The Annals of Statistics
2021, Vol. 49, No. 1, 100–128
https://doi.org/10.1214/20-AOS1949

TRANSFER LEARNING FOR NONPARAMETRIC CLASSIFICATION: MINIMAX RATE AND ADAPTIVE CLASSIFIER

BY T. TONY CAI* AND HONGJI WEI†

*Department of Statistics, The Wharton School, University of Pennsylvania, *tcai@wharton.upenn.edu; †hongjiw@wharton.upenn.edu*

Human learners have the natural ability to use knowledge gained in one setting for learning in a different but related setting. This ability to transfer knowledge from one task to another is essential for effective learning. In this paper, we study transfer learning in the context of nonparametric classification based on observations from different distributions under the posterior drift model, which is a general framework and arises in many practical problems.

We first establish the minimax rate of convergence and construct a rate-optimal two-sample weighted K-NN classifier. The results characterize precisely the contribution of the observations from the source distribution to the classification task under the target distribution. A data-driven adaptive classifier is then proposed and is shown to simultaneously attain within a logarithmic factor of the optimal rate over a large collection of parameter spaces. Simulation studies and real data applications are carried out where the numerical results further illustrate the theoretical analysis. Extensions to the case of multiple source distributions are also considered.

1. Introduction. A key feature of intelligence is the ability to learn from experience. Human learners appear to have the talent to transfer their knowledge gained from one task to another similar but different task. However, in statistical learning, most procedures are designed to solve one single task, or to learn one single distribution based on observations from the same setting. In a wide range of real-world applications, it is important to gain improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned. Transfer learning aims to tackle such a problem. It has attracted increasing attention in machine learning and has been used in many applications. Recent examples include computer vision (Tzeng et al. (2017), Gong et al. (2012)), speech recognition (Huang et al. (2013)), genre classification (Choi et al. (2017)) and also many newly designed algorithms such as Lee et al. (2007), Yao and Doretto (2010). More details about transfer learning can be found in the survey papers (Pan and Yang (2010), Weiss, Khoshgoftaar and Wang (2016)).

Besides significant successes in applications, much recent focus has also been on the theoretical properties of transfer learning. In many practical situations, there are labeled data available from a distribution P, called the source distribution, while a relatively small quantity of labeled or unlabeled data is drawn from a distribution Q, called the target distribution. They are different but to some extent related distributions. The goal is to make statistical inference under Q. A natural questions is: How much information can be transferred from the source distribution P to the target distribution Q, provided a certain level of similarity between the two distributions?

This is quite a general and challenging question. The problem is also known as domain adaptation in the binary classification setting. In domain adaptation, data pairs (X, Y) are

Received April 2019; revised November 2019.
MSC2020 subject classifications. Primary 62F30; secondary 62B10, 62F12.
Key words and phrases. Adaptivity, classification, domain adaptation, minimax rate, transfer learning.

drawn from P and Q defined on $\mathbb{R}^d \times \{0, 1\}$. Data from the source distribution P can be informative about the target distribution Q if the two distributions are similar. Several type of assumptions have been proposed and studied previously in the literature, such as divergence bounds, covariate shift and posterior drift. The first line of work in the literature measures the similarity by the divergence between P and Q. Generalization bounds are derived on unlabeled testing data from the target distribution Q after training by the data from the source distribution P (Ben-David et al. (2007), Blitzer et al. (2008), Mansour, Mohri and Rostamizadeh (2009)). These bounds are general and can be applied to any two distributions, but for more structured source and target distributions those bounds are not suitable. Another line of work imposes some structural assumptions on P and Q such as covariate shift and posterior drift. Covariate shift assumes that the conditional distributions of Y given X are the same under P and Q, that is, $P_{Y|X} = Q_{Y|X}$, but the marginal distributions P_X and Q_X can be different. Such a setting typically arises when the same study/survey is carried out in different populations. For example, when constructing a classifier for a certain disease, source data may be generated from clinical studies, but the goal is to classify people drawn from the general public. The task becomes challenging due to the difference between the two populations. Transfer learning under covariate shift has been studied in previous work such as Kpotufe and Martinet (2018), Shimodaira (2000), Sugiyama et al. (2008).

In the present paper, we study transfer learning under the posterior drift model, where it is assumed that $P_X \approx Q_X$ but $P_{Y|X}$ and $Q_{Y|X}$ can highly differ. To be more specific, suppose there are two data generating distributions P and Q on $\Omega \times \{0, 1\}$, where $\Omega \subset [0, 1]^d$. We observe n_P independent and identically distributed (i.i.d.) samples $(X_1^P, Y_1^P), \ldots, (X_{n_P}^P, Y_{n_P}^P)$ drawn from a source distribution P, and n_Q i.i.d. samples $(X_1^Q, Y_1^Q), \ldots, (X_{n_Q}^Q, Y_{n_Q}^Q)$ drawn from a target distribution Q. The data points from the distributions P and Q are also mutually independent. For each data point (X, Y), the d-dimensional vector X is regarded as covariates (features) of a certain object, while Y is a (noisy) binary label indicating to which of the two classes this object belongs. The goal is to make classification under the target distribution Q: Given the observed data, construct a classifier $\hat{f} : \Omega \to \{0, 1\}$ which minimizes the classification risk under the target distribution Q:

$$R(\hat{f}) \triangleq \mathbb{P}_{(X,Y)\sim Q}\big(Y \neq \hat{f}(X)\big).$$

Here, $\mathbb{P}_{(X,Y)\sim Q}(\cdot)$ means the probability under the distribution Q.

In binary classification, the regression functions are defined as

$$\eta_P(x) \triangleq P(Y = 1|X = x) \quad \text{and} \quad \eta_Q(x) \triangleq Q(Y = 1|X = x),$$

which can be used to represent the conditional distributions $P_{Y|X}$ and $Q_{Y|X}$. In classification, Y can be regarded as an unknown parameter predicted by X, so from this perspective we refer to P_X and Q_X as the class "prior" probabilities and $\eta_P(x)$ and $\eta_Q(x)$ as the class "posterior" probabilities associated with P and Q, respectively (Scott (2019)). We say a "posterior drift" happens when P_X and Q_X have the same support with bounded densities, but $\eta_P(x)$ and $\eta_Q(x)$ are highly different.

Posterior drift is a general framework and arises in many applications, where one collects data from different populations. Here are three examples.

- *Crowdsourcing.* Crowdsourcing is a distributed model for large-scale problem-solving and experimentation such as image classification, video annotation and translation (Yuen, King and Leung (2011), Karger, Oh and Shah (2011), Zhang et al. (2014)). The tasks are broadcasted to multiple independent workers online in order to collect and aggregate their solutions. In crowdsourcing, many noisy answers/labels are available from a large amount of public workers, while sometimes, more accurate answers/labels may be collected from experienced workers or experts. These expert answers/labels are of higher quality but are relatively

few due to the time or budget constraints. One can view this difference in labeling accuracy as a posterior drift. It is desirable to construct a statistical procedure that incorporates both data sets.

- *Concept drift.* Concept drift is a common phenomenon when the underlying distribution of the data changes over time in a streaming environment (Tsymbal (2004), Gama et al. (2014)). One kind of concept drift is called real concept drift where the posterior class probabilities $P(Y|X)$ changes over time. In this situation, posterior drift exists if data are collected at different time. For example, the incidence rate of a certain disease in certain groups may change over time due to the development of treatments and preventive measures.
- *Data corruption.* Data corruption is ubiquitous in applications, where unexpected error on data occurs during storage, transmission or processing (Menon et al. (2015), van Rooyen and Williamson (2017)). In many settings, one receives data of variable quality—perhaps some small amount of clean data, another amount of slightly corrupted data, yet more that is significantly corrupted, and so on (Crammer, Kearns and Wortman (2006)). Data of variable qualities can be viewed as posterior drift between those data generating distributions, thus better strategies are needed to tackle the problem within the posterior drift framework.

Under the posterior drift model, the main difference between P and Q lies in the regression functions $\eta_P(x)$ and $\eta_Q(x)$. So the relationship between $\eta_P(x)$ and $\eta_Q(x)$, which can be captured by the link function ϕ defined below, is important in characterizing the difficulty of the transfer learning problem. In this work, we propose a new concept called the *relative signal exponent* γ to describe the relationship between $\eta_P(x)$ and $\eta_Q(x)$. Our results show that the relative signal exponent γ plays an important role in the minimax rate of convergence for the excess risk under the posterior drift model.

For conceptual simplicity, we assume $\eta_P(x) = \phi(\eta_Q(x))$ for some strictly increasing link function $\phi(\cdot)$ with $\phi(\frac{1}{2}) = \frac{1}{2}$. Note that this is only a simplified version of our formal model which will be given in Section 2. It is natural to assume ϕ is strictly increasing in the settings where those X that are more likely to be labeled $Y = 1$ under Q are also more likely to be labeled $Y = 1$ under P. The assumption $\phi(\frac{1}{2}) = \frac{1}{2}$ means that those X that are noninformative under Q are the same under P. Formally, for a given relative signal exponent $\gamma > 0$ and a constant $C_\gamma > 0$, we denote by $\Gamma(\gamma, C_\gamma)$ the collection of all distribution pairs (P, Q) satisfying

$$\left(\phi(x) - \frac{1}{2}\right)\left(x - \frac{1}{2}\right) \geq 0 \quad \text{and} \quad \left|\phi(x) - \frac{1}{2}\right| \geq C_\gamma \left|x - \frac{1}{2}\right|^\gamma. \tag{1}$$

The relative signal exponent is a key parameter in capturing the usefulness of the data from the source distribution P for the task of classification under the target distribution Q. The smaller the relative signal exponent, the more information transferable from P to Q.

In this work, we consider transfer learning under the posterior drift model in a nonparametric classification setting. When Q satisfies the margin assumption with the parameter α, defined in Section 2, and $\eta_Q(x)$ belongs to the (β, C_β)-Hölder function class, it is shown that, under the regularity conditions, the minimax optimal rate of convergence is given by

$$\inf_{\hat{f}} \max_{(P,Q)\in\Pi} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}) \asymp \left(n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q\right)^{-\frac{\beta(1+\alpha)}{2\beta+d}}, \tag{2}$$

where n_P and n_Q are number of data drawn from P and Q, respectively, d is the number of features and Π is the posterior drift regime where the distribution pair (P, Q) belongs to the class $\Gamma(\gamma, C_\gamma)$ with the relative signal exponent γ and satisfies some additional regularity conditions. Here, $\mathcal{E}_Q(\hat{f})$ is the excess risk on Q which is defined based on the misclassification error:

$$\mathcal{E}_Q(\hat{f}) = R_Q(\hat{f}) - R_Q(f_Q^*), \tag{3}$$

where

$$f_Q^*(x) = \begin{cases} 0 & \text{if } \eta_Q(x) \le \frac{1}{2}, \\ 1 & \text{otherwise} \end{cases} \tag{4}$$

is the Bayes classifier under Q. The expectation $\mathbb{E}_Z$ in (2) is taken over the random realizations of all the observed data, namely the set Z, defined as

$$Z \triangleq \{(X_1^P, Y_1^P), \ldots, (X_{n_P}^P, Y_{n_P}^P), (X_1^Q, Y_1^Q), \ldots, (X_{n_Q}^Q, Y_{n_Q}^Q)\}. \tag{5}$$

Note that if one only had observations from the target distribution Q, the minimax rate would be $n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}$. Therefore, the additional term $n_P^{\frac{2\beta+d}{2\gamma\beta+d}}$ in the minimax rate (2) quantifies an "effective sample size" for transfer learning from the source distribution P relative to Q, and $\frac{2\beta+d}{2\gamma\beta+d}$ can be viewed as the optimal transfer rate. This result answers one of the main questions in transfer learning: $n_P^{\frac{2\beta+d}{2\gamma\beta+d}}$ is the total amount of information that can be transferred from P to Q, and this quantity depends on the relative signal exponent γ which characterizes the discrepancy between P and Q in posterior drift.

We construct a two-sample weighted K-nearest neighbors (K-NN) classifier and show that it attains the optimal rate given in (2). However, this classifier depends on the parameters α, β and γ, which are typically unknown in practice. In this paper, we also propose a data-driven classifier $\hat{f}_a$ that automatically adapts to the unknown model parameters α, β and γ, with an additional log term on the excess risk bound:

$$\sup_{(P,Q)\in\Pi} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}_a) \lesssim \left(\left(\frac{n_P}{\log(n_P + n_Q)} \right)^{\frac{2\beta+d}{2\gamma\beta+d}} + \frac{n_Q}{\log(n_P + n_Q)} \right)^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

This adaptive procedure is essentially different from either the nonadaptive procedure given in this paper, or any nonparametric classification procedures in the literature. The adaptive classifier is constructed based on the ideas inspired by Lepski's method for nonparametric regression. The construction begins with a small number of the nearest neighbors, and gradually increases the number of the neighbors used to make the decision. The algorithm terminates once an empirical signal-to-noise ratio reaches a delicately designed threshold. It is shown that the resulting data-driven classifier automatically adapts to a wide collection of parameter spaces.

In some applications, there are data available from multiple source distributions. Intuitively, the samples from all source distributions are helpful to the classification task under the target distribution. We also consider transfer learning in this setting under the posterior drift model. Suppose there are multiple source distributions $P_1, \ldots, P_m$ and one target distribution Q, each pair of distributions (P_i, Q) has a relative signal exponent γ_i, $i \in \{1, \ldots, m\}$. The minimax optimal rate of convergence is established and the result quantifies precisely the contributions from the data generated by the individual source distributions. An adaptive procedure is constructed and shown to simultaneously attain the optimal rate up to a logarithmic factor over a large class of parameter spaces.

The rest of the paper is organized as follows. In Section 2, after some basic notation and definitions are introduced, the model for transfer learning under the posterior drift model is proposed in a nonparametric classification setting. In Section 3, we establish the minimax optimal rate by constructing a minimax optimal procedure with guaranteed upper bound and a matching lower bound. In Section 4, a data-driven adaptive classifier is proposed and is shown to adaptively attain the optimal rate of convergence, up to a logarithmic factor. Section 6 investigates the numerical performance of the data driven procedure. In Section 7,

a real data application is carried out to further illustrate the benefit of our method. Section 5 considers transfer learning with multiple source distributions and a brief discussion is given in Section 8. For reasons of space, we prove one main result in Section 9 and provide the proofs of the other results and some technical lemmas in the Supplementary Material (Cai and Wei (2020)).

2. Problem formulation. We introduce in this section the posterior drift model. We begin with notation and basic definitions.

2.1. *Notation and definitions.* For a distribution G, denote by $G(\cdot)$ and $\mathbb{E}_G(\cdot)$, respectively, the probability and expectation under G. Denote by P_X and Q_X the marginal distribution of X under the joint distributions P and Q for (X, Y), respectively. Let $\operatorname{supp}(\cdot)$ denote the support of a probability distribution. Throughout the paper, we write $\|\cdot\|$ to denote the Euclidean norm. We use $\mathbb{I}_{\{\cdot\}}$ to denote the indicator function taking values in $\{0, 1\}$. We define $a \vee b = \max(a, b)$, $a \wedge b = \min(a, b)$, and $\lfloor a \rfloor$ be the maximum integer that is not larger than a. We denote by $B(x, r)$ a Euclidean ball centered at x with radius r. We write $\lambda(\cdot)$ to denote Lebesgue measure of a set in a Euclidean space. We denote by C or c some generic constants not depending on n_P or n_Q that may vary from place to place.

2.2. *Posterior drift in nonparametric classification.* For two distributions P and Q for a random pair (X, Y) taking values in $[0, 1]^d \times \{0, 1\}$, we observe two independent random samples, $(X_1^P, Y_1^P), \ldots, (X_{n_P}^P, Y_{n_P}^P) \overset{\text{i.i.d.}}{\sim} P$ and $(X_1^Q, Y_1^Q), \ldots, (X_{n_Q}^Q, Y_{n_Q}^Q) \overset{\text{i.i.d.}}{\sim} Q$. We shall use P-data and Q-data to refer to the data sets drawn from the distributions P and Q, respectively. We consider the transfer learning problem when there is a posterior drift between P and Q. In the posterior drift model, the covariates/features X are drawn from distributions having the same support with bounded densities, but the response/label Y has different conditional distributions given X between P and Q. The readers should notice that the model we introduced in Section 1 is a special case within the model we will introduce in this section.

The regression functions have been defined informally in the Introduction, now we give a precise definition. Let

$$\eta_P(x) = \begin{cases} P(Y = 1 | X = x) & \text{if } x \in \operatorname{supp}(P_X), \\ \dfrac{1}{2} & \text{otherwise,} \end{cases}$$

$$\eta_Q(x) = \begin{cases} Q(Y = 1 | X = x) & \text{if } x \in \operatorname{supp}(Q_X), \\ \dfrac{1}{2} & \text{otherwise} \end{cases}$$

denote the corresponding regression functions of P and Q. Besides the previous definition (4) of Bayes classifier under the target distribution Q, we can similarly define the Bayes classifier for the source distribution P as

$$f_P^*(x) = \begin{cases} 0 & \text{if } \eta_P(x) \le \dfrac{1}{2}, \\ 1 & \text{otherwise.} \end{cases}$$

Now assume (X^P, Y^P) is a data pair drawn from the distribution P. From the definition, given $X^P = x$, Y^P is more likely to be equal to 1 if $f_P^*(x) = 1$ whereas Y^P is more likely to be equal to 0 if $f_P^*(x) = 0$. It is similar for the distribution Q. Thus informally one can regard $f_P^*(x)$ ($f_Q^*(x)$) as the true label at the covariate value x under the distribution P (Q).

In transfer learning, although the observed data are drawn from two or more different distributions, these distributions are usually related to each other so that all of them are useful for

learning the intrinsic true labels. For instance, in a crowdsourcing survey, although accuracy varies among different workers, their answers should be no worse than random guessing. It is reasonable to assume that the answer is correct with probability at least $\frac{1}{2}$. This means we may reasonably assume that, given the same covariate x, the "true labels" under the distributions P and Q are the same. That is,

$$f^*(x) \triangleq f_P^*(x) = f_Q^*(x) \quad \forall x \in \operatorname{supp}(P_X),$$

which is equivalent to

$$\left(\eta_P(x) - \frac{1}{2}\right)\left(\eta_Q(x) - \frac{1}{2}\right) \geq 0.$$

The definitions and assumptions introduced so far treat the P-data and Q-data symmetrically and interchangeably. But in real applications, usually the two data sets are treated differently. We call P the source distribution and Q the target distribution. The goal is to transfer the knowledge gained from the P-data together with the information contained in the Q-data for constructing an optimal classifier under the target distribution Q.

Intuitively, it is clear that the amount of information that can be transferred from the P-data for the inference under Q depends on the similarity between the distributions P and Q. In this paper, we quantify the similarity by the *relative signal exponent* of P with respect to Q.

DEFINITION 1 (Relative signal exponent). The class $\Gamma(\gamma, C_\gamma)$ with relative signal exponent $\gamma \in (0, \infty)$ and a constant $C_\gamma \in (0, \infty)$ is defined as the set of distribution pairs (P, Q), both supported on $\mathbb{R}^d \times \{0, 1\}$, satisfying $\forall x \in \operatorname{supp}(P_X) \cup \operatorname{supp}(Q_X)$,

$$\left(\eta_P(x) - \frac{1}{2}\right)\left(\eta_Q(x) - \frac{1}{2}\right) \geq 0, \tag{6}$$

$$\left|\eta_P(x) - \frac{1}{2}\right| \geq C_\gamma \left|\eta_Q(x) - \frac{1}{2}\right|^\gamma. \tag{7}$$

REMARK 1. The relative signal exponent γ indicates the signal strength of the P-data relative to the Q-data. Note that $|\eta_Q(x) - \frac{1}{2}|$ is always bounded by $1/2$. So generally speaking, the smaller γ is, the larger the difference between $\eta_P(x)$ and $\frac{1}{2}$, which means the P-data is more informative about $f^*(x)$ and consequently more information can be transferred from the P-data to the Q-data. An illustration of the relative signal exponent is given in Figure 1.

One can see that the above definition of relative signal exponent implies when $|\eta_Q(x) - \frac{1}{2}|$ is large, then $|\eta_P(x) - \frac{1}{2}|$ should be relatively large. This is intuitively true in a wide range of real applications. Taking again the crowdsourcing surveys as an example. If one crowd of workers can answer a question correctly with a larger probability, then for another crowd of workers the accuracy of their answers is also usually larger because this question is likely to be easier.

In addition to the relative signal exponent γ, we also need to define a smoothness parameter of η_Q and characterize its behavior near $1/2$.

DEFINITION 2 (Smoothness). The (β, C_β)-Hölder class of functions $(0 < \beta \leq 1)$, denoted by $\mathcal{H}(\beta, C_\beta)$, is defined as the set of functions $g : \mathbb{R}^d \to \mathbb{R}$ satisfying, for any $x_1, x_2 \in \mathbb{R}^d$,

$$\left|g(x_1) - g(x_2)\right| \leq C_\beta \|x_1 - x_2\|^\beta.$$

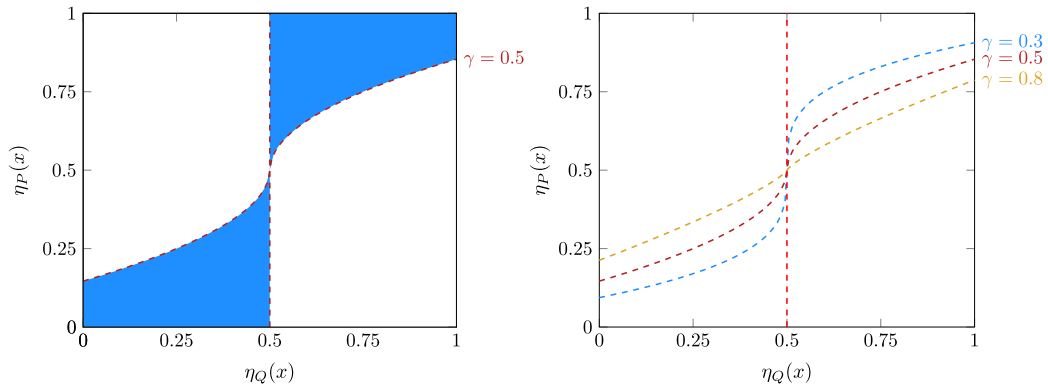


FIG. 1. *Illustration of the relative signal exponent γ. Left panel: feasible region when $\gamma = 0.5$ and $C_\gamma = 0.5$. A pair of distributions (P, Q) has relative signal exponent $\gamma = 0.5$ with $C_\gamma = 0.5$ when $(\eta_P(x), \eta_Q(x))$ falls into the shaded (blue) region for all x in the support. Right panel: feasible region with different choices of γ. Smaller γ implies more information contains in $P_{Y|X}$.*

DEFINITION 3 (Margin assumption). The margin class $\mathcal{M}(\alpha, C_\alpha)$ with $\alpha \geq 0, C_\alpha > 0$ is defined as the set of distributions Q satisfying

$$Q_X\left(\left|\eta_Q(X) - \frac{1}{2}\right| < t\right) \leq C_\alpha t^\alpha.$$

In this paper, we consider the nonparametric classification problem when $\eta_Q(x)$ belongs to a (β, C_β)-Hölder class and Q belongs to a margin class $\mathcal{M}(\alpha, C_\alpha)$. When $Q \in \mathcal{M}(\alpha, C_\alpha)$, we also say that Q satisfies the margin assumption with the parameter α.

REMARK 2. In the main part of our discussion, we focus on the case with $0 < \beta \leq 1$, that is, η belongs to a Hölder function class with smoothness less than or equal to 1. Generally, it is possible to consider more general classes where the smoothness parameter can be larger than 1. The discussion on the model and methods associated with the general smoothness parameter $\beta > 1$ will be deferred to the discussion section.

The margin assumption was first introduced in Audibert and Tsybakov (2007), Tsybakov (2004) to characterize the convergence rate in nonparametric classification. The margin assumption put a constraint on the mass around $\eta_Q(x) \approx \frac{1}{2}$ so that with large probability $\eta_Q(x)$ is either $\frac{1}{2}$ or far from $\frac{1}{2}$. Generally, if an underlying distribution satisfies the margin assumption, then a more accurate classification can be guaranteed.

Another definition is about density constraints on the marginal distributions P_X and Q_X.

DEFINITION 4 (Common support and strong density assumption). (P_X, Q_X) is said to have common support and satisfy strong density assumption with parameter $\mu = (\mu_-, \mu_+), c_\mu > 0, r_\mu > 0$ if both P_X and Q_X are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$, and

$$\Omega \triangleq \operatorname{supp}(P_X) = \operatorname{supp}(Q_X),$$

$$\lambda\big[\Omega \cap B(x, r)\big] \geq c_\mu \lambda\big[B(x, r)\big] \quad \forall 0 < r \leq r_\mu, \forall x \in \Omega,$$

$$\mu_- < \frac{dP_X}{d\lambda}(x) < \mu_+ \quad \forall x \in \Omega,$$

$$\mu_- < \frac{dQ_X}{d\lambda}(x) < \mu_+ \quad \forall x \in \Omega.$$

Define $\mathcal{S}(\mu, c_\mu, r_\mu)$ to be the set of the marginal densities pairs (P_X, Q_X) that have common support and satisfy the strong density assumption with parameter μ, c_μ, r_μ.

REMARK 3. The strong density assumption was first introduced in Audibert and Tsybakov (2007). In this paper, we focus on the scenario that the marginal densities of P_X and Q_X have regular support and are bounded from below and above on the support.

Moreover, note that when Q_X satisfies the strong density assumption, in the regime $\alpha\beta > d$, there is no distribution Q such that the regression function η_Q crosses $\frac{1}{2}$ in the interior of the support Q_X (Audibert and Tsybakov (2007)). So this regime only contains the trivial cases for classification. Therefore, we further assume $\alpha\beta \le d$ in the following discussion.

Given a classifier $\hat{f} : \mathbb{R}^d \to \{0, 1\}$, the excess risk on Q of the classifier $\hat{f}$, defined in equation (3), has a dual representation (Győrfi (1978))

$$\mathcal{E}_Q(\hat{f}) = 2\mathbb{E}_{(X,Y)\sim Q}\left(\left|\eta_Q(X) - \frac{1}{2}\right|\mathbb{I}_{\{\hat{f}(X)\neq f_Q^*(X)\}}\right). \tag{8}$$

A major goal in transfer learning is to construct an empirical decision rule $\hat{f}$ incorporating both the P-data and Q-data, so that the excess risk on Q is minimized. It is interesting to understand when the minimax rate in the transfer learning setting is faster than the optimal rate where only the Q-data is used to construct the decision rule.

Putting the above definitions together, in this paper we consider the posterior drift nonparametric parameter space:

$$\begin{aligned}\Pi(\alpha, C_\alpha, \beta, C_\beta, \gamma, C_\gamma, \mu, c_\mu, r_\mu) = \big\{&(P, Q) : (P, Q) \in \Gamma(\gamma, C_\gamma), Q \in \mathcal{M}(\alpha, C_\alpha),\\ &\eta_Q \in \mathcal{H}(\beta, C_\beta), (P_X, Q_X) \in \mathcal{S}(\mu, c_\mu, r_\mu)\big\}.\end{aligned}$$

In the rest of this paper, we will use the shorthand $\Pi(\alpha, \beta, \gamma, \mu)$ or Π if there is no confusion. The space $\Pi(\alpha, \beta, \gamma, \mu)$ is also called the posterior drift regime with $(\alpha, \beta, \gamma, \mu)$.

3. Minimax rate of convergence. In this section, we establish the minimax rate of convergence for the excess risk on Q for transfer learning under the posterior drift model and propose an optimal procedure using the two-sample weighted K-NN classifier.

The K-NN method has attracted much attention (Cover and Hart (1967), Győrfi (1978), Gadat, Klein and Marteau (2016)) due to its massive practical success and appealing theoretical properties. In the conventional setting where one only has access to the Q-data and there is no P-data, with a suitable choice of the neighborhood size k, the K-NN classifier can achieve the minimax rate of convergence for the excess risk on Q (Gadat, Klein and Marteau (2016)). The K-NN classifier is generated in two steps:

Step 1: For any given x to be classified, one can estimate $\eta_Q(x)$ by taking the empirical mean of the response variables (Y) according to its k nearest covariates (X). Formally, define $X_{(i)}^Q(x)$ be the ith nearest covariates to x among $X_1^Q, \ldots, X_{n_Q}^Q$ and $Y_{(i)}^Q(x)$ is its corresponding response (label). The estimate $\hat{\eta}_Q(x)$ is given by

$$\hat{\eta}_Q(x) = \frac{1}{k}\sum_{i=1}^{k} Y_{(i)}^Q(x).$$

Step 2: The class label for x is estimated by the plug-in rule:

$$\hat{f}(x) = \mathbb{I}_{\{\hat{\eta}_Q(x) > \frac{1}{2}\}}.$$

In transfer learning, one also has access to the P-data in addition to the Q-data, the P-data can be used to help the classification task under the target distribution Q and should be taken into consideration. To accommodate the existing K-NN methods, we should take the empirical mean of not only the k-nearest response variables from the Q-data, but also some nearest response variables from the P-data. In addition, when taking the average, data from the different distributions should have different weights because the signal strength varies between the two distributions. To make the classification at $x \in [0,1]^d$, a new strategy called the two-sample weighted K-NN classifier is summarized as follows:

Step 1: Define $X_{(i)}^P(x)$ to be the ith nearest covariates to x among $X_1^P, \ldots, X_{n_P}^P$ and $Y_{(i)}^P(x)$ is its corresponding response. $X_{(i)}^Q(x)$ and $Y_{(i)}^Q(x)$ can be defined likewise. Construct the two-sample weighted K-NN estimator

$$\hat{\eta}_{\mathrm{NN}}(x) = \frac{w_P \sum_{i=1}^{k_P} Y_{(i)}^P(x) + w_Q \sum_{i=1}^{k_Q} Y_{(i)}^Q(x)}{w_P k_P + w_Q k_Q},$$

where the number of neighbors k_P and k_Q and the weights w_P and w_Q will be specified later.

Step 2: The class label for x is estimated by the plug-in rule:

$$\hat{f}_{\mathrm{NN}}(x) = \mathbb{I}_{\{\hat{\eta}_{\mathrm{NN}}(x) > \frac{1}{2}\}}.$$

The final decision rule $\hat{f}_{\mathrm{NN}}(x)$, which is generated by both the P-data and Q-data, is called the *two-sample weighted K-NN classifier.* An illustration of the two-sample weighted K-NN classifier is given in Figure 2.

The performance of the two-sample weighted K-NN classifier $\hat{f}_{\mathrm{NN}}(x)$ clearly depends on the choice of (k_P, k_Q, w_P, w_Q). The next theorem gives a set of choices of (k_P, k_Q, w_P, w_Q) and a provable upper bound on the excess risk, which gives a guarantee for the performance of the two-sample weighted K-NN classifier with these specific choices of (k_P, k_Q, w_P, w_Q).

THEOREM 1 (Upper bound). *Let $\hat{f}_{\mathrm{NN}}$ be the two-sample weighted K-NN classifier with* $w_Q = (n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q)^{-\frac{\beta}{2\beta+d}}$, $w_P = (n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q)^{-\frac{\gamma\beta}{2\beta+d}}$, $k_Q = \lfloor n_Q (n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q)^{-\frac{d}{2\beta+d}} \rfloor$, *and*

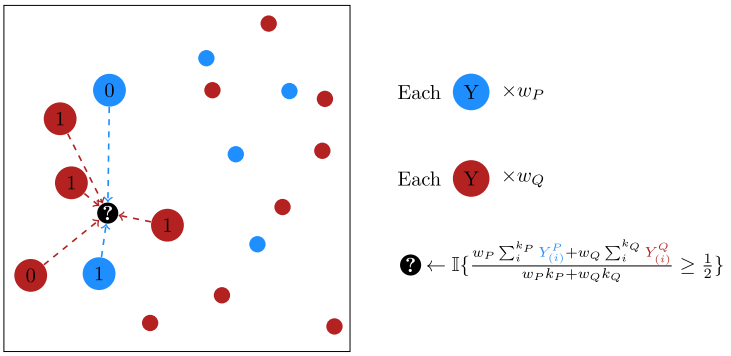


FIG. 2. *An illustration of the two-sample weighted K-NN classifier. (X^P, Y^P) are shown by the blue points and (X^Q, Y^Q) are shown by the red points. For each point in the graph, the coordinates represent its two-dimensional covariates X while the number marked inside the point represents its label Y. To classify the black point (x) located in middle of the graph, by calculation we get (say) $k_P = 2$ and $k_Q = 4$. Then we find k_P nearest neighbors from P-data and k_Q nearest neighbors from Q-data. Finally, we calculate their weighted mean to make the final classification.*

$k_P = \lfloor n_P (n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q)^{-\frac{d}{2\beta+d}} \rfloor$. *Then*

$$\sup_{(P,Q)\in\Pi} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}_{\mathrm{NN}}) \le C\big(n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q\big)^{-\frac{\beta(1+\alpha)}{2\beta+d}}$$

for some constant $C > 0$ not depending on n_P or n_Q.

The following lower bound result shows that the two-sample weighted K-NN classifier $\hat{f}_{\mathrm{NN}}$ given in Theorem 1 is in fact rate optimal.

THEOREM 2 (Lower bound). *There exists a constant $c > 0$ not depending on n_P or n_Q such that*

$$\inf_{\hat{f}} \sup_{(P,Q)\in\Pi} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}) \ge c\big(n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q\big)^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

The proof of Theorem 1 will be given in Section 9, which is based on the general techniques for proving K-NN methods, for instance; see Gadat, Klein and Marteau (2016), Samworth (2012). In the literature of classical nonparametric classification problem, the focus was mainly on bias-variance trade-off. Under posterior drift model, we further extend the general techniques to the two-sample setting, where the weights and the number of neighbors are carefully selected to make the best combination of information. The proof of Theorem 2 is given in the Supplementary Material (Cai and Wei (2020)), using the same general scheme as in (Audibert and Tsybakov (2007), Kpotufe and Martinet (2018)). Theorems 1 and 2 together establish the minimax rate of convergence for transfer learning under the posterior drift model,

$$\inf_{\hat{f}} \sup_{(P,Q)\in\Pi} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}) \asymp \big(n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q\big)^{-\frac{\beta(1+\alpha)}{2\beta+d}}. \tag{9}$$

We make a few remarks on the minimax rate of convergence.

- Based on the minimax rate given in (9), it is easy to see that, in terms of the classification accuracy, the contribution from the P-data is substantial when $n_P^{\frac{2\beta+d}{2\gamma\beta+d}} \gg n_Q$, and the contribution is not significant otherwise.
- Comparing the convergence rates (9) with (10), the minimax rate for transfer learning under the posterior drift model is the same as if one had a sample of size $n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q$ from the distribution Q in the conventional setting. Therefore, one can intuitively view $n_P^{\frac{2\beta+d}{2\gamma\beta+d}}$ as the "effective sample size" of the P-data for the classification task under Q. The exponent $\frac{2\beta+d}{2\gamma\beta+d}$ here can be regarded as the *transfer rate*. The smaller the relative signal exponent γ is, the larger $\frac{2\beta+d}{2\gamma\beta+d}$ is, and more information is transferred from the P-data. This transfer rate provides a quantitative answer to the question posed in the Introduction: How much information can be transferred from the source distribution P to the target distribution Q? It is also interesting to note that, when $\gamma < 1$, $\frac{2\beta+d}{2\gamma\beta+d} > 1$, which implies that in this case an observation from P is more valuable than an observation from Q for the classification problem.
- In the transfer learning literature, much attention has been on an interesting special case where there is no data from the target distribution Q at all, that is, $n_Q = 0$ (Mansour, Mohri and Rostamizadeh (2009), Blitzer et al. (2008)). This case arises when a classifier has been trained based on the data drawn from the source distribution P, and one wishes to

generalize the classifier to unlabeled testing data drawn from the target distribution Q. Our results show that generalization is possible in the posterior drift framework and the optimal rate of convergence is

$$\inf_{\hat{f}} \sup_{(P,Q)\in\Pi} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}) \asymp n_P^{-\frac{\beta(1+\alpha)}{2\gamma\beta+d}}.$$

• It is worth noting that in the conventional setting with access to the Q-data only, the minimax rate, which is given in Audibert and Tsybakov (2007), would be

$$\inf_{\hat{f}} \sup_{(P,Q)\in\Pi} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}) \asymp n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}, \tag{10}$$

which is a special case of (9) with $n_P = 0$. This rate can be achieved by the K-NN classifier given above with the choice of $k \asymp n_Q^{\frac{2\beta}{2\beta+d}}$.

4. Data-driven adaptive classifier. In the previous section, we have established the minimax optimal rate over the parameter space $\Pi(\alpha, \beta, \gamma, \mu)$ for transfer learning under the posterior drift model. This rate can be achieved by the two-sample weighted K-NN classifier given in Theorem 1. A major drawback of this classifier is that it requires the prior knowledge of β and γ, which is typically unavailable in practice. An interesting and practically important question is whether it is possible to construct a data-driven adaptive decision rules that can achieve the same rate of convergence, while automatically adapt to a wide collection of the parameter spaces $\Pi(\alpha, \beta, \gamma, \mu)$.

In nonparametric regression, Lepski's method (Lepski (1991, 1992, 1993)) is a well-known approach for the construction of a data driven estimator that adapts to the unknown smoothness parameter β by screening from a small bandwidth to larger bandwidths with delicately designed stopping rules. This procedure can be used for nonparametric classification in the conventional setting where only Q-data is available and only adaptation to one smoothness parameter β is needed. For the readers' convenience, we include this construction in Section 9. The transfer learning setting is more challenging: we need to adapt to bothparameters β and γ. In this section, we modify Lepski's method in our context and introduce a new stopping rule and show that the resulting classifier adapts to all unknown parameters.

Now we develop a data-driven procedure to make classification at a specific point $x \in [0, 1]^d$. The construction combines all data points from the P-data and the Q-data together and finds nearest neighbors among all the data. Denote by $X_{(i)}(x)$ the ith nearest data point to x in the combined set $\{X_1^P, \ldots, X_{n_P}^P\} \cup \{X_1^Q, \ldots, X_{n_Q}^Q\}$. Similar to Lepski's method, we begin with a small number of nearest neighbors, and gradually increase the number of neighbors used to make the decision. One more nearest neighbor is added in each step. At the kth step, there are k nearest neighbors $X_{(1)}(x), \ldots, X_{(k)}(x)$ among all the points in the combined set $\{X_1^P, \ldots, X_{n_P}^P\} \cup \{X_1^Q, \ldots, X_{n_Q}^Q\}$. Suppose among these k nearest neighbors there are $k_P^{(k)}$ points from the P-data and $k_Q^{(k)}$ points from the Q-data. Heuristically, given these k nearest neighbors, one can obtain a weighted K-NN estimate as

$$\hat{\eta}^{(k)}(x, w_P, w_Q) = \frac{w_P \sum_{i=1}^{k_P^{(k)}} Y_{(i)}^P(x) + w_Q \sum_{i=1}^{k_Q^{(k)}} Y_{(i)}^Q(x)}{w_P k_P^{(k)} + w_Q k_Q^{(k)}}.$$

If β and γ are known, one can calculate the optimal choice of the weights w_P and w_Q for a two-sample weighted K-NN classifier. To construct an adaptive procedure, we need to

find a data driven method for choosing the weights w_P and w_Q. Define the "variance" of $\hat{\eta}^{(k)}(x, w_P, w_Q)$ as

$$v^{(k)}(w_P, w_Q) = \frac{w_P^2 k_P^{(k)} + w_Q^2 k_Q^{(k)}}{(w_P k_P^{(k)} + w_Q k_Q^{(k)})^2}.$$

For a given k, we call the maximum value of the ratio between $(\hat{\eta}^{(k)}(x, w_P, w_Q) - \frac{1}{2})^2$ and the "variance" $v^{(k)}(w_P, w_Q)$ as the signal-to-noise ratio index $\hat{r}^{(k)}$:

$$\hat{r}^{(k)} = \max_{w_P, w_Q} \frac{(\hat{\eta}^{(k)}(x, w_P, w_Q) - \frac{1}{2})^2}{v^{(k)}(w_P, w_Q)}.$$

The algorithm is terminated when $\hat{r}^{(k)} > (d+3)\log(n_P + n_Q)$, and the corresponding w_P and w_Q are chosen as the maximizers of $\frac{(\hat{\eta}^{(k)}(x, w_P, w_Q) - \frac{1}{2})^2}{v^{(k)}(w_P, w_Q)}$. If the algorithm does not terminate at any step, the optimal k can be alternatively chosen by the maximizer of $\hat{r}^{(k)}$. That is, we choose $k = k^*$ with

$$k^* = \begin{cases} \min\{k : \hat{r}^{(k)} > (d+3)\log(n_P + n_Q)\} & \text{if } \max_k \hat{r}^{(k)} > (d+3)\log(n_P + n_Q), \\ \operatorname{argmax}_k \hat{r}^{(k)} & \text{otherwise} \end{cases} \tag{11}$$

and choose $(w_P, w_Q) = (w_P^*, w_Q^*)$ with

$$(w_P^*, w_Q^*) = \operatorname*{argmax}_{(w_P, w_Q)} \frac{(\hat{\eta}^{(k^*)}(x, w_P, w_Q) - \frac{1}{2})^2}{v^{(k^*)}(w_P, w_Q)}.$$

The data driven adaptive classifier is then defined as

$$\hat{f}_a(x) = \mathbb{I}_{\{\hat{\eta}^{(k^*)}(x, w_P^*, w_Q^*) \geq \frac{1}{2}\}}.$$

REMARK 4. The choice of $(d+3)\log(n_P + n_Q)$ as the threshold in the stopping rule (11) is important and requires some explanation. Roughly speaking, this is due to the fact that the maximum fluctuation of $\hat{\eta}^{(k)}(x, w_P, w_Q)$ is bounded by $\sqrt{(d+3)\log(n_P + n_Q) v^{(k)}(w_P, w_Q)}$ with high probability, which will be shown in Lemma 5 with a suitable change of parameter (stated in the Supplementary Material (Cai and Wei (2020))). Thus, when $\hat{r}^{(k)} > (d+3)\log(n_P + n_Q)$, $\hat{\eta}^{(k)}(x, w_P, w_Q) > \frac{1}{2}$ indicates $\mathbb{E}\hat{\eta}^{(k)}(x, w_P, w_Q) > \frac{1}{2}$, which suggests $f^*(x) = 1$, and vice versa.

The procedure is summarized in Algorithm 1 where the above procedure is simplified by using the actual closed-form expression for $\hat{r}^{(k)}$ and $\hat{f}_a(x)$. An illustration of the above adaptive procedure is given in Figure 3.

We investigate the theoretical properties of this data-driven classifier $\hat{f}_a$ in terms of both global and local adaptivity. The theoretical analysis shows that the proposed classifier is, both globally and locally, adaptive to the unknown smoothness and relative signal exponent.

4.1. *Global adaptivity.* Note that $\hat{f}_a$ is a data-driven classifier. The following theorem gives an upper bound for the excess risk under Q.

THEOREM 3. *Let $n = n_P + n_Q$. There exists a constant $C > 0$ not depending on n_P or n_Q such that*

$$\sup_{(P,Q)\in\Pi} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}_a) \leq C\left(\left(\frac{n_P}{\log n}\right)^{\frac{2\beta+d}{2\gamma\beta+d}} + \frac{n_Q}{\log n}\right)^{-\frac{\beta(1+\alpha)}{2\beta+d}}. \tag{12}$$

Algorithm 1 The data driven procedure

Input: $x \in \mathrm{supp}(Q_X)$.

for $k = 1, \ldots, (n_P + n_Q - 1), (n_P + n_Q)$ **do**

Find k nearest covariates to x among all covariates in data $\{X_1^P, X_2^P, \ldots, X_{n_P}^P\} \cup \{X_1^Q, X_2^Q, \ldots, X_{n_Q}^Q\}$. Suppose among those k nearest neighbors $X_{(1)}(x), X_{(2)}(x), \ldots, X_{(k)}(x)$ there are $k_P^{(k)}$ covariates from P-data and $k_Q^{(k)}$ covariates from Q-data.

Compute $k_P^{(k)}$ nearest neighbor estimate in P-data (if $k_P^{(k)} = 0$, set $\hat{\eta}_P^{(k)} \leftarrow \frac{1}{2}$)

$$\hat{\eta}_P^{(k)} \leftarrow \frac{1}{k_P^{(k)}} \sum_{i=1}^{k_P^{(k)}} Y_{(i)}^P(x)$$

and $k_Q^{(k)}$ nearest neighbor estimate in P-data (if $k_Q^{(k)} = 0$, set $\hat{\eta}_Q^{(k)} \leftarrow \frac{1}{2}$)

$$\hat{\eta}_Q^{(k)} \leftarrow \frac{1}{k_Q^{(k)}} \sum_{i=1}^{k_Q^{(k)}} Y_{(i)}^P(x).$$

Let $\hat{r}^{(k)}$ be the signal-to-noise ratio index calculated by

$$\hat{r}^{(k)} \leftarrow \begin{cases} k_P^{(k)}\left(\hat{\eta}_P^{(k)} - \frac{1}{2}\right)^2 + k_Q^{(k)}\left(\hat{\eta}_Q^{(k)} - \frac{1}{2}\right)^2 & \text{if } \mathrm{sign}\left(\hat{\eta}_P^{(k)} - \frac{1}{2}\right) = \mathrm{sign}\left(\hat{\eta}_Q^{(k)} - \frac{1}{2}\right), \\ \max\left(k_P^{(k)}\left(\hat{\eta}_P^{(k)} - \frac{1}{2}\right)^2, k_Q^{(k)}\left(\hat{\eta}_Q^{(k)} - \frac{1}{2}\right)^2\right) & \text{if } \mathrm{sign}\left(\hat{\eta}_P^{(k)} - \frac{1}{2}\right) \neq \mathrm{sign}\left(\hat{\eta}_Q^{(k)} - \frac{1}{2}\right). \end{cases}$$

Define the intermediate classifier by

$$\hat{f}^{(k)}(x) \leftarrow \mathbb{I}_{\{\sqrt{k_P^{(k)}}(\eta_P^{(k)} - \frac{1}{2}) + \sqrt{k_Q^{(k)}}(\eta_Q^{(k)} - \frac{1}{2}) \geq 0\}}.$$

if $\hat{r}^{(k)}(x) > (d+3)\log(n_P + n_Q)$ **then**

Stop and output $\hat{f}_a(x) \leftarrow f^{(k)}(x)$.

Output $\hat{f}_a(x) \leftarrow \hat{f}^{(k_m)}(x)$ where $k_m = \mathrm{argmax}_k \hat{r}^{(k)}$.

The proof of Theorem 3 is given in the Supplementary Material (Cai and Wei (2020)).

Comparing the rate of convergence in (12) for the adaptive classifier $\hat{f}_a$ with the minimax rate (9), the data driven classifier $\hat{f}_a$ simultaneously achieves within a logarithmic factor of the minimax optimal rate over a large collection of parameter spaces.

REMARK 5. If only the Q-data is available and Lepski's method is applied, then the following upper bound on the excess risk under Q holds:

$$\sup_{(P,Q)\in\Pi} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}_L) \leq C \cdot \left(\frac{n_Q}{\log n_Q}\right)^{-\frac{\beta(1+\alpha)}{2\beta+d}}. \tag{13}$$

One can verify that by setting $n_P = 0$, our new adaptive procedure is exactly equivalent to Lepski's method (Algorithm 3), while the rates of convergence for the two methods also coincide.

4.2. *Local adaptivity.* In practice, one might be focused on classifying a given observation x_0, and thus especially interested in the accuracy of a classifier at a specific point x_0. Interestingly, the weights w_P and w_Q, the number k of neighbors of the proposed classifier

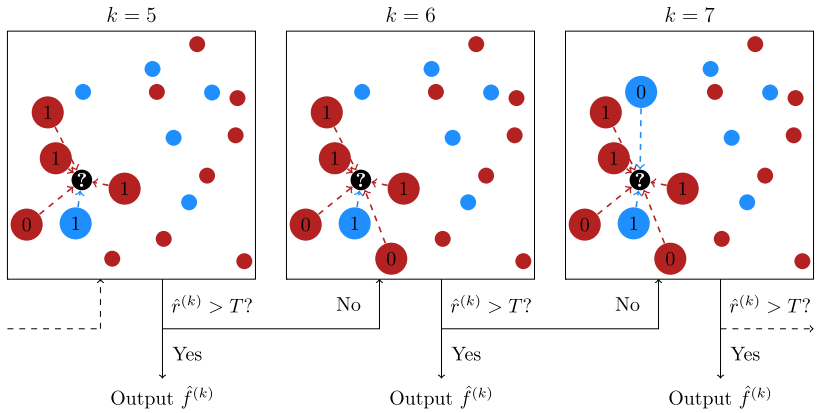


FIG. 3. *An illustration of the adaptive procedure given in Algorithm 1. See Figure 2 for interpretation of the graph. Here, we shorthand the threshold* $T = (d+3)\log(n_P + n_Q)$. *In each step, we evaluate* $r^{(k)}$ *and compare it to the threshold* R. *If* $r^{(k)} > T$, *then output* $\hat{f}^{(k)}$ *generated in current step; if* $r^{(k)} \le T$, *go to next step and add one more nearest neighbor.*

$\hat{f}_a(x)$ are all locally selected and calculated based on samples in a neighborhood of x. It is of practical interest to investigate the local adaptivity of the proposed classifier.

In order to study the local behavior of the classifier $\hat{f}_a$ at a given point x_0, we need to extent the definitions for the posterior drift model to their local versions. First, we define the local excess risk on Q at a point x_0.

DEFINITION 5. For any $x_0 \in \Omega$ and a classifier $\hat{f} : \Omega \to \{0, 1\}$, define the classification risk at x_0 on distribution Q for $\hat{f}$ as

$$R(\hat{f}, x_0) = \mathbb{P}_{(X,Y)\sim Q}\big(Y \neq \hat{f}(x_0) | X = x_0\big).$$

Further, define the local excess risk at x_0 on distribution Q for $\hat{f}$ as

$$\mathcal{E}_Q(\hat{f}, x_0) = R(\hat{f}, x_0) - R\big(f_Q^*, x_0\big).$$

Next, we give a formal definition for local smoothness $\beta_0 = \beta(x_0)$ and local relative signal exponent $\gamma_0 = \gamma(x)$.

DEFINITION 6. A function $g : \mathbb{R}^d \to \mathbb{R}$ has local Hölder smoothness β_0 $(0 < \beta_0 \le 1)$ at point $x_0 \in \mathbb{R}^d$ if there exists $r > 0$ and $C_\beta > 0$ such that for any $x' \in B(x_0, r)$,

$$|g(x') - g(x_0)| \le C_\beta \|x' - x_0\|^\beta.$$

DEFINITION 7. A pair of distributions (P, Q), both supported on $\Omega \times \{0, 1\}$, are defined to have local relative signal exponent γ_0 at a point $x_0 \in \Omega$, if there exists $r > 0$ and $C_\gamma > 0$ such that for any $x \in B(x_0, r)$,

$$\left(\eta_P(x) - \frac{1}{2}\right)\left(\eta_Q(x) - \frac{1}{2}\right) \ge 0,$$

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

The definitions of local smoothness and local relative signal exponent are similar to their global versions, except we only consider in a small neighborhood of x_0. Based on the above definitions, the local adaptivity of $\hat{f}_a$ at x_0 is characterized as follows.

THEOREM 4. *Suppose the distributions (P, Q) are both supported on $\Omega \times \{0, 1\}$ and a point $x_0 \in \Omega$. Suppose the following hold*:

1. *(P, Q) have local relative signal exponent γ_0 at x_0;*
2. *η_Q has local Hölder smoothness β_0 at x_0;*
3. *$(P_X, Q_X) \in \mathcal{S}(\mu, c_\mu, r_\mu)$, that is, P_X and Q_X have common support and satisfy the strong density assumption.*

Let $n = n_P + n_Q$. There exists a constant $C > 0$ such that

$$\mathbb{E}_Z \mathcal{E}_Q(\hat{f}_a, x_0) \le C\left(\left(\frac{n_P}{\log n}\right)^{-\frac{2\beta_0+d}{2\gamma_0\beta_0+d}} + \frac{n_Q}{\log n}\right)^{-\frac{\beta_0}{2\beta_0+d}}. \tag{14}$$

The proof of Theorem 4 is provided in the Supplementary Material (Cai and Wei (2020)).

REMARK 6. Under the same setting as in Theorem 4, when β_0 and γ_0 are known, the local minimax rate of convergence is

$$\inf_{\hat{f}} \sup_{(P,Q)} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}_a, x_0) \asymp \left(n_P^{-\frac{2\beta_0+d}{2\gamma_0\beta_0+d}} + n_Q\right)^{-\frac{\beta_0}{2\beta_0+d}},$$

where the supremum is taken over all distribution pairs (P, Q) satisfying conditions 1, 2, 3 stated in Theorem 4. This minimax rate can be achieved by the same construction as the minimax classifier in Section 3 (using local parameters β_0, γ_0 instead of global parameters β, γ). As a result, Theorem 4 shows that $\hat{f}_a$ also achieves within a logarithmic factor of the local minimax optimal rate. In other words, $\hat{f}_a$ adapts to local smoothness and local signal relative exponent.

REMARK 7. For simplicity, the present paper focuses on the posterior drift model, which is somewhat restrictive since the relation between P and Q is described by a signal parameter γ. However, because $\hat{f}_a$ is adaptive to the local signal relative exponent, it can make nearly optimal classification under heterogeneity where γ varies. In other words, $\hat{f}_a$ works optimally even when P is stronger than Q in some places and weaker than Q elsewhere.

REMARK 8. Note that there is also a dual representation of $\mathbb{E}_Z \mathcal{E}_Q(\hat{f}_a, x_0)$:

$$\mathbb{E}_Z \mathcal{E}_Q(\hat{f}_a, x_0) = 2\left|\eta_Q(x_0) - \frac{1}{2}\right| \mathbb{P}_Z\big(\hat{f}_a(x_0) \ne f_Q^*(x_0)\big).$$

Theorem 4 can be interpreted as follows. For any point x_0, the classifier $\hat{f}_a$ performs well (i.e., the accuracy of $\hat{f}_a$ is bounded away from $1/2$) when

$$\left|\eta_Q(x_0) - \frac{1}{2}\right| \ge C\left(\left(\frac{n_P}{\log n}\right)^{-\frac{2\beta_0+d}{2\gamma_0\beta_0+d}} + \frac{n_Q}{\log n}\right)^{-\frac{\beta_0}{2\beta_0+d}}$$

for some constant $C > 0$. Other than the sample sizes n_P and n_Q, the rate only depends on the local smoothness β_0 and local relative signal exponent γ_0. Also, it is optimal up to a logarithmic factor. The result thus shows that $\hat{f}_a$ is adaptive to the local smoothness and local relative signal exponent.

5. Multiple source distributions. We have so far focused on transfer learning with one source distribution P and one target distribution Q. In practice, data may be generated from more than one source distribution. In this section, we generalize our methods to treat transfer learning in the setting where multiple source distributions are available.

We consider a model where there are several source distributions with different relative signal exponents with respect to the target distribution Q. Suppose there are $n_{P_1}, \ldots, n_{P_m}$, and n_Q i.i.d. data points generated from the source distributions $P_1, \ldots, P_m$, and the target

distribution Q, respectively,

$$(X_1^{P_1}, Y_1^{P_1}), \ldots, (X_{n_{P_1}}^{P_1}, Y_{n_{P_1}}^{P_1}) \overset{\text{i.i.d.}}{\sim} P_1,$$
$$\vdots$$
$$(X_1^{P_m}, Y_1^{P_m}), \ldots, (X_{n_{P_m}}^{P_m}, Y_{n_{P_m}}^{P_m}) \overset{\text{i.i.d.}}{\sim} P_m,$$
$$(X_1^{Q}, Y_1^{Q}), \ldots, (X_{n_Q}^{Q}, Y_{n_Q}^{Q}) \overset{\text{i.i.d.}}{\sim} Q$$

and all the samples are independent. The goal is to make classification under the target distribution Q. Similar as before, it is intuitively clear that how useful the data from the source distributions P_i, $i \in [m]$, to the classification task under Q depends on the relationship between each P_i and Q. The definition of the relative signal exponent needs to be extended to accommodate the multiple source distributions. It is natural to consider the situation where each source distribution P_i and the target distribution Q have a relative signal exponent. This motivates the following definition of the vectorized relative signal exponent when there are multiple source distributions.

DEFINITION 8. Suppose the distributions $P_1, \ldots, P_m$, and Q are supported on $\mathbb{R}^d \times \{0, 1\}$. Define the class $\Gamma(\boldsymbol{\gamma}, C_{\boldsymbol{\gamma}})$ with the 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 $(P_1, \ldots, P_m, Q)$ that satisfy, for each $i \in [m]$, (P_i, Q) belongs to the class $\Gamma(\gamma_i, C_i)$ with the relative signal exponent γ_i.

Similar as in Section 2, adding the regularity conditions on Q including the smoothness, margin assumption and strong density assumption, we define the parameter space in the multiple source distributions setting as follows:

$$\begin{aligned}&\Pi(\alpha, C_\alpha, \beta, C_\beta, \boldsymbol{\gamma}, C_{\boldsymbol{\gamma}}, \mu, c_\mu, r_\mu)\\ &\quad = \big\{(P_1, \ldots, P_m, Q) : (P_1, \ldots, P_m, Q) \in \Gamma(\boldsymbol{\gamma}, C_{\boldsymbol{\gamma}}),\\ &\qquad Q \in \mathcal{M}(\alpha, C_\alpha), \eta_Q \in \mathcal{H}(\beta, C_\beta), (P_{i,X}, Q_X) \in \mathcal{S}(\mu, c_\mu, r_\mu) \text{ for all } i \in [m]\big\}.\end{aligned}$$

We will simply denote $\Pi(\alpha, C_\alpha, \beta, C_\beta, \boldsymbol{\gamma}, C_{\boldsymbol{\gamma}}, \mu, c_\mu, r_\mu)$ by Π or $\Pi(\alpha, \beta, \boldsymbol{\gamma}, \mu)$ if there is no confusion.

In this section, we establish the minimax optimal rate of convergence and propose an adaptive classifier which simultaneously achieves the optimal rate of convergence within a logarithmic factor over a wide collection of the parameter spaces. The proofs are similar to those for Theorems 1, 2 and 3 in the one source distribution setting. For reasons of space, we omit the proofs.

5.1. *Minimax rate of convergence.* We begin with the construction of a minimax rate optimal classifier $\hat{f}_{\text{NN}}$ in the case of multiple source distributions. The classifier is an extension of the two-sample weighted K-NN classifier given in Section 3. It incorporates the information contained in the data drawn from the source distributions P_i, $i \in [m]$, as well as the data drawn from the target distribution Q. The detailed steps are as follows.

Step 1: Compute the weights $w_{P_1}, \ldots, w_{P_m}$, and w_Q by

$$w_{P_i} = \left(n_Q + \sum_{i=1}^m n_{P_i}^{\frac{2\beta+d}{2\gamma_i\beta+d}}\right)^{-\frac{\gamma_i\beta}{2\beta+d}} \quad \text{for all } i \in [m],$$

$$w_Q = \left(n_Q + \sum_{i=1}^m n_{P_i}^{\frac{2\beta+d}{2\gamma_i\beta+d}}\right)^{-\frac{\beta}{2\beta+d}}.$$

Compute the numbers of neighbors $k_{P_1}, \ldots, k_{P_m}, k_Q$ by

$$k_{P_i} = \left\lfloor n_{P_i} \left(n_Q + \sum_{i=1}^{m} n_{P_i}^{\frac{2\beta+d}{2\gamma_i\beta+d}} \right)^{-\frac{d}{2\beta+d}} \right\rfloor \quad \text{for all } i \in [m]$$

$$k_Q = \left\lfloor n_Q \left(n_Q + \sum_{i=1}^{m} n_{P_i}^{\frac{2\beta+d}{2\gamma_i\beta+d}} \right)^{-\frac{d}{2\beta+d}} \right\rfloor.$$

Step 2: Define $X_{(j)}^{P_i}(x)$ to be the jth nearest data point to x among $X_1^{P_i}, \ldots, X_{n_{P_i}}^{P_i}$ and $Y_{(j)}^{P_i}(x)$ is its corresponding response (label). Likewise, let $X_{(j)}^{Q}(x)$ be the jth data point to x among $X_1^{Q}, \ldots, X_{n_Q}^{Q}$ and $Y_{(j)}^{Q}(x)$ is its corresponding response (label). Define the weighted K-NN estimator

$$\hat{\eta}_{\mathrm{NN}}(x) = \frac{w_Q \sum_{j=1}^{k_Q} Y_{(j)}^{Q}(x) + \sum_{i=1}^{m} (w_{P_i} \sum_{j=1}^{k_{P_i}} Y_{(j)}^{P_i}(x))}{w_Q k_Q + \sum_{i=1}^{m} w_{P_i} k_{P_i}}.$$

This estimator takes weighted average of k_{P_i} nearest neighbors from the data points drawn from P_i, each with weight w_{P_i}, and k_Q nearest neighbors from the data points drawn from Q, each with weight w_Q.

Step 3: The final classifier is then defined as

$$\hat{f}_{\mathrm{NN}}(x) = \mathbb{I}_{\{\hat{\eta}_{\mathrm{NN}}(x) > \frac{1}{2}\}}.$$

We now analyze the theoretical properties of the classifier $\hat{f}_{\mathrm{NN}}$. Theorem 5 gives an upper bound for the excess risk $\mathcal{E}_Q(\hat{f}_{\mathrm{NN}})$, while Theorem 6 provides a matching lower bound on the excess risk. These two theorems together establish the minimax rate of convergence and show that $\hat{f}_{\mathrm{NN}}$ attains the optimal rate. In the following theorems, the expectation $\mathbb{E}$ is taken over random realization of all data drawn from source and target distributions.

THEOREM 5 (Upper bound). *There exists a constant $C > 0$ not depending on n_P or n_Q, such that*

$$\sup_{(P_1,\ldots,P_m,Q)\in\Pi(\alpha,\beta,\boldsymbol{\gamma},\mu)} \mathbb{E}\mathcal{E}_Q(\hat{f}_{\mathrm{NN}}) \le C \left(n_Q + \sum_{i=1}^{m} n_{P_i}^{\frac{2\beta+d}{2\gamma_i\beta+d}} \right)^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

THEOREM 6 (Lower bound). *There exists a constant $c > 0$ not depending on n_P or n_Q, such that*

$$\inf_{\hat{f}} \sup_{(P_1,\ldots,P_m,Q)\in\Pi(\alpha,\beta,\boldsymbol{\gamma},\mu)} \mathbb{E}\mathcal{E}_Q(\hat{f}) \ge c \left(n_Q + \sum_{i=1}^{m} n_{P_i}^{\frac{2\beta+d}{2\gamma_i\beta+d}} \right)^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

Theorems 5 and 6 together yield the minimax optimal rate for transfer learning with multiple source distributions:

$$\inf_{\hat{f}} \sup_{(P_1,\ldots,P_m,Q)\in\Pi(\alpha,\beta,\boldsymbol{\gamma},\mu)} \mathbb{E}\mathcal{E}_Q(\hat{f}) \asymp \left(n_Q + \sum_{i=1}^{m} n_{P_i}^{\frac{2\beta+d}{2\gamma_i\beta+d}} \right)^{-\frac{\beta(1+\alpha)}{2\beta+d}}. \tag{15}$$

As discussed in Section 3, here $n_{P_i}^{\frac{2\beta+d}{2\gamma_i\beta+d}}$ can be viewed as the effective sample size for data drawn from the source distribution P_i when the information in this sample is transferred to help the classification task under the target distribution Q. Even when there are multiple

source distributions, the transfer rate associated with P_i remains to be $\frac{2\beta+d}{2\gamma_i\beta+d}$, which is not affected by the presence of the data drawn from the other source distributions.

5.2. *Adaptive classifier.* Again, the minimax classifier is not practical as it depends on the parameters $\boldsymbol{\gamma}$ and μ which are typically unknown. It is desirable to construct a data driven classifier that does not rely on the knowledge of the model parameters. A similar adaptive data-driven classifier can be developed. The detailed steps are summarized in Algorithm 2.

It is clear from the construction that the classifier $\hat{f}_a$ is a data-driven decision rule. Theorem 7 below provides a theoretical guarantee for the excess risk of $\hat{f}_a$ under the target distri-

Algorithm 2 The data driven classifier

Input: $x \in \operatorname{supp}(Q_X)$.

for $k = 1, \ldots, (n_Q + \sum_{i=1}^m n_{P_i} - 1), (n_Q + \sum_{i=1}^m n_{P_i})$ **do**

Find k nearest neighbors $X_{(1)}(x), \ldots, X_{(k)}(x)$ to x among all the covariates $\{X_j^Q : j \in [n_Q]\} \cup \bigcup_{i=1}^m \{X_j^{P_i} : j \in [n_{P_i}]\}$. Suppose $k_{P_i}^{(k)}$ of them are from the distribution P_i, $i = 1, \ldots, m$, and $k_Q^{(k)}$ of them are from Q. That is, the k nearest neighbors are partitioned into $m+1$ parts according to which distribution they are drawn from.

For each $i \in [m]$, Compute the K-NN estimate for η_{P_i} (if $k_{P_i}^{(k)} = 0$, set $\hat{\eta}_{P_i}^{(k)} \leftarrow \frac{1}{2}$)

$$\hat{\eta}_{P_i}^{(k)}(x) \leftarrow \frac{1}{k_{P_i}^{(k)}} \sum_{j=1}^{k_{P_i}^{(k)}} Y_{(j)}^{P_i}(x)$$

and nearest neighbor estimate for η_Q (if $k_Q^{(k)} = 0$, set $\hat{\eta}_Q^{(k)} \leftarrow \frac{1}{2}$)

$$\hat{\eta}_Q^{(k)} \leftarrow \frac{1}{k_Q^{(k)}} \sum_{i=1}^{k_Q^{(k)}} Y_{(i)}^{P}(x).$$

Compute the positive signal-to-noise index

$$\hat{r}_+^{(k)} \leftarrow \mathbb{I}_{\{\eta_Q^{(k)} \geq \frac{1}{2}\}} k_Q^{(k)} \left(\eta_Q^{(k)} - \frac{1}{2}\right)^2 + \sum_{i=1}^m \mathbb{I}_{\{\eta_{P_i}^{(k)} \geq \frac{1}{2}\}} k_{P_i}^{(k)} \left(\eta_{P_i}^{(k)} - \frac{1}{2}\right)^2$$

and negative signal-to-noise index

$$\hat{r}_-^{(k)} \leftarrow \mathbb{I}_{\{\eta_Q^{(k)} < \frac{1}{2}\}} k_Q^{(k)} \left(\eta_Q^{(k)} - \frac{1}{2}\right)^2 + \sum_{i=1}^m \mathbb{I}_{\{\eta_{P_i}^{(k)} < \frac{1}{2}\}} k_{P_i}^{(k)} \left(\eta_{P_i}^{(k)} - \frac{1}{2}\right)^2.$$

Let $\hat{r}^{(k)}$ be the signal-to-noise ratio index calculated by

$$\hat{r}^{(k)} \leftarrow \max\{\hat{r}_+^{(k)}, \hat{r}_-^{(k)}\}.$$

Define the classifier

$$\hat{f}^{(k)}(x) \leftarrow \mathbb{I}_{\{\hat{r}_+^{(k)} \geq \hat{r}_-^{(k)}\}}.$$

if $\hat{r}^{(k)} > (d+3)\log(n_Q + \sum_{i=1}^m n_{P_i})$ **then**

Stop and output $\hat{f}_a(x) \leftarrow f^{(k)}(x)$.

Output $\hat{f}_a(x) \leftarrow \hat{f}^{(k_m)}(x)$ where $k_m = \operatorname{argmax}_k \hat{r}^{(k)}$.

bution Q. In view of the optimal rate given in (15), Theorem 7 shows that $\hat{f}_a$ is adaptively nearly optimal over a wide range of parameter spaces.

THEOREM 7. *Let $n = n_Q + \sum_{i=1}^m n_{P_i}$. There exists a constant $C > 0$ such that for $\Pi = \Pi(\alpha, \beta, \boldsymbol{\gamma}, \mu)$,*

$$\sup_{(P_1,\ldots,P_m,Q)\in\Pi} \mathbb{E}\mathcal{E}_Q(\hat{f}_a) \le C \cdot \left(\frac{n_Q}{\log n} + \sum_{i=1}^m \left(\frac{n_{P_i}}{\log n} \right)^{\frac{2\beta+d}{2\gamma_i\beta+d}} \right)^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

6. Numerical studies. In this section, we carry out simulation studies to further illustrate the performance of the adaptive transfer learning procedure. Numerical comparisons with the existing methods are given. The simulation results are consistent with the theoretical predictions.

For all simulation experiments in this section, the data is generated under the posterior drift model with $d = 2$. The distributions (P, Q) used to generate data is specified as following:

1. Marginal distributions: $P_X = Q_X$ are both uniform distribution on the square $\Omega = [-1, 1]^2$.
2. Regression functions: η_Q and η_P are defined as

$$\eta_Q(x) = 0.5 + p\,\mathrm{sign}(x_1)\big(|x_1|\max\{0, 1 - |x_2|\}\big)^\beta$$

and

$$\eta_P(x) = 0.5 + p\,\mathrm{sign}(x_1)\big(|x_1|\max\{0, 1 - |x_2|\}\big)^{\gamma\beta},$$

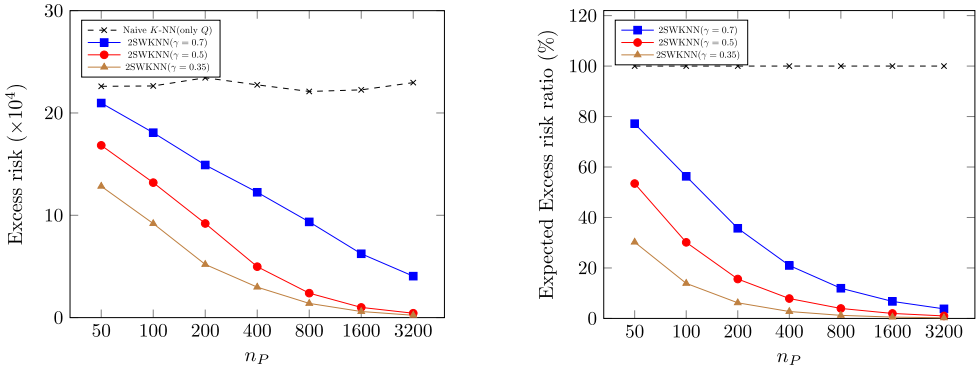
where $x = (x_1, x_2) \in [-1, 1]^2$, p, β and γ are parameters that may vary in different simulation studies.

According to the above construction, both η_P and η_Q take the maximum values at $(1, 0)$ and the minimum values at $(-1, 0)$, and equal to 0.5 when $x_1 = 0$. it can be easily verified that $\eta_Q \in \mathcal{H}(\beta, C_\beta)$ with some $C_\beta > 0$, $(P, Q) \in \Gamma(\gamma, 1)$, Q satisfies the margin assumption with $\alpha = 0.99/\beta$, and P_X and Q_X have the common support and bounded densities.

In the following experiments, we focus on evaluating the average excess risk at a random test sample x drawn uniformly from the square $\Omega = [-1, 1]^2$, given n_P data generated from P and n_Q data generated from Q.

6.1. *Minimax nonadaptive classifier.* For this particular distribution pair (P, Q), theoretically, the minimax rate of convergence for the excess risk can be achieved via the two-sample weighted K-NN classifier when we are able to make use of model parameters β, γ. In the following simulation, we fix $p = 0.03$, $n_Q = 1000$ and $\beta = 1$. By comparing the proposed nonadaptive classifier with a naive K-NN classifier on just the Q-data, we evaluate the improvement on the excess risk under different values of γ and n_P.

During the experiment, we generated datasets with choices of the relative signal exponent $\gamma \in \{0.7, 0.5, 0.35\}$ and number of P-data n_P varying from 50 to 3200. The excess risk of the two-sample weighted K-NN classifier and the naive K-NN method are illustrated in Figure 4(a). Meanwhile, a planer plot is given in Figure 4(b) to illustrate the expected ratio of the excess risk between the two methods based on our developed theory (Theorem 1). One can clearly see how the transfer rates play a role in the experiments with different relative signal exponent γ. The empirical performance and our theoretical prediction are matched to some extent.

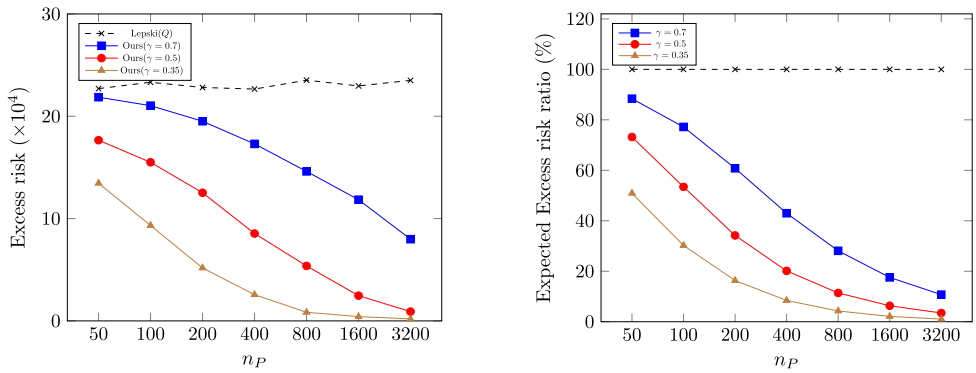


(a) Experimental results (b) Theoretical prediction

FIG. 4. *Left*: *Experiments on nonadaptive methods. We operate the naive K-NN method on only Q-data (dashed line) and our two-sample weighted K-NN classifier on different datasets. The datasets are generated with relative signal exponent* $\gamma = 0.7, 0.5, 0.35$, *respectively. Right*: *based on our theory* (*Theorem* 1), *the expected ratio of excess risk between the two methods we operate in the experiment.*

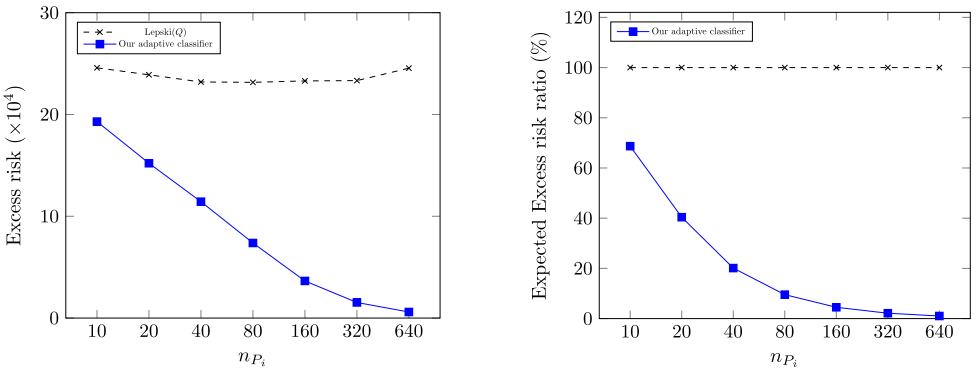
6.2. *Adaptive classifier.* We also compare the proposed adaptive classifier with the existing methods to see whether its numerical performance matches its theoretical guarantees. Lepski's method is a good competitor as it is also adaptive to the smoothness parameter β. Following a similar routine as in the previous experiments, we compare the excess risk between our proposed classifier and Lepski's method applying only the Q-data to evaluate the improvement we may gain empirically.

Fix $p = 0.03$ and $\beta = 1$; we generated $n_Q = 1000$ data from the target distribution Q, and $n_P \in \{50, 100, 200, 400, 800, 1600, 3200\}$ data from the source distribution P with different choices of relative signal exponent $\gamma \in \{0.7, 0.5, 0.35\}$. Results of the numerical experiments are shown in Figure 5(a). A figure of the expected improvement on excess risk, calculated according to Theorem 3, is also available in Figure 5(b). In both figures, the curve looks like a reversed "S" shape when γ is large, whereas a curve of exponential decrease appears when γ is small. Therefore, it is justified that the simulation results are consistent with the theoretical predictions.



(a) Experimental results (b) Theoretical prediction

FIG. 5. *Left*: *Experiments on adaptive methods. We operate the naive Lepski method on only Q-data (dashed line) and our adaptive classifier on different datasets. The datasets are generated with relative signal exponent* $\gamma = 0.7, 0.5, 0.35$, *respectively. Right*: *based on our theory* (*Theorem* 3), *the expected ratio of excess risk between the two methods used in the experiment.*



(a) Experimental results (b) Theoretical prediction

FIG. 6. *Left*: *Experiments on transfer learning from multiple source distributions. We apply the naive Lepski method on only Q-data* (*dashed line*) *and our adaptive classifier for multiple source distributions. Right*: *based on our theory* (*Theorem* 5), *the expected ratio of excess risk between the two methods we operate in the experiment.*

6.3. *Multiple source distributions.* Other than involving only a single source distribution during the previous numerical studies, it is also worthwhile to see whether we can gain desired improvement as our theory predicts when there are multiple source distributions. We only illustrate in this subsection the performance of our adaptive classifier applying to multiple source distributions (Algorithm 2).

Different from the previous simulation studies, in this subsection we generate data from three different source distributions P_1, P_2, P_3 and one target distribution Q. In a similar vein, the distributions (P_1, P_2, P_3, Q) are specified as following:

1. Marginal distributions: we set $P_{1,X} = P_{2,X} = P_{3,X} = Q_X$ to be all uniformly distributed on the square area $\Omega = [-1, 1]^2$.
2. Regression functions: we set η_Q and $\eta_{P_1}, \eta_{P_2}, \eta_{P_3}$ as

$$\eta_Q(x) = 0.5 + p\,\mathrm{sign}(x_1)\big(|x_1|\max\{0, 1 - |x_2|\}\big)^{\beta}$$

and

$$\eta_{P_i}(x) = 0.5 + p\,\mathrm{sign}(x_1)\big(|x_1|\max\{0, 1 - |x_2|\}\big)^{\gamma_i\beta} \quad i = 1, 2, 3,$$

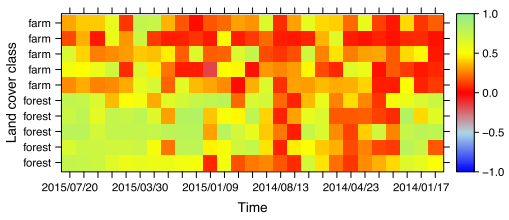
where $x = (x_1, x_2) \in [-1, 1]^2$, p, β and $\gamma_1, \gamma_2, \gamma_3$ are parameters that will be specified later.

In the simulation, we fix $p = 0.03$, $\beta = 1$ and $\gamma_1 = 0.35$, $\gamma_2 = 0.5$, $\gamma_3 = 0.7$, and we always set $n_{P_1} = n_{P_2} = n_{P_3}$. We compare the average excess risk of the two classifiers: our proposed adaptive classifier and the Lepski's procedure with only Q-data involved. By varying number of data drawing from source distributions, we can clearly see an improvement when applying transfer learning methods.

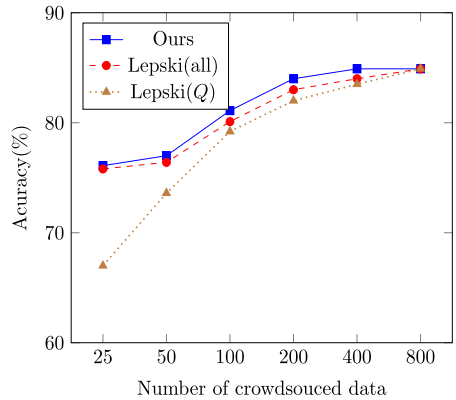
The excess risk of the two methods during the experiments are illustrated in Figure 6(a). Also, we calculate the expected ratio between the two methods according to the theory we developed in Theorem 5. Again, the empirical performance and our theoretical prediction are similar to some extent.

For reasons of space, additional simulation results on different choices of β are given in the Supplementary Material (Cai and Wei (2020)).

7. Application to crowdsourced mapping data. To illustrate the proposed adaptive classifier, we consider in this section an application based on the crowdsourced mapping data (Johnson and Iizuka (2016)). Land use/land cover maps derived from remotely-sensed



(a) Illustration of the dataset.



(b) Accuracy of different classifiers.

FIG. 7. (a) *Illustration of the dataset. Each row represents one of a land cover class* (*farm or forest*) *and corresponding NDVI values of a pixel from remotely-sensed imagery in* 2014–2015. (b) *Accuracy of the three methods on the crowdsourced mapping data with different numbers of crowdsourced data involved. Blue*: *The proposed adaptive classifier. Red*: *Lepski's method using combined data. Brown*: *Lepski's method using only crowdsourced data.*

imagery are important for geographic studies. This dataset contains Landsat time-series satellite imagery information on given pixels and their corresponding land cover class labels (farm, forest, water, etc.) obtained from multiple sources. The goal is to make classification of land cover classes based on NDVI (normalized difference vegetation index) values of those remotely-sensed imagery from the years 2014–2015. In this paper, we focus on classification of two specific classes: farm and forest.

Within this dataset, there are two kinds of label sources, given the NVDI values of the images: (1) crowdsourced georeferenced polygons with land cover labels obtained from OpenStreetMap; (2) accurately labeled data by experts. Although crowdsourced data are massive, free and public, the labels contain various types of errors due to user mislabels or outdated images. Whereas the expert labels are almost accurate, but they are usually too expensive to obtain a large volume. The challenge is to accurately combine the information contained in the two datasets to minimize the classification error.

As in Section 6.2, we apply three methods to make the classification: (1) our proposed adaptive procedure; (2) Lepski's method with all data involved where we do not distinguish data from different sources; (3) Lepski's method with only the crowdsourced data. We use $n_P = 50$ accurately labeled data, and change the number of involved crowdsourced data from $n_Q = 25$ to $n_Q = 800$. We use other 166 accurately labeled data to evaluate the classification accuracy of the three methods mentioned above.

Figure 7(b) shows the accuracy of the three methods with different numbers of crowdsourced data involved. As more and more crowdsourced data are used, the amount of information contained in the crowdsourced data gradually increases, and the relative contribution from the accurately labeled data gradually decreases. The proposed adaptive classifier consistently outperforms the naive Lepski's method, especially when the number of the crowdsourced data is between 100 and 400, because in these cases the adaptive classifier can significantly increase the accuracy by better leveraging the information gained from both distributions.

8. Discussion. We studied in this paper transfer learning under the posterior drift model and established the minimax rate of convergence. The optimal rate quantifies precisely the amount of information in the P-data that can be transferred to help classification under the target distribution Q. A delicately designed data-driven adaptive classifier was also constructed

and shown to be, both globally and locally, adaptive to the unknown smoothness and relative signal exponent. It is simultaneously within a log factor of the optimal rate over a large collection of parameter spaces.

The results and techniques developed in this paper serve as a starting point for the theoretical analysis of other transfer learning problems. For example, in addition to classification, it is also of significant interest to characterize the relationship between the source distribution and the target distribution, so that the data from the source distribution P can help in other statistical problems under the target distribution Q. Examples include regression, hypothesis testing and construction of confidence sets. We will investigate these transfer learning problems in the future.

Within the posterior drift framework of this paper, some of the technical assumptions can be relaxed to a certain extent. For the smoothness parameter β, we focused on the case $0 < \beta \leq 1$. It is possible to consider more general classes where β can be larger than 1, with strengthened relative signal exponent assumptions on the higher order derivatives of $\eta_P(x)$ and $\eta_Q(x)$. When $\beta > 2$, the problem might be solved with a carefully designed weighted K-NN classifier, as was introduced in Samworth (2012). Construction of such a weighted K-NN method is involved and we leave it as future work. For the marginal distributions P_X and Q_X, other than the strong density assumption, there are also weaker regularity conditions introduced in the literature; see, for example, Gadat, Klein and Marteau (2016), Kpotufe and Martinet (2018). Similar results on the minimax rate of convergence can be established under these different regularity conditions. The minimax and adaptive procedures should also be suitably modified.

Also, in complementary work, Kpotufe and Martinet (2018) studied K-NN classifiers for transfer learning in the *covariate shift* framework where the marginal distributions P_X and Q_X are allowed to differ significantly. It is interesting to consider nonparametric classification under both *covariate shift* and *posterior drift*. In such a setting, besides the relative signal exponent γ, one also assumes (P, Q) have *transfer-exponent* $\tau \geq 0$ such that

$$\forall x, r \in (0, \Delta_{\mathcal{X}}], \quad P_X\big(B(x,r)\big) \geq Q_X\big(B(x,r)\big) \cdot C_\tau \left(\frac{r}{\Delta_{\mathcal{X}}}\right)^\tau,$$

and Q_X is (C_d, d)-doubling, as is defined in Definitions 3 and 6 in Kpotufe and Martinet (2018). The detailed analysis appears to be quite involved, we only make some conjectures here based on our preliminary calculations and leave the rigorous proofs and further investigations for future work. Our calculations indicate that the optimal rate of convergence for the excess risk on Q under both covariate shift (transfer-exponent τ) and posterior drift (relative signal exponent γ) should be

$$\inf_{\hat{f}} \sup_{(P,Q)} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}) \asymp \big(n_P^{\frac{2\beta+d}{2\gamma\beta+\tau+d}} + n_Q\big)^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

An additional transfer-exponent τ appears in the denominator of the transfer rate $\frac{2\beta+d}{2\gamma\beta+\tau+d}$. The above optimal rate can be achieved by two-sample weighted K-NN classifier (proposed in our work) with proper choices of w_P, w_Q, k_P and k_Q. In addition, our proposed classifier $\hat{f}_a$ should be nearly optimal adaptive classifier (up to a logrithmic term) in a sense that

$$\sup_{(P,Q)} \mathbb{E}_Z \mathcal{E}_Q(\hat{f}) \lesssim \left(\left(\frac{n_P}{\log n}\right)^{\frac{2\beta+d}{2\gamma\beta+\tau+d}} + \frac{n_Q}{\log n}\right)^{-\frac{\beta(1+\alpha)}{2\beta+d}},$$

where $n = n_P + n_Q$.

Algorithm 3 Lepski's method (Lepski and Spokoiny (1997))

Input: n labeled samples $(X_i, Y_i) \in \mathbb{R}^d \times \{0,1\}$, $i \in [n]$, and a point $x \in \mathbb{R}^d$ to be classified.
Set $\eta_0^- \leftarrow -\infty$ and $\eta_0^+ \leftarrow +\infty$.
for $k = 1, \ldots, (n_P + n_Q - 1), (n_P + n_Q)$ **do**
Find k nearest neighbor estimates $\hat{\eta}_k(x) = \frac{1}{k}\sum_{i=1}^k Y_{(i)}$, where $Y_{(i)}$ denote the label to ith nearest covariates to x.
Set $\eta_k^- \leftarrow \eta_{k-1}^- \vee (\hat{\eta}_k(x) - \sqrt{\frac{d+3}{k}}\log n)$.
Set $\eta_k^+ \leftarrow \eta_{k-1}^+ \wedge (\hat{\eta}_k(x) + \sqrt{\frac{d+3}{k}}\log n)$.
if $\eta_k^- > \frac{1}{2}$ or $\eta_k^+ < \frac{1}{2}$ **then**
Stop and output $\hat{f}_L(x) \leftarrow \mathbb{I}_{\{\hat{\eta}_k(x) \geq \frac{1}{2}\}}$.
Output $\hat{f}_L(x) \leftarrow \mathbb{I}_{\{\hat{\eta}_n(x) \geq \frac{1}{2}\}}$.

9. Proofs. We prove Theorem 1 in this section and leave the proofs of other theorems and additional technical lemmas in the Supplementary Material (Cai and Wei (2020)). For readers' convenience, we begin by stating Lepski's method for nonparametric classification in the conventional setting where there are only the Q-data.

9.1. *Lepski's method.* Algorithm 3 is a version of Lepski's method in nonparametric classification. We state the algorithm here for reference.

9.2. *Proof of Theorem 1.* First, we define some new notation for convenience. In the proof, we use $\zeta_Q(x) = |\eta_Q(x) - \frac{1}{2}|$ and $\zeta_P(x) = |\eta_P(x) - \frac{1}{2}|$ to denote the signal strength. Let $\bar{Y}^Q_{(1:k_Q)}(x) := \frac{1}{k_Q}\sum_{i=1}^{k_Q} Y^Q_{(i)}(x)$ and $\bar{Y}^P_{(1:k_P)}(x) := \frac{1}{k_P}\sum_{i=1}^{k_P} Y^P_{(i)}(x)$ denote the average of k_Q nearest neighbors in Q-data and k_P nearest neighbors in P-data, respectively. We will sometime omit x in the notation such as $X^Q_{(i)}(x)$, $X^P_{(i)}(x)$ if there is no confusion in the context. We also use the shorthand $X^Q_{1:n_Q}$ to denote the whole set of the Q-data covariates $\{X^Q_1, \ldots, X^Q_{n_Q}\}$, and similarly, $X^P_{1:n_P}$ denotes $\{X^P_1, \ldots, X^P_{n_P}\}$. We define $\mathbb{E}_{Y|X}(\cdot) = \mathbb{E}(\cdot|X^Q_{1:n_Q}, X^P_{1:n_P})$ to denote the expectation conditional on the covariates of all data, and $\mathbb{E}$ is the expectation taken over random realization of all data (the same as $\mathbb{E}_Z$ we defined before). Also, in the following proofs we always assume $(P, Q) \in \Pi(\alpha, \beta, \gamma, \mu)$ so we will not state this assumption again in the lemmas.

Before proving the theorem, we first state three useful lemmas. The first Lemma 1 provides a high probability uniform bound on the distance between any point and its kth nearest neighbor.

LEMMA 1 (K-NN distance bound). *There exists a constant $C_D > 0$ such that, with probability at least $1 - C_D\frac{n_Q}{k_Q}\exp(-\frac{k_Q}{6})$, for all $x \in \Omega$,*

$$\|X^Q_{(k_Q)}(x) - x\| \leq C_D\left(\frac{k_Q}{n_Q}\right)^{\frac{1}{d}}. \tag{16}$$

And with probability at least $1 - C_D\frac{n_P}{k_P}\exp(-\frac{k_P}{6})$, for all $x \in \Omega$,

$$\|X^P_{(k_P)}(x) - x\| \leq C_D\left(\frac{k_P}{n_P}\right)^{\frac{1}{d}}. \tag{17}$$

Let E_Q denote the event that inequality (16) holds for all $x \in \Omega$ and let E_P denote (17) holds for all $x \in \Omega$. It follows from Lemma 1 that

$$\mathbb{P}(E_Q) \geq 1 - C_D \frac{n_Q}{k_Q} \exp\left(-\frac{k_Q}{6}\right) \quad \text{and} \quad \mathbb{P}(E_P) \geq 1 - C_D \frac{n_P}{k_P} \exp\left(-\frac{k_P}{6}\right).$$

Lemma 2 points out that when the signal is sufficiently strong, bias of $\bar{Y}^Q(x)$ and $\bar{Y}^P(x)$ will not be too large to overwhelm the signal.

LEMMA 2 (Bias bound). *There exist constants $c_b, C_b > 0$ such that:*
If a point $x \in \Omega$ satisfies $\zeta_Q(x) \geq 2C_\beta \|X^Q_{(k_Q)}(x) - x\|^\beta$, then we have

$$\mathbb{E}_{Y|X}\big(\bar{Y}^Q_{(1:k_Q)}(x)\big) - \frac{1}{2} \geq c_b \zeta_Q(x) \quad \text{if } f^*(x) = 1, \tag{18}$$

$$\mathbb{E}_{Y|X}\big(\bar{Y}^Q_{(1:k_Q)}(x)\big) - \frac{1}{2} \leq -c_b \zeta_Q(x) \quad \text{if } f^*(x) = 0. \tag{19}$$

If a point $x \in \Omega$ satisfies $\zeta_Q(x) \geq 2C_\beta \|X^P_{(k_P)}(x) - x\|^\beta$, then we have

$$\mathbb{E}_{Y|X}\big(\bar{Y}^P_{(1:k_P)}(x)\big) - \frac{1}{2} \geq c_b \zeta_Q(x)^\gamma \quad \text{if } f^*(x) = 1, \tag{20}$$

$$\mathbb{E}_{Y|X}\big(\bar{Y}^P_{(1:k_P)}(x)\big) - \frac{1}{2} \leq -c_b \zeta_Q(x)^\gamma \quad \text{if } f^*(x) = 0. \tag{21}$$

Hence, if a point $x \in \Omega$ satisfies $\zeta_Q(x) \geq C_b(\max\{\frac{k_Q}{n_Q}, \frac{k_P}{n_P}\})^{\frac{\beta}{d}}$, then:

- *Under the event E_Q, we have*

$$\mathbb{E}_{Y|X}\big(\bar{Y}^Q_{(1:k_Q)}(x)\big) - \frac{1}{2} \geq c_b \zeta_Q(x) \quad \text{if } f^*(x) = 1, \tag{22}$$

$$\mathbb{E}_{Y|X}\big(\bar{Y}^Q_{(1:k_Q)}(x)\big) - \frac{1}{2} \leq -c_b \zeta_Q(x) \quad \text{if } f^*(x) = 0. \tag{23}$$

- *Under the event E_P, we have*

$$\mathbb{E}_{Y|X}\big(\bar{Y}^P_{(1:k_P)}(x)\big) - \frac{1}{2} \geq c_b \zeta_Q(x)^\gamma \quad \text{if } f^*(x) = 1, \tag{24}$$

$$\mathbb{E}_{Y|X}\big(\bar{Y}^P_{(1:k_P)}(x)\big) - \frac{1}{2} \leq -c_b \zeta_Q(x)^\gamma \quad \text{if } f^*(x) = 0. \tag{25}$$

Lemma 3 gives a bound on the probability of misclassification at certain covariates x.

LEMMA 3 (Misclassification bound). *Let C_b and c_b be the constants defined in Lemma 2. If $\zeta_Q(x) \geq C_b(\max\{\frac{k_Q}{n_Q}, \frac{k_P}{n_P}\})^{\frac{\beta}{d}}$, then:*

- *Under the event E_Q, we have*

$$\mathbb{P}_{Y|X}\big(\hat{f}_{\mathrm{NN}}(x) \neq f^*_Q(x)\big) \leq \exp\left(-2\frac{[(c_b w_Q k_Q \zeta_Q(x) - w_P k_P) \vee 0]^2}{k_P w_P^2 + k_Q w_Q^2}\right).$$

- *Under the event E_P, we have*

$$\mathbb{P}_{Y|X}\big(\hat{f}_{\mathrm{NN}}(x) \neq f^*_Q(x)\big) \leq \exp\left(-2\frac{[(c_b w_P k_P \zeta_Q(x)^\gamma - w_Q k_Q) \vee 0]^2}{k_P w_P^2 + k_Q w_Q^2}\right).$$

- *Under the event $E_P \cap \mathbb{E}_Q$, we have*

$$\mathbb{P}_{Y|X}\big(\hat{f}_{\mathrm{NN}}(x) \neq f_Q^*(x)\big) \leq \exp\bigg(-2c_b^2 \frac{(w_P k_P \zeta_Q(x)^\gamma + w_Q k_Q \zeta_Q(x))^2}{k_P w_P^2 + k_Q w_Q^2}\bigg).$$

Given the three lemmas above, the remain proof generally follows that of Lemma 3.1 in Audibert and Tsybakov (2007). Let $\delta = (n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q)^{-\frac{\beta}{2\beta+d}}$. When w_P, w_Q, k_P, k_Q are given as in Theorem 1, we have

$$w_Q = \delta, \qquad w_P = \delta^\gamma, \qquad k_Q = \lfloor n_Q \delta^{\frac{d}{\beta}} \rfloor, \qquad k_P = \lfloor n_P \delta^{\frac{d}{\beta}} \rfloor. \tag{26}$$

We will approximate $k_Q = n_Q \delta^{\frac{d}{\beta}}$ and $k_P = n_P \delta^{\frac{d}{\beta}}$ in the following proof because one can easily show such an approximation only results in changing the constant factor in the upper bound.

The following lemma gives a bound for the local misclassification risk when the parameters in the weighted K-NN estimator are properly chosen.

LEMMA 4. *Using w_P, w_Q, k_P, k_Q defined in Theorem 1 to construct a weighted K-NN estimator $\hat{f}_{\mathrm{NN}}$. Then there exist constants $c_1, C_1 > 0$ such that, with probability at least $1 - 2(n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q)^{-\frac{\beta(1+\alpha)}{2\beta+d}}$, for all x we have*

$$\mathbb{P}_{Y|X}\big(\hat{f}_{\mathrm{NN}}(x) \neq f_Q^*(x)\big) \leq C_1 \exp\bigg(-c_1 \bigg(\frac{\zeta_Q(x)}{\delta}\bigg)^{1\wedge\gamma}\bigg). \tag{27}$$

Let E_0 be the event that inequality (27) holds for all x. Lemma 4 implies

$$\mathbb{P}(E_0) \geq 1 - 2\big(n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q\big)^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

Consider the disjoint sets $\mathcal{A}_j \subset \Omega$, $j = 0, 1, 2, \ldots$ defined as

$$\begin{aligned} \mathcal{A}_0 &:= \{x \in \Omega : 0 < \zeta_Q(x) \leq \delta\}, \\ \mathcal{A}_j &:= \{x \in \Omega : 2^{j-1}\delta < \zeta_Q(x) \leq 2^j \delta\} \quad \text{for } j \geq 1. \end{aligned}$$

Note that by the margin assumption, for all j,

$$Q_X(A_j) \leq Q_X\bigg(\bigg|\eta_Q - \frac{1}{2}\bigg| \leq 2^j\delta\bigg) \leq C_\alpha 2^{\alpha j}\delta^\alpha.$$

Based on the partition $A_0, A_1, \ldots$ and the dual representation of $\mathcal{E}_Q(\hat{f})$ shown in (8), we have a decomposition of $\mathcal{E}_Q(\hat{f}_{\mathrm{NN}})$:

$$\begin{aligned} \mathcal{E}_Q(\hat{f}_{\mathrm{NN}}) &= 2\mathbb{E}_{X\sim Q_X}\bigg(\bigg|\eta_Q(X) - \frac{1}{2}\bigg|\mathbb{I}_{\{\hat{f}_{\mathrm{NN}}(X) \neq f_Q^*(X)\}}\bigg) \\ &= 2\sum_{j=0}^{\infty} \mathbb{E}_{X\sim Q_X}\big(\zeta_Q(X)\mathbb{I}_{\{\hat{f}_{\mathrm{NN}}(X) \neq f_Q^*(X)\}}\mathbb{I}_{\{X\in A_j\}}\big). \end{aligned}$$

For $j = 0$, $\mathbb{E}_{X\sim Q_X}(\zeta_Q(X)\mathbb{I}_{\{\hat{f}_{\mathrm{NN}}(X) \neq f_Q^*(X)\}}\mathbb{I}_{\{X\in A_0\}}) \leq \delta \cdot Q_X(A_0) \leq C_\alpha \delta^{\alpha+1}$.

Under the event E_0, $2^{j-1}\delta < \zeta(x) \le 2^j\delta$ for $x \in A_j$ and $j > 1$. Inequality (27) now yields

$$\begin{aligned}
&\mathbb{E}_{Y|X}\mathbb{E}_{X\sim Q_X}\big(\zeta_Q(X)\mathbb{I}_{\{\hat{f}_{\mathrm{NN}}(X)\neq f_Q^*(X)\}}\mathbb{I}_{\{X\in A_j\}}\big)\\
&\quad=\mathbb{E}_{X\sim Q_X}\big(\zeta_Q(X)\mathbb{P}_{Y|X}\big(\hat{f}_{\mathrm{NN}}(X)\neq f_Q^*(X)\big)\mathbb{I}_{\{X\in A_j\}}\big)\\
&\quad\le 2^j\delta\cdot C_1\exp\big(-c_1\cdot 2^{(j-1)\cdot(1\wedge\gamma)}\big)\cdot Q_X(A_j)\\
&\quad\le C_\alpha C_1\big[2^{(1+\alpha)j}\exp\big(-c_1\cdot 2^{(j-1)\cdot(1\wedge\gamma)}\big)\big]\delta^{\alpha+1}.
\end{aligned}$$

Combining these summands together yields

$$\begin{aligned}
\mathbb{E}_{Y|X}\mathcal{E}_Q(\hat{f}_{\mathrm{NN}}) &= 2\sum_{j=0}^{\infty}\mathbb{E}_{Y|X}\mathbb{E}_{X\sim Q_X}\big(\zeta_Q(X)\mathbb{I}_{\{\hat{f}_{\mathrm{NN}}(X)\neq f_Q^*(X)\}}\mathbb{I}_{\{X\in A_j\}}\big)\\
&\le 2C_\alpha\left(1+C_1\sum_{j=0}^{\infty}\big[2^{(1+\alpha)j}\exp\big(-c_1\cdot 2^{(k-1)\cdot(1\wedge\gamma)}\big)\big]\right)\delta^{1+\alpha}\\
&\le C\delta^{1+\alpha},
\end{aligned}$$

where the last step follows from the fact that $\sum_{j=0}^{\infty}[2^{(1+\alpha)j}\exp(-c_1\cdot 2^{(k-1)\cdot(1\wedge\gamma)})]$ converges when $\gamma > 0$. Finally, it follows from Lemma 4 that

$$\mathbb{P}\big(E_0^c\big) \le 2\big(n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q\big)^{-\frac{\beta(1+\alpha)}{2\beta+d}}.$$

Applying the trivial bound $\mathcal{E}_Q(\hat{f}_{\mathrm{NN}}) \le 1$ when E_0^c occurs, we have

$$\begin{aligned}
\mathbb{E}\mathcal{E}_Q(\hat{f}_{\mathrm{NN}}) &= \mathbb{E}\big(\mathbb{E}_{Y|X}\mathcal{E}_Q(\hat{f}_{\mathrm{NN}})\big)\\
&\le \mathbb{E}\big(\mathbb{E}_{Y|X}\mathcal{E}_Q(\hat{f}_{\mathrm{NN}})|E_0\big)\mathbb{P}(E_0) + \mathbb{E}\big(\mathbb{E}_{Y|X}\mathcal{E}_Q(\hat{f}_{\mathrm{NN}})|E_0^c\big)\mathbb{P}\big(E_0^c\big)\\
&\le C\big(n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q\big)^{-\frac{\beta(1+\alpha)}{2\beta+d}}\cdot 1 + 1\cdot 2\big(n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q\big)^{-\frac{\beta(1+\alpha)}{2\beta+d}}\\
&= (C+2)\big(n_P^{\frac{2\beta+d}{2\gamma\beta+d}} + n_Q\big)^{-\frac{\beta(1+\alpha)}{2\beta+d}}.
\end{aligned}$$

□

Acknowledgments. The research was supported in part by NSF Grant DMS-1712735 and NIH Grants R01-GM129781 and R01-GM123056.

SUPPLEMENTARY MATERIAL

Supplement to "Transfer learning for nonparametric classification: Minimax rate and adaptive classifier" (DOI: 10.1214/20-AOS1949SUPP; .pdf). In this supplementary material, we provide additional simulation results, proofs for Theorems 2, 3 and 4, and proofs for technical Lemmas 1, 2, 3 and 4. The proofs of Theorems 5, 6 and 7 are similar, and thus omitted.

REFERENCES

AUDIBERT, J.-Y. and TSYBAKOV, A. B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.* **35** 608–633. MR2336861 https://doi.org/10.1214/009053606000001217

BEN-DAVID, S., BLITZER, J., CRAMMER, K. and PEREIRA, F. (2007). Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems* 137–144.

BLITZER, J., CRAMMER, K., KULESZA, A., PEREIRA, F. and WORTMAN, J. (2008). Learning bounds for domain adaptation. In *Advances in Neural Information Processing Systems* 129–136.

CAI, T. T. and WEI, H. (2020). Supplement to "Transfer learning for nonparametric classification: Minimax rate and adaptive classifier." https://doi.org/10.1214/20-AOS1949SUPP

CHOI, K., FAZEKAS, G., SANDLER, M. and CHO, K. (2017). Transfer learning for music classification and regression tasks. Preprint. Availabe at arXiv:1703.09179.

COVER, T. M. and HART, P. E. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13** 21–27.

CRAMMER, K., KEARNS, M. and WORTMAN, J. (2006). Learning from data of variable quality. In *Advances in Neural Information Processing Systems* 219–226.

GADAT, S., KLEIN, T. and MARTEAU, C. (2016). Classification in general finite dimensional spaces with the k-nearest neighbor rule. *Ann. Statist.* **44** 982–1009. MR3485951 https://doi.org/10.1214/15-AOS1395

GAMA, J., ŽLIOBAITĖ, I., BIFET, A., PECHENIZKIY, M. and BOUCHACHIA, A. (2014). A survey on concept drift adaptation. *ACM Comput. Surv.* **46** Art. ID 44.

GONG, B., SHI, Y., SHA, F. and GRAUMAN, K. (2012). Geodesic flow kernel for unsupervised domain adaptation. In 2012 *IEEE Conference on Computer Vision and Pattern Recognition* 2066–2073. IEEE, Piscataway, NJ.

GYŐRFI, L. (1978). On the rate of convergence of nearest neighbor rules. *IEEE Trans. Inf. Theory* **24** 509–512. MR0501595 https://doi.org/10.1109/TIT.1978.1055898

HUANG, J.-T., LI, J., YU, D., DENG, L. and GONG, Y. (2013). Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In 2013 *IEEE International Conference on Acoustics, Speech and Signal Processing* 7304–7308. IEEE, Piscataway, NJ.

JOHNSON, B. A. and IIZUKA, K. (2016). Integrating OpenStreetMap crowdsourced data and Landsat time-series imagery for rapid land use/land cover (LULC) mapping: Case study of the Laguna de Bay area of the Philippines. *Appl. Geogr.* **67** 140–149.

KARGER, D. R., OH, S. and SHAH, D. (2011). Budget-optimal crowdsourcing using low-rank matrix approximations. In 2011 49*th Annual Allerton Conference on Communication, Control, and Computing* (*Allerton*) 284–291. IEEE, Piscataway, NJ.

KPOTUFE, S. and MARTINET, G. (2018). Marginal singularity, and the benefits of labels in covariate-shift. Preprint. Availabe at arXiv:1803.01833.

LEE, S.-I., CHATALBASHEV, V., VICKREY, D. and KOLLER, D. (2007). Learning a meta-level prior for feature relevance from multiple related tasks. In *Proceedings of the* 24*th International Conference on Machine Learning* 489–496. ACM, New York.

LEPSKI, O. V. and SPOKOINY, V. G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.* **25** 2512–2546. MR1604408 https://doi.org/10.1214/aos/1030741083

LEPSKI, O. V. (1991). On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 454–466.

LEPSKI, O. V. (1992). Asymptotically minimax adaptive estimation. I: Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36** 682–697.

LEPSKI, O. V. (1993). Asymptotically minimax adaptive estimation. II. Schemes without optimal adaptation: Adaptive estimators. *Theory Probab. Appl.* **37** 433–448.

MANSOUR, Y., MOHRI, M. and ROSTAMIZADEH, A. (2009). Domain adaptation: Learning bounds and algorithms. In *Proceedings of the* 22*nd Annual Conference on Learning Theory* (*COLT* 2009), Montréal, Canada.

MENON, A., VAN ROOYEN, B., ONG, C. S. and WILLIAMSON, B. (2015). Learning from corrupted binary labels via class-probability estimation. In *International Conference on Machine Learning* 125–134.

PAN, S. J. and YANG, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22** 1345–1359.

SAMWORTH, R. J. (2012). Optimal weighted nearest neighbour classifiers. *Ann. Statist.* **40** 2733–2763. MR3097618 https://doi.org/10.1214/12-AOS1049

SCOTT, C. (2019). A generalized Neyman–Pearson criterion for optimal domain adaption. In *Algorithmic Learning Theory* 2019. *Proc. Mach. Learn. Res.* (*PMLR*) **98** 738–761. MR3932867

SHIMODAIRA, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Statist. Plann. Inference* **90** 227–244. MR1795598 https://doi.org/10.1016/S0378-3758(00)00115-4

SUGIYAMA, M., NAKAJIMA, S., KASHIMA, H., BUENAU, P. V. and KAWANABE, M. (2008). Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems* 1433–1440.

TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. MR2051002 https://doi.org/10.1214/aos/1079120131

TSYMBAL, A. (2004). The problem of concept drift: Definitions and related work. Computer Science Department, Trinity College Dublin.

TZENG, E., HOFFMAN, J., SAENKO, K. and DARRELL, T. (2017). Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 7167–7176.

VAN ROOYEN, B. and WILLIAMSON, R. C. (2017). A theory of learning with corrupted labels. *J. Mach. Learn. Res.* **18** Art. ID 228. MR3845527

WEISS, K., KHOSHGOFTAAR, T. M. and WANG, D. (2016). A survey of transfer learning. *J. Big Data* **3** Art. ID 9.

YAO, Y. and DORETTO, G. (2010). Boosting for transfer learning with multiple sources. In 2010 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 1855–1862. IEEE, Piscataway, NJ.

YUEN, M.-C., KING, I. and LEUNG, K.-S. (2011). A survey of crowdsourcing systems. In 2011 *IEEE Third International Conference on Privacy, Security, Risk and Trust and* 2011 *IEEE Third International Conference on Social Computing* 766–773. IEEE, Piscataway, NJ.

ZHANG, Y., CHEN, X., ZHOU, D. and JORDAN, M. I. (2014). Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. In *Advances in Neural Information Processing Systems* 1260–1268.