Detecting Local Closeness with Weighted L_2^2 Divergence

Zhun Deng, Jie Ding, Enmao Diao, Vahid Tarokh

Abstract

In this paper, we propose a novel method to measure the local closeness of two unknown distributions, using observed data only. To that end, we propose a nonparametric estimator of a distance between two density functions, which is in the form of L_2^2 distance weighted by a kernel function. Such design aims to amplify the difference between distributions by focusing on a specific data domain. The proposed distance and its estimator can also be applied when a data analyst has prior on where in the data domain to focus on, and (passively) measure the corresponding differences. It can also be applied to test the equality of distributions, with a power of statistical hypothesis test better than that of some popular methods. The idea is to apply a pre-screening step to (actively) choose the optimal weight functions, and then test. It is especially helpful when a data analyst has no prior knowledge about the underlying distributions. Rigorous theoretical analysis is provided that guarantees asymptotic performance.

I. Introduction

Identifying whether two sets of data come from the same distribution is theoretically interesting and practically useful. Example applications are the identification of significant gender disparity within different classes of incomes [1], detection of potential change of volatilities in stock markets [2], and the assessment of examination tests [3]. To be more specific, let F and G be two distributions on $\mathbb{R}^d$ that are absolute continuous with respect to certain measure μ. Let f and g denote their corresponding Radon-Nikodym derivatives with respect to measure μ: $f = dF/d\mu$, $g = dG/d\mu$. We are interested in testing the null hypothesis $\mathcal{H}_0$: $f = g$ almost everywhere against $\mathcal{H}_1$: $f \neq g$ on a set of positive measure. The challenge has been tackled by several previous works. For example, [4] proposed a kernel method based multivariate two-sample test, which advocates the use of fixed bandwidths. After that, [5] proposed another consistent test statistic based on the kernel integrated square difference and applying a central limit theorem for degenerate U-statistic. The previous methodology has been extended to weak dependent processes [6].

However, classical hypothesis tests for equality of distributions cannot capture specific local information, they are "global" in the sense that each observation is equally weighted in the test statistics. Suppose that the two underlying distributions whose densities are very close in most of the data domain, but differ only in a small regime (see Fig. 1 for illustration). Then a global test may fail to reject the null hypothesis that the two distributions are the same. If we can capture local dissimilarity instead, by putting more weights on more distinct regimes, we may be able to achieve much higher statistical test power (in the sense of reducing the probability of missed detection, given a fixed false alarm rate).

In view of the above discussions, we need a test that can capture local information, but not completely ignore the information carried by other regimes. We propose to use the following weighted L_2^2 divergence, also referred to as weighted integrated square difference. In this paper, we shall consider μ as the Lebesgue measure, and $w(x)$ as an almost everywhere positive weight function.

$$\begin{aligned} D(f,g;w) &= \int (f(x)-g(x))^2 w(x) d\mu(x) \\ &= \int f(x)w(x)dF(x) - \int g(x)w(x)dF(x) - \int f(x)w(x)dG(x) + \int g(x)w(x)dG(x). \end{aligned}$$

This work is supported by Defense Advanced Research Projects Agency (DARPA) grant numbers W911NF-14-1-0508 and N66001-15-C-4028.

Z. Deng, J. Ding, E. Diao, and V. Tarokh are with the John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, 02138 USA e-mail: (jieding@fas.harvard.edu).

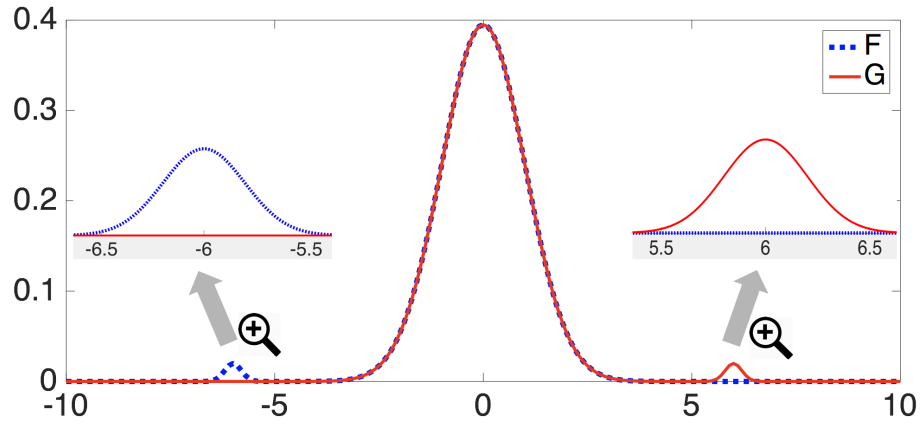


Fig. 1: An illustration of the "magnifier", where F (resp. G) is a two-mixture of Gaussian $\mathcal{N}(0,1)$ with weight 0.99 and $\mathcal{N}(-6,0.2)$ (resp. $\mathcal{N}(6,0.2)$) with weight 0.01.

By choosing appropriate weight functions, we can focus on a specific regime of observations, in order to either 1) increase the statistical power of tests (referred to as active weighting), or 2) measure local distance of particular interest (referred to as passive weighting). We shall *focus on active weighting in the following technical sections, and the same theory applies to passive weighting*. Both usages will be demonstrated by numerical experiments. Note that by this definition, $D(f,g;w)$ is 0 if and only if $f=g$ almost everywhere.

The unweighted version with $w=1$ has been extensively studied in literature. A few works have considered estimation of the the unweighted L_2^2 divergence, such as [7] and [8]. This divergence has practical usage. For instance, this divergence is used for discrete distributions in different settings in information retrieval [9], ecology [10], and elsewhere.

The outline of this paper is given as follows. In Section II, we propose a consistent estimator for $D(f,g;w)$ with independent observations coming from distributions F and G. Based on it, we further derive test statistics for hypothesis tests. We also provide theoretical studies on the asymptotic distributions of test statistics under the null, and the rate of convergence. In Section III, we focus on the situation of passive weight, and propose a novel method to choose the optimal parameters to enhance statistical test power. Experimental studies using both synthetic and real data are included in Section IV.

II. Test Statistics

Notation: We let $\|\cdot\|$ denote the Euclidean norm. Throughout the paper, $E(\cdot)$ is used to denote the expectation with respect the joint distribution that generates all the observations. We let $\to$ and $\to_p$ denote respectively deterministic convergence and convergence in probability. We sometimes rewrite the integration $\int_{x\in\mathbb{R}} f(x)dx$ as $\int f$ for brevity.

A. Design of Estimator and Test Statistics

We first introduce the motivation for the design of estimators. Suppose that X_i (resp. Y_i) for $i=1,\ldots,n$ are independent observations from distribution F (resp. G), and they are jointly independent. Since

$$D(f,g;w)=\int f(x)w(x)dG(x)-\int g(x)w(x)dF(x)-\int f(x)w(x)dG(x)+\int g(x)w(x)dG(x),$$

we denote $\Upsilon_f=\int f(x)wdF(x)$, $\Upsilon_g=\int g(x)w(x)dG(x)$, and $\Upsilon_{f,g}=\int g(x)w(x)dF(x)+\int f(x)w(x)dG(x)$, and then estimate each of them.

As for Υ_f, we construct the following estimator,

$$\frac{1}{n(n-1)h^d}\sum_{i=1,i\neq j,}^{n}\sum_{j=1}^{n}K\left(\frac{X_i-X_j}{h}\right)w(X_j),\quad \text{or}\quad \frac{1}{n(n-1)h^d}\sum_{i=1,i\neq j,}^{n}\sum_{j=1}^{n}K\left(\frac{X_i-X_j}{h}\right)w(X_i),$$

where $K : \mathbb{R}^d \to \mathbb{R}^+$ is a symmetric kernel function and $h \in \mathbb{R}^+$ is the bandwidth. We shall prove that it is a consistent estimator under some regularity conditions. The main idea is that for $i \neq j$,

$$\begin{aligned}\frac{1}{h^d} E\left\{ K\left(\frac{X_i - X_j}{h}\right) w(X_j) \right\} &= \frac{1}{h^d} E_{X_j} E_{X_i|X_j} \left\{ K\left(\frac{X_i - X_j}{h}\right) w(X_j) \right\} \\ &= \int\int K(s) w(u) f(u + sh) f(u) ds du,\end{aligned}$$

which converges to Υ_f as $n \to \infty$ under those regularity conditions. It follows that the means of the above two estimators converge to Υ_f. So it remains to prove that the variance goes to zero as $n \to \infty$, for some appropriately chosen bandwidth h. Technical details will be included later.

To guarantee large sample asymptotics, it is helpful to rewrite the kernel-based estimator in the form of U-statistics [11]. We therefore estimate Υ_f, Υ_g by using

$$\begin{aligned}\widehat{\Upsilon}_f &= \frac{1}{2n(n-1)h^d} \sum_{i=1, i\neq j,}^{n} \sum_{j=1}^{n} \left\{ K\left(\frac{X_i - X_j}{h}\right) w(X_i) + K\left(\frac{X_i - X_j}{h}\right) w(X_j) \right\}, \\ \widehat{\Upsilon}_g &= \frac{1}{2n(n-1)h^d} \sum_{i=1, i\neq j,}^{n} \sum_{j=1}^{n} \left\{ K\left(\frac{Y_i - Y_j}{h}\right) w(Y_i) + K\left(\frac{Y_i - Y_j}{h}\right) w(Y_j) \right\}.\end{aligned}$$

As for the bilinear term, we propose an adaptation of the above U-statistics which ensures that $\widehat{\Upsilon}_f + \widehat{\Upsilon}_g - \widehat{\Upsilon}_{f,g}$ is still in the form of U-statistic for $Z_i = (X_i, Y_i)$. Specifically, we use

$$\begin{aligned}\widehat{\Upsilon}_{f,g} = \frac{1}{2n(n-1)h^d} \sum_{i=1, i\neq j,}^{n} \sum_{j=1}^{n} \Bigg\{ & K\left(\frac{X_i - Y_j}{h}\right) w(Y_j) + K(\frac{X_j - Y_i}{h}) w(Y_i) \\ & + K\left(\frac{Y_i - X_j}{h}\right) w(X_j) + K\left(\frac{Y_j - X_i}{h}\right) w(X_i) \Bigg\}\end{aligned}$$

to estimate $\Upsilon_{f,g}$.

Now, we have a promising estimator for $D(f, g; w)$, which is

$$\widehat{D}_n(f, g; w) = \widehat{\Upsilon}_f + \widehat{\Upsilon}_g - \widehat{\Upsilon}_{f,g}.$$

The subscript n is to emphasis the relation between estimator and sample size n. For simplicity, we define

$$\begin{aligned}\Gamma_{ij}^x &= K\left(\frac{X_i - X_j}{h}\right) w(X_j), \quad \Gamma_{ij}^y = K\left(\frac{Y_i - Y_j}{h}\right) w(Y_j), \\ \Gamma_{ij}^{xy} &= K\left(\frac{X_i - Y_j}{h}\right) w(Y_j), \quad \Gamma_{ij}^{yx} = K\left(\frac{Y_i - X_j}{h}\right) w(X_j).\end{aligned}$$

We therefore obtain

$$\widehat{D}_n(f, g; w) = \frac{1}{2n(n-1)h^d} \sum_{i=1, i\neq j,}^{n} \sum_{j=1}^{n} (\Gamma_{ij}^x + \Gamma_{ji}^x + \Gamma_{ij}^y + \Gamma_{ji}^y - \Gamma_{ij}^{xy} - \Gamma_{ji}^{xy} - \Gamma_{ij}^{yx} - \Gamma_{ji}^{yx}). \tag{1}$$

Definition given by (V) is not a standard U-statistic, since 1) it is degenerate (variance goes to zero), and 2) its form changes with n. However, the following result shows that we can treat such statistics as martingale array.

Lemma 1: [12] If $\{Z_i\}_{i=1}^n$ are independent and identical distributed. Assume H_n is a symmetric function i.e $H_n(Z_i, Z_j) = H_n(Z_j, Z_i)$, $E[H_n(Z_1, Z_2)|Z_1] = 0$ a.e. and $E[H_n^2(Z_1, Z_2)] < \infty$ for each n. Define $G_n(x, y) = E[H_n(x, Z)H_n(y, Z)]$, where the expectation is taken with respect to Z, which is assumed to have the same distribution as Z_1. If

$$\frac{E[G_n^2(Z_1, Z_2)] + n^{-1}E[H_n^4(Z_1, Z_2)]}{\{E[H_n^2(Z_1, Z_2)]\}^2} \to 0 \quad \text{as} \quad n \to 0, \tag{2}$$

then $U_n = \sum_{1\leq i<j\leq n} H_n(Z_i, Z_j)$ is asymptotically normal, with zero mean and variance $\frac{1}{2}n^2 E[H_n^2(Z_1, Z_2)]$.

Then, if we denote $Z_i = (X_i, Y_i)$ and $H_n(Z_i, Z_j) = \Gamma^x_{ij} + \Gamma^x_{ji} + \Gamma^y_{ij} + \Gamma^y_{ji} - \Gamma^{xy}_{ij} - \Gamma^{xy}_{ji} - \Gamma^{yx}_{ij} - \Gamma^{yx}_{ji}$. It can be proved that under mild assumptions (to be introduced) and $\mathcal{H}_0$: $p = q$ *a.e.*, the statistic $H_n(Z_i, Z_j)$ satisfies all the conditions in Lemma 1 and

$$(n-1)h^{d/2}\frac{\widehat{D}_n(f,g;w)}{\sigma(f,g;w)} \to_d \mathcal{N}(0,1)$$

where $\to_d$ denotes converge in distribution and $\sigma^2(f,g;w) = 2\int(fw+gw)^2 \int K(x)^2 dx$. With the same spirit, if we denote

$$L^x_{ij} = K\left(\frac{X_i - X_j}{h}\right)w^2(X_j), \quad L^y_{ij} = K\left(\frac{Y_i - Y_j}{h}\right)w^2(Y_j),$$
$$L^{xy}_{ij} = K\left(\frac{X_i - Y_j}{h}\right)w^2(Y_j), \quad L^{yx}_{ij} = K\left(\frac{Y_i - X_j}{h}\right)w^2(X_j),$$

then a consistent estimator for $\sigma^2(f,g;w)$ can be

$$\hat{\sigma}^2_n(f,g;w) = \frac{1}{n(n-1)h^d}\sum_{i=1,i\neq j,}^{n}\sum_{j=1}^{n}(L^x_{ij} + L^x_{ji} + L^y_{ij} + L^y_{ji} + L^{xy}_{ij} + L^{xy}_{ji} + L^{yx}_{ij} + L^{yx}_{ji})\int K^2, \tag{3}$$

where the subscript n indicates the sample size.

Then, the test statistic is given by

$$T_n(f,g) = (n-1)h^{d/2}\frac{\widehat{D}_n(f,g;w)}{\hat{\sigma}_n(f,g;w)}. \tag{4}$$

Note that the test statistic is scale invariant by our design and $\hat{\sigma}^2_n(f,g;w)$ is estimated with respect to a fix function w.

Remark 1 (Computational complexity): Our estimator can be computed in quadratic time, with a compact kernel, several data structures are available to obtain a more efficient computation. For, example in [13], $\widehat{D}_n(f,g;w)$ can be computed in linear time.

Remark 2: It is easy to see that the test statistic is invariant under a constant scaling of the weight function. The flexibility of choosing weight functions makes our test able to concentrate on a specific data domain of interest.

B. Consistency and Asymptotic Normality

In the sequel, we rigorously establish some theoretical analysis. We need some regularity conditions for the rest of the paper.

Assumption 1: $K(\cdot)$ is a bounded and continuous nonvanishing symmetric function, $\int K(x)dx = 1$, and $\int K^2(x)dx < \infty$.

Assumption 2: Bandwidth h is a scalar function of sample size n such that $h \to 0$ and $nh^d \to \infty$ as $n \to \infty$.

Assumption 3: f and g are bounded and continuous (not necessarily differentiable).

Assumption 4: The weight function $w(x)$ is bounded, continuous, and positive almost everywhere with respect to x.

Remark 3: Assumptions 1 and 2 are common conditions for kernel estimation. Assumptions 3 and 4 may be further relaxed, such as replacing boundedness with square integrability.

We next prove the consistency of the estimator $\widehat{D}_n(f,g;w)$.

Theorem 1: Under Assumptions 1-4, we have:

$$E\left\{\left(\widehat{D}_n(f,g;w) - D(f,g;w)\right)^2\right\} = o(1). \tag{5}$$

In particular, $\widehat{D}_n(f,g;w) \to_p D(f,g;w)$ by Markov's inequality. Moreover, $\hat{\sigma}^2_n(f,g;w)$ converges to $\sigma^2(f,g;w)$ in probability.

By Theorem 1, one can see that $\widehat{D}_n(f,g;w)/\hat{\sigma}_n(f,g;w) \to D(f,g;w)/\sigma(f,g;w)$ in probability. So with Assumption , the test statistic $T_n(f,g) = (n-1)h^{d/2}\widehat{D}_n(f,g;w)/\hat{\sigma}_n(f,g;w)$ goes to infinity in probability under $\mathcal{H}_1$.

We conclude that the test statistic is consistent.

Theorem 2: Under the Assumptions 1-4 and under the null hypothesis, the test statistic $T_n(f,g)$ is asymptotically standard normal. And $P(T_n(f,g) > B_n) \to 1$ as $n \to \infty$ for any non-stochastic sequence $B_n = o(nh^{d/2})$ under the alternative hypothesis.

III. Selection of Weight Functions

For a specified family of weight functions $w(x;\theta)$, $\theta \in \Theta$, we want to choose appropriate $\theta \in \Theta$ to boost our test power. For notational simplicity, we shall replace $D(f,g;w)$ with $D(f,g;\theta)$, $\sigma(f,g;w)$ with $\sigma(f,g;\theta)$, $\widehat{D}_n(f,g;w)$ with $\widehat{D}_n(f,g;\theta)$ and $\hat{\sigma}_n(f,g;w)$ with $\hat{\sigma}_n(f,g;\theta)$ to emphasize that we are dealing with a family of weight functions $w(x;\theta)$. All the previous properties stands for a fixed θ. In this section, we consider $w(x)$ in the form of $w(x;\theta)$, $\theta \in \Theta$ where Θ is a parameter space. An example of $w(x;\theta)$ can be in the form of $C\exp(-\beta\|x-\alpha\|^2)$. For parameter $\theta = (\alpha, \beta, C)$, α is a d dimensional vector used as the location parameter, $\beta \in \mathbb{R}^+$ is the "resolution" parameter, and $C > 0$ is a scaling constant. Then $\left\{C\exp(-\beta\|x-\alpha\|^2) : C > 0, \alpha \in \mathbb{R}^d, \beta \in \mathbb{R}^+\right\}$ gives a family of weight functions.

Since by Theorem 2, we know that under $\mathcal{H}_1$, $T_n(f,g) \to \infty$ in probability. Also, by Theorem 1, we know that $T_n(f,g) \sim (n-1)h^{d/2}D(f,g;\theta)/\sigma(f,g;\theta)$. So, it is natural to expect that larger $|D(f,g;\theta)/\sigma(f,g;\theta)|$ gives smaller p-value. Our goal is to find θ that maximizes $|D(f,g;w)/\sigma(f,g;\theta)|$.

We need the following assumption which is an extension of Assumption 5 in view of the parameterized $w(\cdot)$.

Assumption 5: Suppose that Θ is a compact set embedded in Euclidean space $\mathbb{R}^{d'}$ (d' can be different from d). We choose $w(\cdot)$ such that for each $\theta \in \Theta$, $w(x;\theta)$ is a bounded, continuous, and positive function w.r.t $x \in \mathbb{R}^d$, and is differentiable with respect to θ, and the derivative $\partial w/\partial\theta$ satisfies $\sup_{\theta\in\Theta}\|\partial w/\partial\theta\| \le B$ almost everywhere w.r.t $\theta \in \Theta$ for some constant $B > 0$.

In Assumption 5, compactness is assumed to assure the uniform convergence of $\widehat{D}_n(f,g;\theta)$.

A. How to choose θ from data?

For the passive weight, we address a procedure to choose θ, in order to boost the test power. We propose a technique based on data splitting. Intuitively speaking, we use the first part of data (referred to as the "training data") $Z_1, \cdots, Z_k$ ($1 \le k < n$) to find a θ that maximizes $\left|\widehat{D}_k(f,g;\theta)/\widehat{\sigma}_k(f,g;\theta)\right|$, denoted by $\hat{\theta}_k$. We then apply $\hat{\theta}_k$ to the remaining data for test. Fig. 2 gives a schematic illustration of the procedure. Theoretical justifications are given in the following theorems.

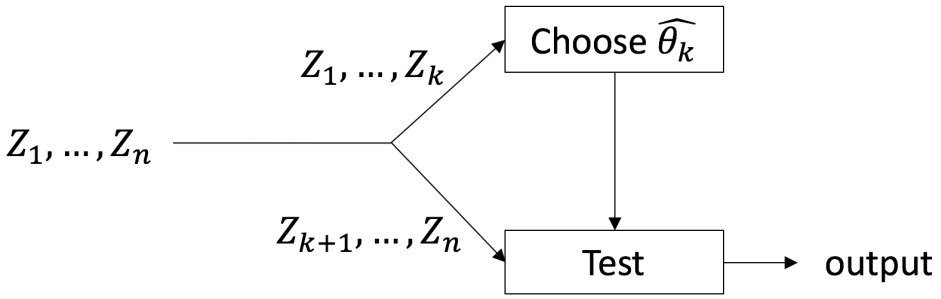


Fig. 2: A diagram of the test procedure.

Theorem 3: Under Assumptions 1-5, if we denote the subset of parameters which maximizes $|D(f,g;\theta)/\sigma(f,g;\theta)|$ as $\Theta^\star$, we can obtain that

$$\frac{D(f,g;\widehat{\theta}_k)}{\sigma(f,g;\widehat{\theta}_k)} \to_p \frac{D(f,g;\theta^\star)}{\sigma(f,g;\theta^\star)}$$

where $\theta^\star \in \Theta^\star$.

Moreover, if the cardinality of $\Theta^\star$ is finite, then, we have

$$d(\widehat{\theta}_k, \Theta^\star) = \inf_{\theta\in\Theta^\star}\|\widehat{\theta}_k - \theta\| \xrightarrow{p} 0.$$

In other words, the maximizer obtained by training data will converge in probability to one of the true maximizers of $|D(f,g;\theta)/\sigma(f,g;\theta)|$.

Then, if we use the rest of the data $Z_{k+1}, Z_{k+2}, \cdots, Z_n$ for testing, we can still obtain the same asymptotic distribution under the null hypothesis. However, in order to have an accurate estimation of the maximizer, we may waste a lot of data (especially when only a small amount of data is given). Then, if we combine the training data and test data together, can we still have good asymptotic properties? The following theorem is of interest in this setting.

Theorem 4: Under Assumptions 1-5, if we fix sample size k of training data $Z_1, Z_2, \cdots, Z_k$, then

$$(n-1)h^{d/2}\frac{\widehat{D}_n(f,g;\hat{\theta}_k)}{\widehat{\sigma}_n(f,g;\hat{\theta}_k)} \to_d \mathcal{N}(0,1)$$

B. Choice of weight function and optimization with penalty

We next focus on the following parametric form of $w(\cdot)$. For $x \in \mathbb{R}^d$, let the weight function be of the form $w(x;\theta) = \exp\big\{-(x-\alpha)^{\mathrm{T}}\Lambda(x-\alpha)\big\}$, where Λ is the diagonal matrix with diagonal elements $\lambda_1, \ldots, \lambda_d$, $\lambda_i \in \mathbb{R}^+$, $\alpha \in \mathbb{R}^d$. Note that $\theta = (\alpha, \Lambda)$ in this case.

From extensive numerical experiments, we find that under the null hypothesis, whenever $\Lambda = O$ (where O denotes the zero matrix), the empirical distribution of p-value is closer to uniform distribution even for medium sample size. Therefore, we propose to maximize

$$M_k(f,g;\theta) = \Big|\frac{\widehat{D}_k(f,g;\theta)}{\widehat{\sigma}_k(f,g;\theta)}\Big| + P_k \cdot \prod_{i=1}^{d}\frac{e^{1/\lambda_i}}{1+e^{1/\lambda_i}}\prod_{i=1}^{d}\frac{e^{1/|\alpha_i|}}{1+e^{1/|\alpha_i|}}$$

where P_k is a penalty that satisfies $P_k + (kh^{d/2}P_k)^{-1} \to 0$ as $k \to \infty$. Since we know that $\widehat{D}_n(f,g;\theta)/\widehat{\sigma}_k(f,g;\theta) = O\big((kh^{d/2})^{-1}\big)$ under the null hypothesis, we have $\Lambda \to O$ as $k \to \infty$. Meanwhile, under the alternative hypothesis, we can still obtain the same result as if there is no penalty. The following theorem is a counterpart of Theorem 4.

Theorem 5: Under Assumptions 1-5, if sample size of training data $k = o(nh^{d/2})$, then $k/(n-k) \to 0$ as $n \to \infty$, and

$$(n-1)h^{d/2}\frac{\widehat{D}_n(f,g;\hat{\theta}_k)}{\widehat{\sigma}_n(f,g;\hat{\theta}_k)} \to_d \mathcal{N}(0,1).$$

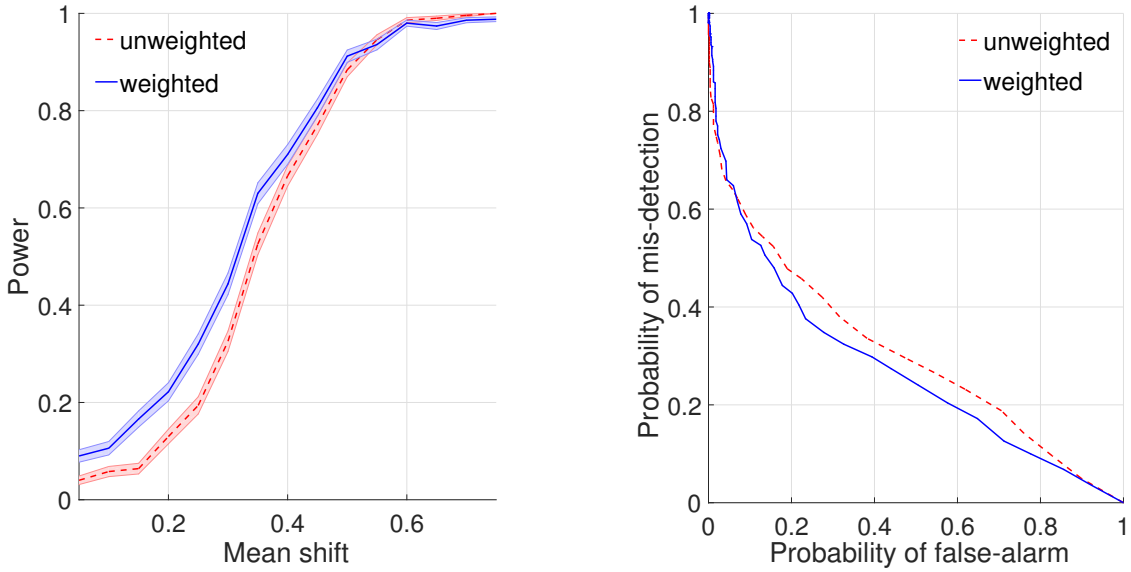
IV. Numerical Experiments

In this section, we provide three experimental studies. In each experiment, we compare our active weighting method (referred to as "weighted") with the method proposed in [5] (referred to as "unweighted"), which used $w(x) = 1$. In the experiments, we choose $n = 300$, training size $k = 100$, bandwidth $h = n^{-0.4}$, and Gaussian class of weight functions. To optimize the choice of (α, λ) of the weighting function more efficiently, we bound α and λ of ours to be $[-10, 10]$ and $[0, 0.3^{-1}]$.

In the first experiment, data are drawn from $\mathcal{N}(0,1)$ and $\mathcal{N}(\mu,1)$, and the null hypothesis is $H_0 : \mu = 0$. We let the mean shift μ vary from 0.15 to 0.75 with step size 0.05. We repeat the experiment 500 times, and summarize the results in Fig. 3. In Fig. 3a, the statistical power (given significance level 0.05) as a function of varying μ is shown. Although they both perform well when the difference is significant, our approach dominates "unweighted" approach when μ is small. In Fig 3b, we plot the receiver operating characteristic (ROC) curve for $\mu = 0.3$. Overall, the weighted approach performs better than unweighted approach.

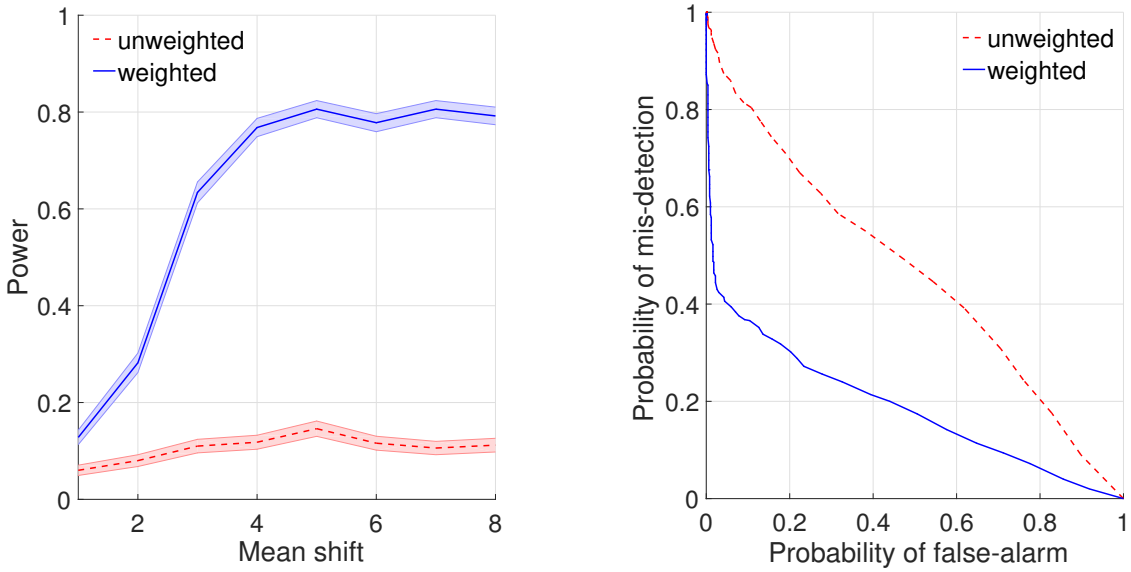
In the second experiment, we consider Gaussian mixture distributions. Data are drawn from $\mathcal{N}(0,1)$ and $0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(\mu,1)$, and the null hypothesis is $H_0 : \mu = 0$. The experiment more clearly demonstrates our focus on the locality difference of distributions. We let the mean shift μ vary from 1 to 8 with step size 1. We repeat the experiment 500 times, and summarize the results in Fig. 4. In Fig. 4a, the statistical power (given significance level 0.05) as a function of varying μ is shown. The unweighted approach has difficulty in distinguishing these two distributions, while the weighted approach substantially improves the performance. In Fig 4b, we plot the ROC curve for $\mu = 7$, where the weighted approach significantly outperforms the unweighted distance.

To intuitively show how our weighted function serves as a magnifier and amplifies the difference between two distributions, we plot in Fig. 5a the contour plot of the test statistic calculated from the testing data. We observe



(a) Statistical power (b) ROC curve

Fig. 3: Experiment 1

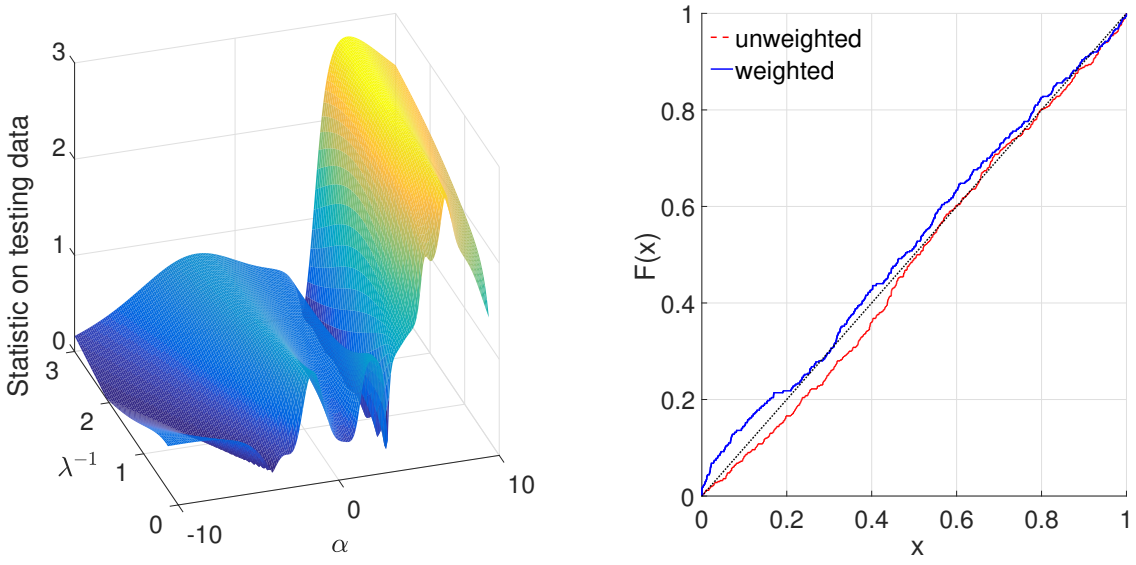


(a) Statistical power (b) ROC curve

Fig. 4: Experiment 2

a bump at around $\mu = 7$, and our optimization can choose α very close to that point. Since the unweighted distance treat everywhere equally, it is hard to detect the small difference in the tail. To show the robustness of our optimization procedure, we also plot in Fig. 5b the empirical cumulative distribution function (CDF) of p-value under the null hypothesis. Both approaches are close to the uniform distribution which shows that our approach can perform well under H_0.

Finally, we recall that instead of (actively searching) the weighting function, data analysts can subjectively choose weighting functions to magnify a specific range of interest in practice. In that case, useful prior information can be incorporated to facilitate scientific discoveries.



(a) Contour plot

(b) CDF of p-value under the null

Fig. 5: Experiment 3

V. Appendix

In the Appendix, we prove all the technical results.

The following lemma establishes asymptotic normality of our test statistics under H_0.

Lemma 2: Under Assumptions 1-4, if we define $Z_i = (X_i, Y_i)$ and $H_n(Z_i, Z_j) = \Gamma^x_{ij} + \Gamma^x_{ji} + \Gamma^y_{ij} + \Gamma^y_{ji} - \Gamma^{xy}_{ij} - \Gamma^{xy}_{ji} - \Gamma^{yx}_{ij} - \Gamma^{yx}_{ji}$, then (2) holds under the null hypothesis. So, we can obtain that $T_n(f,g) = (n-1)h^{d/2}\widehat{D}_n(f,g;w)/\hat{\sigma}_n(f,g;w)$ is asymptotic normal with mean zero and variance 1, where $\widehat{D}_n(f,g;w)$ and $\hat{\sigma}_n(f,g;w)$ is defined by Equations (V) and (3).

Proof of Lemma 2

First, we can break $E[(\widehat{D}_n(f,g;w) - D(f,g;w))^2]$ into two parts since

$$E\left[\left(\widehat{D}_n(f,g;w) - D(f,g;w)\right)^2\right] \leq 2Var(\widehat{D}_n(f,g;w)) + 2\left(E[\widehat{D}_n(f,g;w)] - D(f,g;w)\right)^2$$

For the term $Var(\widehat{D}_n(f,g;w))$, by Hoeffding's decomposition and $E[H_n(Z_1,Z_2)|Z_1] = 0$, we can obtain

$$Var(\widehat{D}_n(f,g;w)) = \binom{n}{2}^{-1} \frac{1}{4h^{2d}} Var(H_n(Z_1,Z_2)) \to 0 \quad \text{as} \quad n \to \infty,$$

by the fact that $Var(H_n(Z_1,Z_2)) = O(h^d)$ and $nh^d \to \infty$ (will be proved in lemma 3) and $h \to 0$ as $n \to 0$.

For another term $\left(E[\widehat{D}_n(f,g;w)] - D(f,g;w)\right)^2$, by triangle inequality, we can obtain

$$|E\widehat{D}_n(f,g;w) - D(f,g;w)| \leq 2\left|E[h^{-d}\Gamma^x_{ij}] - \int f^2 w\right| + 2\left|E[h^{-d}\Gamma^y_{ij}] - \int g^2 w\right| + 4\left|E[h^{-d}\Gamma^{xy}_{ij}] - \int fgw\right|$$

For the term $\left|E[h^{-d}\Gamma^x_{ij}] - \int f^2 w\right|$,

$$\begin{aligned}&\left|\int\int K(s)w(u)f(u+sh)f(u)dsdu-\int\int K(s)w(u)f^2(u)dsdu\right|\\&\leq\int\int K(s)w(u)|f(u+sh)-f(u)|f(u)dsdu\\&\leq B^w\int\int K(s)|f(u+sh)-f(u)|f(u)dsdu\\&\to 0\end{aligned}$$

as $n\to\infty$ by dominated convergence theorem, where $B^w=esssup_{x\in\mathbb{R}^d}w(x)$. Other terms in $E\widehat{D}_n(f,g;\theta)-D^w(f,g;\theta)$ can be proved going to zero by similar technique. Those two steps complete the proof.

Proof of Lemma 3

Here, $H_n(Z_i,Z_j)$ is obviously symmetric. Moreover, under H_0 , $E(\Gamma^x_{ij}-\Gamma^{xy}_{ij}|X_i)=0$. Similarly, we have $E(\Gamma^x_{ji}-\Gamma^{yx}_{ji}|X_i)=0$. Meanwhile, we have $E(\Gamma^y_{ij}-\Gamma^{yx}_{ij}|Y_i)=0$ and $E(\Gamma^y_{ji}-\Gamma^{xy}_{ji}|Y_i)=0$. So, we can see that $E[H_n(Z_i,Z_j)|Z_i]=0$. Then, we prove

$E[H_n^2(Z_1,Z_2)]=O(h^d)$*:* First, we claim that $E(\Gamma^x_{ij})^2=O(h^d)$. That is easy, since by tower property,

$$\begin{aligned}E(\Gamma^x_{ij})^2&=h^d\int\int K^2(s)w^2(u)f(u+sh)f(u)dsdu\\&=h^d\{\int\int K^2(s)w^2(u)f^2(u)dsdu+\int\int K^2(s)w^2(u)f(u)(f(u+sh)-f(u))dsdu\}\end{aligned}$$

It is enough to $h^d\int\int K^2(s)w^2(u)f(u)(f(u+sh)-f(u))dsdu$ is $o(h^d)$, that follows immediately by dominated convergence theorem and boundedness and continuity of f, w.

Secondly, let us prove $E\Gamma^{xy}_{ij}\Gamma^{yx}_{ij}=O(h^d)$. Exact same technique leads to

$$E\Gamma^{xy}_{ij}\Gamma^{yx}_{ij}=h^d\int\int K^2(s)w(u)w(u+sh)f(u+sh)g(u)dsdu=O(h^d)$$

Similarly, we can prove that $E\Gamma^{xy}_{ij}\Gamma^{xy}_{ji}=O(h^{2d})$, $E\Gamma^{x}_{ij}\Gamma^{yx}_{ji}=O(h^{2d})$. These are all the four types of products we need to know, so, we get

$$E[H_n^2(Z_1,Z_2)]=O(h^d)$$

$E[H_n^4(Z_1,Z_2)]=O(h^d)$, $E[G_n^2(Z_1,Z_2)]=O(h^{3d})$*:* Same technique, we omit it here.

So, combine all the facts above we have

$$\left\{E[G_n^2(Z_1,Z_2)]+n^{-1}E[H_n^4(Z_1,Z_2)]\right\}/\left\{E[H_n^2(Z_1,Z_2)]\right\}^2\to 0$$

Thus, the proof is complete.

Proof of Theorem 3

In order to prove theorem 3, we first prove several lemmas.

Lemma 3: Under the Assumptions 1-5, we have a uniform control of convergence of $\widehat{D}_n(f,g;\theta)$, which means $\sup_{\theta\in\Theta}|\widehat{D}_n(f,g;\theta)-D(f,g;\theta)|\to 0$ in probability.

Proof: We have proved that $\sup_{\theta\in\Theta}|E\widehat{D}_n(f,g;\theta)-D(f,g;\theta)|$ goes to zero in the proof of lemma 2. Then, we only need to control $\sup_{\theta\in\Theta}|\widehat{D}_n(f,g;\theta)-E\widehat{D}_n(f,g;\theta)|$.

By Assumption 3, we can easily get Lipschitz's condition, 2almost everywhere. Then, for any ε we can first find $\delta>0$, such that whenever $\|\theta_1-\theta_2\|<\delta$, we have $|w(X_i;\theta_1)-w(X_i;\theta_2)|\leq\varepsilon/(8\int(f+g)^2)$. Then, we find a δ-covering of the space Θ, since Θ is compact, we find k balls that cover Θ. We denote the center of the k balls as θ_i, $1\leq i\leq k$. In order for simplicity, in this part, we also denote $\widehat{D}_n(f,g;\theta)$ as $\widehat{D}^w_n(\theta)$, and $\widehat{D}_n(f,g;\theta_i)$ as $\widehat{D}^w_n(\theta_i)$. Also, if we use $\theta^M(\omega)$ to denote when fix $\omega\in\Omega$, where Ω is the underlying sample

space, the element in Θ that maximizes $|\widehat{D}_n^w(\theta) - E\widehat{D}_n^w(\theta)|$ (the existence of such θ^M can be justified by the compactness of Θ), then $\sup_{\theta\in\Theta}|\widehat{D}_n^w(\theta) - E\widehat{D}_n^w(\theta)| = |\widehat{D}_n(f,g;\theta^M) - E\widehat{D}_n(f,g;\theta^M)|$. Additionally, we denote $\theta^N(\omega) = \{\theta_i : d(\theta^M(\omega),\theta_i) \le d(\theta^M(\omega),\theta_j), i \in \{1,2,\cdots,k\}, j = 1,2,\cdots,k\}$. $\theta^N(\omega)$ may not be unique for fixed ω, if so, we choose θ_i with the smallest indicator. Once, we can prove that as n goes to infinity

$$P\left(\left||\widehat{D}_n(f,g;\theta^M) - E\widehat{D}_n(f,g;\theta^M)| - |\widehat{D}_n(f,g;\theta^N) - E\widehat{D}_n(f,g;\theta^N)|\right| \le \varepsilon/2\right) \xrightarrow{p} 1,$$

we can immediately obtain, as n goes to infinity

$$\begin{aligned}
&P(|\widehat{D}_n(f,g;\theta^M) - E\widehat{D}_n(f,g;\theta^M)| > \epsilon)\\
&= P\Bigg(|\widehat{D}_n(f,g;\theta^M) - E\widehat{D}_n(f,g;\theta^M)| > \varepsilon,\\
&\qquad \left||\widehat{D}_n(f,g;\theta^M) - E\widehat{D}_n(f,g;\theta^M)| - |\widehat{D}_n(f,g;\theta^N) - E\widehat{D}_n(f,g;\theta^N)|\right| \le \varepsilon/2\Bigg) + o_p(1)\\
&\le P\left(\bigcup_i |\widehat{D}_n^w(\theta_i) - E\widehat{D}_n^w(\theta_i)| > \varepsilon/2\right) + o_p(1)\\
&\le \sum_i P\left(|\widehat{D}_n^w(\theta_i) - E\widehat{D}_n^w(\theta_i)| > \varepsilon/2\right) + o_p(1)\\
&\xrightarrow{p} 0
\end{aligned}$$

Now, we are only left to prove

$$(\star)\quad P\left(\left||\widehat{D}_n(f,g;\theta^M) - E\widehat{D}_n(f,g;\theta^M)| - |\widehat{D}_n(f,g;\theta^N) - E\widehat{D}_n(f,g;\theta^N)|\right| \le \varepsilon/2\right) \xrightarrow{p} 1.$$

That can be achieved by

$$\begin{aligned}
(\star) &\ge P\left(|\widehat{D}_n(f,g;\theta^M) - \widehat{D}_n(f,g;\theta^N)| + |E\widehat{D}_n(f,g;\theta^M) - E\widehat{D}_n(f,g;\theta^N)| \le \varepsilon/2\right)\\
&\ge P\left(|\widehat{D}_n(f,g;\theta^M) - \widehat{D}_n(f,g;\theta^N)| \le \varepsilon/4, |E\widehat{D}_n(f,g;\theta^M) - E\widehat{D}_n(f,g;\theta^N)| \le \varepsilon/4\right)
\end{aligned}$$

In order to prove that, by the similar arguments as above, we only need to prove $P\left(|\widehat{D}_n(f,g;\theta^M) - \widehat{D}_n(f,g;\theta^N)| \le \varepsilon/4\right)$ and $P\left(|E\widehat{D}_n(f,g;\theta^M) - E\widehat{D}_n(f,g;\theta^N)| \le \varepsilon/4\right)$ both goes to 1. We denote $K_{ij}^x = K\left(\frac{X_i - X_j}{h}\right)$, $K_{ij}^y = K\left(\frac{Y_i - Y_j}{h}\right)$, $K_{ij}^{xy} = K\left(\frac{X_i - Y_j}{h}\right)$ and $K_{ij}^{yx} = K\left(\frac{Y_i - X_j}{h}\right)$. Then,

$$\begin{aligned}
&P\left(|\widehat{D}_n(f,g;\theta^M) - \widehat{D}_n(f,g;\theta^N)| \le \varepsilon/4\right)\\
&\ge P\left(\frac{1}{2n(n-1)h^d}\sum_{i=1,i\ne j,}^{n}\sum_{j=1}^{n}(K_{ij}^x + K_{ji}^x + K_{ij}^y + K_{ji}^y + K_{ij}^{xy} + K_{ji}^{xy} + K_{ij}^{yx} + K_{ji}^{yx}) \le 2\int (f+g)^2\right)
\end{aligned}$$

which goes to one in probability since

$$\frac{1}{2n(n-1)h^d}\sum_{i=1,i\ne j,}^{n}\sum_{j=1}^{n}(EK_{ij}^x + EK_{ji}^x + EK_{ij}^y + EK_{ji}^y + EK_{ij}^{xy} + EK_{ji}^{xy} + EK_{ij}^{yx} + EK_{ji}^{yx}) \xrightarrow{n} \int (f+g)^2$$

and it variance goes to zero as n goes to infinity. Meanwhile,

$$\begin{aligned}
&P\left(|E\widehat{D}_n(f,g;\theta^M) - E\widehat{D}_n(f,g;\theta^N)| \le \varepsilon/4\right)\\
&\ge P\Bigg(\frac{1}{2n(n-1)h^d}\sum_{i=1,i\ne j,}^{n}\sum_{j=1}^{n}(EK_{ij}^x + EK_{ji}^x + EK_{ij}^y + EK_{ji}^y + EK_{ij}^{xy} + EK_{ji}^{xy} + EK_{ij}^{yx} + EK_{ji}^{yx})\\
&\qquad \le 2\int (f+g)^2\Bigg)
\end{aligned}$$

which also goes to 1 as n goes to infinity. ■

Lemma 4: Under the Assumptions 1-5, we have that

$$\sup_{\theta\in\Theta}|\widehat{D}_n(f,g;\theta)/\widehat{\sigma}_n(f,g;\theta) - D(f,g;\theta)/\sigma(f,g;\theta)| \to 0$$

Proof:

$$\begin{aligned}
&P\left(\sup_{\theta\in\Theta}|\widehat{D}_n(f,g;\theta)/\widehat{\sigma}_n(f,g;\theta) - D(f,g;\theta)/\sigma(f,g;\theta)| \le \varepsilon\right)\\
&\ge P\Big(\sup_{\theta\in\Theta}|\widehat{D}_n(f,g;\theta)/\widehat{\sigma}_n(f,g;\theta) - \widehat{D}_n(f,g;\theta)/\sigma(f,g;\theta)| \le \varepsilon/2,\\
&\sup_{\theta\in\Theta}|\widehat{D}_n(f,g;\theta)/\sigma(f,g;\theta) - D(f,g;\theta)/\sigma(f,g;\theta)| \le \varepsilon/2\Big)
\end{aligned}$$

For $P\left(\sup_{\theta\in\Theta}|\widehat{D}_n(f,g;\theta)/\widehat{\sigma}_n(f,g;\theta) - \widehat{D}_n(f,g;\theta)/\sigma(f,g;\theta)| \le \varepsilon/2\right)$, it is larger or equal to

$$P\left(\sup_{\theta\in\Theta}|\widehat{D}_n(f,g;\theta)|\sup_{\theta\in\Theta}|1/\widehat{\sigma}_n(f,g;\theta) - 1/\sigma(f,g;\theta)| \le \varepsilon/2\right)$$

Let us prove $\sup_{\theta\in\Theta}|\widehat{D}_n(f,g;\theta) - D(f,g;\theta)| \xrightarrow{p} 0$. The proof is similar to lemma 2, by triangular inequality, we only need to prove for $\sup_{\theta\in\Theta}|E\widehat{D}_n(f,g;\theta) - D(f,g;\theta)|$,

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left|\int\int K(s)w(u;\theta)f(u+sh)f(u)dsdu - \int\int K(s)w(u;\theta)f^2(u)dsdu\right|\\
&\le \sup_{\theta\in\Theta}\int\int K(s)w(u;\theta)|f(u+sh)-f(u)|f(u)dsdu\\
&\le B^{w(\cdot,\theta)}\int\int K(s)|f(u+sh)-f(u)|f(u)dsdu\\
&\to 0
\end{aligned}$$

as $n\to\infty$ by dominated convergence theorem, where $B^{w(\cdot,\theta)} = sup_{\theta\in\Theta}esssup_{x\in\mathbb{R}^d}w(x;\theta)$. By exact same technique as proving $\sup_{\theta\in\Theta}|\widehat{D}_n(f,g;\theta) - D(f,g;\theta)| \xrightarrow{p} 0$, we can prove

$$\sup_{\theta\in\Theta}|1/\widehat{\sigma}_n(f,g;\theta) - 1/\sigma(f,g;\theta)| \xrightarrow{p} 0,$$

then we can prove that

$$P\left(\sup_{\theta\in\Theta}|\widehat{D}_n(f,g;\theta)|\sup_{\theta\in\Theta}|1/\widehat{\sigma}_n(f,g;\theta) - 1/\sigma(f,g;\theta)| \le \varepsilon/2\right)$$

■

Lemma 5: Recall that we define the set of parameters which maximizes $D(f,g;w)/\sigma(f,g;\theta)$ as $\Theta^\star$. Then, we have $d(\widehat{\theta}_k,\Theta^\star) = \inf_{\theta\in\Theta^\star} d(\theta_k,\theta)$ goes to zero in probability under Assumptions 1-4, where d is the metric on the parameter space $\Theta^\star$.

Notice that Θ^* is different from $\Theta^\star$ mentioned in theorem 3. The above two lemmas

Proof: The proof is straight forward. The lemma above implies pointwise convergence in probability:

$$|\widehat{D}_n(f,g;\theta)/\sigma_0^n - D^w(f,g;\theta)| \to 0,$$

so we can write $D(f,g;\theta^\star)/\sigma(f,g;\theta^\star) = \widehat{D}_n^w(f,g;\theta^\star)/\sigma_0^n(\theta^\star) + o_p(1)$. If we denote θ^* as a parameter in the subset $\Theta^\star$. Then, we have

$$\begin{aligned}
0 &\le D(f,g;\theta^\star)/\sigma(f,g;\theta^\star) - D^w(f,g;\widehat{\theta}_n)/\sigma_0(f,g;\theta^\star)\\
&= \widehat{D}_n(f,g;\theta^\star)/\widehat{\sigma}_0^n(\theta^\star) + o_p(1) - D^w(f,g;\widehat{\theta}_n)/\sigma_0(f,g;\widehat{\theta}_n)\\
&\le \widehat{D}_n^w(f,g;\widehat{\theta}_n)/\widehat{\sigma}_0^n(\widehat{\theta}_n) + o_p(1) - D^w(f,g;\widehat{\theta}_n)/\sigma_0(f,g;\widehat{\theta}_n)\\
&\le \sup_{\theta\in\Theta}|\widehat{D}_n(f,g;\theta)/\widehat{\sigma}_n(f,g;\theta) - D^w(f,g;\theta)/\sigma(f,g;\theta)| + o_p(1)\\
&\xrightarrow{p} 0.
\end{aligned}$$

The first and third inequality is due to the fact that $\theta^\star$ is the maximizer of $D^w(f,g;\theta)/\sigma(f,g;\theta)$ and $\widehat{\theta}_n$ is the maximizer of $\widehat{D}_n(f,g;\theta)/\widehat{\sigma}_n(f,g;\theta)$. Then, by standard proof of well-separated mode, we can obtain our result. ■

Since the proof of Theorem 4 is a simplified version of Theorem 5, we here only show the proof of Theorem 5.

Proof of theorem 5

First, we use some handy notations.

we use the following notations:

$$K_{ij}^x = K\left(\frac{X_i - X_j}{h}\right), \quad K_{ij}^y = K\left(\frac{Y_i - Y_j}{h}\right),$$
$$K_{ij}^{xy} = K\left(\frac{X_i - Y_j}{h}\right), \quad K_{ij}^{yx} = K\left(\frac{Y_i - X_j}{h}\right).$$

$$K_k = \frac{1}{2n(n-1)h^d} \sum_{i=1, i\neq j,}^{k} \sum_{j=1}^{k} (K_{ij}^x + K_{ji}^x + K_{ij}^y + K_{ji}^y - K_{ij}^{xy} - K_{ji}^{xy} - K_{ij}^{yx} - K_{ji}^{yx}).$$

$$K_{k,n-k} = \frac{1}{2n(n-1)h^d} \sum_{i=1, i\neq j,}^{k} \sum_{j=1}^{n-k} (K_{ij}^x + K_{ji}^x + K_{ij}^y + K_{ji}^y - K_{ij}^{xy} - K_{ji}^{xy} - K_{ij}^{yx} - K_{ji}^{yx})$$
$$+ \frac{1}{2n(n-1)h^d} \sum_{i=1, i\neq j,}^{n-k} \sum_{j=1}^{k} (K_{ij}^x + K_{ji}^x + K_{ij}^y + K_{ji}^y - K_{ij}^{xy} - K_{ji}^{xy} - K_{ij}^{yx} - K_{ji}^{yx}).$$

$$K_{n-k,n-k} = \frac{1}{2n(n-1)h^d} \sum_{i=1, i\neq j,}^{n-k} \sum_{j=1}^{n-k} (K_{ij}^x + K_{ji}^x + K_{ij}^y + K_{ji}^y - K_{ij}^{xy} - K_{ji}^{xy} - K_{ij}^{yx} - K_{ji}^{yx}).$$

Since we know under null we have $essup_x|w(x,\widehat{\theta}_k) - 1| \to 0$, and $\widehat{\sigma}_n(f,g;\widehat{\theta}_k)$ converges to a positive constant, so we only need to prove that

$$nh^{d/2}K_{n-k,n-k}(1 + \frac{nh^{d/2}K_k + nh^{d/2}K_{k,n-k}}{nh^{d/2}K_{n-k,n-k}})$$

is normally distributed as $\mathcal{N}(0, 2\int(f+g)^2 \int K(x)^2 dx)$. It is easy to see that $nh^{d/2}K_{n-k,n-k}$ is distributed as $\mathcal{N}(0, 2\int(f+g)^2 \int K(x)^2 dx)$. Then, by Hoeffding decomposition, we can easily get the variance of $nh^{d/2}K_k + nh^{d/2}K_{k,n-k}$ goes to 0 as n goes to infinity. Since the mean of $nh^{d/2}K_k + nh^{d/2}K_{k,n-k}$ is 0, the proof is complete.

REFERENCES

[1] W. A. Darity Jr and P. L. Mason, "Evidence on discrimination in employment: Codes of color, codes of gender," in *African American Urban Experience: Perspectives from the Colonial Period to the Present*. Springer, 2004, pp. 156–186.

[2] S. Hammoudeh and H. Li, "Sudden changes in volatility in emerging markets: the case of gulf arab stock markets," *International Review of Financial Analysis*, vol. 17, no. 1, pp. 47–63, 2008.

[3] R. E. Bennett, "Formative assessment: A critical review," *Assessment in Education: Principles, Policy & Practice*, vol. 18, no. 1, pp. 5–25, 2011.

[4] N. H. Anderson, P. Hall, and D. M. Titterington, "Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates," *Journal of Multivariate Analysis*, vol. 50, no. 1, pp. 41–54, 1994.

[5] Q. Li, "Nonparametric testing of closeness between two unknown distribution functions," *Econometric Reviews*, vol. 15, no. 3, pp. 261–274, 1996.

[6] Y. Fan and A. Ullah, "On goodness-of-fit tests for weakly dependent processes using kernel method," *Journal of Nonparametric Statistics*, vol. 11, no. 1-3, pp. 337–360, 1999.

[7] B. Póczos and J. G. Schneider, "On the estimation of alpha-divergences." in *AISTATS*, 2011, pp. 609–617.

[8] A. Krishnamurthy, K. Kandasamy, B. Poczos, and L. Wasserman, "Nonparametric estimation of renyi divergence and friends," in *International Conference on Machine Learning*, 2014, pp. 919–927.

[9] C. D. Manning, P. Raghavan, H. Schütze *et al.*, *Introduction to information retrieval*. Cambridge university press Cambridge, 2008, vol. 1, no. 1.

[10] P. Legendre and L. F. Legendre, *Numerical ecology*. Elsevier, 2012, vol. 24.

[11] E. Giné and R. Nickl, "A simple adaptive estimator of the integrated square of a density," *Bernoulli*, pp. 47–61, 2008.
[12] P. Hall, "Central limit theorem for integrated square error of multivariate nonparametric density estimators," *Journal of multivariate analysis*, vol. 14, no. 1, pp. 1–16, 1984.
[13] W. B. March, P. Ram, and A. G. Gray, "Fast euclidean minimum spanning tree: algorithm, analysis, and applications," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 603–612.