Least-Squares Two-Sample Test

Masashi Sugiyama
Tokyo Institute of Technology
and PRESTO, Japan Science and Technology Agency (JST),
2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan.
sugi@cs.titech.ac.jp http://sugiyama-www.cs.titech.ac.jp/˜sugi

Taiji Suzuki
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
s-taiji@stat.t.u-tokyo.ac.jp

Yuta Itoh
Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan.
itoh@sg.cs.titech.ac.jp

Takafumi Kanamori
Nagoya University
Furocho, Chikusaku, Nagoya 464-8603, Japan.
kanamori@is.nagoya-u.ac.jp

Manabu Kimura
Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan.
kimura@sg.cs.titech.ac.jp

Abstract

The goal of the two-sample test (a.k.a. the homogeneity test) is, given two sets of samples, to judge whether the probability distributions behind the samples are the same or not. In this paper, we propose a novel non-parametric method of two-sample test based on a least-squares density ratio estimator. Through various experiments, we show that the proposed method overall produces smaller type-II error (i.e., the probability of judging the two distributions to be the same when they are actually different) than a state-of-the-art method, with slightly larger type-I error (i.e., the probability of judging the two distributions to be different when they are actually the same).

Keywords

two-sample test, homogeneity test, density ratio estimation, unconstrained least-squares importance fitting, Pearson divergence.

1 Introduction

Given two sets of samples, testing whether the probability distributions behind the samples are equivalent or not is a fundamental task in statistical data analysis. This problem is referred to as the *two-sample test* or the *homogeneity test* (Kullback, 1959).

1.1 Motivation of Two-Sample Test

The two-sample test is useful in various practically important learning scenarios. Here we describe some examples.

When learning is performed under non-stationary environment, e.g., in brain-computer interface (Sugiyama et al., 2007) and robot control (Hachiya et al., 2009), testing homogeneity of data generating distributions allows one to determine whether some adaptation scheme should be used or not. When the distributions are not significantly different, one can avoid using data-intensive non-stationarity adaptation techniques, which highly contributes to stabilizing the performance.

When multiple sets of data samples are available for learning, e.g., biological experimental results obtained from different laboratories (Borgwardt et al., 2006), the homogeneity test allows one to make a decision whether all the datasets are analyzed jointly as a single dataset or they should be treated separately.

Similarly, one can use the homogeneity test for deciding whether *multi-task learning* methods (Caruana et al., 1997) are employed or not. The rationale behind multi-task learning is that when several related learning tasks are provided, solving them simultaneously can give better solutions than solving them individually. However, when the tasks are not similar to each other, using multi-task learning techniques can degrade the performance. Thus, it is important to avoid using multi-task learning methods when the tasks are not similar to each other. This may be achieved by testing the homogeneity of datasets.

When several databases containing multiple fields are given, it is useful to identify the correspondence between fields by comparing underlying distributions since this allows one to merge databases (Gretton et al., 2007).

1.2 Methods of Two-Sample Test

The *t-test* (Student, 1908) is a classical method for testing homogeneity, which compares the means of two Gaussian distributions with common variance. Its multi-variate extension also exists (Hotelling, 1951). Although the t-test is a fundamental method for comparing the means, its range of application is limited to Gaussian distributions, which may not be fulfilled in practical applications.

The *Kolmogorov-Smirnov test* and the *Wald-Wolfowitz runs test* are classical non-parametric methods for the two-sample problem; their multi-dimensional variants have also been developed (Bickel, 1969; Friedman & Rafsky, 1979). Since then, different types of non-parametric tests have been studied (Anderson et al., 1994; Li, 1996).

Recently, a non-parametric extension of the t-test called the *maximum mean discrepancy* (MMD) was proposed (Borgwardt et al., 2006; Gretton et al., 2007). MMD compares the means of two distributions in a *universal reproducing kernel Hilbert space* (universal RKHS; Steinwart, 2001)—the Gaussian kernel is a typical example that induces a universal RKHS. MMD does not require a restrictive parametric assumption, so it could be a flexible alternative to the t-test. MMD was experimentally shown to outperform other homogeneity tests such as the *generalized Kolmogorov-Smirnov test* (Friedman & Rafsky, 1979), the *generalized Wald-Wolfowitz test* (Friedman & Rafsky, 1979), the *Hall-Tajvidi test* (Hall & Tajvidi, 2002), and the *Biau-Györfi test* (Biau & Györfi, 2005).

The performance of MMD depends on the choice of universal RKHSs (e.g., the Gaussian width in the case of Gaussian RKHSs). Thus, the universal RKHS should be carefully chosen for obtaining the state-of-the-art performance. The Gaussian RKHS with width set to the median distance between samples has been a popular heuristic in practice (Borgwardt et al., 2006; Gretton et al., 2007). Recently, a novel idea of using the universal RKHS (or the Gaussian widths) yielding the maximum MMD value has been introduced (Sriperumbudur et al., 2009).

1.3 Divergence Estimation

Another approach to the two-sample problem is to evaluate a divergence between two distributions. The divergence-based approach is advantageous in that cross-validation over the divergence functional is available for optimizing tuning parameters in a data-dependent manner. A typical choice of the divergence functional would be the *f-divergences* (Ali & Silvey, 1966; Csiszár, 1967), which includes the *Kullback-Leibler divergence* (Kullback & Leibler, 1951) and the *Pearson divergence* (Pearson, 1900) as special cases.

Various methods for estimating the divergence functional have been studied so far (Darbellay & Vajda, 1999; Wang et al., 2005; Silva & Narayanan, 2007; Pérez-Cruz, 2008). Among them, approaches based on *density ratio estimation* have been shown to be promising both theoretically and experimentally (Sugiyama et al., 2008; Gretton et al., 2009; Kanamori et al., 2009a; Nguyen et al., 2010). So far, a parametric density ratio estimator based on logistic regression (Qin, 1998; Cheng & Chu, 2004) has been applied to the test of homogeneity (Keziou & Leoni-Aubin, 2005).

Although the density ratio estimator based on logistic regression was proved to achieve the smallest asymptotic variance among a class of semi-parametric estimators (Qin, 1998), this theoretical guarantee is valid only when the parametric model is *correctly specified* (i.e., the target density ratio is included in the parametric model at hand). However, when this unrealistic assumption is violated, a divergence-based density ratio estimator (Sugiyama et al., 2008; Nguyen et al., 2010) was shown to perform better (Kanamori et al., 2010).

Among various divergence-based density ratio estimators, a method called *unconstrained least-squares importance fitting* (uLSIF) was demonstrated to be accurate and computationally efficient (Kanamori et al., 2009a). Furthermore, uLSIF was proved to

possess the optimal non-parametric convergence rate and numerical stability (Kanamori et al., 2009b). In this paper, we therefore develop a new method for testing homogeneity based on uLSIF.

Similarly to MMD, our uLSIF-based homogeneity test processes data samples only through kernel functions. Thus, the proposed method can be used for testing the homogeneity of *non-vectorial structured objects* such as strings, trees, and graphs by employing kernel functions defined for such structured data (Lodhi et al., 2002; Duffy & Collins, 2002; Kashima & Koyanagi, 2002; Kondor & Lafferty, 2002; Kashima et al., 2003; Gärtner et al., 2003; Gärtner, 2003). This is an advantage over traditional two-sample tests.

1.4 Organization of This Paper

The rest of this paper is structured as follows. In Section 2, we review the uLSIF method for density ratio estimation. In Section 3, we describe a method of divergence estimation based on uLSIF, and investigate its theoretical properties. In Section 4, we give a two-sample test based on the permutation test (Efron & Tibshirani, 1993), which we call *least-squares two-sample test* (LSTT). We review the MMD method in Section 5, and compare the experimental performance of LSTT with MMD in Section 6. Finally, we conclude in Section 7.

2 Density Ratio Estimation

In this section, we consider the problem of density ratio estimation, and review a method called *unconstrained least-squares importance fitting* (uLSIF; Kanamori et al., 2009a), which will be used in the following sections. Since this section is devoted to reviewing uLSIF, those who are familiar with it may skip this section and directly go to the next section.

2.1 Formulation of Density Ratio Estimation

Suppose we are given a set of samples

$$\mathcal{X} := \{\boldsymbol{x}_i | \boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^n$$

drawn independently from a probability distribution P with density $p(\boldsymbol{x})$, and another set of samples

$$\mathcal{X}' := \{\boldsymbol{x}'_j | \boldsymbol{x}'_j \in \mathbb{R}^d\}_{j=1}^{n'}$$

drawn independently from (possibly) another probability distribution P' with density $p'(\boldsymbol{x})$:

$$\{\boldsymbol{x}_i\}_{i=1}^n \overset{i.i.d.}{\sim} P,$$
$$\{\boldsymbol{x}'_j\}_{j=1}^{n'} \overset{i.i.d.}{\sim} P'.$$

The goal of density ratio estimation is to estimate the density ratio function

$$r(\boldsymbol{x}) := \frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})} \tag{1}$$

from the samples $\mathcal{X}$ and $\mathcal{X}'$, where we assume $p'(\boldsymbol{x}) > 0$ for all $\boldsymbol{x}$.

2.2 Least-Squares Approach to Density Ratio Estimation

Let us model the density ratio function $r(\boldsymbol{x})$ by the following kernel model[1]:

$$\begin{aligned}\widehat{r}(\boldsymbol{x}) &:= \alpha_0 + \sum_{i=1}^{n} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i) \\ &= \boldsymbol{\alpha}^\top \boldsymbol{k}(\boldsymbol{x}),\end{aligned}$$

where

$$\boldsymbol{\alpha} := (\alpha_0, \alpha_1, \ldots, \alpha_{n+1})^\top$$

are parameters to be learned from data samples, $^\top$ denotes the transpose of a matrix or a vector,

$$\boldsymbol{k}(\boldsymbol{x}) := (1, K(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, K(\boldsymbol{x}, \boldsymbol{x}_n))^\top$$

are kernel basis functions. A popular choice of the kernel is the Gaussian function:

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma^2}\right), \tag{2}$$

where σ^2 denotes the Gaussian variance.

We determine the parameter $\boldsymbol{\alpha}$ in the model $\widehat{r}(\boldsymbol{x})$ so that the following squared-error J_0 is minimized:

$$\begin{aligned}J_0(\boldsymbol{\alpha}) &:= \frac{1}{2}\int \left(\widehat{r}(\boldsymbol{x}) - r(\boldsymbol{x})\right)^2 p'(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &= \frac{1}{2}\int \widehat{r}(\boldsymbol{x})^2 p'(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int \widehat{r}(\boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \frac{1}{2}\int r(\boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x},\end{aligned}$$

where the last term is a constant and therefore can be safely ignored. Let us denote the first two terms by J:

$$\begin{aligned}J(\boldsymbol{\alpha}) &:= \frac{1}{2}\int \widehat{r}(\boldsymbol{x})^2 p'(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int \widehat{r}(\boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &= \frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{H} \boldsymbol{\alpha} - \boldsymbol{h}^\top \boldsymbol{\alpha},\end{aligned} \tag{3}$$

[1]We included the constant basis function, 1, in our model, which is different from the original uLSIF paper (Kanamori et al., 2009a). In the context of two-sample test, we empirically found that including the constant basis tends to improve the estimation accuracy since the density ratio function we approximate can be close to constant (i.e., $r(\boldsymbol{x}) \approx 1$) when the two distributions are similar.

where $\boldsymbol{H}$ is the $(n+1) \times (n+1)$ matrix defined by

$$\boldsymbol{H} := \int \boldsymbol{k}(\boldsymbol{x})\boldsymbol{k}(\boldsymbol{x})^\top p'(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$

and $\boldsymbol{h}$ is the $(n+1)$-dimensional vector defined by

$$\boldsymbol{h} := \int \boldsymbol{k}(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

2.3 Empirical Approximation

Since J contains the expectation over unknown densities $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$, we approximate the expectations by empirical averages. Then we obtain

$$\begin{aligned}\widehat{J}(\boldsymbol{\alpha}) &:= \frac{1}{2n'}\sum_{j=1}^{n'} \widehat{r}(\boldsymbol{x}'_j)^2 - \frac{1}{n}\sum_{i=1}^{n} \widehat{r}(\boldsymbol{x}_i) \\ &= \frac{1}{2}\boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \boldsymbol{\alpha}^\top \widehat{\boldsymbol{h}},\end{aligned}$$

where $\widehat{\boldsymbol{H}}$ is the $(n+1) \times (n+1)$ matrix defined by

$$\widehat{\boldsymbol{H}} := \frac{1}{n'}\sum_{j=1}^{n'} \boldsymbol{k}(\boldsymbol{x}'_j)\boldsymbol{k}(\boldsymbol{x}'_j)^\top,$$

and $\widehat{\boldsymbol{h}}$ is the $(n+1)$-dimensional vector defined by

$$\widehat{\boldsymbol{h}} := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{k}(\boldsymbol{x}_i). \tag{4}$$

By including a regularization term, the uLSIF optimization problem is formulated as follows.

$$\widehat{\boldsymbol{\alpha}} := \operatorname*{argmin}_{\boldsymbol{\alpha}} \left[\frac{1}{2}\boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \boldsymbol{\alpha}^\top \widehat{\boldsymbol{h}} + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha}\right], \tag{5}$$

where $\boldsymbol{\alpha}^\top \boldsymbol{\alpha}/2$ is a regularizer and λ (≥ 0) is the regularization parameter that controls the strength of regularization. By taking the derivative of the above objective function with respect to the parameter $\boldsymbol{\alpha}$ and equating it to zero, we can analytically obtain the solution $\widehat{\boldsymbol{\alpha}}$ as

$$\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_{n+1})^{-1}\widehat{\boldsymbol{h}}, \tag{6}$$

where $\boldsymbol{I}_{n+1}$ is the $(n+1)$-dimensional identity matrix. Finally, the density ratio estimator $\widehat{r}(\boldsymbol{x})$ is given by

$$\widehat{r}(\boldsymbol{x}) := \widehat{\boldsymbol{\alpha}}^\top \boldsymbol{k}(\boldsymbol{x}).$$

Thanks to the analytic-form expression, uLSIF is computationally more efficient than alternative density ratio estimators which involve non-linear optimization (Qin, 1998; Cheng & Chu, 2004; Huang et al., 2007; Sugiyama et al., 2008; Nguyen et al., 2010). It was theoretically shown that uLSIF possesses the optimal non-parametric convergence rate and optimal numerical stability (Kanamori et al., 2009b).

2.4 Model Selection by Cross-Validation

The practical performance of uLSIF depends on the choice of the kernel function (the kernel width σ in the case of Gaussian kernel (2)) and the regularization parameter λ. Model selection of uLSIF is possible based on *cross-validation* with respect to the error criterion J defined by Eq.(3) (Kanamori et al., 2009a).

More specifically, each of the sample sets $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{n}$ and $\mathcal{X}' = \{\boldsymbol{x}'_j\}_{j=1}^{n'}$ is divided into M disjoint sets[2] $\{\mathcal{X}_m\}_{m=1}^{M}$ and $\{\mathcal{X}'_m\}_{m=1}^{M}$. Then an uLSIF solution $\widehat{r}_m(\boldsymbol{x})$ is obtained using $\mathcal{X}\backslash\mathcal{X}_m$ and $\mathcal{X}'\backslash\mathcal{X}'_m$ (i.e., all samples without $\mathcal{X}_m$ and $\mathcal{X}'_m$), and its J-value for the hold-out samples $\mathcal{X}_m$ and $\mathcal{X}'_m$ is computed as

$$\widehat{J}_m^{\mathrm{CV}} := \frac{1}{2|\mathcal{X}'_m|} \sum_{\boldsymbol{x}' \in \mathcal{X}'_m} \widehat{r}_m(\boldsymbol{x}')^2 - \frac{1}{|\mathcal{X}_m|} \sum_{\boldsymbol{x} \in \mathcal{X}_m} \widehat{r}_m(\boldsymbol{x}),$$

where $|\mathcal{X}|$ denotes the number of elements in the set $\mathcal{X}$. This procedure is repeated for $m = 1, \ldots, M$, and the average of $\widehat{J}_m^{\mathrm{CV}}$ over all m is computed as

$$\widehat{J}^{\mathrm{CV}} := \frac{1}{M} \sum_{m=1}^{M} \widehat{J}_m^{\mathrm{CV}}.$$

Finally, the model (the kernel width σ and the regularization parameter λ in the current setup) that minimizes $\widehat{J}^{\mathrm{CV}}$ is chosen as the most suitable one.

3 Divergence Estimation

In this section, we describe a divergence estimator based on uLSIF, and investigate its theoretical properties.

[2] $M = 5$ seems to be a popular choice (Hastie et al., 2001). We also follow this 'rule-of-thumb' choice in this paper.

3.1 Formulation of Divergence Estimation

Let us consider the *Pearson divergence* (Pearson, 1900) from P to P' as a discrepancy measure between P and P', which is defined and expressed as follows:

$$\begin{aligned}\mathrm{PE}(P,P') &:= \frac{1}{2}\int\left(\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}-1\right)^2 p'(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &= \frac{1}{2}\int r(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int r(\boldsymbol{x})p'(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \frac{1}{2},\end{aligned} \tag{7}$$

where $r(\boldsymbol{x})$ is the density ratio function defined by

$$r(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}.$$

$\mathrm{PE}(P,P')$ vanishes if and only if $P = P'$. The Pearson divergence is a squared-loss variant of the *Kullback-Leibler divergence* (Kullback & Leibler, 1951), and is an instance of the *f-divergences*, which are also known as the *Csiszár f-divergences* (Csiszár, 1967) or the *Ali-Silvey distances* (Ali & Silvey, 1966).

3.2 uLSIF-based Pearson Divergence Estimation

Approximating the expectations in Eq.(7) by empirical averages and replacing the density ratio function $r(\boldsymbol{x})$ by an uLSIF-based estimator $\widehat{r}(\boldsymbol{x})$, we have the following Pearson divergence estimator:

$$\begin{aligned}\widehat{\mathrm{PE}}(\mathcal{X},\mathcal{X}') &:= \frac{1}{2n}\sum_{i=1}^{n}\widehat{r}(\boldsymbol{x}_i) - \frac{1}{n'}\sum_{j=1}^{n'}\widehat{r}(\boldsymbol{x}'_j) + \frac{1}{2}\\ &= \frac{1}{2}\widehat{\boldsymbol{\alpha}}^\top\widehat{\boldsymbol{h}} - \widehat{\boldsymbol{\alpha}}^\top\widehat{\boldsymbol{h}}' + \frac{1}{2},\end{aligned} \tag{8}$$

where $\widehat{\boldsymbol{\alpha}}$ is given by Eq.(6), $\widehat{\boldsymbol{h}}$ is defined by Eq.(4), and $\widehat{\boldsymbol{h}}'$ is the $(n+1)$-dimensional vector defined by

$$\widehat{\boldsymbol{h}}' := \frac{1}{n'}\sum_{j=1}^{n'}\boldsymbol{k}(\boldsymbol{x}'_j).$$

Note that $\widehat{\mathrm{PE}}(\mathcal{X},\mathcal{X}')$ can take a negative value, although the true $\mathrm{PE}(P,P')$ is non-negative by definition. Thus, the estimation accuracy of $\widehat{\mathrm{PE}}(\mathcal{X},\mathcal{X}')$ can be improved by taking its positive part by rounding up a negative estimate to zero. However, we do not employ this rounding-up strategy here since we are interested in the relative *ranking* of the divergence estimates, as explained in Section 4.1.

3.3 Theoretical Properties

Here, let us theoretically investigate asymptotic properties of the uLSIF-based Pearson divergence estimator $\widehat{\mathrm{PE}}(\mathcal{X}, \mathcal{X}')$. More specifically, we show the asymptotic convergence rate of our non-parametric estimator $\widehat{\mathrm{PE}}(\mathcal{X}, \mathcal{X}')$ to the true $\mathrm{PE}(P, P')$.

Since the derivation of the convergence rate is highly technical, we defer all the technical details in Appendix A. Here, we focus on explaining the insight we can gain from our theoretical analysis.

Theorem 1. *Under the technical assumptions described in Appendix A, we have*

$$|\widehat{\mathrm{PE}}(\mathcal{X}, \mathcal{X}') - \mathrm{PE}(P, P')| = \mathcal{O}_p\left(\left(\frac{\log \overline{n}}{\overline{n}}\right)^{\frac{2}{2+\gamma}} + C\left(\frac{\log \overline{n}}{\overline{n}}\right)^{\frac{1}{2+\gamma}}\right), \tag{9}$$

where

$$C := \sqrt{\int (r(\boldsymbol{x}) - 1)^2 p'(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}}. \tag{10}$$

$\mathcal{O}_p$ *denotes the asymptotic order in probability,* $\overline{n} := \min(n, n')$*, and* γ *(*$0 < \gamma < 1$*) is a constant determined by the kernel function* $K(\cdot, \cdot)$*.*

The above theorem means that the convergence rate of $\widehat{\mathrm{PE}}(\mathcal{X}, \mathcal{X}')$ to $\mathrm{PE}(P, P')$ is $\left(\frac{\log \overline{n}}{\overline{n}}\right)^{\frac{1}{2+\gamma}}$ in general. However, when the two distributions P and P' are the same, $r(\boldsymbol{x}) = 1$ and thus $C = 0$ (see Eq.(10)). Then, the $\mathcal{O}_p\left(\left(\frac{\log \overline{n}}{\overline{n}}\right)^{\frac{1}{2+\gamma}}\right)$-term in Eq.(9) disappears, and therefore our estimator possesses an even faster convergence rate $\mathcal{O}_p\left(\left(\frac{\log \overline{n}}{\overline{n}}\right)^{\frac{2}{2+\gamma}}\right)$.

4 Least-Squares Two-Sample Test

Theoretical properties of our Pearson divergence estimator $\widehat{\mathrm{PE}}(\mathcal{X}, \mathcal{X}')$ have been elucidated above. In this section, we propose a two-sample test based on $\widehat{\mathrm{PE}}(\mathcal{X}, \mathcal{X}')$. We first describe a basic procedure of our two-sample test, and study its theoretical properties. Then we illustrate its behavior using toy datasets, and discuss practical issues for improving the performance.

4.1 Permutation Test with Finite Samples

Our two-sample test is based on the *permutation test* (Efron & Tibshirani, 1993).

We first run the uLSIF-based Pearson divergence estimation procedure using the original datasets $\mathcal{X}$ and $\mathcal{X}'$, and obtain a Pearson divergence estimate $\widehat{\mathrm{PE}}(\mathcal{X}, \mathcal{X}')$. Next, we randomly permute the $|\mathcal{X} \cup \mathcal{X}'|$ samples, and assign the first $|\mathcal{X}|$ samples to a set $\widetilde{\mathcal{X}}$ and the remaining $|\mathcal{X}'|$ samples to another set $\widetilde{\mathcal{X}}'$. Then we run the uLSIF-based Pearson

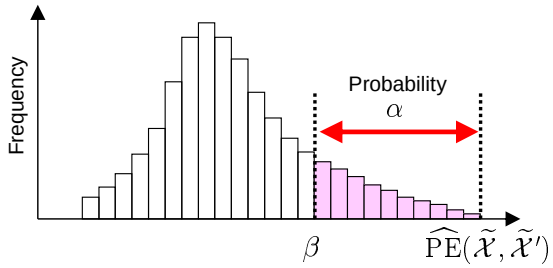


Figure 1: The role of the variables α and β in Theorem 2.

divergence estimation procedure again using the randomly shuffled datasets $\widetilde{\mathcal{X}}$ and $\widetilde{\mathcal{X}}'$, and obtain a Pearson divergence estimate $\widehat{\mathrm{PE}}(\widetilde{\mathcal{X}},\widetilde{\mathcal{X}}')$. Since $\widetilde{\mathcal{X}}$ and $\widetilde{\mathcal{X}}'$ can be regarded as being drawn from the same distribution, $\widehat{\mathrm{PE}}(\widetilde{\mathcal{X}},\widetilde{\mathcal{X}}')$ would take a value close to zero. This random shuffling procedure is repeated many times, and the distribution of $\widehat{\mathrm{PE}}(\widetilde{\mathcal{X}},\widetilde{\mathcal{X}}')$ under the null-hypothesis (i.e., the two distributions are the same) is constructed. Finally, the p-value is approximated by evaluating the relative ranking of $\widehat{\mathrm{PE}}(\mathcal{X},\mathcal{X}')$ in the distribution of $\widehat{\mathrm{PE}}(\widetilde{\mathcal{X}},\widetilde{\mathcal{X}}')$.

We refer to this procedure as the *least-squares two-sample test* (LSTT).

4.2 Theoretical Properties

Here, we investigate theoretical properties of the above permutation procedure under the null-hypothesis $P = P'$.

Theorem 2. *Suppose* $|\mathcal{X}| = |\mathcal{X}'|$*, and let* F *be the distribution function of* $\widehat{\mathrm{PE}}(\widetilde{\mathcal{X}},\widetilde{\mathcal{X}}')$*. Let*

$$\beta := \sup\{t \in \mathbb{R} \mid F(t) \leq 1-\alpha\}$$

be the upper 100α*-percentile point of* F *(see Figure 1). If* $P = P'$*, we have*

$$\mathrm{Prob}\left(\widehat{\mathrm{PE}}(\mathcal{X},\mathcal{X}') > \beta\right) \leq \alpha,$$

where '$\mathrm{Prob}(e)$' *denotes the probability of an event* e*.*

A proof of Theorem 2 is provided in Appendix B.

Theorem 2 means that, for a given significance level[3] α, the probability that $\widehat{\mathrm{PE}}(\mathcal{X},\mathcal{X}')$ exceeds β is at most α when $P = P'$. Thus, when the null hypothesis is correct, it will be properly accepted with a specified probability.

[3]Conventionally, $\alpha = 0.01$ or 0.05 is used.

4.3 Numerical Examples

Let the number of samples be $n = n' = 500$, and

$$\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{n} \overset{i.i.d.}{\sim} P = N(0,1),$$
$$\mathcal{X}' = \{\boldsymbol{x}'_j\}_{j=1}^{n'} \overset{i.i.d.}{\sim} P' = N(\mu, \sigma^2),$$

where $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. We consider the following four setups:

(a) $(\mu, \sigma) = (0, 1.3)$: P' has larger standard deviation than P,

(b) $(\mu, \sigma) = (0, 0.7)$: P' has smaller standard deviation than P,

(c) $(\mu, \sigma) = (0.3, 1)$: P and P' have different means,

(d) $(\mu, \sigma) = (0, 1)$: P and P' are the same.

Histograms of $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{n}$ and $\mathcal{X}' = \{\boldsymbol{x}'_j\}_{j=1}^{n'}$ for the above four cases are depicted in Figure 2. Examples of randomly shuffled samples $\widetilde{\mathcal{X}}$ are also plotted at the bottom, where $\widetilde{\mathcal{X}}$ is thought to follow $\frac{1}{2}N(0,1) + \frac{1}{2}N(\mu, \sigma^2)$. Since $\widetilde{\mathcal{X}}'$ has a similar histogram to $\widetilde{\mathcal{X}}$, its plot is omitted.

Figure 3 depicts histograms of $\widehat{\mathrm{PE}}(\widetilde{\mathcal{X}}, \widetilde{\mathcal{X}}')$ (i.e., shuffled datasets), showing that the profiles of the null distribution (i.e., the two distributions are the same) are rather similar to each other for the four cases. The values of $\widehat{\mathrm{PE}}(\mathcal{X}, \mathcal{X}')$ (i.e., the original datasets) are also plotted in Figure 3 using the '$\times$'-symbol on the horizontal axis, showing that the p-values tends to be small when $P \neq P'$ and the p-value is large when $P = P'$. This is desirable behavior as a hypothesis test.

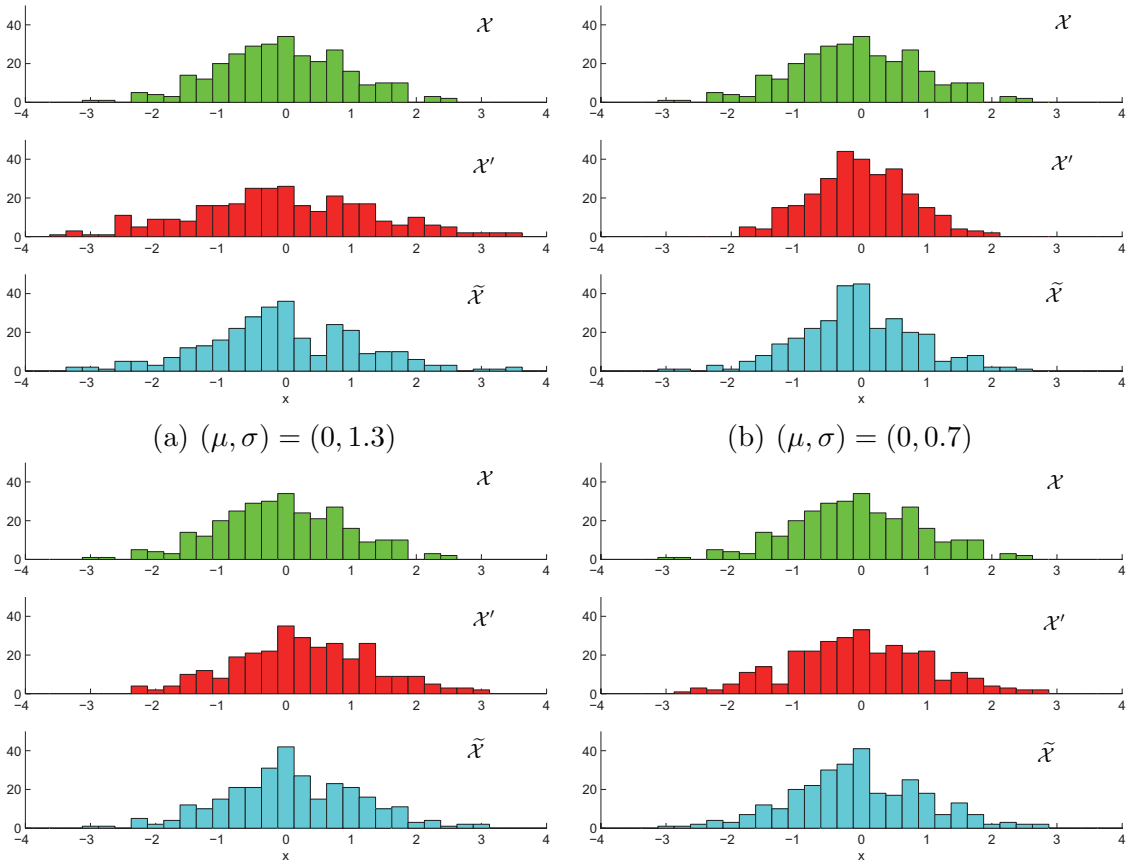
Figure 4 depicts the mean and standard deviation of p-values over 100 runs as functions of the sample size n $(= n')$, indicated by 'plain'. The graphs show that, when $P \neq P'$, the p-values tend to decrease as n increases. On the other hand, when $P = P'$, the p-values are almost unchanged and kept to relatively large values.

Figure 5 depicts the rate of accepting the null hypothesis (i.e., $P = P'$) over 100 runs when the significance level is set to 0.05 (i.e., the rate of p-values larger than 0.05). The graphs show that, when $P \neq P'$, the null hypothesis tends to be more frequently rejected as n increases. On the other hand, when $P = P'$, the null hypothesis is almost always accepted. Thus, the proposed test was shown to work properly for these toy datasets.

4.4 Choice of Numerator/Denominator Densities

In our test procedure, we are using uLSIF for estimating the density ratio function $r(\boldsymbol{x})$:

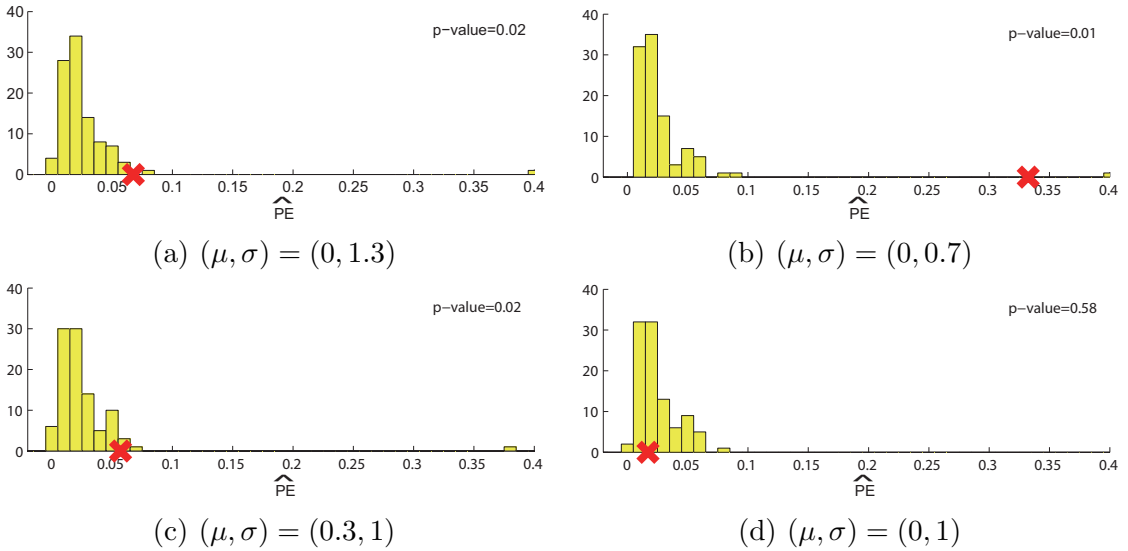
$$r(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}.$$



(a) $(\mu, \sigma) = (0, 1.3)$ (b) $(\mu, \sigma) = (0, 0.7)$

(c) $(\mu, \sigma) = (0.3, 1)$ (d) $(\mu, \sigma) = (0, 1)$

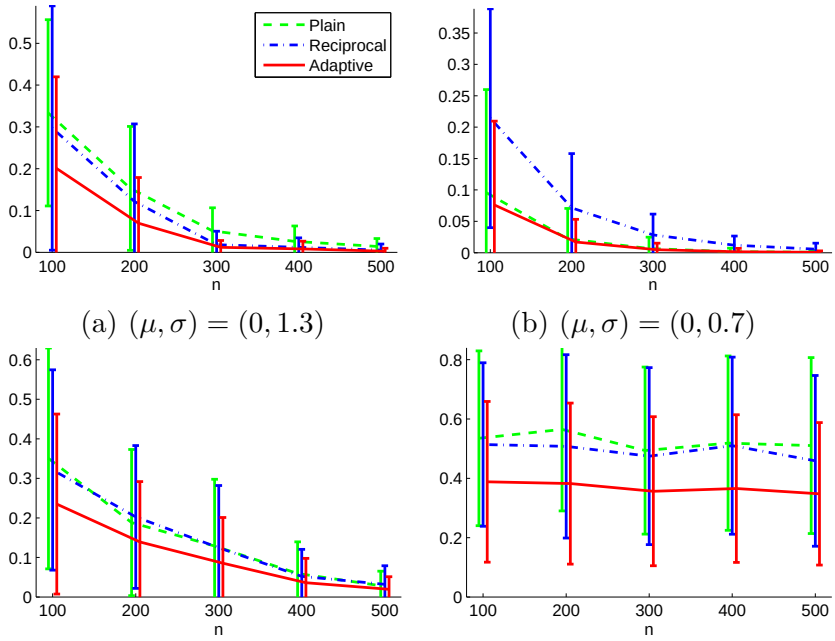
Figure 2: Histograms of original samples $\mathcal{X} \sim N(0, 1)$ and $\mathcal{X}' \sim N(\mu, \sigma^2)$, and the shuffled samples (which are thought to follow $\widetilde{\mathcal{X}} \sim \frac{1}{2}N(0, 1) + \frac{1}{2}N(\mu, \sigma^2)$) for the toy datasets.



(a) $(\mu, \sigma) = (0, 1.3)$ (b) $(\mu, \sigma) = (0, 0.7)$

(c) $(\mu, \sigma) = (0.3, 1)$ (d) $(\mu, \sigma) = (0, 1)$

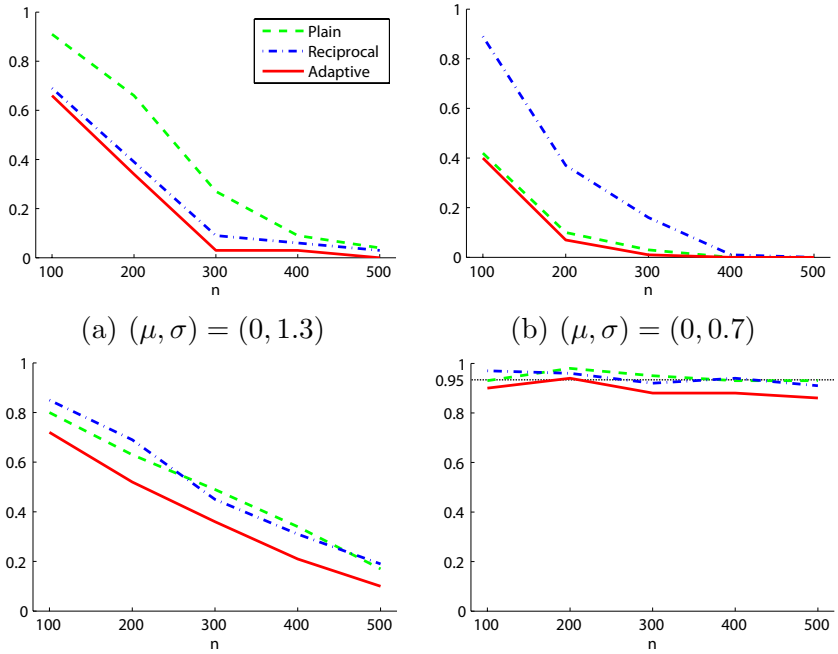
Figure 3: Histograms of $\widehat{\mathrm{PE}}(\widetilde{\mathcal{X}}, \widetilde{\mathcal{X}}')$ (i.e., shuffled datasets) for the toy datasets. '×' indicates the value of $\widehat{\mathrm{PE}}(\mathcal{X}, \mathcal{X}')$ (i.e., the original datasets).



(a) $(\mu, \sigma) = (0, 1.3)$ (b) $(\mu, \sigma) = (0, 0.7)$

(c) $(\mu, \sigma) = (0.3, 1)$ (d) $(\mu, \sigma) = (0, 1)$

Figure 4: Mean and standard deviation of p-values for the toy datasets.



(a) $(\mu, \sigma) = (0, 1.3)$ (b) $(\mu, \sigma) = (0, 0.7)$

(c) $(\mu, \sigma) = (0.3, 1)$ (d) $(\mu, \sigma) = (0, 1)$

Figure 5: The rate of accepting the null hypothesis (i.e., $P = P'$) for the toy datasets under the significance level 0.05.

By definition, the *reciprocal* of the density ratio $r(\boldsymbol{x})$,

$$\frac{1}{r(\boldsymbol{x})} = \frac{p'(\boldsymbol{x})}{p(\boldsymbol{x})},$$

is also a density ratio function, assuming that $p(\boldsymbol{x}) > 0$ for all $\boldsymbol{x}$. This means that we can use uLSIF in two ways, either estimating the original density ratio $r(\boldsymbol{x})$ or its reciprocal $1/r(\boldsymbol{x})$.

To illustrate this difference, we also carried out the same experiments as Section 4.3 by swapping $\mathcal{X}$ and $\mathcal{X}'$. The obtained p-values and the acceptance rate are also plotted in Figure 4 and Figure 5 as 'reciprocal'. In the experiments, we prefer to have smaller p-values when $P \neq P'$ and larger p-values when $P = P'$. The graphs show that, when $(\mu, \sigma) = (0, 1.3)$, estimating the inverted density ratio gives slightly smaller p-values and a significantly lower acceptance rate. On the other hand, when $(\mu, \sigma) = (0, 0.7)$, reciprocal estimation yields larger p-values and a significantly higher acceptance rate. When $(\mu, \sigma) = (0.3, 1)$ and $(\mu, \sigma) = (0, 1)$, the 'plain' and 'reciprocal' methods result in similar p-values and thus similar acceptance rates. These experimental results imply that, if we *adaptively* choose the plain and reciprocal approaches, the performance of homogeneity test may be improved.

Figure 4 showed that, when $P = P'$ (i.e., $(\mu, \sigma) = (0, 1)$), the p-values are large enough to reject the null hypothesis for both the plain and reciprocal approaches. Thus, the *type-I error* (the probability of rejecting correct null-hypotheses, i.e., two distributions are judged to be different when they are actually the same) would be sufficiently small for both approaches, as illustrated in Figure 5. Based on this observation, we propose to choose a smaller p-value between the plain and reciprocal approaches. This allows us to reduce the *type-II error* (the probability of accepting incorrect null-hypotheses, i.e., two distributions are judged to be the same when they are actually different), and thus the *power* of the test can be enhanced.

The experimental results of this adaptive method are also included in Figure 4 and Figure 5 as 'adaptive'. The results show that p-values obtained by the adaptive method are smaller than those obtained by the plain and reciprocal approaches, providing significant performance improvement when $P \neq P'$. On the other hand, smaller p-values can be problematic when $P = P'$ since the acceptance rate can be lowered. However, as the experimental results show, the p-values are still large enough to accept the null hypothesis and thus there is no critical performance degradation in this illustrative example.

A pseudo-code of the 'adaptive' LSTT method is summarized in Figure 6 and Figure 7. Although the permutation test process is computationally intensive, it can be easily parallelized using multi-processors/cores.

A MATLAB® implementation of LSTT is available from

http://sugiyama-www.cs.titech.ac.jp/~sugi/software/LSTT/

Input: Two sets of samples $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{n}$ and $\mathcal{X}' = \{\boldsymbol{x}'_j\}_{j=1}^{n'}$
Output: p-value $\widehat{p}$

$p_0 \longleftarrow \widehat{\mathrm{PE}}(\mathcal{X}, \mathcal{X}')$;
$p'_0 \longleftarrow \widehat{\mathrm{PE}}(\mathcal{X}', \mathcal{X})$;
For $t = 1, \ldots, T$
  Randomly split $\mathcal{X} \cup \mathcal{X}'$ into $\widetilde{\mathcal{X}}$ of size $|\mathcal{X}|$ and $\widetilde{\mathcal{X}}'$ of size $|\mathcal{X}'|$;
  $p_t \longleftarrow \widehat{\mathrm{PE}}(\widetilde{\mathcal{X}}, \widetilde{\mathcal{X}}')$;
  $p'_t \longleftarrow \widehat{\mathrm{PE}}(\widetilde{\mathcal{X}}', \widetilde{\mathcal{X}})$;
End
$p \longleftarrow \dfrac{1}{T}\displaystyle\sum_{t=1}^{T} I(p_t > p_0)$;
$p' \longleftarrow \dfrac{1}{T}\displaystyle\sum_{t=1}^{T} I(p'_t > p'_0)$;
$\widehat{p} \longleftarrow \min(p, p')$;

Figure 6: Pseudo code of LSTT. Pseudo code of $\widehat{\mathrm{PE}}(\mathcal{X}, \mathcal{X}')$ is given in Figure 7. $I(c)$ denotes the indicator function, i.e., $I(c) = 1$ if the condition c is true; otherwise $I(c) = 0$. When $|\widetilde{\mathcal{X}}| = |\widetilde{\mathcal{X}}'|$ (i.e., $n = n'$), $p'_t \longleftarrow \widehat{\mathrm{PE}}(\widetilde{\mathcal{X}}', \widetilde{\mathcal{X}})$ may be replaced by $p'_t \longleftarrow p_t$ since switching $\mathcal{X}$ and $\mathcal{X}'$ does not essentially affect the estimation of the Pearson divergence.

5 Maximum Mean Discrepancy

Maximum mean discrepancy (MMD; Borgwardt et al., 2006; Gretton et al., 2007) is a state-of-the-art method of homogeneity test. In this section, we review the definition of MMD and explain its basic properties. In the next section, the proposed LSTT is experimentally compared with MMD.

MMD is an *integral probability metric* (Müller, 1997) defined as

$$\mathrm{MMD}(\mathcal{H}, P, P') := \sup_{f \in \mathcal{H}} \left[\int f(\boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int f(\boldsymbol{x}) p'(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right], \tag{11}$$

where $\mathcal{H} : \mathbb{R}^d \to \mathbb{R}$ is some function class. When $\mathcal{H}$ is a unit ball in a *universal reproducing kernel Hilbert space* (universal RKHS; Steinwart, 2001 defined on a compact metric space, then $\mathrm{MMD}(\mathcal{H}, P, P')$ vanishes if and only if $P = P'$. Gaussian RKHSs are examples of the universal RKHS.

Let $K(\boldsymbol{x}, \boldsymbol{x}')$ be a reproducing kernel function. Then the reproducing property (Aronszajn, 1950) allows one to extract the value of a function $f \in \mathcal{H}$ at a point $\boldsymbol{x}$ as

$$f(\boldsymbol{x}) = \langle f(\cdot), K(\boldsymbol{x}, \cdot) \rangle_{\mathcal{H}}, \tag{12}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the inner product in the RKHS $\mathcal{H}$. Let $\|\cdot\|_{\mathcal{H}}$ be the norm in the

Input: Two sets of samples $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{n}$ and $\mathcal{X}' = \{\boldsymbol{x}'_j\}_{j=1}^{n'}$
Output: Pearson divergence estimate $\widehat{\mathrm{PE}}(\mathcal{X}, \mathcal{X}')$

Randomly split $\mathcal{X}$ into $\{\mathcal{X}_m\}_{m=1}^{M}$ and $\mathcal{X}'$ into $\{\mathcal{X}'_m\}_{m=1}^{M}$;
For each candidate of Gaussian width σ
  For $m = 1, \ldots, M$
    % $\boldsymbol{k}_\sigma(\boldsymbol{x}) = (1, K_\sigma(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, K_\sigma(\boldsymbol{x}, \boldsymbol{x}_n))^\top$
    % $K_\sigma(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2\sigma^2}\right)$
    $\widehat{\boldsymbol{G}}_m \longleftarrow \sum_{\boldsymbol{x}' \in \mathcal{X}'_m} \boldsymbol{k}_\sigma(\boldsymbol{x}') \boldsymbol{k}_\sigma(\boldsymbol{x}')^\top$;
    $\widehat{\boldsymbol{g}}_m \longleftarrow \sum_{\boldsymbol{x} \in \mathcal{X}_m} \boldsymbol{k}_\sigma(\boldsymbol{x})$;
  End
  For each candidate of regularization parameter λ
    For $m = 1, \ldots, M$
      $\widehat{\boldsymbol{\alpha}}_m \longleftarrow \left(\frac{1}{|\mathcal{X}' \backslash \mathcal{X}'_m|} \sum_{m' \neq m} \widehat{\boldsymbol{G}}_{m'} + \lambda \boldsymbol{I}_{n+1}\right)^{-1} \left(\frac{1}{|\mathcal{X} \backslash \mathcal{X}_m|} \sum_{m' \neq m} \widehat{\boldsymbol{g}}_{m'}\right)$;
      $\widehat{J}_m^{\mathrm{CV}}(\sigma, \lambda) \longleftarrow \frac{1}{2|\mathcal{X}'_m|} \widehat{\boldsymbol{\alpha}}_m^\top \widehat{\boldsymbol{G}}_m \widehat{\boldsymbol{\alpha}}_m - \frac{1}{|\mathcal{X}_m|} \widehat{\boldsymbol{\alpha}}_m^\top \widehat{\boldsymbol{g}}_m$;
    End
    $\widehat{J}^{\mathrm{CV}}(\sigma, \lambda) \longleftarrow \frac{1}{M} \sum_{m=1}^{M} \widehat{J}_m^{\mathrm{CV}}(\sigma, \lambda)$;
  End
End
$(\widehat{\sigma}, \widehat{\lambda}) \longleftarrow \operatorname{argmin}_{(\sigma,\lambda)} \widehat{J}^{\mathrm{CV}}(\sigma, \lambda)$;
$\widehat{\boldsymbol{h}} \longleftarrow \frac{1}{|\mathcal{X}|} \sum_{\boldsymbol{x} \in \mathcal{X}} \boldsymbol{k}_{\widehat{\sigma}}(\boldsymbol{x})$;
$\widehat{\boldsymbol{\alpha}} \longleftarrow \left(\frac{1}{|\mathcal{X}'|} \sum_{\boldsymbol{x}' \in \mathcal{X}'} \boldsymbol{k}_{\widehat{\sigma}}(\boldsymbol{x}') \boldsymbol{k}_{\widehat{\sigma}}(\boldsymbol{x}')^\top + \widehat{\lambda} \boldsymbol{I}_{n+1}\right)^{-1} \widehat{\boldsymbol{h}}$;
$\widehat{\mathrm{PE}}(\mathcal{X}, \mathcal{X}') \longleftarrow \frac{1}{2} \widehat{\boldsymbol{\alpha}}^\top \widehat{\boldsymbol{h}} - \widehat{\boldsymbol{\alpha}}^\top \left(\frac{1}{|\mathcal{X}'|} \sum_{\boldsymbol{x}' \in \mathcal{X}'} \boldsymbol{k}_{\widehat{\sigma}}(\boldsymbol{x}')\right) + \frac{1}{2}$;

Figure 7: Pseudo code of uLSIF-based Pearson divergence estimator.

RKHS $\mathcal{H}$. Then, one can explicitly express MMD in terms of the kernel function as

$$\begin{aligned}\mathrm{MMD}(\mathcal{H},P,P') &= \sup_{\|f\|_{\mathcal{H}}\leq 1}\left[\int \langle f(\cdot),K(\boldsymbol{x},\cdot)\rangle_{\mathcal{H}}\, p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int \langle f(\cdot),K(\boldsymbol{x},\cdot)\rangle_{\mathcal{H}}\, p'(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right]\\ &= \sup_{\|f\|_{\mathcal{H}}\leq 1}\left\langle f(\cdot), \int K(\boldsymbol{x},\cdot)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int K(\boldsymbol{x},\cdot)p'(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right\rangle_{\mathcal{H}}\\ &= \left\|\int K(\boldsymbol{x},\cdot)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int K(\boldsymbol{x},\cdot)p'(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right\|_{\mathcal{H}},\end{aligned}$$

where the *Cauchy-Schwarz inequality* (Bachman & Narici, 2000) was used in the last equality. Furthermore, by using

$$K(\boldsymbol{x},\boldsymbol{x}') = \langle K(\boldsymbol{x},\cdot),K(\boldsymbol{x}',\cdot)\rangle_{\mathcal{H}},$$

the squared MMD can be expressed as

$$\begin{aligned}\mathrm{MMD}^2(\mathcal{H},P,P') &= \left\|\int K(\boldsymbol{x},\cdot)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int K(\boldsymbol{x},\cdot)p'(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right\|_{\mathcal{H}}^2\\ &= \iint K(\boldsymbol{x},\boldsymbol{x}')p(\boldsymbol{x})p(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}' + \iint K(\boldsymbol{x},\boldsymbol{x}')p'(\boldsymbol{x})p'(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}'\\ &\quad - 2\iint K(\boldsymbol{x},\boldsymbol{x}')p(\boldsymbol{x})p'(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}'.\end{aligned}$$

The above expression allows one to immediately obtain an empirical estimator—with the i.i.d. samples $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^n$ following $p(\boldsymbol{x})$ and $\mathcal{X}' = \{\boldsymbol{x}'_j\}_{j=1}^{n'}$ following $p'(\boldsymbol{x})$, a consistent estimator of $\mathrm{MMD}^2(\mathcal{H},P,P')$ is given as

$$\begin{aligned}\widehat{\mathrm{MMD}^2}(\mathcal{H},\mathcal{X},\mathcal{X}') &:= \frac{1}{n^2}\sum_{i,i'=1}^{n} K(\boldsymbol{x}_i,\boldsymbol{x}_{i'}) + \frac{1}{n'^2}\sum_{j,j'=1}^{n'} K(\boldsymbol{x}'_j,\boldsymbol{x}'_{j'})\\ &\quad - \frac{2}{nn'}\sum_{i=1}^{n}\sum_{j=1}^{n'} K(\boldsymbol{x}_i,\boldsymbol{x}'_j).\end{aligned}$$

By the same permutation test procedure as the one described in Section 4.1, one can compute p-values for $\widehat{\mathrm{MMD}^2}(\mathcal{H},\mathcal{X},\mathcal{X}')$. Furthermore, an asymptotic distribution of $\widehat{\mathrm{MMD}^2}(\mathcal{H},\mathcal{X},\mathcal{X}')$ under $P = P'$ can be explicitly obtained (Borgwardt et al., 2006; Gretton et al., 2007). This allows one to compute the p-values without resorting to the computationally-intensive permutation procedure, which is an advantage of MMD over LSTT.

$\widehat{\mathrm{MMD}^2}(\mathcal{H},\mathcal{X},\mathcal{X}')$ depends on the choice of the universal RKHS $\mathcal{H}$. In the original MMD papers (Borgwardt et al., 2006; Gretton et al., 2007), the Gaussian RKHS with width set to the median distance between samples was used, which is a popular heuristic in the kernel method community (Schölkopf & Smola, 2002). Recently, an idea of using the

universal RKHS yielding the maximum MMD value has been introduced (Sriperumbudur et al., 2009). In the experiments in the next section, we use this maximum-MMD technique for choosing the universal RKHS, which we confirmed to work better than the 'median' heuristic.

6 Experiments

In this section, we report experimental results comparing the performance of the proposed LSTT (Section 4) with that of the state-of-the-art MMD (Section 5).

6.1 IDA Benchmark Datasets

In the first set of experiments, we used binary classification datasets taken from the *IDA repository* (Rätsch et al., 2001). For each dataset, we randomly split all the positive training samples into two disjoint sets, $\mathcal{X}$ and $\mathcal{X}'$ with $|\mathcal{X}| = |\mathcal{X}'|$.

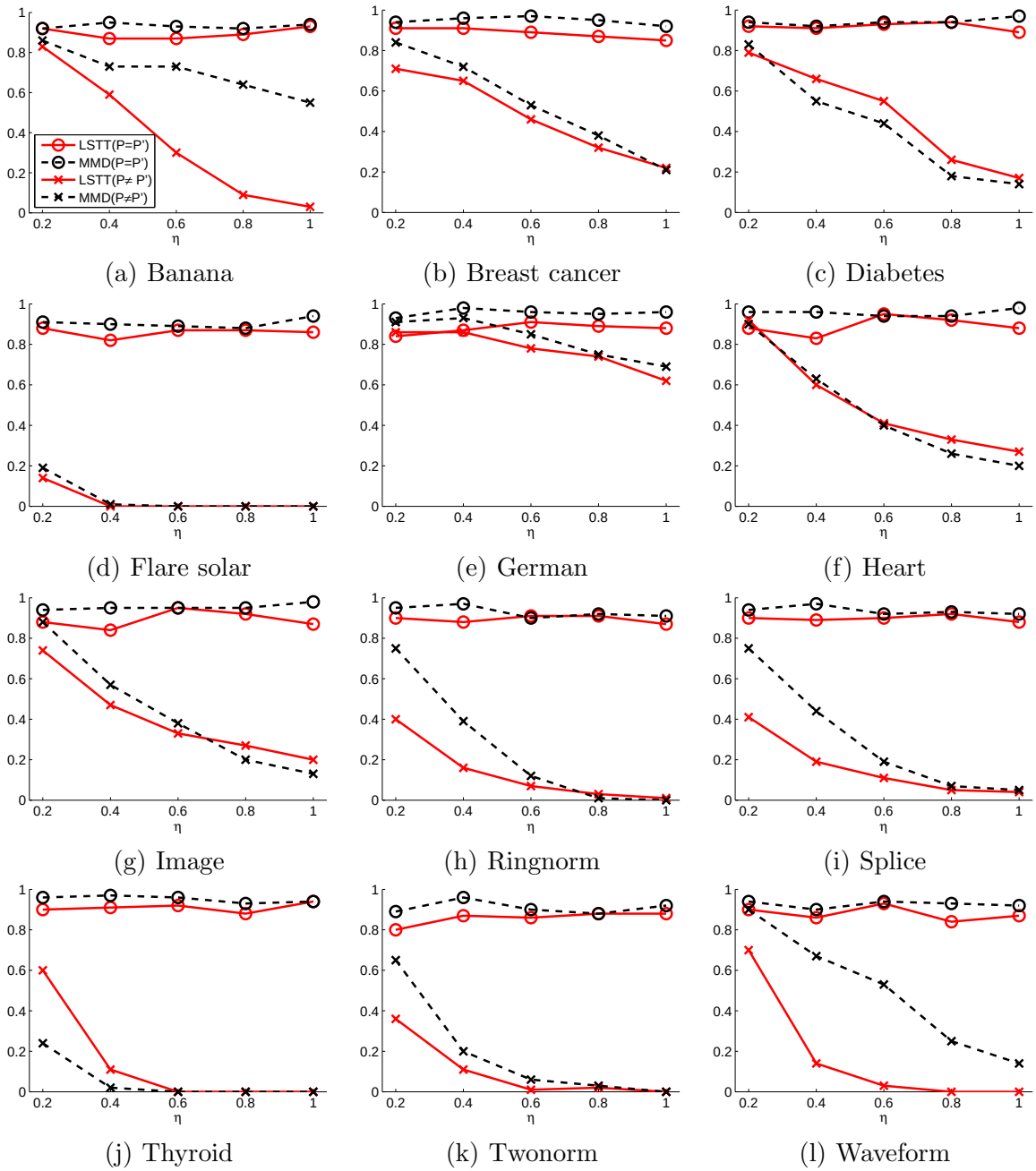
We first investigated whether the tests can correctly accept the null hypotheses (i.e., $\mathcal{X}$ and $\mathcal{X}'$ follow the same distribution). We used the Gaussian kernel both for LSTT and MMD. The Gaussian width and the regularization parameter in LSTT were determined by 5-fold cross-validation (see Section 2.4). The Gaussian width in MMD was chosen so that the MMD value is maximized (see Section 5). Since the permutation test procedures in LSTT and MMD are exactly the same, we are purely comparing the performance of the MMD and LSTT criteria in this experiment.

We investigated the rate of accepting the null hypothesis as functions of the relative sample size η for the significance level 0.05. The relative sample size η means that we used samples of size $\eta|\mathcal{X}|$ and $\eta|\mathcal{X}'|$ for homogeneity test. The experimental results are plotted in Figure 8 by lines with '$\circ$'-symbols. The results show that both methods almost always accepted the null hypothesis correctly, meaning that the type-I error is small enough for both MMD and LSTT. However, MMD seems to perform slightly better than LSTT in terms of the type-I error.

Next, we replaced a fraction of samples in the set $\mathcal{X}'$ by randomly chosen negative training samples. Thus, while $\mathcal{X}$ contains only positive training samples, $\mathcal{X}'$ includes both positive and negative training samples. The experimental results are also plotted in Figure 8 by lines with '$\times$'-symbols. The results show that LSTT tended to correctly reject the null hypothesis more frequently than MMD for the 'banana', 'ringnorm', 'splice', 'twonorm', and 'waveform' datasets. MMD worked better than LSTT for the 'thyroid' dataset, and the two methods were comparable to each other for the other datasets. Overall, LSTT compares favorably with MMD in terms of the type-II error.

6.2 USPS Hand-Written Digit Dataset

In the second sets of experiments, we used the *USPS hand-written digit* dataset provided by U.S. Postal Service (Hastie et al., 2001). Each digit image (representing an integer in



(a) Banana (b) Breast cancer (c) Diabetes

(d) Flare solar (e) German (f) Heart

(g) Image (h) Ringnorm (i) Splice

(j) Thyroid (k) Twonorm (l) Waveform

Figure 8: The rate of accepting the null hypothesis (i.e., $P = P'$) for IDA datasets under the significance level 0.05. η indicates the relative sample size we used in the experiments.

$\{0, 1, 2, \ldots, 9\}$) consists of 256 ($= 16 \times 16$) pixels, each of which takes a value between -1 to $+1$ representing its intensity level in gray-scale.

We formed two sets of samples as follows: one consists of 500 samples randomly chosen from class c ($\in \{0, 1, 2, \ldots, 9\}$), while the other consists of $500(1 - \delta)$ samples randomly chosen from class c and 500δ samples randomly chosen from another class c' ($\neq c$), where δ is the contamination rate. The goal is to test whether the two sets of samples are drawn from the same distribution or not for various contamination rates.

Table 1 shows the number of times LSTT or MMD incorrectly rejected the null hypothesis over 10 runs when the null hypothesis is correct (i.e., $\delta = 0$, meaning that the two distributions are the same). Thus, the smaller the number is, the better the performance is. The significance level was set to 0.05. The format 'l/m' in the table means that LSTT and MMD rejected the null hypothesis l and m times, respectively. The results show that both LSTT and MMD almost always accepted the correct null hypothesis successfully.

Next, we compared the performance of LSTT and MMD when the contamination rate was increased as $\delta = 0.02, 0.04, 0.06, \ldots, 0.2$. Table 2 shows the number of times LSTT or MMD rejected the null hypothesis with a lower contamination rate δ. The format '$l/t/m$' in the table means that LSTT rejected the null hypothesis with a lower contamination rate δ than MMD l times, and vice versa for m times. t denotes the number of times the smallest δ that LSTT and MMD rejected the null hypothesis was the same. The significance level was set to 0.05. The results show that LSTT tended to reject the null hypothesis with low contamination rate δ.

6.3 Brown Corpus Dataset with Tree Kernels

In the last set of experiments, we compared the performance of LSTT and MMD using natural language datasets.

We used the *Brown corpus* dataset[4], which is a carefully compiled selection of current American English. The Brown corpus consists of a million words sampled from 15 genres such as news and religion, and it is accompanied with *part-of-speech* tags, which represent relationship with adjacent and related words in a phrase, sentence, or paragraph. We converted the Brown corpus data to *dependency tree* representation by the *MaltParser*[5].

We prepared two sets of dependency trees as follows: one consists of 1000 samples taken from the 'news' category, and the other consists of $1000(1 - \delta)$ samples taken from the 'news' category and 1000δ samples taken from the 'romance' category, where δ is the contamination rate. The goal is to test whether the two sets of samples were drawn from the same distribution or not for various contamination rates.

We computed the *labeled ordered tree kernel* (Kashima & Koyanagi, 2002) between two dependency trees, which counts the number of sub-trees common to both trees. Then

[4]The Brown corpus dataset can be downloaded by using the *Natural Language Toolkit*, which contains open source Python modules, linguistic data, and documentation for research and development in natural language processing and text analysis. The Natural Language Toolkit is available from http://www.nltk.org/.

[5]The MaltParser is available from http://maltparser.org/.

Table 1: The experimental results for the USPS datasets. The number of times LSTT or MMD incorrectly rejected the null hypothesis over 10 runs when the null hypothesis was correct (i.e., the two distributions are the same). c in the table denotes the target class. The format 'l/m' means that LSTT and MMD rejected the null hypothesis l and m times, respectively. The significance level was set to 0.05.

c	0	1	2	3	4	5	6	7	8	9
	0/1	1/0	1/0	0/0	0/1	1/0	1/0	0/1	0/0	1/0

Table 2: The experimental results for the USPS datasets. The number of times LSTT or MMD rejected the null hypothesis with a smaller contamination rate. c denotes the target class and c' denotes the contamination class. The format '$l/t/m$' means that LSTT/MMD rejected the null hypothesis with a smaller contamination rate than MMD/LSTT l/m times, while the smallest contamination rate that LSTT and MMD rejected the null hypothesis was the same t times. The significance level was set to 0.05. The numbers are boldfaced if they are larger than or equal to 5.

$c\backslash c'$	0	1	2	3	4	5	6	7	8	9
0	–	**6**/2/2	**6**/3/1	**5**/1/4	**6**/2/2	**5**/2/3	**7**/2/1	**6**/1/3	**7**/0/3	**6**/3/1
1	**5**/4/1	–	4/3/3	3/2/**5**	3/1/**6**	3/2/**5**	3/4/3	3/1/**6**	2/3/**5**	3/1/**6**
2	2/**6**/2	2/2/**6**	–	2/1/**7**	3/4/3	2/1/**7**	4/0/**6**	2/0/**8**	3/1/**6**	**5**/2/3
3	**7**/3/0	4/4/2	**9**/1/0	–	**6**/4/0	1/3/**6**	**10**/0/0	1/**5**/4	**9**/1/0	4/**5**/1
4	**8**/1/1	3/2/**5**	**7**/2/1	**8**/0/2	–	**7**/1/2	**7**/2/1	2/3/**5**	**8**/1/1	4/2/4
5	**6**/3/1	**7**/2/1	**9**/1/0	4/2/4	**5**/4/1	–	**6**/3/1	4/3/3	**8**/1/1	**7**/3/0
6	**6**/2/2	**8**/2/0	**9**/1/0	**8**/1/1	**8**/1/1	**6**/2/2	–	**8**/2/0	**8**/2/0	**9**/1/0
7	**7**/2/1	**6**/2/2	**7**/1/2	**7**/1/2	**7**/0/3	**6**/2/2	**7**/1/2	–	**7**/1/2	**7**/0/3
8	**5**/3/2	3/1/**6**	**7**/1/2	**5**/2/3	3/4/3	4/3/3	**7**/1/2	3/2/**5**	–	**5**/1/4
9	**8**/1/1	**6**/3/1	**8**/0/2	**9**/0/1	4/2/4	**8**/0/2	**8**/1/1	1/1/**8**	**9**/0/1	–

the kernel values were directly fed into the LSTT and MMD algorithms. The labeled ordered tree kernel contains the *decay factor* parameter γ ($0 < \gamma \leq 1$), which controls the weights for large sub-trees (Collins & Duffy, 2002). We computed kernel values for $\gamma = 0.1, 0.4, 0.7$, and chose the one that minimized the cross-validation score in the case of LSTT and the one that maximized the MMD value in the case of MMD.

We first investigated the number of times LSTT or MMD incorrectly rejected the null hypothesis when the null hypothesis was correct (i.e., $\delta = 0$, meaning that the two distributions are the same). Thus, the smaller the number is, the better the performance is. The significance level was set to 0.05. The results were that LSTT rejected the correct null hypothesis 30 times out of 100 runs, while MMD rejected the correct null hypothesis only 8 times. Thus MMD gave smaller type-I error.

Next, we compared the performance of LSTT and MMD when the contamination rate was increased as $\delta = 0.05, 0.1, 0.15, \ldots, 0.35$. The significance level was set to 0.05. The results were that LSTT rejected the null hypothesis with a lower contamination rate δ

than MMD 60 times out of 100 runs, while MMD rejected the null hypothesis with a lower contamination rate δ than MMD only 18 times; The smallest δ that LSTT and MMD rejected the null hypothesis was the same 22 times. This means that LSTT tended to reject the null hypothesis with low contamination rate δ.

7 Conclusions

We proposed a novel method of testing homogeneity called the *least-squares two-sample test* (LSTT). Through various experiments, we overall confirmed that LSTT tends to produce smaller type-II error than the state-of-the-art MMD method, with slightly larger type-I error.

The performance of LSTT relies on the accuracy of density ratio estimation. We adopted *unconstrained least-squares importance fitting* (uLSIF; Kanamori et al., 2009a) since it possesses the optimal non-parametric convergence rate and optimal numerical stability (Kanamori et al., 2009b). uLSIF is computationally highly efficient thanks to the analytic-form solution, which is an attractive feature in the computationally-demanding permutation test procedure. Nevertheless, the permutation test procedure is still time consuming, so speedup is an important future research topic.

We have elucidated the convergence rate of our uLSIF-based Pearson divergence estimator. We further showed that our uLSIF-based Pearson divergence estimator even achieves a faster convergence rate when the two distributions are the same. An important future study along this line of research is to elucidate the asymptotic distribution of the LSTT estimator so that homogeneity testing can be carried out analytically.

Based on the uLSIF estimator $\widehat{r}(\boldsymbol{x})$, we constructed a *consistent* Pearson divergence estimator given by

$$\widehat{\mathrm{PE}}(\mathcal{X},\mathcal{X}') := \frac{1}{2n}\sum_{i=1}^{n}\widehat{r}(\boldsymbol{x}_i) - \frac{1}{n'}\sum_{j=1}^{n'}\widehat{r}(\boldsymbol{x}'_j) + \frac{1}{2}.$$

On the other hand, it is possible to construct different consistent estimators, e.g.,

$$\widehat{\mathrm{PE}}'(\mathcal{X},\mathcal{X}') := \frac{1}{2n}\sum_{i=1}^{n}\widehat{r}(\boldsymbol{x}_i) - \frac{1}{2},$$

$$\widehat{\mathrm{PE}}''(\mathcal{X},\mathcal{X}') := -\frac{1}{2n'}\sum_{j=1}^{n'}\widehat{r}(\boldsymbol{x}'_j)^2 + \frac{1}{n}\sum_{i=1}^{n}\widehat{r}(\boldsymbol{x}_i) - \frac{1}{2}.$$

$\widehat{\mathrm{PE}}'(\mathcal{X},\mathcal{X}')$ would be the simplest estimator, while $\widehat{\mathrm{PE}}''(\mathcal{X},\mathcal{X}')$ can be obtained as the *Legendre-Fenchel dual* of the Pearson divergence (Nguyen et al., 2010). Investigating theoretical and experimental performance of these variants in terms of accuracy and computational efficiency is left open as a future work.

Recently, novel approaches to density ratio estimation for high-dimensional problems have been explored (Sugiyama et al., 2010; Yamada et al., 2010; Sugiyama et al., 2011).

In our future work, we would like to incorporate these new ideas into the framework of LSTT and see how the test performance can be improved.

Acknowledgment

MS was supported by SCAT, AOARD, and the JST PRESTO program. TS was supported by MEXT Grant-in-Aid for Young Scientists (B) 22700289. TK was supported by MEXT Grant-in-Aid for Young Scientists 20700251. MK was supported by the JST PRESTO program.

A Proof of Theorem 1

In this section, we prove Theorem 1. For simplicity we consider a situation where $n = n'$. Even if $n \neq n'$, the following proof is valid for $n := \min(n, n')$.

For arbitrary function f, let

$$P_n f := \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i), \qquad P'_n f := \frac{1}{n}\sum_{j=1}^{n} f(\boldsymbol{x}'_j),$$
$$Pf := \mathrm{E}_{\boldsymbol{x}\sim p'}[f(\boldsymbol{x})], \qquad P'f := \mathrm{E}_{\boldsymbol{x}'\sim p}[f(\boldsymbol{x}')].$$

Let $\mathcal{G}$ be a reproducing kernel Hilbert space (RKHS) corresponding to a kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, and the estimated density ratio $\widetilde{g}$ is defined as the minimizer of the following minimization problem:

$$\widetilde{g} := \operatorname*{arg\,min}_{g\in\mathcal{G}} \frac{1}{2n}\sum_{j=1}^{n} g(\boldsymbol{x}'_j)^2 - \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{x}_i) + \frac{\lambda_n}{2}\|g\|_{\mathcal{G}}^2.$$

The estimated Pearson divergence $\widehat{\mathrm{PE}}$ is computed as

$$\widehat{\mathrm{PE}}(\mathcal{X}, \mathcal{X}') = \frac{1}{2}P_n\widetilde{g} - P'_n\widetilde{g} + \frac{1}{2}.$$

By Mercer's theorem, the kernel $K(\boldsymbol{x}, \boldsymbol{x}')$ has the following spectrum decomposition with respect to p':

$$K(\boldsymbol{x}, \boldsymbol{x}') = \sum_{k=1}^{\infty} e_k(\boldsymbol{x})\mu_k e_k(\boldsymbol{x}'),$$

where $\{e_k\}_{k=1}^{\infty}$ is an orthogonal system in $L_2(p')$, i.e., $Pe_k^2 = 1$ and $P(e_k e_{k'}) = 0$ for $k \neq k'$. Define $\mathcal{N}(\lambda)$ as

$$\mathcal{N}(\lambda) := \sum_{k=1}^{\infty} \frac{\mu_k}{\mu_k + \lambda}.$$

We assume the following conditions:

- $\sup_{\boldsymbol{x}\in\mathbb{R}^d} K(\boldsymbol{x},\boldsymbol{x}) \leq 1$,
- The constant function 1 is contained in $\mathcal{G}$: $1 \in \mathcal{G}$,
- The true density ratio p/p' is contained in $\mathcal{G}$: $p/p' = g^* \in \mathcal{G}$,
- There exists a constant $0 < \gamma < 1$ such that the spectrum μ_k of the kernel decays as $\mu_k \leq ck^{-\frac{2}{\gamma}}$, where c is a positive constant.

Then we obtain the following theorem and lemma (these are more precise versions of Theorem 1).

Theorem 1'. *Under the assumption described above, for* $\lambda_n = \left(\frac{\log n}{n}\right)^{2/(2+\gamma)}$, *we have*

$$|\widehat{\mathrm{PE}}(\mathcal{X},\mathcal{X}') - \mathrm{PE}(P,P')| = \mathcal{O}_p\left(\left(\frac{\log n}{n}\right)^{\frac{2}{2+\gamma}} + \sqrt{P'(g^*-1)^2}\left(\frac{\log n}{n}\right)^{\frac{1}{2+\gamma}}\right).$$

Lemma 1. *Suppose that the assumption described above hold and*

$$n \geq \frac{64\log^2(12/\eta)\mathcal{N}(\lambda_n)}{\lambda_n}. \tag{13}$$

Then we have

$$\begin{aligned}
&|\widehat{\mathrm{PE}}(\mathcal{X},\mathcal{X}') - \mathrm{PE}(P,P')| \\
&\leq 8\log(12/\eta)^2\left(\frac{16}{n^2\lambda_n} + \frac{(\|g^*\|_{\mathcal{G}} + \sqrt{\|g^*\|_{\mathcal{G}}} + \|g^*\|_{\mathcal{G}}^{\frac{3}{2}})\mathcal{N}(\lambda_n)}{n}\right) \\
&+ \log(12/\eta)\left(\frac{4\|g^*\|_{\mathcal{G}}}{n} + (\|g^*\|_{\mathcal{G}} + \|g^*\|_{\mathcal{G}}^{\frac{3}{2}})\sqrt{\frac{\lambda_n\mathcal{N}(\lambda_n)}{n}}\right) \\
&+ \frac{3}{2}\lambda_n C_{n,\eta} \\
&+ \log(12/\eta)\left(\frac{4\|g^*-1\|_\infty}{n} + \sqrt{\frac{P'(g^*-1)^2}{n}} + \sqrt{\frac{P(g^*-1)^2}{n}}\right) \\
&+ \frac{1}{2}\sqrt{P'(g^*-1)^2}\sqrt{128\log^2(12/\eta)\left(\frac{8}{n^2\lambda_n} + \frac{(\|g^*\|_{\mathcal{G}} + \|g^*\|_{\mathcal{G}}^2)\mathcal{N}(\lambda_n)}{n}\right) + 2\lambda_n\|g^*\|_{\mathcal{G}}^2},
\end{aligned} \tag{14}$$

with probability at least $1-\eta$, *where*

$$C_{n,\eta} = \frac{\|1\|_{\mathcal{G}}^2}{1+\lambda_n\|1\|_{\mathcal{G}}^2}\left\{8\left(\|g^*\|_{\mathcal{G}}^2 + 8\log^2(12/\eta)\left(\frac{4}{\lambda_n^2 n^2} + \frac{\mathcal{N}(\lambda_n)\|g^*\|_{\mathcal{G}}}{\lambda_n n}\right)\right) + 1\right\}.$$

Before proving the lemma, we introduce the following proposition that is a part of Proposition 2 in Caponnetto and de Vito (2007).

Proposition 1. *Let ξ be a random variable taking values in a real separable Hilbert space $\mathcal{K}$ on a probability space $(\Omega, \mathcal{F}, P)$. Assume that there are two positive constants L and σ such that*

$$\|\xi\|_{\mathcal{K}} \leq \frac{L}{2} \quad a.s., \tag{15}$$

$$\mathrm{E}[\|\xi\|_{\mathcal{K}}^2] \leq \sigma^2. \tag{16}$$

Then, for all $n \geq 1$ and $0 < \eta < 1$, it holds that

$$\mathrm{Prob}_{(\omega_1,\ldots,\omega_n)\sim P^n}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\xi(\omega_i) - \mathrm{E}[\xi]\right\|_{\mathcal{K}} \leq 2\left(\frac{L}{n} + \frac{\sigma}{\sqrt{n}}\right)\log\frac{2}{\eta}\right] \geq 1 - \eta. \tag{17}$$

Proof of Lemma 1. First we define some notation. Let $K_{\boldsymbol{x}}$ be an element of $\mathcal{G}$ such that

$$\langle K_{\boldsymbol{x}}, f\rangle = f(\boldsymbol{x})$$

for $f \in \mathcal{G}$ and $\boldsymbol{x} \in \mathbb{R}^d$, i.e., $K_{\boldsymbol{x}}(\cdot) = K(\boldsymbol{x}, \cdot)$ as an element of $\mathcal{G}$. We define $T_{p'} : \mathcal{G} \to \mathcal{G}$ as

$$\langle g, T_{p'} f\rangle = \mathrm{E}_{\boldsymbol{x}'\sim p'}[g(\boldsymbol{x}')f(\boldsymbol{x}')],$$

for $f, g \in \mathcal{G}$. Similarly we define $\widehat{T}_{p'} : \mathcal{G} \to \mathcal{G}$ as

$$\langle g, \widehat{T}_{p'} f\rangle = \frac{1}{n}\sum_{j=1}^{n} g(\boldsymbol{x}'_j)f(\boldsymbol{x}'_j).$$

Note that $T_{p'} = \mathrm{E}_{\boldsymbol{x}'\sim p'}[K_{\boldsymbol{x}'}K^{\circ}_{\boldsymbol{x}'}]$ where $K^{\circ}_{\boldsymbol{x}}$ is the adjoint of $K_{\boldsymbol{x}}$. Let $\phi_k := \sqrt{\mu_k}e_k$. Then $\{\phi_k\}_{k=1}^{\infty}$ is a complete orthonormal system in the RKHS $\mathcal{G}$, and $T_{p'}$ can be represented as

$$T_{p'} = \sum_{k=1}^{\infty}\phi_k\mu_k\phi_k^{\circ}.$$

Let $h_1, \widehat{h}_1, h_2, \widehat{h}_2 \in \mathcal{G}$ be

$$h_1 := \mathrm{E}_{\boldsymbol{x}'\sim p'}[K_{\boldsymbol{x}'}], \quad \widehat{h}_1 = \frac{1}{n}\sum_{j=1}^{n} K_{\boldsymbol{x}'_j},$$

$$h_2 := \mathrm{E}_{\boldsymbol{x}\sim p}[K_{\boldsymbol{x}}] = \mathrm{E}_{\boldsymbol{x}'\sim p'}[K_{\boldsymbol{x}'}g^*(\boldsymbol{x}')] = \mathrm{E}_{\boldsymbol{x}'\sim p'}[K_{\boldsymbol{x}'}\langle K_{\boldsymbol{x}'}, g^*\rangle_{\mathcal{G}}] = T_{p'}g^*, \quad \widehat{h}_2 = \frac{1}{n}\sum_{i=1}^{n} K_{\boldsymbol{x}_i}.$$

Note that $\mathrm{E}[\widehat{h}_1] = h_1$ and $\mathrm{E}[\widehat{h}_2] = h_2$, and

$$\langle h_1, f\rangle = P'f, \quad \langle \widehat{h}_1, f\rangle = P'_n f, \quad \langle h_2, f\rangle = Pf, \quad \langle \widehat{h}_2, f\rangle = P_n f. \tag{18}$$

It can be easily checked that

$$\widetilde{g} = (\widehat{T}_{p'} + \lambda_n)^{-1}\widehat{h}_2.$$

Here we define

$$g_{\lambda_n} = (T_{p'} + \lambda_n)^{-1} h_2.$$

The difference between $\widehat{\mathrm{PE}}(\mathcal{X}, \mathcal{X}')$ and $\mathrm{PE}(P, P')$ is expanded as

$$\begin{aligned}
&\widehat{\mathrm{PE}}(\mathcal{X}, \mathcal{X}') - \mathrm{PE}(P, P') \\
&= \frac{1}{2}(P_n \widetilde{g} - P g^*) - (P'_n \widetilde{g} - P' g^*) \\
&= \frac{1}{2}[(P_n - P)(\widetilde{g} - g^*) + P(\widetilde{g} - g^*) + (P_n - P) g^*] - (P'_n \widetilde{g} - 1). \qquad (19)
\end{aligned}$$

Since $P(\widetilde{g} - g^*)$ is bounded as

$$\begin{aligned}
|P(\widetilde{g} - g^*)| &= |P'(\widetilde{g} - g^*)| + |P'((g^* - 1)(\widetilde{g} - g^*))| \\
&\leq |P'(\widetilde{g} - g^*)| + \sqrt{P'(g^* - 1)^2} \sqrt{P'(\widetilde{g} - g^*)^2} \\
&= |(P' - P'_n)(\widetilde{g} - g^*) + P'_n \widetilde{g} - P' g^* + (P' - P'_n) g^*| \\
&\quad + \sqrt{P'(g^* - 1)^2} \sqrt{P'(\widetilde{g} - g^*)^2} \\
&\leq |(P' - P'_n)(\widetilde{g} - g^*)| + |P'_n \widetilde{g} - 1| + |(P' - P'_n) g^*| \\
&\quad + \sqrt{P'(g^* - 1)^2} \sqrt{P'(\widetilde{g} - g^*)^2}, \qquad (20)
\end{aligned}$$

Eq.(19) indicates

$$\begin{aligned}
|\widehat{\mathrm{PE}}(\mathcal{X}, \mathcal{X}') - \mathrm{PE}(P, P')| \leq & \frac{1}{2}|(P'_n - P')(\widetilde{g} - g^*)| + \frac{1}{2}|(P_n - P)(\widetilde{g} - g^*)| \\
& + \frac{3}{2}|P'_n \widetilde{g} - 1| \\
& + \frac{1}{2}|(P'_n - P') g^*| + \frac{1}{2}|(P_n - P) g^*| \\
& + \frac{1}{2}\sqrt{P'(g^* - 1)^2} \sqrt{P'(\widetilde{g} - g^*)^2}. \qquad (21)
\end{aligned}$$

Step 1. Bounding $(P'_n - P')(\widetilde{g} - g^*)$

$$\begin{aligned}
&(P'_n - P')(\widetilde{g} - g^*) \\
&= \langle \widehat{h}_1 - h_1, (\widehat{T}_{p'} + \lambda_n)^{-1} \widehat{h}_2 - g^* \rangle \\
&= \langle \widehat{h}_1 - h_1, (\widehat{T}_{p'} + \lambda_n)^{-1} (\widehat{h}_2 - h_2) + (\widehat{T}_{p'} + \lambda_n)^{-1} h_2 - (T_{p'} + \lambda_n)^{-1} h_2 + (T_{p'} + \lambda_n)^{-1} h_2 - g^* \rangle \\
&= \langle \widehat{h}_1 - h_1, (\widehat{T}_{p'} + \lambda_n)^{-1} (\widehat{h}_2 - h_2) \rangle + \langle \widehat{h}_1 - h_1, (\widehat{T}_{p'} + \lambda)^{-1} (T_{p'} - \widehat{T}_{p'}) (T_{p'} + \lambda_n)^{-1} h_2 \rangle \\
&\quad + \langle \widehat{h}_1 - h_1, (T_{p'} + \lambda_n)^{-1} h_2 - g^* \rangle \\
&= \underbrace{\langle \widehat{h}_1 - h_1, (\widehat{T}_{p'} + \lambda_n)^{-1} (\widehat{h}_2 - h_2) \rangle}_{\text{(1-a)}} + \underbrace{\langle \widehat{h}_1 - h_1, (\widehat{T}_{p'} + \lambda_n)^{-1} (T_{p'} - \widehat{T}_{p'}) (T_{p'} + \lambda_n)^{-1} h_2 \rangle}_{\text{(1-b)}} \\
&\quad - \underbrace{\langle \widehat{h}_1 - h_1, (T_{p'} + \lambda)^{-1} \lambda_n g^* \rangle}_{\text{(1-c)}}, \qquad (22)
\end{aligned}$$

where in the last inequality we used the relation $T_{p'}g^* = h_2$.

Let $\|\cdot\|_{\mathcal{L}(\mathcal{G})}$ be the operator norm of the bounded linear operator from $\mathcal{G}$ to $\mathcal{G}$. Then

$$\begin{aligned}
&\|(T_{p'}+\lambda)^{\frac{1}{2}}(\widehat{T}_{p'}+\lambda_n)^{-1}(T_{p'}+\lambda_n)^{\frac{1}{2}}\|_{\mathcal{L}(\mathcal{G})} \\
&= \left\|[(T_{p'}+\lambda_n)^{-\frac{1}{2}}(\widehat{T}_{p'}+\lambda_n - T_{p'} - \lambda_n + T_{p'} + \lambda_n)(T_{p'}+\lambda_n)^{-\frac{1}{2}}]^{-1}\right\|_{\mathcal{L}(\mathcal{G})} \\
&= \left\|[(T_{p'}+\lambda_n)^{-\frac{1}{2}}(\widehat{T}_{p'} - T_{p'})(T_{p'}+\lambda_n)^{-\frac{1}{2}} + I]^{-1}\right\|_{\mathcal{L}(\mathcal{G})}. \qquad (23)
\end{aligned}$$

We define $\mathcal{A}_1$ as follows:

$$\mathcal{A}_1 = \left\{ \left\|(T_{p'} - \widehat{T}_{p'})(T_{p'}+\lambda_n)^{-1}\right\|_{\mathcal{L}(\mathcal{G})} \leq \frac{1}{2} \right\}.$$

Caponnetto and de Vito (2007) showed that under the event $\mathcal{A}_1$,

$$\left\|(T_{p'}+\lambda_n)^{-\frac{1}{2}}(\widehat{T}_{p'} - T_{p'})(T_{p'}+\lambda_n)^{-\frac{1}{2}}\right\|_{\mathcal{L}(\mathcal{G})} \leq \frac{1}{2},$$

and the probability of $\mathcal{A}_1$ is at least $1-\eta/6$ under the condition Eq.(13). Therefore we obtain

$$\begin{aligned}
&\|(T_{p'}+\lambda_n)^{\frac{1}{2}}(\widehat{T}_{p'}+\lambda_n)^{-1}(T_{p'}+\lambda_n)^{\frac{1}{2}}\|_{\mathcal{L}(\mathcal{G})} \\
&= \left\|[(T_{p'}+\lambda_n)^{-\frac{1}{2}}(\widehat{T}_{p'} - T_{p'})(T_{p'}+\lambda_n)^{-\frac{1}{2}} + I]^{-1}\right\|_{\mathcal{L}(\mathcal{G})} \\
&\leq 2 \qquad (24)
\end{aligned}$$

on the event $\mathcal{A}_1$.

Bounding (1-a):

$$\begin{aligned}
&\langle \widehat{h}_1 - h_1, (\widehat{T}_{p'}+\lambda_n)^{-1}(\widehat{h}_2 - h_2)\rangle \\
&\leq \langle \widehat{h}_1 - h_1, (T_{p'}+\lambda_n)^{-\frac{1}{2}}[(T_{p'}+\lambda_n)^{\frac{1}{2}}(\widehat{T}_{p'}+\lambda_n)^{-1}(T_{p'}+\lambda_n)^{\frac{1}{2}}](T_{p'}+\lambda_n)^{-\frac{1}{2}}(\widehat{h}_2 - h_2)\rangle \\
&\leq \left\|(T_{p'}+\lambda_n)^{-\frac{1}{2}}(\widehat{h}_1 - h_1)\right\|_{\mathcal{G}} \left\|(T_{p'}+\lambda_n)^{\frac{1}{2}}(\widehat{T}_{p'}+\lambda_n)^{-1}(T_{p'}+\lambda_n)^{\frac{1}{2}}\right\|_{\mathcal{L}(\mathcal{G})} \\
&\quad \times \left\|(T_{p'}+\lambda_n)^{-\frac{1}{2}}(\widehat{h}_2 - h_2)\right\|_{\mathcal{G}}.
\end{aligned}$$

According to Eq.(24), we have

$$\left\|(T_{p'}+\lambda_n)^{\frac{1}{2}}(\widehat{T}_{p'}+\lambda_n)^{-1}(T_{p'}+\lambda_n)^{\frac{1}{2}}\right\|_{\mathcal{L}(\mathcal{G})} \leq 2$$

on the event $\mathcal{A}_1$.

Let $\xi : \mathbb{R}^d \to \mathcal{G}$ be the random variable

$$\xi(\boldsymbol{x}') = (T_{p'}+\lambda_n)^{-\frac{1}{2}} K_{\boldsymbol{x}'}.$$

Then

$$\begin{aligned}
(T_{p'}+\lambda_n)^{-\frac{1}{2}}(\widehat{h}_1-h_1) &= (P'_n-P')\xi \\
\|\xi\|_{\mathcal{G}} &= \sqrt{K^{\circ}_{\boldsymbol{x}'}(T_{p'}+\lambda_n)^{-1}K_{\boldsymbol{x}'}} \leq \sqrt{\lambda_n^{-1}}, \\
\mathrm{E}_{\boldsymbol{x}'\sim p'}[\|\xi\|^2_{\mathcal{G}}] &= \mathrm{E}_{\boldsymbol{x}'\sim p'}[K^{\circ}_{\boldsymbol{x}}(T_{p'}+\lambda_n)^{-1}K_{\boldsymbol{x}'}] \\
&= \mathrm{E}_{\boldsymbol{x}'\sim p'}\left[\sum_{k=1}^{\infty}\frac{\mu_k}{\mu_k+\lambda_n}e_k(\boldsymbol{x}')^2\right] \\
&= \sum_{k=1}^{\infty}\frac{\mu_k}{\mu_k+\lambda_n} \\
&= \mathcal{N}(\lambda_n).
\end{aligned}$$

Therefore, by Proposition 1, we have

$$\|(T_{p'}+\lambda_n)^{-\frac{1}{2}}(\widehat{h}_1-h_1)\|_{\mathcal{G}} = \|(P'_n-P')\xi\|_{\mathcal{G}} \leq 2\log(12/\eta)\left(\frac{2}{n\sqrt{\lambda_n}}+\sqrt{\frac{\mathcal{N}(\lambda_n)}{n}}\right), \quad (25)$$

with probability $1-\eta/6$. We define $\mathcal{A}_2$ as the event where the above inequality holds:

$$\mathcal{A}_2 := \left\{\|(T_{p'}+\lambda_n)^{-\frac{1}{2}}(\widehat{h}_1-h_1)\|_{\mathcal{G}} \leq 2\log(12/\eta)\left(\frac{2}{n\sqrt{\lambda_n}}+\sqrt{\frac{\mathcal{N}(\lambda_n)}{n}}\right)\right\}. \quad (26)$$

One can obtain a similar bound for $\|(T_{p'}+\lambda_n)^{-\frac{1}{2}}(\widehat{h}_2-h_2)\|_{\mathcal{G}}$. In fact, using

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{x}'\sim p}\left[\sum_{k=1}^{\infty}\frac{\mu_k}{\mu_k+\lambda_n}e_k(\boldsymbol{x}')^2\right] &\leq \mathrm{E}_{\boldsymbol{x}'\sim p'}\left[g^*(\boldsymbol{x}')\sum_{k=1}^{\infty}\frac{\mu_k}{\mu_k+\lambda_n}e_k(\boldsymbol{x}')^2\right] \\
&\leq \|g^*\|_{\mathcal{G}}\mathrm{E}_{\boldsymbol{x}'\sim p'}\left[\sum_{k=1}^{\infty}\frac{\mu_k}{\mu_k+\lambda_n}e_k(\boldsymbol{x}')^2\right] = \|g^*\|_{\mathcal{G}}\mathcal{N}(\lambda_n), \quad (27)
\end{aligned}$$

instead of Eq.(25), one can show that, by Proposition 1,

$$\|(T_{p'}+\lambda_n)^{-\frac{1}{2}}(\widehat{h}_2-h_2)\|_{\mathcal{G}} \leq 2\log(12/\eta)\left(\frac{2}{n\sqrt{\lambda_n}}+\sqrt{\frac{\|g^*\|\mathcal{N}(\lambda_n)}{n}}\right), \quad (28)$$

with probability $1-\eta/6$. We define $\mathcal{A}_3$ as the event where the above inequality holds:

$$\mathcal{A}_3 := \left\{\|(T_{p'}+\lambda_n)^{-\frac{1}{2}}(\widehat{h}_2-h_2)\|_{\mathcal{G}} \leq 2\log(12/\eta)\left(\frac{2}{n\sqrt{\lambda_n}}+\sqrt{\frac{\|g^*\|\mathcal{N}(\lambda_n)}{n}}\right)\right\}. \quad (29)$$

Combining Eqs.(24), (25), and (28), we can show that the term (a) is bounded as

$$|\langle\widehat{h}_1-h_1,(\widehat{T}_{p'}+\lambda_n)^{-1}(\widehat{h}_2-h_2)\rangle| \leq 16\log(12/\eta)^2\left(\frac{4}{n^2\lambda_n}+\frac{\sqrt{\|g^*\|}\mathcal{N}(\lambda_n)}{n}\right), \quad (30)$$

on the event $\mathcal{A}_1\cap\mathcal{A}_2\cap\mathcal{A}_3$.

Bounding (1-b):

$$\begin{aligned}
&\langle \widehat{h}_1 - h_1, (\widehat{T}_{p'} + \lambda_n)^{-1}(T_{p'} - \widehat{T}_{p'})(T_{p'} + \lambda_n)^{-1} h_2 \rangle \\
&\leq \|(T_{p'} + \lambda_n)^{-\frac{1}{2}}(\widehat{h}_1 - h_1)\|_{\mathcal{G}} \|(T_{p'} + \lambda_n)^{\frac{1}{2}}(\widehat{T}_{p'} + \lambda_n)^{-1}(T_{p'} + \lambda_n)^{\frac{1}{2}}\|_{\mathcal{L}(\mathcal{G})} \\
&\quad \times \|(T_{p'} + \lambda_n)^{-\frac{1}{2}}(T_{p'} - \widehat{T}_{p'})(T_{p'} + \lambda_n)^{-1} h_2\|_{\mathcal{G}}.
\end{aligned}$$

We have already obtained bounds for $\|(T_{p'}+\lambda)^{-\frac{1}{2}}(\widehat{h}_1 - h_1)\|$ and $\|(T_{p'}+\lambda)^{\frac{1}{2}}(\widehat{T}_{p'}+\lambda)^{-1}(T_{p'}+\lambda)^{\frac{1}{2}}\|$ in Eq.(25) and Eq.(24):

$$\|(T_{p'} + \lambda_n)^{-\frac{1}{2}}(\widehat{h}_1 - h_1)\|_{\mathcal{G}} \leq 2\log(12/\eta)\left(\frac{2}{n\sqrt{\lambda_n}} + \sqrt{\frac{\mathcal{N}(\lambda_n)}{n}}\right), \tag{31}$$

$$\|(T_{p'} + \lambda_n)^{\frac{1}{2}}(\widehat{T}_{p'} + \lambda_n)^{-1}(T_{p'} + \lambda_n)^{\frac{1}{2}}\|_{\mathcal{L}(\mathcal{G})} \leq 2, \tag{32}$$

on the event $\mathcal{A}_1 \cap \mathcal{A}_2$.

Let $\xi : \mathbb{R}^d \to \mathcal{G}$ be the random variable such as

$$\xi(\boldsymbol{x}) = (T_{p'} + \lambda_n)^{-\frac{1}{2}} K_{\boldsymbol{x}} K_{\boldsymbol{x}}^{\circ} (T_{p'} + \lambda_n)^{-1} h_2.$$

Then we have

$$\begin{aligned}
\|\xi(\boldsymbol{x})\|_{\mathcal{G}} &= \|(T_{p'} + \lambda_n)^{-\frac{1}{2}} K_{\boldsymbol{x}} K_{\boldsymbol{x}}^{\circ} (T_{p'} + \lambda_n)^{-1} T_{p'} g^*\|_{\mathcal{G}} \\
&\leq \|(T_{p'} + \lambda_n)^{-\frac{1}{2}}\|_{\mathcal{L}(\mathcal{G})} \|K_{\boldsymbol{x}} K_{\boldsymbol{x}}^{\circ}\|_{\mathcal{L}(\mathcal{G})} \|(T_{p'} + \lambda_n)^{-1} T_{p'}\|_{\mathcal{L}(\mathcal{G})} \|g^*\|_{\mathcal{G}} \\
&\leq \lambda_n^{-\frac{1}{2}} \|g^*\|_{\mathcal{G}},
\end{aligned}$$

where we used the relation

$$\begin{aligned}
\|(K_{\boldsymbol{x}} K_{\boldsymbol{x}}^{\circ}) h\|_{\mathcal{G}} &= \|K_{\boldsymbol{x}} \langle K_{\boldsymbol{x}}, h \rangle_{\mathcal{G}}\|_{\mathcal{G}} = \langle K_{\boldsymbol{x}}, h \rangle_{\mathcal{G}} \|K_{\boldsymbol{x}}\|_{\mathcal{G}} \\
&\leq \|h\|_{\mathcal{G}} \|K_{\boldsymbol{x}}\|_{\mathcal{G}}^2 = \|h\|_{\mathcal{G}} K(\boldsymbol{x}, \boldsymbol{x}) \leq \|h\|_{\mathcal{G}}
\end{aligned}$$

for all $h \in \mathcal{G}$. Then,

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{x}' \sim p'}[\|\xi(\boldsymbol{x}')\|_{\mathcal{G}}^2] &= \mathrm{E}_{\boldsymbol{x}' \sim p'}[\|(T_{p'} + \lambda_n)^{-\frac{1}{2}} K_{\boldsymbol{x}'} K_{\boldsymbol{x}'}^{\circ} (T_{p'} + \lambda_n)^{-1} T_{p'} g^*\|_{\mathcal{G}}^2] \\
&\leq \mathrm{E}_{\boldsymbol{x}' \sim p'}\left[\|(T_{p'} + \lambda_n)^{-1} K_{\boldsymbol{x}'} K_{\boldsymbol{x}'}^{\circ}\|_{\mathcal{L}(\mathcal{G})}\right] \|K_{\boldsymbol{x}'}^{\circ} K_{\boldsymbol{x}'}\|_{\mathcal{L}(\mathcal{G})} \|(T_{p'} + \lambda_n)^{-1} T_{p'} g^*\|_{\mathcal{G}}^2 \\
&\leq \mathrm{E}_{\boldsymbol{x}' \sim p'}\left[\mathrm{tr}\left((T_{p'} + \lambda_n)^{-1} K_{\boldsymbol{x}'} K_{\boldsymbol{x}'}^{\circ}\right)\right] \|g^*\|_{\mathcal{G}}^2 \\
&= \mathrm{tr}\left((T_{p'} + \lambda_n)^{-1} T_{p'}\right) \|g^*\|_{\mathcal{G}}^2 \\
&= \mathcal{N}(\lambda_n) \|g^*\|_{\mathcal{G}}^2.
\end{aligned}$$

Therefore, by Proposition 1, we obtain

$$\begin{aligned}
&\|(T_{p'} + \lambda_n)^{-\frac{1}{2}}(T_{p'} - \widehat{T}_{p'})(T_{p'} + \lambda_n)^{-1} h_2\|_{\mathcal{G}} \\
&\leq 2\log(12/\eta)\left(\frac{2}{n\sqrt{\lambda_n}} + \sqrt{\frac{\|g^*\|^2 \mathcal{N}(\lambda_n)}{n}}\right),
\end{aligned} \tag{33}$$

with probability $1-\eta/6$. We define $\mathcal{A}_4$ as the event where the above inequality holds

$$\mathcal{A}_4 := \Bigg\{ \|(T_{p'}+\lambda_n)^{-\frac{1}{2}}(T_{p'}-\widehat{T}_{p'})(T_{p'}+\lambda_n)^{-1}h_2\|_{\mathcal{G}} \\ \leq 2\log(12/\eta)\left(\frac{2}{n\sqrt{\lambda_n}}+\sqrt{\frac{\|g^*\|^2\mathcal{N}(\lambda_n)}{n}}\right)\Bigg\}.$$

Combining Eqs.(25), (24), and (33), the term (1-b) is bounded as

$$|\langle \widehat{h}_1-h_1, (\widehat{T}_{p'}+\lambda_n)^{-1}(T_{p'}-\widehat{T}_{p'})(T_{p'}+\lambda_n)^{-1}h_2\rangle| \\ \leq 16\log(12/\eta)^2\left(\frac{4}{n^2\lambda_n}+\frac{\|g^*\|\mathcal{N}(\lambda_n)}{n}\right),$$

on the event $\mathcal{A}_1\cap\mathcal{A}_2\cap\mathcal{A}_4$.

Bounding (1-c): We have

$$\langle \widehat{h}_1-h_1, (T_{p'}+\lambda_n)^{-1}\lambda_n g^*\rangle = \langle (T_{p'}+\lambda_n)^{-\frac{1}{2}}(\widehat{h}_1-h_1), (T_{p'}+\lambda_n)^{-\frac{1}{2}}\lambda_n g^*\rangle \\ \leq \|(T_{p'}+\lambda_n)^{-\frac{1}{2}}(\widehat{h}_1-h_1)\|_{\mathcal{G}}\|(T_{p'}+\lambda_n)^{-\frac{1}{2}}\sqrt{\lambda_n}g^*\|_{\mathcal{G}}\sqrt{\lambda_n}. \tag{34}$$

Notice that Eq.(25) gives

$$\|(T_{p'}+\lambda_n)^{-\frac{1}{2}}(\widehat{h}_1-h_1)\|_{\mathcal{G}} \leq 2\log(12/\eta)\left(\frac{2}{n\sqrt{\lambda_n}}+\sqrt{\frac{\mathcal{N}(\lambda_n)}{n}}\right), \tag{35}$$

on the event $\mathcal{A}_2$. This and $\|(T_{p'}+\lambda_n)^{-\frac{1}{2}}\sqrt{\lambda_n}g^*\|_{\mathcal{G}} \leq \|g^*\|_{\mathcal{G}}$ give

$$|\langle \widehat{h}_1-h_1, (T_{p'}+\lambda_n)^{-1}\lambda_n g^*\rangle| \leq 2\log(12/\eta)\left(\frac{2\|g^*\|_{\mathcal{G}}}{n}+\|g^*\|_{\mathcal{G}}\sqrt{\frac{\mathcal{N}(\lambda_n)\lambda_n}{n}}\right), \tag{36}$$

on the event $\mathcal{A}_2$.

Combining the bounds of (1-a), (1-b), and (1-c):

$$
\begin{aligned}
&|(P_n' - P')(\widetilde{g} - g^*)| \\
&\leq 16\log(12/\eta)^2 \left(\frac{4}{n^2\lambda_n} + \frac{\sqrt{\|g^*\|_{\mathcal{G}}}\mathcal{N}(\lambda_n)}{n}\right) + 16\log(12/\eta)^2 \left(\frac{4}{n^2\lambda_n} + \frac{\|g^*\|_{\mathcal{G}}\mathcal{N}(\lambda_n)}{n}\right) \\
&\quad + 2\log(12/\eta)\left(\frac{2\|g^*\|_{\mathcal{G}}}{n} + \|g^*\|_{\mathcal{G}}\sqrt{\frac{\mathcal{N}(\lambda_n)\lambda_n}{n}}\right) \\
&= 16\log(12/\eta)^2 \left(\frac{8}{n^2\lambda_n} + \frac{(\|g^*\|_{\mathcal{G}} + \sqrt{\|g^*\|_{\mathcal{G}}})\mathcal{N}(\lambda_n)}{n}\right) \\
&\quad + 2\log(12/\eta)\left(\frac{2\|g^*\|_{\mathcal{G}}}{n} + \|g^*\|_{\mathcal{G}}\sqrt{\frac{\mathcal{N}(\lambda_n)\lambda_n}{n}}\right),
\end{aligned}
$$

on the event $\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3 \cap \mathcal{A}_4$.

Step 2. Bounding $|(P_n - P)(\widetilde{g} - g^*)|$

As in Eq.(22), we have

$$
\begin{aligned}
&(P_n - P)(\widetilde{g} - g^*) \\
&= \langle \widehat{h}_2 - h_2, (\widehat{T}_{p'} + \lambda_n)^{-1}(\widehat{h}_2 - h_2)\rangle + \langle \widehat{h}_2 - h_2, (\widehat{T}_{p'} + \lambda_n)^{-1}(T_{p'} - \widehat{T}_{p'})(T_{p'} + \lambda_n)^{-1}h_2\rangle \\
&\quad + \langle \widehat{h}_2 - h_2, (T_{p'} + \lambda_n)^{-1}\lambda_n g^*\rangle.
\end{aligned}
$$

Using Eq.(28) instead of Eq.(25), on the event $\mathcal{A}_1 \cap \mathcal{A}_3 \cap \mathcal{A}_4$, each term is bounded as

$$
\begin{aligned}
&|\langle \widehat{h}_2 - h_2, (\widehat{T}_{p'} + \lambda_n)^{-1}(\widehat{h}_2 - h_2)\rangle| \\
&\qquad \leq 16\log(12/\eta)^2 \left(\frac{4}{n^2\lambda_n} + \frac{\|g^*\|_{\mathcal{G}}\mathcal{N}(\lambda_n)}{n}\right), \\
&|\langle \widehat{h}_2 - h_2, (\widehat{T}_{p'} + \lambda_n)^{-1}(T_{p'} - \widehat{T}_{p'})(T_{p'} + \lambda_n)^{-1}h_2\rangle| \\
&\qquad \leq 16\log(12/\eta)^2 \left(\frac{4}{n^2\lambda_n} + \frac{\|g^*\|_{\mathcal{G}}^{\frac{3}{2}}\mathcal{N}(\lambda_n)}{n}\right), \\
&|\langle \widehat{h}_2 - h_2, (T_{p'} + \lambda_n)^{-1}\lambda g^*\rangle| \\
&\qquad \leq 2\log(12/\eta)\left(\frac{2\|g^*\|_{\mathcal{G}}}{n} + \|g^*\|_{\mathcal{G}}^{\frac{3}{2}}\sqrt{\frac{\mathcal{N}(\lambda_n)\lambda_n}{n}}\right).
\end{aligned}
$$

Then we obtain the following bound:

$$\begin{aligned}
&|(P_n - P)(\widetilde{g} - g^*)| \\
&\leq 16 \log(12/\eta)^2 \left(\frac{8}{n^2 \lambda_n} + \frac{(\|g^*\|_{\mathcal{G}}^{\frac{3}{2}} + \|g^*\|_{\mathcal{G}}) \mathcal{N}(\lambda_n)}{n} \right) \\
&\quad + 2 \log(12/\eta) \left(\frac{2\|g^*\|_{\mathcal{G}}}{n} + \|g^*\|_{\mathcal{G}}^{\frac{3}{2}} \sqrt{\frac{\mathcal{N}(\lambda_n)}{n}} \right),
\end{aligned}$$

on the event $\mathcal{A}_1 \cap \mathcal{A}_3 \cap \mathcal{A}_4$.

Step 3. Bounding $|P_n' \widetilde{g} - 1|$

We decompose $\widetilde{g}$ as

$$\widetilde{g} = \widehat{u} + \widehat{\beta}, \tag{37}$$

where $\widehat{u} \perp 1$ in $\mathcal{G}$ and $\widehat{\beta}$ is a constant function. Then one can easily show that

$$\widehat{\beta} = \frac{1 - P_n' \widehat{u}}{1 + \lambda_n \|1\|_{\mathcal{G}}}.$$

Therefore

$$P_n' \widetilde{g} = P_n' \widehat{u} + \frac{1 - P_n' \widehat{u}}{1 + \lambda_n \|1\|_{\mathcal{G}}} = 1 + \lambda_n \frac{\|1\|_{\mathcal{G}}^2}{1 + \lambda_n \|1\|_{\mathcal{G}}^2} (P_n' \widehat{u} - 1). \tag{38}$$

If we can show that $\widehat{u}$ is bounded (i.e., $\mathcal{O}_p(1)$), then $P_n' \widetilde{g} - 1 = \mathcal{O}_p(\lambda_m)$. To show that, we bound $\|\widetilde{g}\|$ because

$$\|\widehat{u}\|_\infty \leq \|\widehat{u}\|_{\mathcal{G}} \leq \sqrt{\|\widehat{u}\|_{\mathcal{G}}^2 + \|\widehat{\beta}\|_{\mathcal{G}}^2} = \|\widetilde{g}\|_{\mathcal{G}}.$$

We have

$$\begin{aligned}
\|\widetilde{g}\|_{\mathcal{G}}^2 &= \langle \widehat{h}_2, (\widehat{T}_{p'} + \lambda_n)^{-2} \widehat{h}_2 \rangle \\
&= \langle (T_{p'} + \lambda_n)^{-1} \widehat{h}_2, [(T_{p'} + \lambda_n)(\widehat{T}_{p'} + \lambda_n)^{-2} (T_{p'} + \lambda_n)] (T_{p'} + \lambda_n)^{-1} \widehat{h}_2 \rangle
\end{aligned}$$

Here

$$(T_{p'} + \lambda_n)(\widehat{T}_{p'} + \lambda_n)^{-1} = (I - (T_{p'} - \widehat{T}_{p'})(T_{p'} + \lambda_n)^{-1})^{-1} \tag{39}$$

and on the event $\mathcal{A}_1$ with the condition Eq.(13), we have

$$\|(T_{p'} - \widehat{T}_{p'})(T_{p'} + \lambda_n)^{-1}\|_{\mathcal{L}(\mathcal{G})} \leq \frac{1}{2}.$$

Hence

$$\|(T_{p'}+\lambda_n)(\widehat{T}_{p'}+\lambda_n)^{-1}\|_{\mathcal{L}(\mathcal{G})} \leq 2 \tag{40}$$

on the event $\mathcal{A}_1$ with the condition Eq.(13).

We have that

$$\begin{aligned}\|(T_{p'}+\lambda_n)^{-1}(\widehat{h}_2-h_2)\|_{\mathcal{G}} &\leq \lambda_n^{-\frac{1}{2}}\|(T_{p'}+\lambda_n)^{-\frac{1}{2}}(\widehat{h}_2-h_2)\|_{\mathcal{G}} \\ &\leq 2\log(12/\eta)\left(\frac{2}{\lambda_n n}+\sqrt{\frac{\mathcal{N}(\lambda_n)\|g^*\|_{\mathcal{G}}}{\lambda_n n}}\right),\end{aligned} \tag{41}$$

on the event $\mathcal{A}_3$. Hence Eqs.(40) and (41) and

$$\|(T_{p'}+\lambda_n)^{-1}h_2\| = \|(T_{p'}+\lambda_n)^{-1}T_{p'}g^*\| \leq \|g^*\|_{\mathcal{G}}$$

give

$$\|\widetilde{g}\|^2 \leq 8\left(\|g^*\|_{\mathcal{G}}^2+8\log^2(12/\eta)\left(\frac{4}{\lambda_n^2 n^2}+\frac{\mathcal{N}(\lambda_n)\|g^*\|_{\mathcal{G}}}{\lambda_n n}\right)\right), \tag{42}$$

on the event $\mathcal{A}_3$.

Therefore, Eqs.(38) and (42) give

$$\begin{aligned}|P_n'\widetilde{g}-1| &= \lambda_n\frac{\|1\|_{\mathcal{G}}^2}{1+\lambda_n\|1\|_{\mathcal{G}}^2}|P_n'\widehat{u}-1| \\ &\leq \lambda_n\frac{\|1\|_{\mathcal{G}}^2}{1+\lambda_n\|1\|_{\mathcal{G}}^2}\left\{8\left(\|g^*\|_{\mathcal{G}}^2+8\log^2(12/\eta)\left(\frac{4}{\lambda_n^2 n^2}+\frac{\mathcal{N}(\lambda_n)\|g^*\|_{\mathcal{G}}}{\lambda_n n}\right)\right)+1\right\} \\ &=: \lambda_n C_{n,\eta},\end{aligned} \tag{43}$$

on the event $\mathcal{A}_3$.

Step 4. Bounding $P'(\widetilde{g}-g^*)^2$

Decompose $\widetilde{g}-g^*$ as

$$\widetilde{g}-g^* = (\widetilde{g}-g_{\lambda_n})+(g_{\lambda_n}-g^*).$$

The first term is evaluated as follows:

$$\begin{aligned}\widetilde{g}-g_{\lambda_n} &= (\widehat{T}_{p'}+\lambda_n)^{-1}\widehat{h}_2-(T_{p'}+\lambda_n)^{-1}h_2 \\ &= (\widehat{T}_{p'}+\lambda_n)^{-1}\left\{(\widehat{h}_2-h_2)+(T_{p'}-\widehat{T}_{p'})(T_{p'}+\lambda_n)^{-1}h_2\right\}.\end{aligned} \tag{44}$$

Thus

$$P'(\widetilde{g}-g_{\lambda_n})^2 = \left\| \sqrt{T_{p'}}(\widehat{T}_{p'}+\lambda_n)^{-1}\left\{(\widehat{h}_2-h_2)+(T_{p'}-\widehat{T}_{p'})(T_{p'}+\lambda_n)^{-1}h_2\right\}\right\|_{\mathcal{G}}^2$$
$$\leq 2\Bigg\{ \underbrace{\|\sqrt{T_{p'}}(\widehat{T}_{p'}+\lambda_n)^{-1}(\widehat{h}_2-h_2)\|_{\mathcal{G}}^2}_{\text{(4-a)}}$$
$$+ \underbrace{\|\sqrt{T_{p'}}(\widehat{T}_{p'}+\lambda_n)^{-1}(T_{p'}-\widehat{T}_{p'})(T_{p'}+\lambda_n)^{-1}h_2\|_{\mathcal{G}}^2}_{\text{(4-b)}} \Bigg\}. \tag{45}$$

Bounding (4-a): We have

$$\begin{aligned}&\|\sqrt{T_{p'}}(\widehat{T}_{p'}+\lambda_n)^{-1}(\widehat{h}_2-h_2)\|_{\mathcal{G}}\\ &\leq \|\sqrt{T_{p'}}(T_{p'}+\lambda_n)^{-\frac{1}{2}}\|_{\mathcal{L}(\mathcal{G})}\|(T_{p'}+\lambda_n)^{\frac{1}{2}}(\widehat{T}_{p'}+\lambda_n)^{-1}(T_{p'}+\lambda_n)^{\frac{1}{2}}\|_{\mathcal{G}}\\ &\quad\times\|(T_{p'}+\lambda_n)^{-\frac{1}{2}}(\widehat{h}_2-h_2)\|_{\mathcal{G}}.\end{aligned}$$

It is obvious that

$$\|\sqrt{T_{p'}}(T_{p'}+\lambda_n)^{-\frac{1}{2}}\|_{\mathcal{G}}\leq 1. \tag{46}$$

By Eq.(24),

$$\|(T_{p'}+\lambda_n)^{\frac{1}{2}}(\widehat{T}_{p'}+\lambda_n)^{-1}(T_{p'}+\lambda_n)^{\frac{1}{2}}\|_{\mathcal{G}}\leq 2 \tag{47}$$

on the event $\mathcal{A}_1$. By Eq.(28),

$$\|(T_{p'}+\lambda_n)^{-\frac{1}{2}}(\widehat{h}_2-h_2)\|_{\mathcal{G}}\leq 2\log(12/\eta)\left(\frac{2}{n\sqrt{\lambda_n}}+\sqrt{\frac{\|g^*\|_{\mathcal{G}}\mathcal{N}(\lambda_n)}{n}}\right), \tag{48}$$

on the event $\mathcal{A}_3$.

Combining Eqs.(46), (24), and (28), we have

$$\|\sqrt{T_{p'}}(\widehat{T}_{p'}+\lambda_n)^{-1}(\widehat{h}_2-h_2)\|_{\mathcal{G}}\leq 4\log(12/\eta)\left(\frac{2}{n\sqrt{\lambda_n}}+\sqrt{\frac{\|g^*\|_{\mathcal{G}}\mathcal{N}(\lambda_n)}{n}}\right), \tag{49}$$

on the event $\mathcal{A}_1\cap\mathcal{A}_3$.

Bounding (4-b): We have

$$\begin{aligned}&\|\sqrt{T_{p'}}(\widehat{T}_{p'}+\lambda_n)^{-1}(T_{p'}-\widehat{T}_{p'})(T_{p'}+\lambda_n)^{-1}h_2\|_{\mathcal{G}}\\ &\leq \|\sqrt{T_{p'}}(T_{p'}+\lambda_n)^{-\frac{1}{2}}\|_{\mathcal{L}(G)}\|(T_{p'}+\lambda_n)^{\frac{1}{2}}(\widehat{T}_{p'}+\lambda_n)^{-1}(T_{p'}+\lambda_n)^{\frac{1}{2}}\|_{\mathcal{L}(G)}\\ &\quad\times\|(T_{p'}+\lambda_n)^{-\frac{1}{2}}(T_{p'}-\widehat{T}_{p'})(T_{p'}+\lambda_n)^{-1}h_2\|_{\mathcal{G}}.\end{aligned}$$

By Eq.(33), we have

$$\|(T_{p'}+\lambda_n)^{-\frac{1}{2}}(T_{p'}-\widehat{T}_{p'})(T_{p'}+\lambda_n)^{-1}h_2\|_{\mathcal{G}} \leq 2\log(12/\eta)\left(\frac{2}{n\sqrt{\lambda_n}}+\sqrt{\frac{\|g^*\|_{\mathcal{G}}^2\mathcal{N}(\lambda_n)}{n}}\right),$$

on the event $\mathcal{A}_4$. Thus Eqs.(46), (24), and (33) indicate

$$\begin{aligned}&\|\sqrt{T_{p'}}(\widehat{T}_{p'}+\lambda_n)^{-1}(T_{p'}-\widehat{T}_{p'})(T_{p'}+\lambda_n)^{-1}h_2\|_{\mathcal{G}}\\ &\leq 4\log(12/\eta)\left(\frac{2}{n\sqrt{\lambda_n}}+\sqrt{\frac{\|g^*\|_{\mathcal{G}}^2\mathcal{N}(\lambda_n)}{n}}\right),\end{aligned} \tag{50}$$

on the event $\mathcal{A}_1\cap\mathcal{A}_4$.

Combining the bounds of (4-a) and (4-b): Substituting Eqs.(50) and (49) to Eq.(45), we have

$$\begin{aligned}P'(\widetilde{g}-g_{\lambda_n})^2 \leq 64\Bigg\{&\log^2(12/\eta)\left(\frac{4}{n^2\lambda_n}+\frac{\|g^*\|_{\mathcal{G}}^2\mathcal{N}(\lambda_n)}{n}\right)\\ &+\log^2(12/\eta)\left(\frac{4}{n^2\lambda_n}+\frac{\|g^*\|_{\mathcal{G}}\mathcal{N}(\lambda_n)}{n}\right)\Bigg\},\end{aligned} \tag{51}$$

on the event $\mathcal{A}_1\cap\mathcal{A}_4$.

On the other hand, $P'(g_{\lambda_n}-g^*)^2$ is bounded as

$$\begin{aligned}P'(g_{\lambda_n}-g^*)^2 &= \|\sqrt{T_{p'}}((T_{p'}+\lambda_n)^{-1}h_2-g^*)\|_{\mathcal{G}}^2 = \|\sqrt{T_{p'}}(T_{p'}+\lambda_n)^{-1}(h_2-(T_{p'}+\lambda_n)g^*)\|_{\mathcal{G}}^2\\ &= \|\sqrt{T_{p'}}(T_{p'}+\lambda_n)^{-1}\lambda_n g^*)\|_{\mathcal{G}}^2 \leq \|(T_{p'}+\lambda_n)^{-\frac{1}{2}}\lambda_n g^*)\|_{\mathcal{G}}^2 \leq \lambda_n\|g^*\|_{\mathcal{G}}^2.\end{aligned} \tag{52}$$

By Eqs.(51) and (52), $P'(\widetilde{g}-g^*)^2$ is bounded as

$$\begin{aligned}&P'(\widetilde{g}-g^*)^2\\ &\leq 2(P'(\widetilde{g}-g_{\lambda_n})^2+P'(g_{\lambda_n}-g^*)^2)\\ &\leq 128\log^2(12/\eta)\left(\frac{8}{n^2\lambda_n}+\frac{(\|g^*\|_{\mathcal{G}}+\|g^*\|_{\mathcal{G}}^2)\mathcal{N}(\lambda_n)}{n}\right)+2\lambda_n\|g^*\|_{\mathcal{G}}^2,\end{aligned}$$

on the event $\mathcal{A}_1\cap\mathcal{A}_4$.

Step 5. Bounding $|(P_n'-P')(g^*-1)|$ and $|(P_n-P)(g^*-1)|$

By Proposition 1, we have the following bound

$$|(P_n'-P')(g^*-1)| \leq 2\log(12/\eta)\left(\frac{2\|g^*-1\|_\infty}{n}+\sqrt{\frac{P'(g^*-1)^2}{n}}\right), \tag{53}$$

with probability $1-\eta/6$. Similarly we have

$$|(P_n - P)(g^* - 1)| \leq 2\log(12/\eta)\left(\frac{2\|g^*-1\|_\infty}{n} + \sqrt{\frac{P(g^*-1)^2}{n}}\right), \tag{54}$$

with probability $1-\eta/6$.

We define $\mathcal{A}_5$ and $\mathcal{A}_6$ as the events where the above inequalities hols:

$$\mathcal{A}_5 := \left\{|(P'_n - P')(g^* - 1)| \leq 2\log(12/\eta)\left(\frac{2\|g^*-1\|_\infty}{n} + \sqrt{\frac{P'(g^*-1)^2}{n}}\right)\right\}, \tag{55}$$

$$\mathcal{A}_6 := \left\{|(P_n - P)(g^* - 1)| \leq 2\log(12/\eta)\left(\frac{2\|g^*-1\|_\infty}{n} + \sqrt{\frac{P(g^*-1)^2}{n}}\right)\right\}. \tag{56}$$

Step 6. Combining the bounds of Step 1 to 5.

Finally we obtain

$$\begin{aligned}
&|\widehat{\mathrm{PE}}(\mathcal{X},\mathcal{X}') - \mathrm{PE}(P,P')| \\
&\leq 8\log(12/\eta)^2\left(\frac{8}{n^2\lambda_n} + \frac{(\|g^*\|_{\mathcal{G}} + \sqrt{\|g^*\|_{\mathcal{G}}})\mathcal{N}(\lambda_n)}{n}\right) \\
&\quad + \log(12/\eta)\left(\frac{2\|g^*\|_{\mathcal{G}}}{n} + \|g^*\|_{\mathcal{G}}\sqrt{\frac{\lambda_n\mathcal{N}(\lambda_n)}{n}}\right) \\
&\quad + 8\log(12/\eta)^2\left(\frac{8}{n^2\lambda_n} + \frac{(\|g^*\|_{\mathcal{G}}^{\frac{3}{2}} + \|g^*\|_{\mathcal{G}})\mathcal{N}(\lambda_n)}{n}\right) \\
&\quad + \log(12/\eta)\left(\frac{2\|g^*\|_{\mathcal{G}}}{n} + \|g^*\|_{\mathcal{G}}^{\frac{3}{2}}\sqrt{\frac{\lambda_n\mathcal{N}(\lambda_n)}{n}}\right) \\
&\quad + \frac{3}{2}\lambda_n C_{n,\eta} \\
&\quad + \log(12/\eta)\left(\frac{4\|g^*-1\|_\infty}{n} + \sqrt{\frac{P'(g^*-1)^2}{n}} + \sqrt{\frac{P(g^*-1)^2}{n}}\right) \\
&\quad + \frac{1}{2}\sqrt{P'(g^*-1)^2}\sqrt{128\log^2(12/\eta)\left(\frac{8}{n^2\lambda_n} + \frac{(\|g^*\|_{\mathcal{G}} + \|g^*\|_{\mathcal{G}}^2)\mathcal{N}(\lambda_n)}{n}\right) + 2\lambda_n\|g^*\|_{\mathcal{G}}^2},
\end{aligned}$$

on the event $\bigcap_{\ell=1}^6 \mathcal{A}_\ell$ the probability of which is at least $1-\eta$. □

Proof of Theorem 1'. By Proposition 3 in Caponnetto and de Vito (2007), we obtain

$$\mathcal{N}(\lambda) \leq \frac{2c}{2-\gamma}\lambda^{-\frac{\gamma}{2}},$$

where c is the constant appears in the assumption. Then substituting the above inequality and $\lambda_n = \left(\frac{\log n}{n}\right)^{\frac{2}{2+\gamma}}$ to Eq.(14), we can see that there is a constant K depending on $c, \gamma, \|g\|_{\mathcal{G}}$ such that

$$\begin{aligned}
&|\widehat{\mathrm{PE}}(\mathcal{X}, \mathcal{X}') - \mathrm{PE}(P, P')| \\
&\leq K \Bigg\{ \left(\log(12/\eta)^2 + \log(12/\eta) + 1\right) \left(\frac{\log n}{n}\right)^{\frac{2}{2+\gamma}} \\
&\qquad + \log(12/\eta) \left(\sqrt{\frac{P'(g^* - 1)^2}{n}} + \frac{\|g^* - 1\|_\infty}{n}\right) \\
&\qquad + \sqrt{P'(g^* - 1)^2} \sqrt{(\log(12/\eta)^2 + 1)} \left(\frac{\log n}{n}\right)^{\frac{1}{2+\gamma}} \Bigg\}, \qquad (57)
\end{aligned}$$

with probability at least $1 - \eta$ under the condition Eq.(13). The condition Eq.(13) is satisfied for sufficiently large n. Therefore Eq.(57) implies that

$$|\widehat{\mathrm{PE}}(\mathcal{X}, \mathcal{X}') - \mathrm{PE}(P, P')| = \mathcal{O}_p\left(\left(\frac{\log n}{n}\right)^{\frac{2}{2+\gamma}} + \sqrt{P'(g^* - 1)^2} \left(\frac{\log n}{n}\right)^{\frac{1}{2+\gamma}}\right).$$

□

B Proof of Theorem 2

In this section, we prove Theorem 2. Here, for being more precise, we rewrite Theorem 2 as follow.

Theorem 2'. *Let $\widetilde{F}_n(\cdot|\mathcal{X} \cup \mathcal{X}')$ be the distribution function of $\widehat{\mathrm{PE}}(\widetilde{\mathcal{X}}, \widetilde{\mathcal{X}}')$ given $\mathcal{X} \cup \mathcal{X}'$. Let*

$$\widetilde{q}(\mathcal{X} \cup \mathcal{X}') = \sup\{x \in \mathbb{R} \mid \widetilde{F}_n(x|\mathcal{X} \cup \mathcal{X}') \leq 1 - \alpha\}$$

be the upper 100α-percentile point. Then, if the null hypothesis is true (i.e., $P = P'$),

$$\mathrm{Prob}\left(\widehat{\mathrm{PE}}(\mathcal{X}, \mathcal{X}') > \widetilde{q}(\mathcal{X} \cup \mathcal{X}')\right) \leq \alpha.$$

Proof. Since the samples $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{x}'_i\}_{i=1}^n$ are distributed i.i.d. and $P = P'$, they are *exchangeable*, i.e., the distribution of $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{2n}) = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_n)$ is same as that of $(\boldsymbol{y}_{\tau(1)}, \ldots, \boldsymbol{y}_{\tau(2n)})$ for any permutation τ on $\{1, \ldots, 2n\}$. This means that the distribution function $\widetilde{F}_n(\cdot \mid \mathcal{S})$ is the same as that of $\mathrm{PE}(\mathcal{X}, \mathcal{X}')$ conditioned on $\mathcal{S} = \mathcal{X} \cup \mathcal{X}'$. Then, we have

$$\begin{aligned}
\mathrm{Prob}\left(\widehat{\mathrm{PE}}(\mathcal{X}, \mathcal{X}') > \widetilde{q}(\mathcal{X} \cup \mathcal{X}')\right) &= \mathrm{E}_{\mathcal{X} \cup \mathcal{X}'}\left[\mathrm{Prob}\left(\widehat{\mathrm{PE}}(\mathcal{X}, \mathcal{X}') > \widetilde{q}(\mathcal{X} \cup \mathcal{X}') \mid \mathcal{X} \cup \mathcal{X}'\right)\right] \\
&= \mathrm{E}_{\mathcal{X} \cup \mathcal{X}'}\left[1 - \widetilde{F}_n\left(\widetilde{q}(\mathcal{X} \cup \mathcal{X}') \mid \mathcal{X} \cup \mathcal{X}'\right)\right] \\
&\leq \alpha,
\end{aligned}$$

which concludes the proof. □

References

Ali, S. M., & Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, *28*, 131–142.

Anderson, N., Hall, P., & Titterington, D. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, *50*, 41–54.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, *68*, 337–404.

Bachman, G., & Narici, L. (2000). *Functional analysis*. Mineola, NY, USA: Dover Publications.

Biau, G., & Györfi, L. (2005). On the asymptotic properties of a nonparametric l1-test statistic of homogeneity. *IEEE Transactions on Information Theory*, *51*, 3965–3973.

Bickel, P. (1969). A distribution free version of the Smirnov two sample test in the p-variate case. *The Annals of Mathematical Statistics*, *40*, 1–23.

Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., & Smola, A. J. (2006). Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, *22*, e49–e57.

Caponnetto, A., & de Vito, E. (2007). Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, *7*, 331–368.

Caruana, R., Pratt, L., & Thrun, S. (1997). Multitask learning. *Machine Learning*, *28*, 41–75.

Cheng, K. F., & Chu, C. K. (2004). Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, *10*, 583–604.

Collins, M., & Duffy, N. (2002). Convolution kernels for natural language. *Advances in Neural Information Processing Systems 14* (pp. 625–632). Cambridge, MA: MIT Press.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, *2*, 229–318.

Darbellay, G. A., & Vajda, I. (1999). Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, *45*, 1315–1321.

Duffy, N., & Collins, M. (2002). Convolution kernels for natural language. *Advances in Neural Information Processing Systems 14* (pp. 625–632). Cambridge, MA: MIT Press.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Friedman, J., & Rafsky, L. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, *7*, 697–717.

Gärtner, T. (2003). A survey of kernels for structured data. *SIGKDD Explorations*, *5*, S268–S275.

Gärtner, T., Flach, P., & Wrobel, S. (2003). On graph kernels: Hardness results and efficient alternatives. *Proceedings of the Sixteenth Annual Conference on Computational Learning Theory* (pp. 129–143).

Gretton, A., Borgwardt, K. M., M.Rasch, Schölkopf, B., & Smola, A. J. (2007). A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt and T. Hoffman (Eds.), *Advances in neural information processing systems 19*, 513–520. Cambridge, MA: MIT Press.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., & Schölkopf, B. (2009). Covariate shift by kernel mean matching. In J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer and N. Lawrence (Eds.), *Dataset shift in machine learning*, chapter 8, 131–160. Cambridge, MA: MIT Press.

Hachiya, H., Akiyama, T., Sugiyama, M., & Peters, J. (2009). Adaptive importance sampling for value function approximation in off-policy reinforcement learning. *Neural Networks*, *22*, 1399–1410.

Hall, P., & Tajvidi, N. (2002). Permutation tests for equality of distributions in highdimensional settings. *Biometrika*, *89*, 359–374.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Hotelling, H. (1951). A generalized t test and measure of multivariate dispersion. *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability* (pp. 23–41). Berkeley, CA., USA: University of California Press.

Huang, J., Smola, A., Gretton, A., Borgwardt, K. M., & Schölkopf, B. (2007). Correcting sample selection bias by unlabeled data. In B. Schölkopf, J. Platt and T. Hoffman (Eds.), *Advances in neural information processing systems 19*, 601–608. Cambridge, MA, USA: MIT Press.

Kanamori, T., Hido, S., & Sugiyama, M. (2009a). A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, *10*, 1391–1445.

Kanamori, T., Suzuki, T., & Sugiyama, M. (2009b). *Condition number analysis of kernel-based density ratio estimation* (Technical Report). arXiv.

Kanamori, T., Suzuki, T., & Sugiyama, M. (2010). Theoretical analysis of density ratio estimation. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, *E93-A*, 787–798.

Kashima, H., & Koyanagi, T. (2002). Kernels for semi-structured data. *Proceedings of the Nineteenth International Conference on Machine Learning* (pp. 291–298). San Francisco, CA: Morgan Kaufmann.

Kashima, H., Tsuda, K., & Inokuchi, A. (2003). Marginalized kernels between labeled graphs. *Proceedings of the Twentieth International Conference on Machine Learning* (pp. 321–328). San Francisco, CA: Morgan Kaufmann.

Keziou, A., & Leoni-Aubin, S. (2005). Test of homogeneity in semiparametric two-sample density ratio models. *Comptes Rendus Mathématique*, *340*, 905–910.

Kondor, R. I., & Lafferty, J. (2002). Diffusion kernels on graphs and other discrete input spaces. *Proceedings of the Nineteenth International Conference on Machine Learning* (pp. 315–322).

Kullback, S. (1959). *Information theory and statistics*. New York: Wiley.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79–86.

Li, Q. (1996). Nonparametric testing of closeness between two unknown distribution functions. *Econometric Reviews*, *15*, 261–274.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., & Watkins, C. (2002). Text classification using string kernels. *Journal of Machine Learning Research*, *2*, 419–444.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, *29*, 429–443.

Nguyen, X., Wainwright, M. J., & Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*. to appear.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, *50*, 157–175.

Pérez-Cruz, F. (2008). Kullback-Leibler divergence estimation of continuous distributions. *Proceedings of IEEE International Symposium on Information Theory* (pp. 1666–1670). Nice, France.

Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, *85*, 619–639.

Rätsch, G., Onoda, T., & Müller, K.-R. (2001). Soft margins for adaboost. *Machine Learning*, *42*, 287–320.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels*. Cambridge, MA: MIT Press.

Silva, J., & Narayanan, S. (2007). Universal consistency of data-driven partitions for divergence estimation. *Proceedings of IEEE International Symposium on Information Theory* (pp. 2021–2025). Nice, France.

Sriperumbudur, B., Fukumizu, K., Gretton, A., Lanckriet, G., & Schölkopf, B. (2009). Kernel choice and classifiability for rkhs embeddings of probability distributions. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams and A. Culotta (Eds.), *Advances in neural information processing systems 22*, 1750–1758. Cambridge, MA: MIT Press.

Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, *2*, 67–93.

Student (1908). The probable error of a mean. *Biometrika*, *6*, 1–25.

Sugiyama, M., Kawanabe, M., & Chui, P. L. (2010). Dimensionality reduction for density ratio estimation in high-dimensional spaces. *Neural Networks*, *23*, 44–59.

Sugiyama, M., Krauledat, M., & Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, *8*, 985–1005.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., & Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, *60*, 699–746.

Sugiyama, M., Yamada, M., von Bünau, P., Suzuki, T., Kanamori, T., & Kawanabe, M. (2011). Direct density-ratio estimation with dimensionality reduction via least-squares hetero-distributional subspace search. *Neural Networks*.

Wang, Q., Kulkarmi, S. R., & Verdú, S. (2005). Divergence estimation of contiunous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, *51*, 3064–3074.

Yamada, M., Sugiyama, M., Wichern, G., & Simm, J. (2010). Direct importance estimation with a mixture of probabilistic principal component analyzers. *IEICE Transactions on Information and Systems*, *E93-D*, 2846–2849.