquad\qquad \mathbb{P}_0^X \in \mathcal{M}(\alpha, C_\alpha), (P_0^X, P_1^X, \ldots, P_m^X) \in \mathcal{S}(\mu, c_\mu, r_\mu)\}
\end{aligned}$$

For the rest of the paper we will use the shorthand $\Pi(\alpha, \beta, \boldsymbol{\gamma}, \mu)$ or Π if there is no confusion.

3. Main Results

In this section, we present the key findings of our paper, where we establish the minimax rate of convergence for transfer learning under differential privacy constraints, specifically addressing the nonparametric classification problem.We divide our results into two subsections: Section 3.1 covers the homogeneous case, while Section 3.2 addresses the general heterogeneous case.

3.1. Minimax Rates under Source Homogeneity

To derive meaningful and interpretable insights from our minimax rate, we first examine the scenario where the source servers are exchangeable in terms of the distributed classification problem under transfer learning and privacy constraints. This homogeneous scenario is characterized by equal sample sizes ($n_j = n$), privacy parameters ($\varepsilon_j = \varepsilon$, $\delta_j = \delta$) and transfer exponents ($\gamma_j = \gamma$) for all $j = 1, \dots, m$.

THEOREM 3.1. *Suppose $n_j = n, \varepsilon_j = \varepsilon, \delta_j = \delta$ and $\gamma_j = \gamma$ for all $j = 1, \dots, m$ and assume that $\delta = o((nm)^{-1})$. Then the minimax rate for the excess risk satisfies*

$$\inf_{\widehat{f} \in \mathcal{M}(\boldsymbol{\varepsilon}, \boldsymbol{\delta})} \sup_{(P_0, \dots, P_m) \in \Pi} \mathcal{E}_0(\widehat{f}) \asymp \Bigg[L_N \Bigg\{ \left(n_0^{\frac{1}{2\beta+d}} \wedge (n_0^2 \varepsilon_0^2)^{\frac{1}{2\beta+2d}} \right) + \left((mn)^{\frac{1}{2\beta\gamma+d}} \wedge (mn^2\varepsilon^2)^{\frac{1}{2\beta\gamma+2d}} \right) \Bigg\}^{-\beta(1+\alpha)} \wedge 1 \Bigg]$$

for a sequence L_N of order at most $(\log((\delta \wedge \delta_0)^{-1}))^{\frac{\beta(1+\alpha)}{2\beta(\gamma \wedge 1)+d}}$.

Note that the minimax rate depends on the sum of quantities: the first of which determines the minimax rate for the problem using only the target data, while the second corresponds to the minimax rate for the problem using solely the source data from the source servers. Some remarks are in order regarding the minimax rate obtained of Theorem 3.1 in some special settings:

REMARK 3.1 (PRIVATE CLASSIFICATION WITH NO TRANSFER). When the number of servers $m = 0$ or there is no source data, i.e., $n = 0$ for all source servers we obtain the single server minimax classification rate under the (ε, δ)-DP constraint, given by

$$\inf_{\widehat{f} \in \mathcal{M}(\boldsymbol{\varepsilon}, \boldsymbol{\delta})} \sup_{(P_0, \dots, P_m) \in \Pi} \mathcal{E}_0(\widehat{f}) \asymp \left[L_N \left(n_0^{\frac{1}{2\beta+d}} \wedge (n_0^2 \varepsilon_0^2)^{\frac{1}{2\beta+2d}} \right)^{-\beta(1+\alpha)} \wedge 1 \right].$$

REMARK 3.2 (NON-PRIVATE TRANSFER LEARNING). When the privacy requirements are not stringent, the tradeoff is completely characterized by a comparison between the non-private rates of the target and the source. In particular, if $\varepsilon > \left(m^{d/2} n^{-\beta\gamma} \right)^{\frac{1}{2\beta\gamma+d}}$, we find the non-private transfer learning rates

$$\inf_{\widehat{f} \in \mathcal{M}(\boldsymbol{\varepsilon}, \boldsymbol{\delta})} \sup_{(P_0, \dots, P_m) \in \Pi} \mathcal{E}_0(\widehat{f}) \asymp \left(n_0 + (mn)^{\frac{2\beta+d}{2\beta\gamma+d}} \right)^{-\frac{\beta(1+\alpha)}{2\beta+d}}$$

which coincides with the results of Cai and Wei [2021], in the transfer homogeneous regime with equal source sample sizes.

In the current setting, we further emphasize how the requirement of privacy leads to a worsening of the rate if the same amount of data is distributed across a larger number of servers. Since the sources are all equivalent in terms of data quality (as quantified by γ), traditional knowledge suggests a rate depending on the pooled source sample size mn. In the m-server non-private transfer learning setup of Cai and Wei [2021], that is indeed

the case, as also shown by the non-private rate of $(mn)^{-\frac{\beta(1+\alpha)}{2\beta\gamma+2d}}$ appearing in Theorem 3.1. Note however that the private source rate is given by $m^{\frac{\beta(1+\alpha)}{2\beta\gamma+2d}}(mn\varepsilon)^{-\frac{\beta(1+\alpha)}{\beta\gamma+d}}$, so that if the pooled sample size mn remains fixed, the rate worsens as m, the number of servers, increases. This phenomenon is reminiscent of the minimax rate behavior observed in Cai et al. [2023] in the non-private setting.

In order to further our understanding about interplay between the transfer exponent γ and privacy parameters we restrict our attention to the case where the target server has same privacy budget, i.e., $\varepsilon_0 = \varepsilon$, $\delta_0 = \delta$ and the number of target samples is between n and mn. Other sample size regimes can be described similarly. As is clear from Theorem 3.1, the minimax rate of decay for the excess risk is given in this case by:

$$\mathcal{E}_0(\widehat{f}) \asymp \left[L_N \left\{ \left(n_0^{-\frac{\beta(1+\alpha)}{2\beta+d}} \vee (n_0^2\varepsilon^2)^{-\frac{\beta(1+\alpha)}{2\beta+2d}} \right) \bigwedge \left((mn)^{-\frac{\beta(1+\alpha)}{2\beta\gamma+d}} \vee (mn^2\varepsilon^2)^{-\frac{\beta(1+\alpha)}{2\beta\gamma+2d}} \right) \right\} \wedge 1 \right]. \tag{4}$$

Figure 1 illustrates the different regimes showing which of the four terms on the right hand side of (4) ends up determining the overall rate of excess risk.

We will refer to the four terms on the right hand side in (4) as the non-private target rate (NP_t), the private target rate (P_t), the non-private source rate (NP_s), and the private source rate (P_s) respectively. Depending on the value of the common privacy parameter ε and the transfer exponent γ, the overall rate will be determined by one of these four rates, as demonstrated by the following table. Corollary 3.2 formally states the results of Table 1 along with the endpoints of the various transfer and privacy regimes.

Table 1. Minimax rate of excess risk at different transfer and privacy regimes. See Corollary 3.2.

<table>
<tr><th>Transfer \ Privacy</th><th>$\varepsilon \in (0, \varepsilon^{(1)}]$</th><th>$\varepsilon \in (\varepsilon^{(1)}, \varepsilon^{(2)}]$</th><th>$\varepsilon \in (\varepsilon^{(2)}, \varepsilon^{(3)}]$</th><th>$\varepsilon \in (\varepsilon^{(3)}, 1]$</th></tr>
<tr><td>$\gamma \in (0, 1]$</td><td rowspan="3">1</td><td>$\begin{cases} \mathrm{P}_t & \text{if } \varepsilon \le \varepsilon^{(11)} \\ \mathrm{P}_s & \text{if } \varepsilon > \varepsilon^{(11)} \end{cases}$</td><td colspan="2">$\mathrm{NP}_s$</td></tr>
<tr><td>$\gamma \in (1, \gamma^{(*)}]$</td><td rowspan="2">$\begin{cases} \mathrm{P}_s & \text{if } \varepsilon \le \varepsilon^{(11)} \\ \mathrm{P}_t & \text{if } \varepsilon > \varepsilon^{(11)} \end{cases}$</td><td>$\begin{cases} \mathrm{NP}_t & \text{if } \varepsilon \le \varepsilon^{(21)} \\ \mathrm{P}_s & \text{if } \varepsilon > \varepsilon^{(21)} \end{cases}$</td><td>$\mathrm{NP}_s$</td></tr>
<tr><td>$\gamma \in (\gamma^{(*)}, \infty)$</td><td colspan="2">$\mathrm{NP}_t$</td></tr>
</table>

Corollary 3.2. *Suppose $n_j = n, \gamma_j = \gamma \, \forall 1 \le j \le m$, $n \le n_0 \le mn$, and equal privacy budget $\varepsilon_j = \varepsilon$, $\delta_j = \delta \, \forall 0 \le j \le m$. Further assume that $\delta = o((mn)^{-1})$. Then the minimax rate for the excess risk are as given in Table 1 with the various regimes characterized by the following endpoints:*

(a) $\gamma^{(*)} = \frac{1}{2\beta}\left[\frac{(2\beta+d)\log mn}{\log n_0} - d\right]$.

(b) $\varepsilon^{(1)} = (\sqrt{mn})^{-1} \wedge n_0^{-1}$; $\varepsilon^{(2)} = n_0^{-\frac{\beta}{2\beta+d}}$; $\varepsilon^{(3)} = \left(m^{d/2} n^{-\beta\gamma}\right)^{\frac{1}{2\beta\gamma+d}}$.

$$(c)\ \varepsilon^{(11)} = \begin{cases} \left[(\sqrt{m}n)^{\beta+d} n_0^{-(\beta\gamma+d)}\right]^{\frac{1}{\beta(\gamma-1)}} & \text{if } \gamma \neq 1, \\ \varepsilon^{(1)} & \text{if } \gamma = 1,\ n_0 \leq \sqrt{mn^2}, \\ \varepsilon^{(2)} & \text{if } \gamma = 1,\ n_0 > \sqrt{mn^2}. \end{cases}$$

$$(d)\ \varepsilon^{(21)} = (\sqrt{m}n)^{-1} n_0^{\frac{\beta\gamma+d}{\beta+d}}.$$

As we increase the value of ε, we observe interesting phenomena characterized by distinct phase transitions. Not surprisingly, the rates behave differently based on whether γ is small — where the quality of the source data is relatively better than the target data — versus when γ is larger. The two transition points in γ are at $\gamma = 1$ and at $\gamma^{(*)} = \frac{1}{2\beta}\left[\frac{(2\beta+d)\log mn}{\log n_0} - d\right]$. The different rates can be described based on ultra-high, high, moderate, and low privacy regimes.

First, in the ultra-high privacy regime where $0 < \varepsilon \leq \varepsilon^{(1)}$, the privacy requirements are so severe that no classifier has disappearing excess risk in this regime, and a random guess is the best one can do. Next, we move to the high privacy regime of $\varepsilon^{(1)} < \varepsilon \leq \varepsilon^{(2)}$, where the private rates of both the target and the source dominate over their non-private counterparts. Another interesting phenomenon emerges based on whether γ is smaller than or greater than one. If $\gamma < 1$ and the privacy requirement is high, the target private rates appear for very small ε, followed by the source private rates. This is because when $\gamma < 1$, the source data are of better quality and hence for sufficiently small ε, the noise added to the target dominates over the noise for the sources. The pattern reverses when $\gamma > 1$ where the source private rates appear first.

Further increasing ε, we arrive at the moderate privacy regime where $\varepsilon^{(2)} < \varepsilon \leq \varepsilon^{(3)}$. For small γ, characterized by $\gamma \leq 1$, the high quality source data points prove to be particularly beneficial. This results in the non-private source rate dominating over the non-private target rate, as well as the noises added for privatizing the source and the target. In contrast, when γ is very large, in particular $\gamma > \gamma^{(*)}$, the source data are particularly poor and not useful, so that the target non-private rate therefore dominates over all the other contenders. A more interesting picture emerges for moderate γ given by $1 < \gamma \leq \gamma^{(*)}$. In this regime of intermediate transfer and medial privacy, we find the non-private target rate first, followed by private source rates.

Finally for $\varepsilon > \varepsilon^{(3)}$, the low privacy scenario emerges. Here the effect of extraneous noise for privacy is not at all significant. We therefore only find the non-private rates based on the transfer efficacy. For $\gamma \leq \gamma^{(*)}$, the source data are more useful, resulting in NP_s governing the overall rate. As expected, for $\gamma > \gamma^{(*)}$ the relatively poorer source data do not contribute to the transfer. Consequently the non-private rates from the target becomes dominant, reflecting the diminished relevance of the source data in such scenarios.

In the rest of this subsection, we describe another specialized setting where we allow one of the source servers to be public. To demonstrate the effect of publicly available data, we take $m = 2$ sources, with one private and one public source server. The minimax rate for excess risk is then given by the following corollary:

COROLLARY 3.3. *Suppose that $\gamma_1 = \gamma_2 = \gamma$, $\varepsilon_1 = \infty$, $\varepsilon_0 = \varepsilon_2 = \varepsilon$ and $n_2 > n_0^{\frac{2\beta\gamma+d}{2\beta+d}}$, $\delta_0 \vee \delta_2 = o(n_2^{-1})$. Then the minimax rate for the excess risk satisfies the following.*

(a) If $n_1 > n_2$, $\inf_{\widehat{f}\in\mathcal{M}(\boldsymbol{\varepsilon},\boldsymbol{\delta})} \sup_{(P_0,\ldots,P_m)\in\Pi} \mathcal{E}_0(\widehat{f}) \asymp n_1^{-\frac{\beta(1+\alpha)}{2\beta\gamma+d}}$.

(b) If $n_0^{\frac{2\beta\gamma+d}{2\beta+d}} \le n_1 \le n_2$, *then*

$$\inf_{\widehat{f}\in\mathcal{M}(\boldsymbol{\varepsilon},\boldsymbol{\delta})} \sup_{(P_0,\ldots,P_m)\in\Pi} \mathcal{E}_0(\widehat{f}) \asymp \begin{cases} L_N n_1^{-\frac{\beta(1+\alpha)}{2\beta\gamma+d}} & \text{if } \varepsilon \le n_2^{-1} n_1^{\frac{\beta\gamma+d}{2\beta\gamma+d}} \\ L_N (n_2^2\varepsilon^2)^{-\frac{\beta(1+\alpha)}{2\beta\gamma+2d}} & \text{if } n_2^{-1} n_1^{\frac{\beta\gamma+d}{2\beta\gamma+d}} < \varepsilon \le L_N n_2^{-\frac{\beta\gamma}{2\beta\gamma+d}} \\ L_N n_2^{-\frac{\beta(1+\alpha)}{2\beta\gamma+d}} & \text{if } n_2^{-\frac{\beta\gamma+d}{2\beta\gamma+d}} < \varepsilon \le 1. \end{cases}$$

(c) If $n_1 \le n_0^{\frac{2\beta\gamma+d}{2\beta+d}} \le n_2$, *then*

$$\inf_{\widehat{f}\in\mathcal{M}(\boldsymbol{\varepsilon},\boldsymbol{\delta})} \sup_{(P_0,\ldots,P_m)\in\Pi} \mathcal{E}_0(\widehat{f})$$
$$\asymp \begin{cases} L_N n_1^{-\frac{\beta(1+\alpha)}{2\beta\gamma+d}} & \text{if } \varepsilon \le \widetilde{n}^{-\beta} \\ \left[L_N \left\{ \left(n_0^{\frac{1}{2\beta+d}} \wedge (n_0^2\varepsilon^2)^{\frac{1}{2\beta+2d}} \right) + \left(n_2^{\frac{1}{2\beta\gamma+d}} \wedge (n_2^2\varepsilon^2)^{\frac{1}{2\beta\gamma+2d}} \right) \right\}^{-\beta(1+\alpha)} \wedge 1 \right] & \text{otherwise,} \end{cases}$$

where $\widetilde{n} = n_0^{\frac{1}{2\beta+d}} \wedge n_2^{\frac{\gamma}{2\beta\gamma+d}}$. *Here* L_N *is a sequence of order at most* $(\log((\delta_0 \wedge \delta_2)^{-1}))^{\frac{\beta(1+\alpha)}{2\beta(\gamma\wedge 1)+d}}$.

The above corollary demonstrates three different rates based on n_1, the size of the public source server, with respect to the private source and the target. When n_1 is larger than n_2, the public source has a significantly large number of samples, enough for the non-private rates from this server to dominate over the rest. Next, when n_1 is relatively large with respect to the target sample size, but still smaller than the private source server, we find three different rates. In the high privacy regime of very small ε, the availability of public data ensures that we do not have to pay an additional price of private rates by adding noise. Next, the private rates of the source appear: the reason being that the private source has higher number of samples, and in this intermediate privacy regime, even after adding additional noise, the private source turns out to be more useful than the public server. For even lower privacy, the extraneous noise level diminishes further and the non-private rate from source server 2 appears. The final case is when the public data size n_1 is relatively smaller than the sample size in the two privacy constrained servers. In the regime of very high privacy, the public data is still effective, and enables one to avoid the private rates. For moderate and low privacy, the tradeoff is characterized by the relative signal strengths and privacy requirements of the private source and target servers. The exact rates would be similar to Table 1 in this scenario.

3.2. Minimax Rates in General Setting

We now turn our attention to the general case where the sample sizes n_j, transfer exponents γ_j, privacy parameters $(\varepsilon_j, \delta_j)$ are all allowed to vary for $0 \le j \le m$. Our main result, captured in Theorem 3.4, quantifies the rate. The homogeneous case described earlier can be thought of as a special case of this vastly more general setting.

THEOREM 3.4. *Let $r \in \mathbb{R}_+$ be the solution to the following equation:*

$$(n_0 \wedge n_0^2\varepsilon_0^2 r^d) r^{2\beta+d} + \sum_{j=1}^{m} (n_j \wedge n_j^2\varepsilon_j^2 r^d) r^{2\beta\gamma_j+d} = 1 \tag{5}$$

The minimax rate for excess risk is given by

$$\inf_{\widehat{f}\in\mathcal{M}(\boldsymbol{\varepsilon},\boldsymbol{\delta})} \sup_{(P_0,\ldots,P_m)\in\Pi(\alpha,\beta,\boldsymbol{\gamma},\mu)} \mathcal{E}_0(\widehat{f}) \asymp \left(L_N r^{\beta(1+\alpha)} \wedge 1\right). \tag{6}$$

whenever $\sum_j n_j\delta_j \to 0$, for a sequence L_N of order at most $(-\log(\delta_{\min}))^{\frac{\beta(1+\alpha)}{2\beta\gamma_{\min}+d}}$.

It is important to note that (5) always yields a positive solution, since the left-hand side of (5) is a strictly increasing continuous function of r, ranging from 0 to ∞ as r varies from 0 to ∞. Thus the function of r defined on the left side of (5) must take the value 1 somewhere in $\mathbb{R}_+$. We now provide a brief commentary on the derived result, starting with a comparison with non-private transfer learning.

REMARK 3.3 (GENERAL NON-PRIVATE TRANSFER LEARNING). When the privacy budget is large, setting $\varepsilon_j = \infty$ for $j = 1, \ldots, m$ in (5) we recover the non-private rate

$$\inf_{\widehat{f}\in\mathcal{M}(\boldsymbol{\varepsilon},\boldsymbol{\delta})} \sup_{(P_0,\ldots,P_m)\in\Pi(\alpha,\beta,\boldsymbol{\gamma},\mu)} \mathcal{E}_0(\widehat{f}) \asymp \left(n_0 + \sum_{j=1}^{m} n_j^{\frac{2\beta+d}{2\beta\gamma_j+d}}\right)^{-\frac{\beta(1+\alpha)}{2\beta+d}}$$

when m is fixed. This coincides with Theorems 5 and 6 of Cai and Wei [2021]. Note however that Theorem 3.4 allows the number of servers, m, to grow to ∞.

In the most general case, the minimax optimal rate of convergence exhibited in the transfer learning problem under distributed privacy is determined by r, which has implicit dependencies on various factors, including the number of servers, privacy parameters, transfer exponents, and sample sizes. The value of r determines the radius within which local methods can share information. In Section 4.1 we find that when using kernel estimators, the optimal bandwidth choice is innately connected to r, and in fact differs from the solution to (5) by at most a logarithmic factor. To further interpret the role of r, note that (5) quantifies the exact contribution of each server to the entire classification procedure. For each $j \in \{0, \ldots, m\}$, the variance of the local estimator for the j^{th} server is determined by the sum of two quantities: the inverse of the sample size, and the variance of the additional noise required for privacy. The term $n_j \wedge n_j^2\varepsilon_j^2 r^d$ then appears as a quantity proportional to the inverse of this variance and determines the *precision* of the j^{th} server.

4. Minimax Optimal Classification Procedure

In this section, we propose an optimal classifier and establish the minimax rate of convergence. In the first subsection, we develop a nonparametric classifier for the target population that appropriately utilizes information from the sources while satisfying privacy requirements for each server. In the second subsection, we prove a minimax lower bound, demonstrating that our classifier achieves the minimax rate of convergence in the distributed private transfer learning context.

4.1. Classifier

We now describe a classifier for transfer learning with distributed privacy. Our method has three main steps. First, we use a kernel estimator to estimate $(\eta_j(x) - \frac{1}{2})g(x)$ for $j = 0, 1, \ldots, m$. Second, then use a convex combination of these estimators, where the weights are designed to borrow strength from the source servers under the transfer learning setup. The third step is to add a Gaussian noise to the weighted kernel estimator to satisfy privacy requirements. Our classifier is given by the sign of the noise perturbed weighted estimator.

Consider a kernel $K(t)$ supported on $[-1, 1]^d$ with the following properties:

(a) $\int K(t)dt = 1$.

(b) $K(\cdot)$ is L_K-Lipschitz.

(c) $\max_{t\in[-1,1]^d} K(t) \le c_K$.

(d) $\min\{K(t) : \|t\| \le 1/2\} \ge b_K$.

(e) K is positive definite.

Let $X = x_0$ be the test point we wish to classify. Suppose first that $x_0 \in [0, 1]^d$. Then for the source samples we compute

$$T_h^{(j)}(x_0) := \frac{1}{n_j h^d} \sum_{i=1}^{n_j} \left(Y_i^{(j)} - \frac{1}{2}\right) K\left(\frac{X_i^{(j)} - x_0}{h}\right) \tag{7}$$

for $0 \le j \le m$. To satisfy privacy requirements we follow the framework of Hall et al. [2013] and a Gaussian process to the above estimator. Note that $T_h^{(j)}$ belong to the RKHS $\mathcal{K}$ given by linear combination $\{\sum_i \theta_i K((X_i - x)/h) : \theta_i \in \mathbb{R}\}$. For two functions $f = \sum_i \theta_i K((X_i - x)/h)$ and $g = \sum_i \tau_i K((X_i - x)/h)$, their inner product under $\mathcal{K}$ is given by

$$\langle f, g\rangle_{\mathcal{K}} = \sum_i \sum_j \theta_i \tau_j K\left(\frac{X_i - X_j}{h}\right).$$

Let $T_h^{(j)\prime}$ be the versions of $T_h^{(j)}$ with $(X_1^{(j)}, Y_1^{(j)})$ replaced by $(X_1^{(j)\prime}, Y_1^{(j)\prime})$ for $j = 0, 1, \ldots, m$. Then the RKHS norm of $T_h^{(j)}(\cdot) - T_h^{(j)\prime}(\cdot)$ can be bounded by

$$\|T_h^{(j)} - T_h^{(j)\prime}\|_{\mathcal{K}} \le \frac{\sqrt{c_K}}{n_{P_j} h^d} \quad \text{and} \quad \|T_h^{(0)} - T_h^{(0)\prime}\|_{\mathcal{K}} \le \frac{\sqrt{c_K}}{n_Q h^d}. \tag{8}$$

Let us define $(m + 1)$ independent mean zero Gaussian processes $\xi^{(j)}(\cdot)$ with covariance kernels

$$\mathrm{Cov}(\xi^j(s), \xi^{(j)}(t)) = K\left(\frac{s - t}{h}\right) \text{ for } s, t \in [0, 1].$$

and

$$\widetilde{\xi}_h(\cdot) = \frac{\sqrt{2c_K \log(2/\delta_j)}}{n_j \varepsilon_j h^d} \xi^{(j)}(\cdot) \text{ for } j = 0, 1, \ldots, m. \tag{9}$$

We then release

$$\left\{T_h^{(j)}(x_0) + \widetilde{\xi}_h^{(j)}(x_0)\right\}_{j=0}^{m}. \tag{10}$$

The next proposition asserts that the transcripts described above matches the required distributed privacy requirements, and its proof follows by results from Hall et al. [2013].

PROPOSITION 4.1. *For any $h \in [0,1]$ the transcripts $\{T_h^{(j)}(x_0) + \widetilde{\xi}_h^{(j)}(x_0) : 0 \le j \le m\}$ described above satisfies $(\varepsilon_j, \delta_j)$ differential privacy distributed across servers $j \in \{0, 1 \ldots, m\}$.*

The optimal bandwidth choice h_{opt} is given by the solution to (5). To account for the additional δ factor for approximate privacy, we now define $h_{opt,\delta}$ which is the solution to:

$$(n_0 \wedge n_0^2\varepsilon_0^2 r^d) r^{2\beta+d} + \sum_{j=1}^{m} (n_j \wedge n_j^2\varepsilon_j^2 r^d) r^{2\beta\gamma_j + d} = \log\left(\frac{2}{\delta_{\min}}\right) \tag{11}$$

Let us now define the weights

$$\begin{aligned} v_j &= (n_j \wedge n_j^2\varepsilon_j^2 h_{opt,\delta}^d) h_{opt,\delta}^{\gamma_j \beta} \quad \text{for} \quad j = 0, \ldots, m \\ u_j &= \frac{v_j}{\sum_{j=0}^{m} v_P} \quad \text{for} \quad j = 0, \ldots, m. \end{aligned} \tag{12}$$

With these weights, the target server computes a weighted average of $\{T_h^{(j)}(x_0) + \widetilde{\xi}_h^{(j)}(x_0) : 0 \le j \le m\}$ as follows:

$$\widetilde{T}_h(x_0) := u_0 \left(T_h^{(0)}(x_0) + \widetilde{\xi}^{(0)}(x_0)\right) + \sum_{j=1}^{m} u_j \left(T_h^{(j)}(x_0) + \widetilde{\xi}^{(j)}(x_0)\right). \tag{13}$$

Finally our classifier is given by

$$\widehat{f}(x_0) := \mathbb{1}(\widetilde{T}_{h_{opt,\delta}}(x_0) \ge 0) \tag{14}$$

where $h_{opt,\delta}$ is the solution to (11). The following theorem provides an upper bound for the excess risk of this classifier.

THEOREM 4.2. *Let r be the solution to (5). Let $\widehat{f}$ be the classifier defined in (14) based on the weighted kernel estimator (13). Then,*

$$\sup_{(P_0,\ldots,P_m) \in \Pi(\alpha,\beta,\boldsymbol{\gamma},\mu)} \mathcal{E}_0(\widehat{f}) \le C_* r^{\beta(1+\alpha)} \left(\log\left(\frac{1}{\delta_{\min}}\right)\right)^{\frac{\beta(1+\alpha)}{2\beta\gamma_{\min}+d}}$$

where C_* is a constant depending on $L, d, \alpha, \beta, \gamma_j$, while $\gamma_{\min} = \min\{1, \gamma_1, \ldots, \gamma_m\}$ and $\delta_{\min} = \min\{\delta_0, \ldots, \delta_m\}$.

4.2. Minimax Lower Bounds

The above theorem bounds the error rate of our kernel based classifier. Alongside the upper bound above, in this subsection we derive the minimax lower bound on the excess risk, to establish that our kernel based classifier is minimax optimal up to logarithmic factors.

We introduce a general data processing inequality which extends the findings presented in Cai et al. [2023]. This new result provides a bound on the total variation (TV) distance between the push forward measures of the transcripts $\mathbb{P}_{\sigma}^{\widetilde{T}}$ and $\mathbb{P}_{\sigma'}^{\widetilde{T}}$, utilizing the TV distance between their underlying distributions. Such an inequality may be of independent interest beyond the current setting due to its broader applicability.

LEMMA 1. *For any subset $\mathcal{S} \subseteq \{0, \ldots, m\}$, the TV distance is bounded as follows:*

$$\mathrm{TV}\left(\mathbb{P}_{\sigma}^{\widetilde{T}}, \mathbb{P}_{\sigma'}^{\widetilde{T}}\right) \leq \sqrt{2}\sqrt{\sum_{j\in S} \bar{\varepsilon}_j \left(e^{\bar{\varepsilon}_j} - 1\right) + \sum_{j\in S^c} n_j \mathrm{KL}(P_{j,\sigma}, P_{j,\sigma'})} + 4\sum_{j\in S} e^{\bar{\varepsilon}_j} n_j \delta_j \rho_j, \tag{15}$$

where $\bar{\varepsilon}_j = 6n_j\varepsilon_j\rho_j$ and $\rho_j = \mathrm{TV}(P_{j,\sigma}, P_{j,\sigma'})$.

A notable aspect of this lemma, in contrast to previous results in Cai et al. [2023], is the inclusion of interactions between the target and source servers. Consequently, all the information about the transcripts from the source servers is encapsulated within the final transcript $\widetilde{T}$ computed by the target server. This approach allows for heterogeneity not only in sample sizes and privacy budgets but also in the data distributions across different servers.

In constructing the lower bound, we formulate a family of distributions $P_{j,\sigma}$ for each $j = 1, \ldots, m$, where σ is a vector of $\{+1, -1\}^M$ and M is suitably chosen. We apply Assouad's Lemma to bound the total variation distance between the push forward measures of the transcripts $\mathbb{P}_{\sigma}^{\widetilde{T}}$ and $\mathbb{P}_{\sigma'}^{\widetilde{T}}$ for σ and σ' differing in one entry, as stipulated by Lemma 1. The rest of the construction of the lower bound crucially depends on the selection of the set $\mathcal{S}$. This choice corresponds to the index of servers where the privacy cost significantly outweighs the non-private risk. The major contributions to the upper bound for the TV distance are $\bar{\varepsilon}_j \left(e^{\bar{\varepsilon}_j} - 1\right)$ from the privacy-stringent servers, and the KL divergence for the samples on non-stringent servers, $n_j \mathrm{KL}(P_{j,\sigma}, P_{j,\sigma'})$. The optimal balance of contributions is achieved when $\mathcal{S} = \{j : \varepsilon_j \leq (r^d n_j)^{-1/2}\}$.

The following theorem establishes the fundamental cost of privacy for the nonparametric classification problem in the distributed privacy setting.

THEOREM 4.3. *Suppose δ_j's are such that $\sum_j n_j\delta_j = o(1)$, then there exists a $c > 0$ not depending on n_j for $j = 0, \ldots, m$ such that*

$$\inf_{\widehat{f}\in\mathcal{M}(\boldsymbol{\varepsilon},\boldsymbol{\delta})} \sup_{(P_0,\ldots,P_m)\in\Pi} \mathcal{E}_0(\widehat{f}) \geq c r^{\beta(1+\alpha)}$$

where r is the solution to (5).

A comparison with the upper bound on the excess risk obtained from Theorem 4.2 now establishes the rate optimality, up to a logarithmic factor, of our kernel-based differentially private distributed classifier.

5. Data-driven Adaptive Classifier

In practice, the smoothness and transfer exponent parameters are unknown, making it challenging to select the correct bandwidth h. To address this, we will use an estimator based on the Lepski method to choose h from a grid of possible values. While choosing the exact optimal h is infeasible, this method adapts to the unknown parameters β and $\gamma_1, \dots, \gamma_m$.

This section is divided into two subsections. First, we consider the transfer homogeneous case, where the source populations have the same relative transfer exponent $\gamma_j = \gamma$ for $j = 1, \dots, m$ with respect to the target. Here we allow m to grow at an appropriately slow rate as n increases. In the second subsection, we address the general, heterogeneous case with transfer exponents $(\gamma_1, \dots, \gamma_m)$ with the important restriction that the number of sources m is finite.

To choose the best candidate bandwidth, we define a grid of possible choices for h as:

$$\mathcal{H} = \{2^{-j} : j = 0, 1, \dots, (\log n_*)/d\}, \text{ where } n_* = \sum_{j=0}^{m} n_j \wedge n_j^2 \varepsilon_j^2.$$

Let $\Delta^m = \{w : w_i \in [0,1], \sum_{i=0}^m w_i = 1\}$ denote the m-dimensional simplex. For a weight vector $w = (w_0, w_1, \dots, w_m) \in \Delta^m$ we define

$$\widetilde{T}(x_0, h, w) := \sum_{j=0}^{m} w_j T_h^{(j)}(x_0) + \sum_{j=0}^{m} w_j \frac{\sqrt{2c_K \log(2|\mathcal{H}|/\delta_j)}|\mathcal{H}|}{n_j \varepsilon_j h^d} \xi^{(j)}(x_0) \text{ for } h, w \in [0,1] \quad (16)$$

where $T_h^{(j)}(\cdot)$ is as defined in (7) and $\xi^{(j)}(\cdot)$ are independent mean zero Gaussian processes with covariance kernel $K(\cdot/h)$.

5.1. Adaptation under Source Homogeneity

In this subsection we consider the sources to have transfer homogeneity, i.e., every source has identical transfer exponent $\gamma_j = \gamma$ for $j = 1, \dots, m$. It is then intuitive to weigh the estimators $T_h^{(j)}$ and the noise $\xi^{(j)}$ with the weights proportional to $n_j \wedge n_j^2 \varepsilon_j^2 h^d$ for the sources. More specifically, this gives the restricted set of weights:

$$\mathcal{W}(h) := \left\{ (w_0, w_1, \dots, w_m) : \; w_0 \in [0,1], \, w_j = \frac{(1-w_0)u_j(h)}{\sum_{j=1}^m u_j(h)} \text{ for } j = 1, \dots, m \right\} \quad (17)$$

where $u_j(h) := n_j \wedge n_j^2 (\varepsilon_j/|\mathcal{H}|)^2 h^d$. Note that the privacy requirements dictate that $\mathcal{W}$ depends on h. The rationale behind this set of weights comes from (12) restricted to the transfer homogeneous setting. We determine the deviation of $\widetilde{T}(x_0, h, w)$ around its expectation through

$$\begin{aligned} v_0(h, w) = w_0^2 &\left(\frac{c_K g_{\max}}{3 n_0 h^d} + \frac{2c_K^2 \log(2|\mathcal{H}|/\delta_0)}{n_0^2 (\varepsilon_0/|\mathcal{H}|)^2 h^{2d}} \right) \\ &+ (1-w_0)^2 \left(\sum_{j=1}^{m} \frac{(u_j(h))^2}{(\sum_{j=1}^m u_j(h))^2} \left(\frac{c_K g_{\max}}{3 n_j h^d} + \frac{2c_K^2 \log(2|\mathcal{H}|/\delta_j)}{n_j^2 (\varepsilon_j/|\mathcal{H}|)^2 h^{2d}} \right) \right). \end{aligned} \quad (18)$$

The above metric measures the scale of noise present in $\widetilde{T}(x_0, h, w)$. We therefore define the signal to noise ratio as

$$\widehat{\rho}_0(h) = \max_{w \in \mathcal{W}(h)} \frac{(\widetilde{T}(x_0, h, w))^2}{v_0(h, w)}. \tag{19}$$

Then the adaptive choice of h in the transfer homogeneous setting is given by

$$h_0 = \begin{cases} \min\{h \in \mathcal{H} : \widehat{\rho}_0(h) > 4.5\log(2n_*|\mathcal{H}|)\} & \text{if } \max_{h \in \mathcal{H}} \widehat{\rho}_0(h) > 4.5\log(2n_*|\mathcal{H}|) \\ \text{argmax}_h \widehat{\rho}_0(h) & \text{otherwise.} \end{cases}$$

Define

$$w_{*(0)} = \text{argmax}_{w \in \mathcal{W}(h_0)} \frac{(\widetilde{T}(x_0, h_0, w))^2}{v_0(h_0, w)}.$$

The adaptive classifier is now defined as

$$\widehat{f}_0(x) := \mathbb{1}(\widetilde{T}(x, h_0, w_{*(0)}) > 0). \tag{20}$$

Algorithm 1 summarizes the steps of the adaptive procedure.

Algorithm 1 Data-adaptive mechanism for bandwidth selection under Source Homogeneity

1: **Input:** test point: $x_0 \in [0,1]^d$; data: $\{(X_i^{(j)}, Y_i^{(j)}) : 1 \le i \le n_j, j \in \{0, 1, \ldots, m\}\}$; privacy parameters: $\{(\varepsilon_j, \delta_j) : j \in \{0, 1, \ldots, m\}\}$; kernel: $K(\cdot)$.
2: Compute the grid of bandwidths: $\mathcal{H} = \{2^{-j} : j = 0, 1, \ldots, \lfloor \log(n_*)/d \rfloor\}$.
3: **for** h in $\mathcal{H}$ **do**
4: Compute the set of weights $\mathcal{W}(h)$ following (17).
5: Compute the signal-to-noise ratio:

$$\widehat{\rho}_0(h) = \max_{w \in \mathcal{W}(h)} \frac{(\widetilde{T}(x_0, h, w))^2}{v_0(h, w)}$$

where $\widetilde{T}(x_0, h, w)$ and $v_0(h, w)$ are as defined in (16) and (18) respectively.
6: Choose the bandwidth as

$$h_0 = \begin{cases} \min\{h \in \mathcal{H} : \widehat{\rho}_0(h) > 4.5\log(2n_*|\mathcal{H}|)\} & \text{if } \max_{h \in \mathcal{H}} \widehat{\rho}_0(h) > 4.5\log(2n_*|\mathcal{H}|) \\ \text{argmax}_{h \in \mathcal{H}} \widehat{\rho}_0(h) & \text{otherwise.} \end{cases}$$

7: Choose the best weight as

$$w_{*(0)} = \text{argmax}_{w \in \mathcal{W}(h_0)} \frac{(\widetilde{T}(x_0, h_0, w))^2}{v_0(h_0, w)}.$$

8: **Output:** $\widehat{f}_0(x) := \mathbb{1}(\widetilde{T}(x, h_0, w_0) > 0)$.

The following theorem states the excess risk of the adaptive classifier in terms of the regression function parameter α, β, the transfer exponent γ and the privacy constraints.

THEOREM 5.1. *Let r be the solution to (5) with $\gamma_j = \gamma$ for $j = 1, \dots, m$. Let $\widehat{f}_0$ be the data adaptive classifier defined in (20). Then,*

$$\sup_{(P_0,\dots,P_m)\in\Pi(\alpha,\beta,\gamma,\mu)} \mathcal{E}_0(\widehat{f}_0) \leq C'_* r^{\beta(1+\alpha)} \left[\left(\log(n_*|\mathcal{H}|) \log\left(2|\mathcal{H}|/\delta_{\min}\right)\right)^{\frac{\beta(1+\alpha)}{2\beta(1\wedge\gamma)+d}} \vee |\mathcal{H}|^{\frac{2\beta(1+\alpha)}{d}} \right]$$

where C'_* is a constant depending on $m, L, d, \alpha, \beta, \gamma$, while $\delta_{\min} = \min\{\delta_0, \dots, \delta_m\}$.

It is instructive to compare the above rate with the rate from Theorem 4.2. Since $|\mathcal{H}| = \text{polylog}(n_*)$ it follows that the excess risk of the adaptive estimator is worse by a multiplicative factor of $\text{polylog}(n_*)$). This is due to two reasons. Firstly, a factor of $\log n_*$ is expected as a cost of adaptation, and can be found in the transfer learning classification setup considered in Cai and Wei [2021]. Secondly, to ensure that the $(\varepsilon_j, \delta_j)$ privacy requirements are satisfied throughout the adaptation procedure, we require that for every $h \in \mathcal{H}$ our estimators are $(\varepsilon_j/|\mathcal{H}|, \delta_j/|\mathcal{H}|)$ differentially private. This contributes an extra factor of $|\mathcal{H}|$ to the rate obtained in Theorem 5.1.

An important special case of the above theorem is the server homogeneous case, where sample sizes and the privacy parameters are the same for every server, i.e., $n_j = n$, $\varepsilon_j = \varepsilon$ and $\delta_j = \delta$ for $j = 0, 1, \dots, m$. The following corollary describes this special case.

COROLLARY 5.2. *Let r be the solution to (5) with $n_j = n, \varepsilon_j = \varepsilon, \delta_j = \delta$ and $\gamma_j = \gamma$ for all $j = 1, \dots, m$. Let $\widehat{f}_0$ be the data adaptive classifier defined in (20). Then,*

$$\sup_{(P_0,\dots,P_m)\in\Pi(\alpha,\beta,\gamma,\mu)} \mathcal{E}_0(\widehat{f}_0) \leq C'_* \Bigg[L_N^{(ada)} \Bigg\{ \left(n_0^{\frac{1}{2\beta+d}} \wedge (n_0^2\varepsilon_0^2)^{\frac{1}{2\beta+2d}} \right) + \left((mn)^{\frac{1}{2\beta\gamma+d}} \wedge (mn^2\varepsilon^2)^{\frac{1}{2\beta\gamma+2d}} \right) \Bigg\}^{-\beta(1+\alpha)} \wedge 1 \Bigg]$$

where $L_N^{(ada)}$ is given by $\left[\left(\log(n_*|\mathcal{H}|) \log\left(2|\mathcal{H}|/\delta\right)\right)^{\frac{\beta(1+\alpha)}{2\beta(1\wedge\gamma)+d}} \vee |\mathcal{H}|^{\frac{2\beta(1+\alpha)}{d}} \right]$, and C'_* is a constant depending on $m, L, d, \alpha, \beta, \gamma$.

A comparison with Theorem 3.1 shows that this rate is minimax optimal up to a factor polynomial in logarithmic terms.

5.2. General Adaptation for Multiple Sources

We now shift to the general setting where we no longer constrain $\gamma_1, \dots, \gamma_m$ to be all equal. Note that for optimal estimation (as in Theorem 4.2), one requires knowledge of potentially m many different parameters $\gamma_1, \dots, \gamma_m$. The adaptation procedure therefore requires optimizing over all possible weights in $w \in \Delta^m$. When m increases with n, the adaptation to this growing number of parameters necessarily worsens the rate of decay for the excess risk. We will not delve further into issue and focus instead on the case where m is finite and does not increase with n.

In this general case we must consider all possible weight vectors $w \in \Delta^m$. As in (18) earlier we compute an approximate variance of $\widetilde{T}(x_0, h, w)$ as

$$v(h, w) = \sum_{j=0}^{m} w_j^2 \left(\frac{c_K g_{\max}}{3n_j h^d} + \frac{2c_K^2 \log(2|\mathcal{H}|/\delta_j)}{n_j^2(\varepsilon_j/|\mathcal{H}|)^2 h^{2d}} \right). \tag{21}$$

Let us define the signal-to-noise ratio index $\widehat{\rho}(h)$:

$$\widehat{\rho}(h) = \max_{w \in \Delta^m} \frac{(\widetilde{T}(x_0, h, w))^2}{v(h, w)},$$

In this general setting, the adaptive choice of h is given by

$$h_* = \begin{cases} \min\{h \in \mathcal{H} : \widehat{r}(h) > C_* \log(n_* |\mathcal{H}|(m+1))\} & \text{if } \max_{h \in \mathcal{H}} \widehat{\rho}(h) > C_* \log(n_* |\mathcal{H}|(m+1)) \\ \text{argmax}_h \ \widehat{\rho}(h) & \text{otherwise.} \end{cases}$$

where $C_* = 2.25(m+1)$. Defining

$$w_* = \text{argmax}_{w \in \Delta^m} \frac{(T(x_0, h_*, w))^2}{v(h_*, w)}.$$

we obtain the adaptive classifier

$$\widehat{f}_a(x) := \mathbb{1}(T(x, h_*, w_*) > 0). \tag{22}$$

Similar to Algorithm 1 we have Algorithm 2 which presents the adaptive procedure allowing for possible heterogeneity among servers.

Algorithm 2 Data-adaptive mechanism for bandwidth selection in the general case

1: **Input:** test point: $x_0 \in [0,1]^d$; data: $\{(X_i^{(j)}, Y_i^{(j)}) : 1 \le i \le n_j, j \in \{0, 1, \ldots, m\}\}$; privacy parameters: $\{(\varepsilon_j, \delta_j) : j \in \{0, 1, \ldots, m\}\}$; kernel: $K(\cdot)$.
2: Compute the grid of bandwidths: $\mathcal{H} = \{2^{-j} : j = 0, 1, \ldots, \lfloor \log(n_*)/d \rfloor\}$.
3: **for** h in $\mathcal{H}$ **do**
4: Compute the signal-to-noise ratio:

$$\widehat{\rho}(h) = \max_{w \in \Delta^m} \frac{(\widetilde{T}(x_0, h, w))^2}{v(h, w)}$$

where $\widetilde{T}(x_0, h, w)$ and $v(h, w)$ are as defined in (16) and (21) respectively.
5: For $C_* = 2.25(m+1)$ choose the bandwidth as

$$h_* = \begin{cases} \min\{h \in \mathcal{H} : \widehat{\rho}(h) > C_* \log(2n_* |\mathcal{H}|)\} & \text{if } \max_{h \in \mathcal{H}} \widehat{\rho}(h) > C_* \log(2n_* |\mathcal{H}|) \\ \text{argmax}_{h \in \mathcal{H}} \widehat{\rho}(h) & \text{otherwise.} \end{cases}$$

6: Choose the best weight as

$$w_* = \text{argmax}_{w \in \Delta^m} \frac{(\widetilde{T}(x_0, h_0, w))^2}{v(h_0, w)}.$$

7: **Output:** $\widehat{f}_a(x) := \mathbb{1}(\widetilde{T}(x, h_*, w_*) > 0)$.

The following theorem verifies the efficacy of the general adaptive procedure.

THEOREM 5.3. *Let r be the solution to (5). Let $\widehat{f}_a$ be the data adaptive classifier defined in (22). Then,*

$$\sup_{(P_0,\ldots,P_m)\in\Pi(\alpha,\beta,\boldsymbol{\gamma},\mu)} \mathcal{E}_0(\widehat{f}) \leq C'_* r^{\beta(1+\alpha)} \left[(\log(n_*|\mathcal{H}|) \log(2|\mathcal{H}|/\delta_{\min}))^{\frac{\beta(1+\alpha)}{2\beta\gamma_{\min}+d}} \vee |\mathcal{H}|^{\frac{2\beta(1+\alpha)}{d}} \right]$$

where C'_* is a constant depending on $m, L, d, \alpha, \beta, \gamma_j$, while $\gamma_{\min} = \min\{1, \gamma_1, \ldots, \gamma_m\}$ and $\delta_{\min} = \min\{\delta_0, \ldots, \delta_m\}$.

REMARK 5.1. Oftentimes users might find it helpful to restrict the set of weights based on prior knowledge. For example, under source homogeneity we allowed the restricted set $\mathcal{W}(h)$. Similarly it is possible to use ad-hoc choices of weights w, based for example, on the sample proportions for each server. Our algorithm 2 automatically allows these specific choices of weights and can be used to select the bandwidth h to be used with the classifiers weighted with respect to w.

6. Numerical Studies

The data-driven classifier proposed in this paper is easy to implement. In this section, we examine the numerical performance of our methods through various simulated and real data experiments. In the first subsection, we conduct simulation studies to compare our proposed classifier with alternative methods across different parameter settings and varying levels of problem difficulty. In the second subsection, we evaluate the performance of the proposed classifier on a real dataset, verifying its practical effectiveness.

6.1. Simulation Study

We compare the prediction accuracy of our classifier with other existing methods in the literature across various parameter settings.

6.1.1. Simulation Design

In our simulation study, we consider a setup with m source servers and a single target server. All source servers are assumed to have the same data quality. Additionally, for most of the simulation setups we assume that both the source and target servers have an equal number of observations, denoted by n. The privacy budget for each server is given by (ε, δ), where we set $\delta = n^{-2}$.

The data for both target and source servers is generated as follows: the marginal distribution for $\boldsymbol{X}$ on both the target and source servers is supported on a two-dimensional cube, $[0,1]^2$, with a uniform distribution over its support.

The conditional distribution for the target server, given $\boldsymbol{X}$, is defined as:

$$P_T(Y=1|\boldsymbol{X}) = \eta_T(\boldsymbol{X}) = 1 \wedge \left(\frac{1}{2} + \text{sign}\left(\left(x_1 - \frac{1}{2}\right)\left(x_2 - \frac{1}{2}\right) \right) \left| x_1 - \frac{1}{2} \right|^{\frac{1}{4}} \left| x_2 - \frac{1}{2} \right|^{\frac{1}{4}} \right)_+$$

where for $a \in \mathbb{R}$ we write $a_+ = \max\{a, 0\}$. For the source servers, the conditional distribution is defined as:

$$P_S(Y = 1|\boldsymbol{X}) = 1 \wedge \left(\frac{1}{2} + \text{sign}\left(\eta_T(\boldsymbol{X}) - \frac{1}{2}\right)\left|\eta_T(\boldsymbol{X}) - \frac{1}{2}\right|^{\gamma}\right)_+ .$$

The proposed construction for these regression functions satisfies key assumptions, such as smoothness and margin conditions. For a visual representation of these regression functions, see Figure 2.

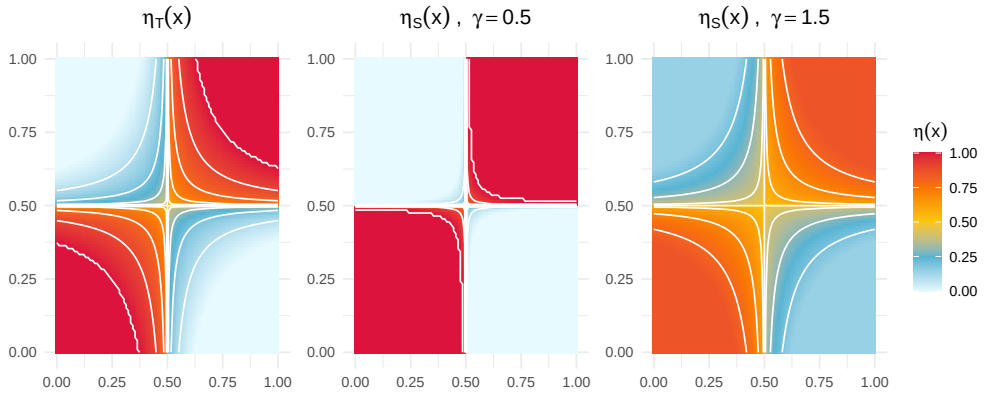

Fig. 2. Regression functions η_T and η_S for $\gamma \in \{0.5, 1.5\}$.

We would be comparing the following methods:

- **Distributed Transfer Learning with kernels (DTK)** is the proposed method with two different choices of kernel, the gaussian kernel and the triangular kernel. We select the best tuning parameters h the bandwidth and, $w \in [0, 1]$ the relative weight of the target data w.r.t the source data. The parameter grid for those are $w \in \left\{\frac{i}{100} : i \in \{0, \ldots, 100\}\right\}$ and $h \in \{2^{-i} : i \in \{1, \ldots, 7\}\}$.
- **Distributed Transfer Learning using histogram (DT-HIST)** is a variant of the method proposed in Berrett and Butucea [2019] adapted to the transfer learning setting with distributed privacy. This method also has the the two parameters w and h, where h denotes the histogram bin size for this estimator. We vary both the parameters on the same grid as **DTK**.
- **Adaptive Distributed Learning with kernels (AdaptDTK)** This is the adaptive version of the DTK classifier, proposed in Section 5, which automatically tunes the hyper-parameters h and w based on a Lepski-style method.

6.1.2. Effect of Source Data

In this section, we examine the impact of source data on classification performance. We compare our transfer learning approach to a method that relies solely on target data. The latter method, referred to as **targetDTK**, serves as our baseline for comparison.

To assess the utility of source data, we plot the best accuracy obtained from each method across a range of hyperparameters, as detailed in Section 6.1.1. This analysis provides insight into the benefits of incorporating source data in our classification process. In Figure 3, we contrast our transfer learning method with a naive approach that doesn't

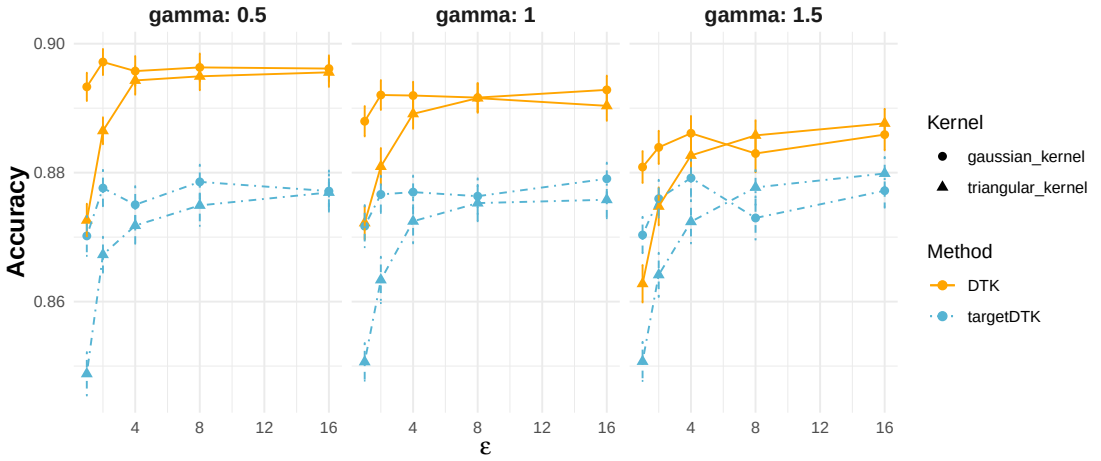

Fig. 3. Effect of source data against ε for $\gamma \in \{0.5, 1, 1.5\}$

leverage source data. To ensure fairness, we limit the scenario to just one additional source server ($m = 1$). Introducing multiple source servers would widen the performance gap between transfer learning methods and non-transfer learning methods. Also we set the number of observations on the source and target server to be $n = 100$, the reason for choosing a comparatively smaller n is for better visualization purposes, for larger n we expect the same phenomena , although the gain in accuracy would be difficult to perceive visually.

Within this setup, we vary two problem parameters, ε and γ. Across various kernel choices and values of ε and γ, our transfer learning method consistently outperforms the alternatives. As privacy constraints relax (i.e., ε increases), performance improves across all methods. Notably, as γ increases, the performance advantage gained from using source servers diminishes. This phenomenon is expected: with higher γ, the quality of source data decreases, consequently reducing the performance gain. Since it is always beneficial to use source data we do not compare the target only method in the further sections.

6.1.3. Comparison of Our Adaptive Classifiers with Other Methods

In this section we compare our adaptive classifier **AdaptDTK** to non-adaptive methods like **DTK** and **DT-HIST** across different ε, γ and m values. It is important to note that both non-adaptive methods are tuned with optimal hyperparameters based on test data, resulting in a potential performance advantage over the adaptive method.

However, our results show that the "cost of adaptation"—the difference in accuracy between the adaptive method and the non-adaptive methods—is relatively small. Despite

the adaptive method's lack of knowledge about oracle hyper-parameters, its performance is only marginally lower compared to methods with oracle tuning. This indicates that our adaptive approach is effective, even without precise hyperparameter tuning.

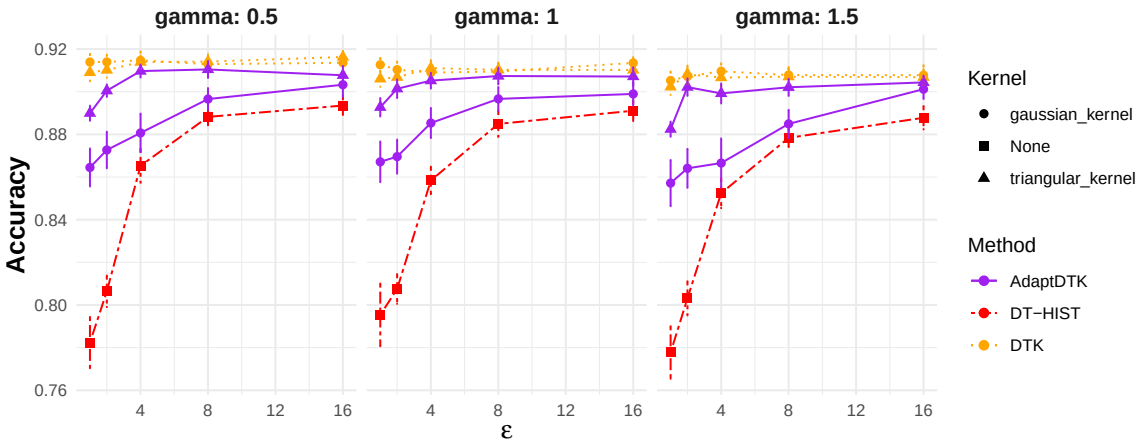

Fig. 4. Accuracy v/s ε for $\gamma \in \{0.5, 1, 1.5\}$

Effect of Privacy Budget: Here, we maintain the same experimental setup as described in Section 6.1.2 but with larger $n = 500$. Our kernel-based method, **AdaptDTK**, along with **DTK** (with oracle hyper-parameter tuning), consistently outperforms **DT-HIST**. Particularly in scenarios with low privacy budgets, the performance gap is significant, indicating that privatizing kernel-based methods empirically yields better results compared to privatizing histogram bin count-based methods.

Regarding the choice of kernel, we note that the triangular kernel demonstrates exceptional adaptability, showing only minimal performance degradation compared to its oracle-tuned counterpart. This difference in performance is likely attributed to the fact that our theoretical adaptive procedure was developed with bounded kernels (like triangular kernels) in mind.

Effect of distributed data: In our experimental setup, we maintain a constant total sample size of 500 across both target and source servers. By varying the number of source servers m from 1 to 20, we delve into the impact of data distribution under a fixed total sample size. Additionally, we explore how the source data quality, represented by γ, and the privacy budget, ε, influence the outcomes. Across all methods, a consistent pattern emerges: as data becomes more distributed, accuracy declines. This observation aligns with our theoretical findings. A notable trend is that performance degradation is more marked at higher γ values. This suggests a disproportionately negative impact on accuracy when the data is more fragmented, especially if the data quality is poorer.

The cost of adaptation also presents intriguing behavior. We note that the performance gap between adaptive and non-adaptive methods widens with increasing m, hinting

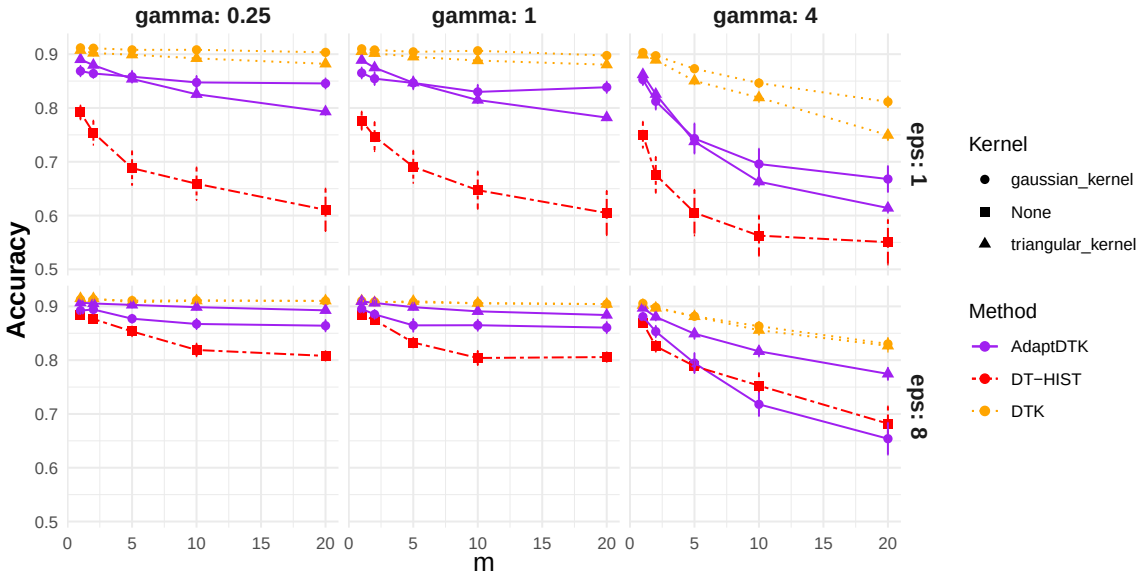

Fig. 5. Accuracy v/s m for $\gamma \in \{0.25, 1, 4\}$

that distributed data may amplify the challenges of adaptation. When it comes to kernel choice in the context of adaptation, distinct behaviors emerge depending on the specific scenario. For instance, in regimes of high privacy with low data distribution (e.g., $m = 1$), the triangular kernel outperforms its Gaussian counterpart in adaptation. Conversely, for larger m values, the Gaussian kernel seems more adept. In scenarios of lower privacy constraints, the triangular kernel consistently outperforms the Gaussian.

From a practical standpoint, the general recommendation is to opt for the Gaussian kernel only when dealing with extensively distributed data (large m) coupled with strict privacy requirements. In all other scenarios, the triangular kernel proves to be more effective.

6.2. Real Data

We now apply our private, distributed, nonparametric classifier to a real dataset to demonstrate its practical merits. We have selected the heart disease dataset from Detrano et al. [1989], publicly available on the UCI Machine Learning Repository. The primary task is to predict the prevalence of heart disease based on 13 covariates, including demographics (age, sex) and clinical measurements (blood pressure, resting heart rate, cholesterol, chest pain prevalence, etc.). This dataset comprises patient data from four hospitals in Cleveland, Hungary, Switzerland, and Long Beach. The distributed nature of the data, along with the presence of sensitive patient information, makes it an ideal example for applying our distributed differentially private classification algorithm.

In order to illustrate the heterogeneity across different servers, we evaluated the performance of our kernel-based classification algorithm in a non-private and non-transfer

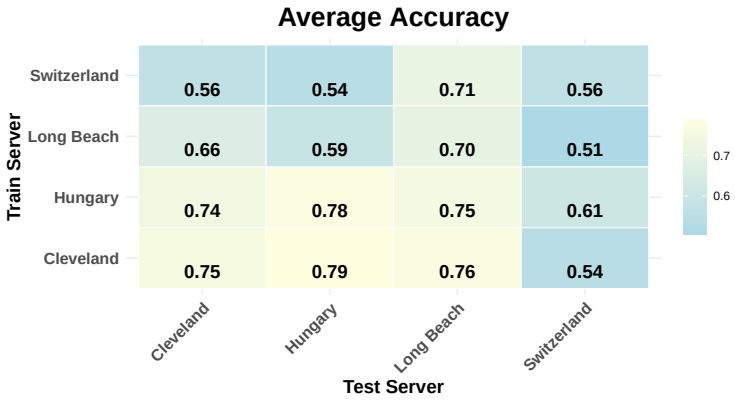

Fig. 6. Accuracy for different train-test server pairs

learning setting. Specifically, we selected 100 random samples from a designated training server and assessed the model's performance not only on a separate test set from the same server but also on data from three additional servers. Figure 6 visualizes these results, clearly showing that the same model exhibits varying levels of performance across different servers.

The analysis reveals significant variations in the classification model's performance, which highlight the inherent heterogeneity in the datasets. Notably, when trained on Cleveland data, the model achieves its highest accuracy on Hungarian data, suggesting some level of similarity between these datasets. In contrast, models trained on Swiss data exhibit markedly lower performance across all servers, including on local data, which points to potential challenges with the Swiss dataset's complexity or representativeness. Meanwhile, the data from Long Beach demonstrates consistent performance across diverse test servers, suggesting its features may possess a higher degree of generalizability. This variability underscores the need for adaptive modeling strategies in distributed systems to manage data heterogeneity and enhance predictive accuracy.

To address these challenges and potentially enhance model performance, we apply our transfer learning models to borrow strength from various sources. We will use the Hungarian hospital as the transfer target, and the rest would be used at source datasets.

6.2.1. Implementation Details

The choice of variables is critical for an illustration of our method, since it is nonparametric and hence extremely susceptible to the curse of dimensionality. Additionally, the data is plagued with missing observations for many of its columns. Among the 13 covariates available, we first picked the 9 predictors which had at least 90% non-missing observations across all the data. Among the rest, we removed `fbs` because it had 60% missingness for Switzerland data. We also removed the categorical variable `restecg` since its three levels lead to a higher dimensionality, while existing studies suggest its lower importance in predicting heart disease. Our chosen set of variables is therefore of size 7, and consists of `age`, `sex`, `cp` (chest pain type), `exang` (exercise induced angina, yes or no), `thalach` (max-

imum heart rate), `oldpeak` (ST depression induced by exercise), and `trestbps` (resting blood pressure). We further excluded the patients who had missing observations for these variables and set aside a test set of size 150 from our target. This resulted in $n_0 = 142$ patients for the target (Hungary), while the sources had $n_1 = 303$, $n_2 = 141$, and $n_3 = 116$ patients for the hospitals in Cleveland, Long Beach and Switzerland respectively.

In order to better place the problem in our theoretical setting, we scale each of the chosen covariates to range between 0 and 0.5. Another important distinction is the centering for our estimators. An initial exploratory analysis for our data reveals that the sources have different degrees of disease prevalence: 47%, 36%, 93%, and 79% for Cleveland, Hungary, Switzerland, and Long Beach respectively. To alleviate this issue, we re-weight the summands in our kernel estimator by source-specific mean of the disease status.

As the true transfer parameters are unknown, we opt for the data-adaptive procedures specified in Section 5. We will compare the following adaptation variations, each incorporating different amounts of auxiliary information:

(a) **AdaptAll**: We choose the best bandwidth and weights following Section 5.2.

(b) **AdaptTar**: The best bandwidth is chosen based solely on the target estimator. That is, the weights are zero for each source server.

(c) **AdaptSamp**: The weights are taken to be proportional to the sample sizes for each server.

(d) **AdaptHomog**: The adaptation procedure under source homogeneity (see Section 5.1) is used to choose the best bandwidth and weights.

We compare accuracy of the five different adaptation techniques against the common privacy parameter ε. We also compare the different adaptation procedures on another popular performance metric called the F1 score which is the harmonic mean of precision and recall. Figure 7 shows the average accuracy and F1 score obtained by replicating our procedure 200 times on different test and train folds.

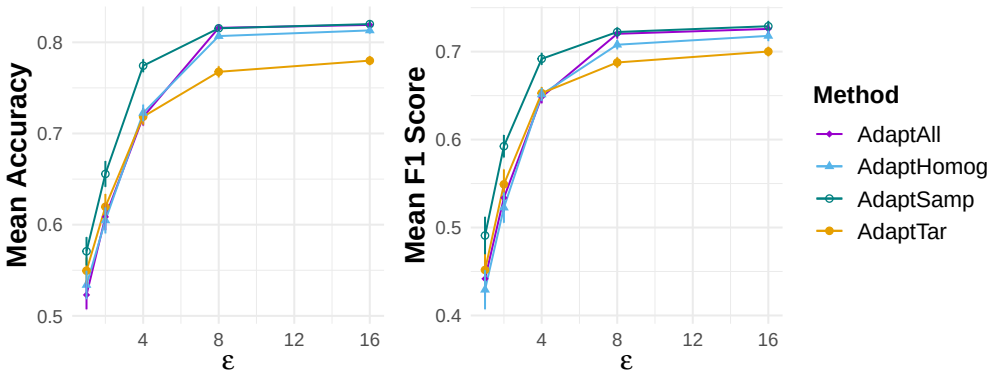

Fig. 7. Accuracy and F1 score values for different privacy levels

It is evident from Figure 7 that both the accuracy and the F1 score values improve for all the classifiers as the privacy requirements become less stringent. Among the five classifiers, **AdaptTar**, the one depending only on the target shows the poorest performance, thus highlighting the benefits of transfer learning in this problem. On the other hand, the sample size weighted adaptive classifier, denoted **AdaptSamp** is clearly the best among the five. The all-adaptive classifier, **AdaptAll**, automatically adjusts to the unknown weights and attains the same accuracy and F1 score values as **AdaptSamp** for higher values of ε. Finally the **AdaptHomog** is strictly worse than our estimator possibly indicating the inherent heterogeneity among the source servers. Overall, we find a prediction accuracy of 81% and an F1 score of 73% in predicting heart disease from the hospitals dataset, showing the effectiveness of our data-adaptive distributed private transfer learning classification mechanism.

7. Discussion

In this paper, we establish the minimax misclassification rate in a heterogeneous distributed setting with varying sample sizes, privacy parameters, and data distributions across servers under the posterior drift model. Our results precisely characterize the effects of privacy constraints, source sample sizes, and target sample size.

We rigorously quantify the trade-offs between data heterogeneity and variations in privacy budgets. By analyzing these trade-offs through minimax optimality, we clarify how differences in data distribution across servers impact the overall learning process and model performance. Our findings highlight the critical balance needed between maintaining privacy and effectively leveraging distributed data. The construction of the minimax and data-driven adaptive classifiers addresses excess risk and encapsulates the inherent trade-offs introduced by differential privacy. The impact of privacy constraints can be mitigated, to some extent, by leveraging larger datasets. The nuanced behavior of these classifiers, specifically how they scale with changes in the privacy constraints and sample sizes, offers a promising avenue for optimizing distributed learning systems. Our results are robust and theoretically justified within the defined heterogeneous settings and the posterior drift model.

Motivated by several practical requirements, it would be of significant interest to consider the transfer learning problem in other models with distributed differential privacy constraints. One natural direction is to consider distributed classification under the covariate shift model with differential privacy constraints. Such an analysis could provide additional insights and methods for a range of potential applications. Another future direction is to study distributed classification under parametric models such as logistic regression. We leave these intriguing problems for future research.

Funding: The research was supported in part by NIH grants R01-GM123056 and R01-GM129781.

References

John M Abowd. The challenge of scientific reproducibility and privacy protection for statistical agencies. *Census Scientific Advisory Committee*, pages 1011–1020, 2016.

Jayadev Acharya, Yuhan Liu, and Ziteng Sun. Discrete distribution estimation under user-level local differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pages 8561–8585. PMLR, 2023.

Arnab Auddy, T. Tony Cai, and Abhinav Chakraborty. Supplement to "Minimax and adaptive nonparametric classification for transfer learning under distributed differential privacy constraints". *Technical Report*, 2024.

Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608 – 633, 2007. doi: 10.1214/009053606000001217. URL https://doi.org/10.1214/009053606000001217.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

Thomas Berrett and Cristina Butucea. Classification under local differential privacy. *arXiv preprint arXiv:1912.04629*, 2019.

T. Tony Cai and Hongji Wei. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1):100 – 128, 2021. doi: 10.1214/20-AOS1949. URL https://doi.org/10.1214/20-AOS1949.

T. Tony Cai, Abhinav Chakraborty, and Lasse Vuursteen. Optimal federated learning for nonparametric regression with heterogeneous distributed differential privacy constraints. *Technical Report*, 2023.

P.K. Chan, W. Fan, A.L. Prodromidis, and S.J. Stolfo. Distributed data mining in credit card fraud detection. *IEEE Intelligent Systems and their Applications*, 14(6):67–74, 1999. doi: 10.1109/5254.809570.

Corinna Cortes, Mehryar Mohri, and Andrés Munoz Medina. Adaptation based on generalized discrepancy. *Journal of Machine Learning Research*, 20(1):1–30, 2019.

Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H Guppy, Stella Lee, and Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, 64(5):304–310, 1989.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

Jianqing Fan, Cheng Gao, and Jason M Klusowski. Robust transfer learning with unreliable source data. *arXiv preprint arXiv:2310.04606*, 2023.

Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton. A unified view of label shift estimation. *Advances in Neural Information Processing Systems*, 33:3290–3300, 2020.

Pascal Germain, Amaury Habrard, Francois Laviolette, and Emilie Morvant. A pac-bayesian approach for domain adaptation with specialization to linear classifiers. In *International conference on machine learning*, pages 738–746. PMLR, 2013.

Wei Guo, Fuzhen Zhuang, Xiao Zhang, Yiqi Tong, and Jin Dong. A comprehensive survey of federated transfer learning: Challenges, methods and applications. *arXiv preprint arXiv:2403.01387*, 2024.

Rob Hall, Alessandro Rinaldo, and Larry Wasserman. Differential privacy for functions and functional data. *The Journal of Machine Learning Research*, 14(1):703–727, 2013.

Ce Ju, Dashan Gao, Ravikiran Mane, Ben Tan, Yang Liu, and Cuntai Guan. Federated transfer learning for eeg signal classification. In *2020 42nd annual international conference of the IEEE engineering in medicine & biology society (EMBC)*, pages 3040–3045. IEEE, 2020.

Samory Kpotufe and Guillaume Martinet. Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics*, 49(6):3299–3323, 2021.

Bertrand Lebichot, Yann-Aël Le Borgne, Liyun He-Guelton, Frederic Oblé, and Gianluca Bontempi. Deep-learning domain adaptation techniques for credit cards fraud detection. In *Recent Advances in Big Data and Deep Learning: Proceedings of the INNS Big Data and Deep Learning Conference INNSBDDL2019, held at Sestri Levante, Genova, Italy 16-18 April 2019*, pages 78–88. Springer, 2020.

Daniel Levy, Ziteng Sun, Kareem Amin, Satyen Kale, Alex Kulesza, Mehryar Mohri, and Ananda Theertha Suresh. Learning with user-level privacy. *Advances in Neural Information Processing Systems*, 34:12466–12479, 2021.

Mengchu Li, Ye Tian, Yang Feng, and Yi Yu. Federated transfer learning with differential privacy. *arXiv preprint arXiv:2403.11343*, 2024.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018.

Ruiqi Liu, Kexuan Li, and Zuofeng Shang. A computationally efficient classification algorithm in posterior drift model: phase transition and minimax adaptivity. *arXiv preprint arXiv:2011.04147*, 2020a.

Yuhan Liu, Ananda Theertha Suresh, Felix Xinnan X Yu, Sanjiv Kumar, and Michael Riley. Learning discrete distributions: user vs item-level privacy. *Advances in Neural Information Processing Systems*, 33:20965–20976, 2020b.

Yuheng Ma and Hanfang Yang. Optimal locally private nonparametric classification with public data. *arXiv preprint arXiv:2311.11369*, 2023.

Subha Maity, Yuekai Sun, and Moulinath Banerjee. Minimax optimal approaches to the label shift problem in non-parametric settings. *Journal of Machine Learning Research*, 23(346):1–45, 2022.

Subha Maity, Diptavo Dutta, Jonathan Terhorst, Yuekai Sun, and Moulinath Banerjee. A linear adjustment-based approach to posterior drift in transfer learning. *Biometrika*, 111(1):31–50, 2024.

Clifton Phua, Vincent Lee, Kate Smith, and Ross Gayler. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*, 2010.

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

Jose Ramon Saura, Ana Reyes-Menendez, and Ferrão Filipe. Comparing data-driven methods for extracting knowledge from user generated content. *Journal of Open Innovation: Technology, Market, and Complexity*, 5(4):74, 2019.

Clayton Scott. A generalized neyman-pearson criterion for optimal domain adaptation. In *Algorithmic Learning Theory*, pages 738–761. PMLR, 2019.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in neural information processing systems*, 20, 2007.

Yongpeng Wang, Hong Yu, Guoyin Wang, and Yongfang Xie. Cross-domain recommendation based on sentiment analysis and latent feature mapping. *Entropy*, 22(4):473, 2020.

Tina Yeung. Local health department adoption of electronic health records and health information exchanges and its impact on population health. *International Journal of Medical Informatics*, 128:1–6, 2019.