Relative Density-Ratio Estimation for Robust Distribution Comparison

Makoto Yamada
Tokyo Institute of Technology
yamada@sg.cs.titech.ac.jp

Taiji Suzuki
The University of Tokyo
s-taiji@stat.t.u-tokyo.ac.jp

Takafumi Kanamori
Nagoya University
kanamori@is.nagoya-u.ac.jp

Hirotaka Hachiya **Masashi Sugiyama**
Tokyo Institute of Technology
{hachiya@sg. sugi@}cs.titech.ac.jp

Abstract

Divergence estimators based on direct approximation of density-ratios without going through separate approximation of numerator and denominator densities have been successfully applied to machine learning tasks that involve distribution comparison such as outlier detection, transfer learning, and two-sample homogeneity test. However, since density-ratio functions often possess high fluctuation, divergence estimation is still a challenging task in practice. In this paper, we propose to use *relative divergences* for distribution comparison, which involves approximation of *relative density-ratios*. Since relative density-ratios are always smoother than corresponding ordinary density-ratios, our proposed method is favorable in terms of the non-parametric convergence speed. Furthermore, we show that the proposed divergence estimator has asymptotic variance *independent* of the model complexity under a parametric setup, implying that the proposed estimator hardly overfits even with complex models. Through experiments, we demonstrate the usefulness of the proposed approach.

1 Introduction

Comparing probability distributions is a fundamental task in statistical data processing. It can be used for, e.g., *outlier detection* [1, 2], *two-sample homogeneity test* [3, 4], and transfer learning [5, 6].

A standard approach to comparing probability densities $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$ would be to estimate a divergence from $p(\boldsymbol{x})$ to $p'(\boldsymbol{x})$, such as the *Kullback-Leibler (KL) divergence* [7]:

$$\mathrm{KL}[p(\boldsymbol{x}), p'(\boldsymbol{x})] := \mathbb{E}_{p(\boldsymbol{x})}\left[\log r(\boldsymbol{x})\right], \quad r(\boldsymbol{x}) := p(\boldsymbol{x})/p'(\boldsymbol{x}),$$

where $\mathbb{E}_{p(\boldsymbol{x})}$ denotes the expectation over $p(\boldsymbol{x})$. A naive way to estimate the KL divergence is to separately approximate the densities $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$ from data and plug the estimated densities in the above definition. However, since density estimation is known to be a hard task [8], this approach does not work well unless a good parametric model is available. Recently, a divergence estimation approach which directly approximates the *density-ratio* $r(\boldsymbol{x})$ without going through separate approximation of densities $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$ has been proposed [9, 10]. Such density-ratio approximation methods were proved to achieve the optimal non-parametric convergence rate in the mini-max sense.

However, the KL divergence estimation via density-ratio approximation is computationally rather expensive due to the non-linearity introduced by the 'log' term. To cope with this problem, another divergence called the *Pearson (PE) divergence* [11] is useful. The PE divergence is defined as

$$\mathrm{PE}[p(\boldsymbol{x}), p'(\boldsymbol{x})] := \tfrac{1}{2}\mathbb{E}_{p'(\boldsymbol{x})}\left[(r(\boldsymbol{x})-1)^2\right].$$

The PE divergence is a squared-loss variant of the KL divergence, and they both belong to the class of the *Ali-Silvey-Csiszár divergences* (which is also known as the *f-divergences*, see [12, 13]). Thus, the PE and KL divergences share similar properties, e.g., they are non-negative and vanish if and only if $p(\boldsymbol{x}) = p'(\boldsymbol{x})$.

Similarly to the KL divergence estimation, the PE divergence can also be accurately estimated based on density-ratio approximation [14]: the density-ratio approximator called *unconstrained least-squares importance fitting* (uLSIF) gives the PE divergence estimator *analytically*, which can be computed just by solving a system of linear equations. The practical usefulness of the uLSIF-based PE divergence estimator was demonstrated in various applications such as outlier detection [2], two-sample homogeneity test [4], and dimensionality reduction [15].

In this paper, we first establish the non-parametric convergence rate of the uLSIF-based PE divergence estimator, which elucidates its superior theoretical properties. However, it also reveals that its convergence rate is actually governed by the 'sup'-norm of the true density-ratio function: $\max_{\boldsymbol{x}} r(\boldsymbol{x})$. This implies that, in the region where the denominator density $p'(\boldsymbol{x})$ takes small values, the density-ratio $r(\boldsymbol{x}) = p(\boldsymbol{x})/p'(\boldsymbol{x})$ tends to take large values and therefore the overall convergence speed becomes slow. More critically, density-ratios can even diverge to infinity under a rather simple setting, e.g., when the ratio of two Gaussian functions is considered [16]. This makes the paradigm of divergence estimation based on density-ratio approximation unreliable.

In order to overcome this fundamental problem, we propose an alternative approach to distribution comparison called *α-relative divergence estimation*. In the proposed approach, we estimate the *α-relative divergence*, which is the divergence from $p(\boldsymbol{x})$ to the *α-mixture density*:

$$q_\alpha(\boldsymbol{x}) = \alpha p(\boldsymbol{x}) + (1-\alpha)p'(\boldsymbol{x}) \quad \text{for } 0 \leq \alpha < 1.$$

For example, the α-relative PE divergence is given by

$$\mathrm{PE}_\alpha[p(\boldsymbol{x}), p'(\boldsymbol{x})] := \mathrm{PE}[p(\boldsymbol{x}), q_\alpha(\boldsymbol{x})] = \tfrac{1}{2}\mathbb{E}_{q_\alpha(\boldsymbol{x})}\left[(r_\alpha(\boldsymbol{x})-1)^2\right], \tag{1}$$

where $r_\alpha(\boldsymbol{x})$ is the *α-relative density-ratio* of $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$:

$$r_\alpha(\boldsymbol{x}) := p(\boldsymbol{x})/q_\alpha(\boldsymbol{x}) = p(\boldsymbol{x})/\Big(\alpha p(\boldsymbol{x}) + (1-\alpha)p'(\boldsymbol{x})\Big). \tag{2}$$

We propose to estimate the α-relative divergence by direct approximation of the *α-relative density-ratio*.

A notable advantage of this approach is that the α-relative density-ratio is always bounded above by $1/\alpha$ when $\alpha > 0$, even when the ordinary density-ratio is unbounded. Based on this feature, we theoretically show that the α-relative PE divergence estimator based on α-relative density-ratio approximation is more favorable than the ordinary density-ratio approach in terms of the non-parametric convergence speed.

We further prove that, under a correctly-specified parametric setup, the asymptotic variance of our α-relative PE divergence estimator does not depend on the model complexity. This means that the proposed α-relative PE divergence estimator hardly overfits even with complex models.

Through experiments on outlier detection, two-sample homogeneity test, and transfer learning, we demonstrate that our proposed α-relative PE divergence estimator compares favorably with alternative approaches.

2 Estimation of Relative Pearson Divergence via Least-Squares Relative Density-Ratio Approximation

Suppose we are given independent and identically distributed (i.i.d.) samples $\{\boldsymbol{x}_i\}_{i=1}^n$ from a d-dimensional distribution P with density $p(\boldsymbol{x})$ and i.i.d. samples $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ from another d-dimensional distribution P' with density $p'(\boldsymbol{x})$. Our goal is to compare the two underlying distributions P and P' only using the two sets of samples $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$.

In this section, we give a method for estimating the α-relative PE divergence based on direct approximation of the α-relative density-ratio.

Direct Approximation of α-Relative Density-Ratios: Let us model the α-relative density-ratio $r_\alpha(\boldsymbol{x})$ (2) by the following kernel model $g(\boldsymbol{x};\boldsymbol{\theta}) := \sum_{\ell=1}^{n} \theta_\ell K(\boldsymbol{x},\boldsymbol{x}_\ell)$, where $\boldsymbol{\theta} := (\theta_1,\ldots,\theta_n)^\top$ are parameters to be learned from data samples, $^\top$ denotes the transpose of a matrix or a vector, and $K(\boldsymbol{x},\boldsymbol{x}')$ is a kernel basis function. In the experiments, we use the Gaussian kernel.

The parameters $\boldsymbol{\theta}$ in the model $g(\boldsymbol{x};\boldsymbol{\theta})$ are determined so that the following expected squared-error J is minimized:

$$\begin{aligned} J(\boldsymbol{\theta}) &:= \tfrac{1}{2}\mathbb{E}_{q_\alpha(\boldsymbol{x})}\left[\left(g(\boldsymbol{x};\boldsymbol{\theta}) - r_\alpha(\boldsymbol{x})\right)^2\right] \\ &= \tfrac{\alpha}{2}\mathbb{E}_{p(\boldsymbol{x})}\left[g(\boldsymbol{x};\boldsymbol{\theta})^2\right] + \tfrac{(1-\alpha)}{2}\mathbb{E}_{p'(\boldsymbol{x})}\left[g(\boldsymbol{x};\boldsymbol{\theta})^2\right] - \mathbb{E}_{p(\boldsymbol{x})}\left[g(\boldsymbol{x};\boldsymbol{\theta})\right] + \mathrm{Const.}, \end{aligned}$$

where we used $r_\alpha(\boldsymbol{x})q_\alpha(\boldsymbol{x}) = p(\boldsymbol{x})$ in the third term. Approximating the expectations by empirical averages, we obtain the following optimization problem:

$$\widehat{\boldsymbol{\theta}} := \mathrm{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^n}\left[\tfrac{1}{2}\boldsymbol{\theta}^\top\widehat{\boldsymbol{H}}\boldsymbol{\theta} - \widehat{\boldsymbol{h}}^\top\boldsymbol{\theta} + \tfrac{\lambda}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta}\right], \tag{3}$$

where a penalty term $\lambda\boldsymbol{\theta}^\top\boldsymbol{\theta}/2$ is included for regularization purposes, and λ (≥ 0) denotes the regularization parameter. $\widehat{\boldsymbol{H}}$ and $\widehat{\boldsymbol{h}}$ are defined as

$$\widehat{H}_{\ell,\ell'} := \tfrac{\alpha}{n}\textstyle\sum_{i=1}^{n} K(\boldsymbol{x}_i,\boldsymbol{x}_\ell)K(\boldsymbol{x}_i,\boldsymbol{x}_{\ell'}) + \tfrac{(1-\alpha)}{n'}\sum_{j=1}^{n'} K(\boldsymbol{x}'_j,\boldsymbol{x}_\ell)K(\boldsymbol{x}'_j,\boldsymbol{x}_{\ell'}), \widehat{h}_\ell := \tfrac{1}{n}\textstyle\sum_{i=1}^{n} K(\boldsymbol{x}_i,\boldsymbol{x}_\ell).$$

It is easy to confirm that the solution of Eq.(3) can be *analytically* obtained as $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{H}} + \lambda\boldsymbol{I}_n)^{-1}\widehat{\boldsymbol{h}}$, where $\boldsymbol{I}_n$ denotes the n-dimensional identity matrix. Finally, a density-ratio estimator is given as

$$\widehat{r}_\alpha(\boldsymbol{x}) := g(\boldsymbol{x};\widehat{\boldsymbol{\theta}}) = \textstyle\sum_{\ell=1}^{n}\widehat{\theta}_\ell K(\boldsymbol{x},\boldsymbol{x}_\ell).$$

When $\alpha = 0$, the above method is reduced to a direct density-ratio estimator called *unconstrained least-squares importance fitting* (uLSIF) [14]. Thus, the above method can be regarded as an extension of uLSIF to the α-relative density-ratio. For this reason, we refer to our method as *relative uLSIF* (RuLSIF).

The performance of RuLSIF depends on the choice of the kernel function (the kernel width in the case of the Gaussian kernel) and the regularization parameter λ. Model selection of RuLSIF is possible based on cross-validation (CV) with respect to the squared-error criterion J.

Using an estimator of the α-relative density-ratio $r_\alpha(\boldsymbol{x})$, we can construct estimators of the α-relative PE divergence (1). After a few lines of calculation, we can show that the α-relative PE divergence (1) is equivalently expressed as

$$\mathrm{PE}_\alpha = -\tfrac{\alpha}{2}\mathbb{E}_{p(\boldsymbol{x})}\left[r_\alpha(\boldsymbol{x})^2\right] - \tfrac{(1-\alpha)}{2}\mathbb{E}_{p'(\boldsymbol{x})}\left[r_\alpha(\boldsymbol{x})^2\right] + \mathbb{E}_{p(\boldsymbol{x})}\left[r_\alpha(\boldsymbol{x})\right] - \tfrac{1}{2} = \tfrac{1}{2}\mathbb{E}_{p(\boldsymbol{x})}\left[r_\alpha(\boldsymbol{x})\right] - \tfrac{1}{2}.$$

Note that the middle expression can also be obtained via *Legendre-Fenchel convex duality* of the divergence functional [17].

Based on these expressions, we consider the following two estimators:

$$\widehat{\mathrm{PE}}_\alpha := -\tfrac{\alpha}{2n}\textstyle\sum_{i=1}^{n}\widehat{r}_\alpha(\boldsymbol{x}_i)^2 - \tfrac{(1-\alpha)}{2n'}\sum_{j=1}^{n'}\widehat{r}_\alpha(\boldsymbol{x}'_j)^2 + \tfrac{1}{n}\sum_{i=1}^{n}\widehat{r}_\alpha(\boldsymbol{x}_i) - \tfrac{1}{2}, \tag{4}$$

$$\widetilde{\mathrm{PE}}_\alpha := \tfrac{1}{2n}\textstyle\sum_{i=1}^{n}\widehat{r}_\alpha(\boldsymbol{x}_i) - \tfrac{1}{2}. \tag{5}$$

We note that the α-relative PE divergence (1) can have further different expressions than the above ones, and corresponding estimators can also be constructed similarly. However, the above two expressions will be particularly useful: the first estimator $\widehat{\mathrm{PE}}_\alpha$ has superior theoretical properties (see Section 3) and the second one $\widetilde{\mathrm{PE}}_\alpha$ is simple to compute.

3 Theoretical Analysis

In this section, we analyze theoretical properties of the proposed PE divergence estimators. Since our theoretical analysis is highly technical, we focus on explaining practical insights we can gain from the theoretical results here; we describe all the mathematical details in the supplementary material.

For theoretical analysis, let us consider a rather abstract form of our relative density-ratio estimator described as

$$\mathrm{argmin}_{g\in\mathcal{G}}\left[\frac{\alpha}{2n}\sum_{i=1}^{n}g(\boldsymbol{x}_i)^2+\frac{(1-\alpha)}{2n'}\sum_{j=1}^{n'}g(\boldsymbol{x}'_j)^2-\frac{1}{n}\sum_{i=1}^{n}g(\boldsymbol{x}_i)+\frac{\lambda}{2}R(g)^2\right], \qquad (6)$$

where $\mathcal{G}$ is some function space (i.e., a statistical model) and $R(\cdot)$ is some regularization functional.

Non-Parametric Convergence Analysis: First, we elucidate the non-parametric convergence rate of the proposed PE estimators. Here, we practically regard the function space $\mathcal{G}$ as an infinite-dimensional *reproducing kernel Hilbert space* (RKHS) [18] such as the Gaussian kernel space, and $R(\cdot)$ as the associated RKHS norm.

Let us represent the complexity of the function space $\mathcal{G}$ by γ ($0<\gamma<2$); the larger γ is, the more complex the function class $\mathcal{G}$ is (see the supplementary material for its precise definition). We analyze the convergence rate of our PE divergence estimators as $\bar{n}:=\min(n,n')$ tends to infinity for $\lambda=\lambda_{\bar{n}}$ under

$$\lambda_{\bar{n}}\to o(1) \ \text{ and } \ \lambda_{\bar{n}}^{-1}=o(\bar{n}^{2/(2+\gamma)}).$$

The first condition means that $\lambda_{\bar{n}}$ tends to zero, but the second condition means that its shrinking speed should not be too fast.

Under several technical assumptions detailed in the supplementary material, we have the following asymptotic convergence results for the two PE divergence estimators $\widehat{\mathrm{PE}}_\alpha$ (4) and $\widetilde{\mathrm{PE}}_\alpha$ (5):

$$\widehat{\mathrm{PE}}_\alpha-\mathrm{PE}_\alpha=\mathcal{O}_p(\bar{n}^{-1/2}c\|r_\alpha\|_\infty+\lambda_{\bar{n}}\max(1,R(r_\alpha)^2)), \qquad (7)$$

$$\widetilde{\mathrm{PE}}_\alpha-\mathrm{PE}_\alpha=\mathcal{O}_p\Big(\lambda_{\bar{n}}^{1/2}\|r_\alpha\|_\infty^{1/2}\max\{1,R(r_\alpha)\}+\lambda_{\bar{n}}\max\{1,\|r_\alpha\|_\infty^{(1-\gamma/2)/2},R(r_\alpha)\|r_\alpha\|_\infty^{(1-\gamma/2)/2},R(r_\alpha)\}\Big), \qquad (8)$$

where $\mathcal{O}_p$ denotes the asymptotic order in probability,

$$c:=(1+\alpha)\sqrt{\mathbb{V}_{p(\boldsymbol{x})}[r_\alpha(\boldsymbol{x})]}+(1-\alpha)\sqrt{\mathbb{V}_{p'(\boldsymbol{x})}[r_\alpha(\boldsymbol{x})]},$$

and $\mathbb{V}_{p(\boldsymbol{x})}$ denotes the variance over $p(\boldsymbol{x})$:

$$\mathbb{V}_{p(\boldsymbol{x})}[f(\boldsymbol{x})]=\int\big(f(\boldsymbol{x})-\int f(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\big)^2p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

In both Eq.(7) and Eq.(8), the coefficients of the leading terms (i.e., the first terms) of the asymptotic convergence rates become smaller as $\|r_\alpha\|_\infty$ gets smaller. Since

$$\|r_\alpha\|_\infty=\Big\|\big(\alpha+(1-\alpha)/r(\boldsymbol{x})\big)^{-1}\Big\|_\infty<\tfrac{1}{\alpha}\quad\text{for }\alpha>0,$$

larger α would be more preferable in terms of the asymptotic approximation error. Note that when $\alpha=0$, $\|r_\alpha\|_\infty$ can tend to infinity even under a simple setting that the ratio of two Gaussian functions is considered [16]. Thus, our proposed approach of estimating the α-relative PE divergence (with $\alpha>0$) would be more advantageous than the naive approach of estimating the plain PE divergence (which corresponds to $\alpha=0$) in terms of the non-parametric convergence rate.

The above results also show that $\widehat{\mathrm{PE}}_\alpha$ and $\widetilde{\mathrm{PE}}_\alpha$ have different asymptotic convergence rates. The leading term in Eq.(7) is of order $\bar{n}^{-1/2}$, while the leading term in Eq.(8) is of order $\lambda_{\bar{n}}^{1/2}$, which is slightly slower (depending on the complexity γ) than $\bar{n}^{-1/2}$. Thus, $\widehat{\mathrm{PE}}_\alpha$ would be more accurate than $\widetilde{\mathrm{PE}}_\alpha$ in large sample cases. Furthermore, when $p(\boldsymbol{x})=p'(\boldsymbol{x})$, $\mathbb{V}_{p(\boldsymbol{x})}[r_\alpha(\boldsymbol{x})]=0$ holds and thus $c=0$ holds. Then the leading term in Eq.(7) vanishes and therefore $\widehat{\mathrm{PE}}_\alpha$ has the even faster convergence rate of order $\lambda_{\bar{n}}$, which is slightly slower (depending on the complexity γ) than $\bar{n}^{-1}$. Similarly, if α is close to 1, $r_\alpha(\boldsymbol{x})\approx1$ and thus $c\approx0$ holds.

When $\bar{n}$ is not large enough to be able to neglect the terms of $o(\bar{n}^{-1/2})$, the terms of $O(\lambda_{\bar{n}})$ matter. If $\|r_\alpha\|_\infty$ and $R(r_\alpha)$ are large (this can happen, e.g., when α is close to 0), the coefficient of the $O(\lambda_{\bar{n}})$-term in Eq.(7) can be larger than that in Eq.(8). Then $\widetilde{\mathrm{PE}}_\alpha$ would be more favorable than $\widehat{\mathrm{PE}}_\alpha$ in terms of the approximation accuracy.

See the supplementary material for numerical examples illustrating the above theoretical results.

Parametric Variance Analysis: Next, we analyze the asymptotic variance of the PE divergence estimator $\widehat{\mathrm{PE}}_\alpha$ (4) under a parametric setup.

As the function space $\mathcal{G}$ in Eq.(6), we consider the following parametric model: $\mathcal{G} = \{g(\boldsymbol{x};\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^b\}$ for a finite b. Here we assume that this parametric model is *correctly specified*, i.e., it includes the true relative density-ratio function $r_\alpha(\boldsymbol{x})$: there exists $\boldsymbol{\theta}^*$ such that $g(\boldsymbol{x};\boldsymbol{\theta}^*) = r_\alpha(\boldsymbol{x})$. Here, we use RuLSIF without regularization, i.e., $\lambda = 0$ in Eq.(6).

Let us denote the variance of $\widehat{\mathrm{PE}}_\alpha$ (4) by $\mathbb{V}[\widehat{\mathrm{PE}}_\alpha]$, where randomness comes from the draw of samples $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$. Then, under a standard regularity condition for the asymptotic normality [19], $\mathbb{V}[\widehat{\mathrm{PE}}_\alpha]$ can be expressed and upper-bounded as

$$\mathbb{V}[\widehat{\mathrm{PE}}_\alpha] = \mathbb{V}_{p(\boldsymbol{x})}\left[r_\alpha - \alpha r_\alpha(\boldsymbol{x})^2/2\right]/n + \mathbb{V}_{p'(\boldsymbol{x})}\left[(1-\alpha)r_\alpha(\boldsymbol{x})^2/2\right]/n' + o(n^{-1}, n'^{-1}) \quad (9)$$

$$\leq \|r_\alpha\|_\infty^2/n + \alpha^2\|r_\alpha\|_\infty^4/(4n) + (1-\alpha)^2\|r_\alpha\|_\infty^4/(4n') + o(n^{-1}, n'^{-1}). \quad (10)$$

Let us denote the variance of $\widetilde{\mathrm{PE}}_\alpha$ by $\mathbb{V}[\widetilde{\mathrm{PE}}_\alpha]$. Then, under a standard regularity condition for the asymptotic normality [19], the variance of $\widetilde{\mathrm{PE}}_\alpha$ is asymptotically expressed as

$$\mathbb{V}[\widetilde{\mathrm{PE}}_\alpha] = \mathbb{V}_{p(\boldsymbol{x})}\left[\left(r_\alpha + (1-\alpha r_\alpha)\mathbb{E}_{p(\boldsymbol{x})}[\nabla g]^\top \boldsymbol{H}_\alpha^{-1}\nabla g\right)/2\right]/n + \mathbb{V}_{p'(\boldsymbol{x})}\left[\left((1-\alpha)r_\alpha\mathbb{E}_{p(\boldsymbol{x})}[\nabla g]^\top \boldsymbol{H}_\alpha^{-1}\nabla g\right)/2\right]/n' + o(n^{-1}, n'^{-1}), \quad (11)$$

where ∇g is the gradient vector of g with respect to $\boldsymbol{\theta}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ and

$$\boldsymbol{H}_\alpha = \alpha\mathbb{E}_{p(\boldsymbol{x})}[\nabla g \nabla g^\top] + (1-\alpha)\mathbb{E}_{p'(\boldsymbol{x})}[\nabla g \nabla g^\top].$$

Eq.(9) shows that, up to $O(n^{-1}, n'^{-1})$, the variance of $\widehat{\mathrm{PE}}_\alpha$ depends only on the true relative density-ratio $r_\alpha(\boldsymbol{x})$, not on the estimator of $r_\alpha(\boldsymbol{x})$. This means that the model complexity does not affect the asymptotic variance. Therefore, *overfitting* would hardly occur in the estimation of the relative PE divergence even when complex models are used. We note that the above superior property is applicable only to relative PE divergence estimation, not to relative density-ratio estimation. This implies that overfitting occurs in relative density-ratio estimation, but the approximation error cancels out in relative PE divergence estimation.

On the other hand, Eq.(11) shows that the variance of $\widetilde{\mathrm{PE}}_\alpha$ is affected by the model $\mathcal{G}$, since the factor $\mathbb{E}_{p(\boldsymbol{x})}[\nabla g]^\top \boldsymbol{H}_\alpha^{-1}\nabla g$ depends on the model in general. When the equality $\mathbb{E}_{p(\boldsymbol{x})}[\nabla g]^\top \boldsymbol{H}_\alpha^{-1}\nabla g(\boldsymbol{x};\boldsymbol{\theta}^*) = r_\alpha(\boldsymbol{x})$ holds, the variances of $\widetilde{\mathrm{PE}}_\alpha$ and $\widehat{\mathrm{PE}}_\alpha$ are asymptotically the same. However, in general, the use of $\widehat{\mathrm{PE}}_\alpha$ would be more recommended.

Eq.(10) shows that the variance $\mathbb{V}[\widehat{\mathrm{PE}}_\alpha]$ can be upper-bounded by the quantity depending on $\|r_\alpha\|_\infty$, which is monotonically lowered if $\|r_\alpha\|_\infty$ is reduced. Since $\|r_\alpha\|_\infty$ monotonically decreases as α increases, our proposed approach of estimating the α-relative PE divergence (with $\alpha > 0$) would be more advantageous than the naive approach of estimating the plain PE divergence (which corresponds to $\alpha = 0$) in terms of the parametric asymptotic variance.

See the supplementary material for numerical examples illustrating the above theoretical results.

4 Experiments

In this section, we experimentally evaluate the performance of the proposed method in two-sample homogeneity test, outlier detection, and transfer learning tasks.

Two-Sample Homogeneity Test: First, we apply the proposed divergence estimator to two-sample homogeneity test.

Given two sets of samples $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P$ and $\mathcal{X}' = \{\boldsymbol{x}'_j\}_{j=1}^{n'} \overset{\text{i.i.d.}}{\sim} P'$, the goal of the two-sample homogeneity test is to test the *null hypothesis* that the probability distributions P and P' are the same against its complementary alternative (i.e., the distributions are different). By using an estimator $\widehat{\mathrm{Div}}$ of some divergence between the two distributions P and P', homogeneity of two distributions can be tested based on the *permutation test* procedure [20].

Table 1: Experimental results of two-sample test. The mean (and standard deviation in the bracket) rate of accepting the null hypothesis (i.e., $P = P'$) for IDA benchmark repository under the significance level 5% is reported. Left: when the two sets of samples are both taken from the positive training set (i.e., the null hypothesis is correct). Methods having the mean acceptance rate 0.95 according to the *one-sample t-test* at the significance level 5% are specified by bold face. Right: when the set of samples corresponding to the numerator of the density-ratio are taken from the positive training set and the set of samples corresponding to the denominator of the density-ratio are taken from the positive training set and the negative training set (i.e., the null hypothesis is not correct). The best method having the lowest mean accepting rate and comparable methods according to the *two-sample t-test* at the significance level 5% are specified by bold face.

			$P = P'$				$P \neq P'$			
Datasets	d	$n = n'$	MMD	LSTT (α = 0.0)	LSTT (α = 0.5)	LSTT (α = 0.95)	MMD	LSTT (α = 0.0)	LSTT (α = 0.5)	LSTT (α = 0.95)
banana	2	100	**.96 (.20)**	**.93 (.26)**	**.92 (.27)**	**.92 (.27)**	.52 (.50)	**.10 (.30)**	**.02 (.14)**	**.17 (.38)**
thyroid	5	19	**.96 (.20)**	**.95 (.22)**	**.95 (.22)**	.88 (.33)	**.52 (.50)**	.81 (.39)	**.65 (.48)**	.80 (.40)
titanic	5	21	**.94 (.24)**	**.86 (.35)**	**.92 (.27)**	**.89 (.31)**	**.87 (.34)**	**.86 (.35)**	**.87 (.34)**	**.88 (.33)**
diabetes	8	85	**.96 (.20)**	.87 (.34)	**.91 (.29)**	.82 (.39)	**.31 (.46)**	**.42 (.50)**	.47 (.50)	.57 (.50)
b-cancer	9	29	.98 (.14)	**.91 (.29)**	**.94 (.24)**	**.92 (.27)**	.87 (.34)	**.75 (.44)**	**.80 (.40)**	**.79 (.41)**
f-solar	9	100	**.93 (.26)**	**.91 (.29)**	**.95 (.22)**	**.93 (.26)**	**.51 (.50)**	.81 (.39)	**.55 (.50)**	**.66 (.48)**
heart	13	38	1.00 (.00)	.85 (.36)	**.91 (.29)**	**.93 (.26)**	.53 (.50)	**.28 (.45)**	**.40 (.49)**	.62 (.49)
german	20	100	.99 (.10)	**.91 (.29)**	**.92 (.27)**	**.89 (.31)**	.56 (.50)	.55 (.50)	**.44 (.50)**	.68 (.47)
ringnorm	20	100	**.97 (.17)**	**.93 (.26)**	**.91 (.29)**	.85 (.36)	**.00 (.00)**	**.00 (.00)**	**.00 (.00)**	**.02 (.14)**
waveform	21	66	.98 (.14)	**.92 (.27)**	**.93 (.26)**	**.88 (.33)**	**.00 (.00)**	**.00 (.00)**	**.02 (.14)**	**.00 (.00)**

When an asymmetric divergence such as the KL divergence [7] or the PE divergence [11] is adopted for two-sample test, the test results depend on the choice of *directions*: a divergence from P to P' or from P' to P. [4] proposed to choose the direction that gives a smaller p-value—it was experimentally shown that, when the uLSIF-based PE divergence estimator is used for the two-sample test (which is called the *least-squares two-sample test*; LSTT), the heuristic of choosing the direction with a smaller p-value contributes to reducing 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), while the increase of 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) is kept moderate.

We apply the proposed method to the binary classification datasets taken from the *IDA benchmark repository* [21]. We test LSTT with the RuLSIF-based PE divergence estimator for $\alpha = 0$, 0.5, and 0.95; we also test the *maximum mean discrepancy* (MMD) [22], which is a kernel-based two-sample test method. The performance of MMD depends on the choice of the Gaussian kernel width. Here, we adopt a version proposed by [23], which automatically optimizes the Gaussian kernel width. The p-values of MMD are computed in the same way as LSTT based on the permutation test procedure.

First, we investigate the rate of accepting the null hypothesis when the null hypothesis is correct (i.e., the two distributions are the same). We split all the positive training samples into two sets and perform two-sample test for the two sets of samples. The experimental results are summarized in the left half of Table 1, showing that LSTT with $\alpha = 0.5$ compares favorably with those with $\alpha = 0$ and 0.95 and MMD in terms of the type-I error.

Next, we consider the situation where the null hypothesis is not correct (i.e., the two distributions are different). The numerator samples are generated in the same way as above, but a half of denominator samples are replaced with negative training samples. Thus, while the numerator sample set contains only positive training samples, the denominator sample set includes both positive and negative training samples. The experimental results are summarized in the right half of Table 1, showing that LSTT with $\alpha = 0.5$ again compares favorably with those with $\alpha = 0$ and 0.95. Furthermore, LSTT with $\alpha = 0.5$ tends to outperform MMD in terms of the type-II error.

Overall, LSTT with $\alpha = 0.5$ is shown to be a useful method for two-sample homogeneity test. See the supplementary material for more experimental evaluation.

Inlier-Based Outlier Detection: Next, we apply the proposed method to outlier detection.

Let us consider an outlier detection problem of finding irregular samples in a dataset (called an "evaluation dataset") based on another dataset (called a "model dataset") that only contains regular samples. Defining the density-ratio over the two sets of samples, we can see that the density-ratio

Table 2: Experimental results of outlier detection. Mean AUC score (and standard deviation in the bracket) over 100 trials is reported. The best method having the highest mean AUC score and comparable methods according to the *two-sample t-test* at the significance level 5% are specified by bold face. The datasets are sorted in the ascending order of the input dimensionality d.

Datasets	d	OSVM ($\nu = 0.05$)	OSVM ($\nu = 0.1$)	RuLSIF ($\alpha = 0$)	RuLSIF ($\alpha = 0.5$)	RuLSIF ($\alpha = 0.95$)
IDA:banana	2	**.668 (.105)**	**.676 (.120)**	.597 (.097)	.619 (.101)	.623 (.115)
IDA:thyroid	5	.760 (.148)	**.782 (.165)**	**.804 (.148)**	**.796 (.178)**	.722 (.153)
IDA:titanic	5	**.757 (.205)**	**.752 (.191)**	**.750 (.182)**	.701 (.184)	.712 (.185)
IDA:diabetes	8	**.636 (.099)**	.610 (.090)	.594 (.105)	.575 (.105)	**.663 (.112)**
IDA:breast-cancer	9	**.741 (.160)**	.691 (.147)	**.707 (.148)**	**.737 (.159)**	**.733 (.160)**
IDA:flare-solar	9	.594 (.087)	.590 (.083)	**.626 (.102)**	**.612 (.100)**	.584 (.114)
IDA:heart	13	.714 (.140)	.694 (.148)	**.748 (.149)**	**.769 (.134)**	.726 (.127)
IDA:german	20	**.612 (.069)**	**.604 (.084)**	**.605 (.092)**	**.597 (.101)**	**.605 (.095)**
IDA:ringnorm	20	**.991 (.012)**	**.993 (.007)**	.944 (.091)	.971 (.062)	**.992 (.010)**
IDA:waveform	21	.812 (.107)	.843 (.123)	**.879 (.122)**	**.875 (.117)**	**.885 (.102)**
Speech	50	.788 (.068)	**.830 (.060)**	.804 (.101)	**.821 (.076)**	**.836 (.083)**
20News ('rec')	100	.598 (.063)	.593 (.061)	.628 (.105)	.614 (.093)	**.767 (.100)**
20News ('sci')	100	.592 (.069)	.589 (.071)	.620 (.094)	.609 (.087)	**.704 (.093)**
20News ('talk')	100	.661 (.084)	.658 (.084)	.672 (.117)	.670 (.102)	**.823 (.078)**
USPS (1 vs. 2)	256	.889 (.052)	**.926 (.037)**	.848 (.081)	.878 (.088)	.898 (.051)
USPS (2 vs. 3)	256	.823 (.053)	.835 (.050)	.803 (.093)	.818 (.085)	**.879 (.074)**
USPS (3 vs. 4)	256	.901 (.044)	.939 (.031)	.950 (.056)	.961 (.041)	**.984 (.016)**
USPS (4 vs. 5)	256	.871 (.041)	.890 (.036)	.857 (.099)	.874 (.082)	**.941 (.031)**
USPS (5 vs. 6)	256	.825 (.058)	.859 (.052)	.863 (.078)	.867 (.068)	**.901 (.049)**
USPS (6 vs. 7)	256	.910 (.034)	.950 (.025)	.972 (.038)	.984 (.018)	**.994 (.010)**
USPS (7 vs. 8)	256	.938 (.030)	.967 (.021)	.941 (.053)	.951 (.039)	**.980 (.015)**
USPS (8 vs. 9)	256	.721 (.072)	.728 (.073)	.721 (.084)	.728 (.083)	**.761 (.096)**
USPS (9 vs. 0)	256	.920 (.037)	.966 (.023)	.982 (.048)	.989 (.022)	**.994 (.011)**

values for regular samples are close to one, while those for outliers tend to be significantly deviated from one. Thus, density-ratio values could be used as an index of the degree of outlyingness [1, 2].

Since the evaluation dataset usually has a wider support than the model dataset, we regard the evaluation dataset as samples corresponding to the denominator density $p'(\boldsymbol{x})$, and the model dataset as samples corresponding to the numerator density $p(\boldsymbol{x})$. Then, outliers tend to have smaller density-ratio values (i.e., close to zero). Thus, density-ratio approximators can be used for outlier detection.

We evaluate the proposed method using various datasets: IDA benchmark repository [21], an in-house French speech dataset, the 20 Newsgroup dataset, and the USPS hand-written digit dataset (the detailed specification of the datasets is explained in the supplementary material).

We compare the *area under the ROC curve* (AUC) [24] of RuLSIF with $\alpha = 0$, 0.5, and 0.95, and *one-class support vector machine (OSVM)* with the Gaussian kernel [25]. We used the *LIBSVM* implementation of OSVM [26]. The Gaussian width is set to the median distance between samples, which has been shown to be a useful heuristic [25]. Since there is no systematic method to determine the tuning parameter ν in OSVM, we report the results for $\nu = 0.05$ and 0.1.

The mean and standard deviation of the AUC scores over 100 runs with random sample choice are summarized in Table 2, showing that RuLSIF overall compares favorably with OSVM. Among the RuLSIF methods, small α tends to perform well for low-dimensional datasets, and large α tends to work well for high-dimensional datasets.

Transfer Learning: Finally, we apply the proposed method to transfer learning.

Let us consider a transductive transfer learning setup where labeled training samples $\{(\boldsymbol{x}_j^{\mathrm{tr}}, y_j^{\mathrm{tr}})\}_{j=1}^{n_{\mathrm{tr}}}$ drawn i.i.d. from $p(y|\boldsymbol{x})p_{\mathrm{tr}}(\boldsymbol{x})$ and unlabeled test samples $\{\boldsymbol{x}_i^{\mathrm{te}}\}_{i=1}^{n_{\mathrm{te}}}$ drawn i.i.d. from $p_{\mathrm{te}}(\boldsymbol{x})$ (which is generally different from $p_{\mathrm{tr}}(\boldsymbol{x})$) are available. The use of *exponentially-weighted importance weighting* was shown to be useful for adaptation from $p_{\mathrm{tr}}(\boldsymbol{x})$ to $p_{\mathrm{te}}(\boldsymbol{x})$ [5]:

$$\min_{f \in \mathcal{F}} \left[\frac{1}{n_{\mathrm{tr}}} \sum_{j=1}^{n_{\mathrm{tr}}} \left(\frac{p_{\mathrm{te}}(\boldsymbol{x}_j^{\mathrm{tr}})}{p_{\mathrm{tr}}(\boldsymbol{x}_j^{\mathrm{tr}})} \right)^{\tau} \mathrm{loss}(y_j^{\mathrm{tr}}, f(\boldsymbol{x}_j^{\mathrm{tr}})) \right],$$

where $f(\boldsymbol{x})$ is a learned function and $0 \leq \tau \leq 1$ is the exponential flattening parameter. $\tau = 0$ corresponds to plain empirical-error minimization which is statistically efficient, while $\tau = 1$ corresponds to importance-weighted empirical-error minimization which is statistically consistent; $0 < \tau < 1$ will give an intermediate estimator that balances the trade-off between statistical efficiency and consistency. τ can be determined by *importance-weighted cross-validation* [6] in a data dependent fashion.

Table 3: Experimental results of transfer learning in human activity recognition. Mean classification accuracy (and the standard deviation in the bracket) over 100 runs for human activity recognition of a new user is reported. We compare the plain *kernel logistic regression* (KLR) without importance weights, KLR with relative importance weights (RIW-KLR), KLR with exponentially-weighted importance weights (EIW-KLR), and KLR with plain importance weights (IW-KLR). The method having the highest mean classification accuracy and comparable methods according to the *two-sample t-test* at the significance level 5% are specified by bold face.

Task	KLR ($\alpha = 0, \tau = 0$)		RIW-KLR ($\alpha = 0.5$)	EIW-KLR ($\tau = 0.5$)	IW-KLR ($\alpha = 1, \tau = 1$)	
Walks vs. run	0.803	(0.082)	**0.889 (0.035)**	**0.882 (0.039)**	**0.882**	**(0.035)**
Walks vs. bicycle	0.880	(0.025)	**0.892 (0.035)**	0.867 (0.054)	0.854	(0.070)
Walks vs. train	0.985	(0.017)	**0.992 (0.008)**	0.989 (0.011)	0.983	(0.021)

However, a potential drawback is that estimation of $r(\boldsymbol{x})$ (i.e., $\tau = 1$) is rather hard, as shown in this paper. Here we propose to use *relative importance weights* instead:

$$\min_{f \in \mathcal{F}} \left[\frac{1}{n_{\mathrm{tr}}} \sum_{j=1}^{n_{\mathrm{tr}}} \frac{p_{\mathrm{te}}(\boldsymbol{x}_j^{\mathrm{tr}})}{(1-\alpha)p_{\mathrm{te}}(\boldsymbol{x}_j^{\mathrm{tr}}) + \alpha p_{\mathrm{tr}}(\boldsymbol{x}_j^{\mathrm{tr}})} \mathrm{loss}(y_j^{\mathrm{tr}}, f(\boldsymbol{x}_j^{\mathrm{tr}})) \right].$$

We apply the above transfer learning technique to *human activity recognition* using accelerometer data. Subjects were asked to perform a specific task such as walking, running, and bicycle riding, which was collected by *iPodTouch*. The duration of each task was arbitrary and the sampling rate was 20Hz with small variations (the detailed experimental setup is explained in the supplementary material). Let us consider a situation where a new user wants to use the activity recognition system. However, since the new user is not willing to label his/her accelerometer data due to troublesomeness, no labeled sample is available for the new user. On the other hand, unlabeled samples for the new user and labeled data obtained from existing users are available. Let labeled training data $\{(\boldsymbol{x}_j^{\mathrm{tr}}, y_j^{\mathrm{tr}})\}_{j=1}^{n_{\mathrm{tr}}}$ be the set of labeled accelerometer data for 20 existing users. Each user has at most 100 labeled samples for each action. Let unlabeled test data $\{\boldsymbol{x}_i^{\mathrm{te}}\}_{i=1}^{n_{\mathrm{te}}}$ be unlabeled accelerometer data obtained from the new user.

The experiments are repeated 100 times with different sample choice for $n_{\mathrm{tr}} = 500$ and $n_{\mathrm{te}} = 200$. The classification accuracy for 800 test samples from the new user (which are different from the 200 unlabeled samples) are summarized in Table 3, showing that the proposed method using relative importance weights for $\alpha = 0.5$ works better than other methods.

5 Conclusion

In this paper, we proposed to use a relative divergence for robust distribution comparison. We gave a computationally efficient method for estimating the relative Pearson divergence based on direct relative density-ratio approximation. We theoretically elucidated the convergence rate of the proposed divergence estimator under non-parametric setup, which showed that the proposed approach of estimating the relative Pearson divergence is more preferable than the existing approach of estimating the plain Pearson divergence. Furthermore, we proved that the asymptotic variance of the proposed divergence estimator is independent of the model complexity under a correctly-specified parametric setup. Thus, the proposed divergence estimator hardly overfits even with complex models. Experimentally, we demonstrated the practical usefulness of the proposed divergence estimator in two-sample homogeneity test, inlier-based outlier detection, and transfer learning tasks.

In addition to two-sample homogeneity test, inlier-based outlier detection, and transfer learning, density-ratios can be useful for tackling various machine learning problems, for example, multi-task learning, independence test, feature selection, causal inference, independent component analysis, dimensionality reduction, unpaired data matching, clustering, conditional density estimation, and probabilistic classification. Thus, it would be promising to explore more applications of the proposed relative density-ratio approximator beyond two-sample homogeneity test, inlier-based outlier detection, and transfer learning.

Acknowledgments
MY was supported by the JST PRESTO program, TS was partially supported by MEXT KAKENHI 22700289 and Aihara Project, the FIRST program from JSPS, initiated by CSTP, TK was partially supported by Grant-in-Aid for Young Scientists (20700251), HH was supported by the FIRST program, and MS was partially supported by SCAT, AOARD, and the FIRST program.

References

[1] A. J. Smola, L. Song, and C. H. Teo. Relative novelty detection. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS2009)*, pages 536–543, 2009.

[2] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori. Statistical outlier detection using direct density ratio estimation. *Knowledge and Information Systems*, 26(2):309–336, 2011.

[3] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, Cambridge, MA, 2007.

[4] M. Sugiyama, T. Suzuki, Y. Itoh, T. Kanamori, and M. Kimura. Least-squares two-sample test. *Neural Networks*, 24(7):735–751, 2011.

[5] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[6] M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, May 2007.

[7] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[8] V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, 1998.

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

[10] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[11] K. Pearson. 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, 1900.

[12] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 28:131–142, 1966.

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

[14] T. Kanamori, S. Hido, and M. Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

[15] T. Suzuki and M. Sugiyama. Sufficient dimension reduction via squared-loss mutual information estimation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS2010)*, pages 804–811, 2010.

[16] C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 442–450. 2010.

[17] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, USA, 1970.

[18] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[19] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

[20] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, NY, 1993.

[21] G. Rätsch, T. Onoda, and K.-R. Müller. Soft margins for adaboost. *Machine Learning*, 42(3):287–320, 2001.

[22] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[23] B. Sriperumbudur, K. Fukumizu, A. Gretton, G. Lanckriet, and B. Schölkopf. Kernel choice and classifiability for RKHS embeddings of probability distributions. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1750–1758. 2009.

[24] A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159, 1997.

[25] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

[26] C.-C. Chang and C.h-J. Lin. *LIBSVM: A Library for Support Vector Machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Supplementary Material

Makoto Yamada
Tokyo Institute of Technology
yamada@sg.cs.titech.ac.jp

Taiji Suzuki
The University of Tokyo
s-taiji@stat.t.u-tokyo.ac.jp

Takafumi Kanamori
Nagoya University
kanamori@is.nagoya-u.ac.jp

Hirotaka Hachiya **Masashi Sugiyama**
Tokyo Institute of Technology
{hachiya@sg. sugi@}cs.titech.ac.jp

Abstract

Divergence estimators based on direct approximation of density-ratios without going through separate approximation of numerator and denominator densities have been successfully applied to machine learning tasks that involve distribution comparison such as outlier detection, transfer learning, and two-sample homogeneity test. However, since density-ratio functions often possess high fluctuation, divergence estimation is still a challenging task in practice. In this paper, we propose to use *relative divergences* for distribution comparison, which involves approximation of *relative density-ratios*. Since relative density-ratios are always smoother than corresponding ordinary density-ratios, our proposed method is favorable in terms of the non-parametric convergence speed. Furthermore, we show that the proposed divergence estimator has asymptotic variance *independent* of the model complexity under a parametric setup, implying that the proposed estimator hardly overfits even with complex models. Through experiments, we demonstrate the usefulness of the proposed approach.

1 Introduction

Comparing probability distributions is a fundamental task in statistical data processing. It can be used for, e.g., *outlier detection* [1, 2], *two-sample homogeneity test* [3, 4], and *transfer learning* [5, 6].

A standard approach to comparing probability densities $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$ would be to estimate a divergence from $p(\boldsymbol{x})$ to $p'(\boldsymbol{x})$, such as the *Kullback-Leibler (KL) divergence* [7]:

$$\mathrm{KL}[p(\boldsymbol{x}), p'(\boldsymbol{x})] := \int \log\left(\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}\right) p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

A naive way to estimate the KL divergence is to separately approximate the densities $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$ from data and plug the estimated densities in the above definition. However, since density estimation is known to be a hard task [8], this approach does not work well unless a good parametric model is available. Recently, a divergence estimation approach which directly approximates the *density ratio*,

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

without going through separate approximation of densities $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$ has been proposed [9, 10]. Such density-ratio approximation methods were proved to achieve the optimal non-parametric convergence rate in the mini-max sense.

However, the KL divergence estimation via density-ratio approximation is computationally rather expensive due to the non-linearity introduced by the 'log' term. To cope with this problem, another

divergence called the *Pearson (PE) divergence* [11] is useful. The PE divergence from $p(\boldsymbol{x})$ to $p'(\boldsymbol{x})$ is defined as

$$\mathrm{PE}[p(\boldsymbol{x}), p'(\boldsymbol{x})] := \frac{1}{2}\int \left(\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})} - 1\right)^2 p'(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

The PE divergence is a squared-loss variant of the KL divergence, and they both belong to the class of the *Ali-Silvey-Csiszár divergences* [which is also known as the *f-divergences*, see 12, 13]. Thus, the PE and KL divergences share similar properties, e.g., they are non-negative and vanish if and only if $p(\boldsymbol{x}) = p'(\boldsymbol{x})$.

Similarly to the KL divergence estimation, the PE divergence can also be accurately estimated based on density-ratio approximation [14]: the density-ratio approximator called *unconstrained least-squares importance fitting* (uLSIF) gives the PE divergence estimator *analytically*, which can be computed just by solving a system of linear equations. The practical usefulness of the uLSIF-based PE divergence estimator was demonstrated in various applications such as outlier detection [2], two-sample homogeneity test [4], and dimensionality reduction [15].

In this paper, we first establish the non-parametric convergence rate of the uLSIF-based PE divergence estimator, which elucidates its superior theoretical properties. However, it also reveals that its convergence rate is actually governed by the 'sup'-norm of the true density-ratio function: $\max_{\boldsymbol{x}} r(\boldsymbol{x})$. This implies that, in the region where the denominator density $p'(\boldsymbol{x})$ takes small values, the density ratio $r(\boldsymbol{x}) = p(\boldsymbol{x})/p'(\boldsymbol{x})$ tends to take large values and therefore the overall convergence speed becomes slow. More critically, density ratios can even diverge to infinity under a rather simple setting, e.g., when the ratio of two Gaussian functions is considered [16]. This makes the paradigm of divergence estimation based on density-ratio approximation unreliable.

In order to overcome this fundamental problem, we propose an alternative approach to distribution comparison called *α-relative divergence estimation*. In the proposed approach, we estimate the quantity called the *α-relative divergence*, which is the divergence from $p(\boldsymbol{x})$ to the *α-mixture density* $\alpha p(\boldsymbol{x}) + (1-\alpha)p'(\boldsymbol{x})$ for $0 \leq \alpha < 1$. For example, the α-relative PE divergence is given by

$$\begin{aligned}\mathrm{PE}_\alpha[p(\boldsymbol{x}), p'(\boldsymbol{x})] &:= \mathrm{PE}[p(\boldsymbol{x}), \alpha p(\boldsymbol{x}) + (1-\alpha)p'(\boldsymbol{x})] \\ &= \frac{1}{2}\int \left(\frac{p(\boldsymbol{x})}{\alpha p(\boldsymbol{x}) + (1-\alpha)p'(\boldsymbol{x})} - 1\right)^2 (\alpha p(\boldsymbol{x}) + (1-\alpha)p'(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}.\end{aligned}$$

We estimate the α-relative divergence by direct approximation of the *α-relative density-ratio*:

$$r_\alpha(\boldsymbol{x}) := \frac{p(\boldsymbol{x})}{\alpha p(\boldsymbol{x}) + (1-\alpha)p'(\boldsymbol{x})}.$$

A notable advantage of this approach is that the α-relative density-ratio is always bounded above by $1/\alpha$ when $\alpha > 0$, even when the ordinary density-ratio is unbounded. Based on this feature, we theoretically show that the α-relative PE divergence estimator based on α-relative density-ratio approximation is more favorable than the ordinary density-ratio approach in terms of the non-parametric convergence speed.

We further prove that, under a correctly-specified parametric setup, the asymptotic variance of our α-relative PE divergence estimator does not depend on the model complexity. This means that the proposed α-relative PE divergence estimator hardly overfits even with complex models.

Through extensive experiments on outlier detection, two-sample homogeneity test, and transfer learning, we demonstrate that our proposed α-relative PE divergence estimator compares favorably with alternative approaches.

The rest of this paper is structured as follows. In Section 2, our proposed relative PE divergence estimator is described. In Section 3, we provide non-parametric analysis of the convergence rate and parametric analysis of the variance of the proposed PE divergence estimator. In Section 4, we experimentally evaluate the performance of the proposed method on various tasks. Finally, in Section 5, we conclude the paper by summarizing our contributions and describing future prospects.

2 Estimation of Relative Pearson Divergence via Least-Squares Relative Density-Ratio Approximation

In this section, we propose an estimator of the relative Pearson (PE) divergence based on least-squares relative density-ratio approximation.

2.1 Problem Formulation

Suppose we are given independent and identically distributed (i.i.d.) samples $\{\boldsymbol{x}_i\}_{i=1}^n$ from a d-dimensional distribution P with density $p(\boldsymbol{x})$ and i.i.d. samples $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ from another d-dimensional distribution P' with density $p'(\boldsymbol{x})$:

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

The goal of this paper is to compare the two underlying distributions P and P' only using the two sets of samples $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$.

For $0 \leq \alpha < 1$, let $q_\alpha(\boldsymbol{x})$ be the *α-mixture density* of $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$:

$$q_\alpha(\boldsymbol{x}) := \alpha p(\boldsymbol{x}) + (1-\alpha)p'(\boldsymbol{x}).$$

Let $r_\alpha(\boldsymbol{x})$ be the *α-relative density-ratio* of $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$:

$$r_\alpha(\boldsymbol{x}) := \frac{p(\boldsymbol{x})}{\alpha p(\boldsymbol{x}) + (1-\alpha)p'(\boldsymbol{x})} = \frac{p(\boldsymbol{x})}{q_\alpha(\boldsymbol{x})}. \tag{1}$$

We define *the α-relative PE divergence* from $p(\boldsymbol{x})$ to $p'(\boldsymbol{x})$ as

$$\mathrm{PE}_\alpha := \frac{1}{2}\mathbb{E}_{q_\alpha(\boldsymbol{x})}\left[(r_\alpha(\boldsymbol{x})-1)^2\right], \tag{2}$$

where $\mathbb{E}_{p(\boldsymbol{x})}[f(\boldsymbol{x})]$ denotes the expectation of $f(\boldsymbol{x})$ under $p(\boldsymbol{x})$:

$$\mathbb{E}_{p(\boldsymbol{x})}[f(\boldsymbol{x})] = \int f(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

When $\alpha = 0$, PE_α is reduced to the ordinary PE divergence. Thus, the α-relative PE divergence can be regarded as a 'smoothed' extension of the ordinary PE divergence.

Below, we give a method for estimating the α-relative PE divergence based on the approximation of the α-relative density-ratio.

2.2 Direct Approximation of α-Relative Density-Ratios

Here, we describe a method for approximating the α-relative density-ratio (1).

Let us model the α-relative density-ratio $r_\alpha(\boldsymbol{x})$ by the following kernel model:

$$g(\boldsymbol{x};\boldsymbol{\theta}) := \sum_{\ell=1}^n \theta_\ell K(\boldsymbol{x},\boldsymbol{x}_\ell),$$

where $\boldsymbol{\theta} := (\theta_1,\ldots,\theta_n)^\top$ are parameters to be learned from data samples, $^\top$ denotes the transpose of a matrix or a vector, and $K(\boldsymbol{x},\boldsymbol{x}')$ is a kernel basis function. In the experiments, we use the Gaussian kernel:

$$K(\boldsymbol{x},\boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2\sigma^2}\right), \tag{3}$$

where σ (> 0) is the kernel width.

The parameters $\boldsymbol{\theta}$ in the model $g(\boldsymbol{x};\boldsymbol{\theta})$ are determined so that the following expected squared-error J is minimized:

$$J(\boldsymbol{\theta}) := \frac{1}{2}\mathbb{E}_{q_\alpha(\boldsymbol{x})}\left[(g(\boldsymbol{x};\boldsymbol{\theta}) - r_\alpha(\boldsymbol{x}))^2\right]$$
$$= \frac{\alpha}{2}\mathbb{E}_{p(\boldsymbol{x})}\left[g(\boldsymbol{x};\boldsymbol{\theta})^2\right] + \frac{(1-\alpha)}{2}\mathbb{E}_{p'(\boldsymbol{x})}\left[g(\boldsymbol{x};\boldsymbol{\theta})^2\right] - \mathbb{E}_{p(\boldsymbol{x})}\left[g(\boldsymbol{x};\boldsymbol{\theta})\right] + \mathrm{Const.},$$

where we used $r_\alpha(\boldsymbol{x})q_\alpha(\boldsymbol{x}) = p(\boldsymbol{x})$ in the third term. Approximating the expectations by empirical averages, we obtain the following optimization problem:

$$\widehat{\boldsymbol{\theta}} := \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^n}\left[\frac{1}{2}\boldsymbol{\theta}^\top\widehat{\boldsymbol{H}}\boldsymbol{\theta} - \widehat{\boldsymbol{h}}^\top\boldsymbol{\theta} + \frac{\lambda}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta}\right], \tag{4}$$

where a penalty term $\lambda\boldsymbol{\theta}^\top\boldsymbol{\theta}/2$ is included for regularization purposes, and λ (≥ 0) denotes the regularization parameter. $\widehat{\boldsymbol{H}}$ is the $n \times n$ matrix with the (ℓ,ℓ')-th element

$$\widehat{H}_{\ell,\ell'} := \frac{\alpha}{n}\sum_{i=1}^{n} K(\boldsymbol{x}_i,\boldsymbol{x}_\ell)K(\boldsymbol{x}_i,\boldsymbol{x}_{\ell'}) + \frac{(1-\alpha)}{n'}\sum_{j=1}^{n'} K(\boldsymbol{x}'_j,\boldsymbol{x}_\ell)K(\boldsymbol{x}'_j,\boldsymbol{x}_{\ell'}). \tag{5}$$

$\widehat{\boldsymbol{h}}$ is the n-dimensional vector with the ℓ-th element

$$\widehat{h}_\ell := \frac{1}{n}\sum_{i=1}^{n} K(\boldsymbol{x}_i,\boldsymbol{x}_\ell).$$

It is easy to confirm that the solution of Eq.(4) can be *analytically* obtained as

$$\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{H}} + \lambda\boldsymbol{I}_n)^{-1}\widehat{\boldsymbol{h}},$$

where $\boldsymbol{I}_n$ denotes the n-dimensional identity matrix. Finally, a density-ratio estimator is given as

$$\widehat{r}_\alpha(\boldsymbol{x}) := g(\boldsymbol{x};\widehat{\boldsymbol{\theta}}) = \sum_{\ell=1}^{n}\widehat{\theta}_\ell K(\boldsymbol{x},\boldsymbol{x}_\ell). \tag{6}$$

When $\alpha = 0$, the above method is reduced to a direct density-ratio estimator called *unconstrained least-squares importance fitting* [uLSIF; 14]. Thus, the above method can be regarded as an extension of uLSIF to the α-relative density-ratio. For this reason, we refer to our method as *relative uLSIF* (RuLSIF).

The performance of RuLSIF depends on the choice of the kernel function (the kernel width σ in the case of the Gaussian kernel) and the regularization parameter λ. Model selection of RuLSIF is possible based on cross-validation with respect to the squared-error criterion J, in the same way as the original uLSIF [14].

2.3 α-Relative PE Divergence Estimation Based on RuLSIF

Using an estimator of the α-relative density-ratio $r_\alpha(\boldsymbol{x})$, we can construct estimators of the α-relative PE divergence (2). After a few lines of calculation, we can show that the α-relative PE divergence (2) is equivalently expressed as

$$\mathrm{PE}_\alpha = -\frac{\alpha}{2}\mathbb{E}_{p(\boldsymbol{x})}\left[r_\alpha(\boldsymbol{x})^2\right] - \frac{(1-\alpha)}{2}\mathbb{E}_{p'(\boldsymbol{x})}\left[r_\alpha(\boldsymbol{x})^2\right] + \mathbb{E}_{p(\boldsymbol{x})}\left[r_\alpha(\boldsymbol{x})\right] - \frac{1}{2}$$
$$= \frac{1}{2}\mathbb{E}_{p(\boldsymbol{x})}\left[r_\alpha(\boldsymbol{x})\right] - \frac{1}{2}.$$

Note that the first line can also be obtained via *Legendre-Fenchel convex duality* of the divergence functional [17].

Based on these expressions, we consider the following two estimators:

$$\widehat{\mathrm{PE}}_\alpha := -\frac{\alpha}{2n}\sum_{i=1}^{n}\widehat{r}(\boldsymbol{x}_i)^2 - \frac{(1-\alpha)}{2n'}\sum_{j=1}^{n'}\widehat{r}(\boldsymbol{x}'_j)^2 + \frac{1}{n}\sum_{i=1}^{n}\widehat{r}(\boldsymbol{x}_i) - \frac{1}{2}, \tag{7}$$

$$\widetilde{\mathrm{PE}}_\alpha := \frac{1}{2n}\sum_{i=1}^{n}\widehat{r}(\boldsymbol{x}_i) - \frac{1}{2}. \tag{8}$$

We note that the α-relative PE divergence (2) can have further different expressions than the above ones, and corresponding estimators can also be constructed similarly. However, the above two expressions will be particularly useful: the first estimator $\widehat{\mathrm{PE}}_\alpha$ has superior theoretical properties (see Section 3) and the second one $\widetilde{\mathrm{PE}}_\alpha$ is simple to compute.

2.4 Illustrative Examples

Here, we numerically illustrate the behavior of RuLSIF (6) using toy datasets. Let the numerator distribution be $P = N(0,1)$, where $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. The denominator distribution P' is set as follows:

(a) $P' = N(0,1)$: P and P' are the same.

(b) $P' = N(0, 0.6)$: P' has smaller standard deviation than P.

(c) $P' = N(0,2)$: P' has larger standard deviation than P.

(d) $P' = N(0.5, 1)$: P and P' have different means.

(e) $P' = 0.95N(0,1) + 0.05N(3,1)$: P' contains an additional component to P.

We draw $n = n' = 300$ samples from the above densities, and compute RuLSIF for $\alpha = 0$, 0.5, and 0.95.

Figure 1 shows the true densities, true density-ratios, and their estimates by RuLSIF. As can be seen from the graphs, the profiles of the true α-relative density-ratios get smoother as α increases. In particular, in the datasets (b) and (d), the true density-ratios for $\alpha = 0$ diverge to infinity, while those for $\alpha = 0.5$ and 0.95 are bounded (by $1/\alpha$). Overall, as α gets large, the estimation quality of RuLSIF tends to be improved since the complexity of true density-ratio functions is reduced.

Note that, in the dataset (a) where $p(\boldsymbol{x}) = p'(\boldsymbol{x})$, the true density-ratio $r_\alpha(\boldsymbol{x})$ does not depend on α since $r_\alpha(\boldsymbol{x}) = 1$ for any α. However, the estimated density-ratios still depend on α through the matrix $\widehat{\boldsymbol{H}}$ (see Eq.(5)).

3 Theoretical Analysis

In this section, we analyze theoretical properties of the proposed PE divergence estimators. More specifically, we provide non-parametric analysis of the convergence rate in Section 3.1, and parametric analysis of the estimation variance in Section 3.2. Since our theoretical analysis is highly technical, we focus on explaining practical insights we can gain from the theoretical results here; we describe all the mathematical details of the non-parametric convergence-rate analysis in Appendix A and the parametric variance analysis in Appendix B.

For theoretical analysis, let us consider a rather abstract form of our relative density-ratio estimator described as

$$\operatorname*{argmin}_{g \in \mathcal{G}} \left[\frac{\alpha}{2n} \sum_{i=1}^{n} g(\boldsymbol{x}_i)^2 + \frac{(1-\alpha)}{2n'} \sum_{j=1}^{n'} g(\boldsymbol{x}'_j)^2 - \frac{1}{n} \sum_{i=1}^{n} g(\boldsymbol{x}_i) + \frac{\lambda}{2} R(g)^2 \right], \tag{9}$$

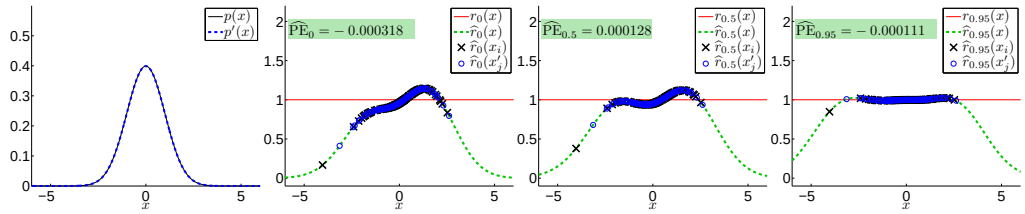
where $\mathcal{G}$ is some function space (i.e., a statistical model) and $R(\cdot)$ is some regularization functional.

3.1 Non-Parametric Convergence Analysis

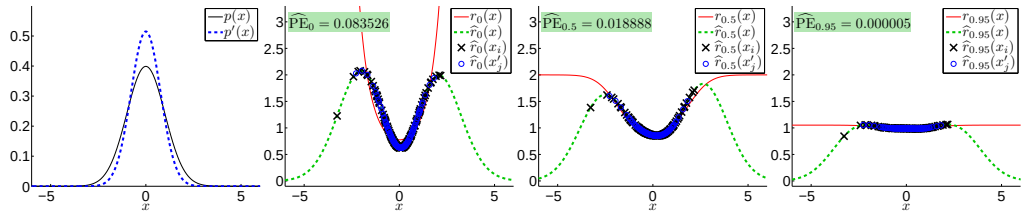
First, we elucidate the non-parametric convergence rate of the proposed PE estimators. Here, we practically regard the function space $\mathcal{G}$ as an infinite-dimensional *reproducing kernel Hilbert space* [RKHS; 18] such as the Gaussian kernel space, and $R(\cdot)$ as the associated RKHS norm.

3.1.1 Theoretical Results

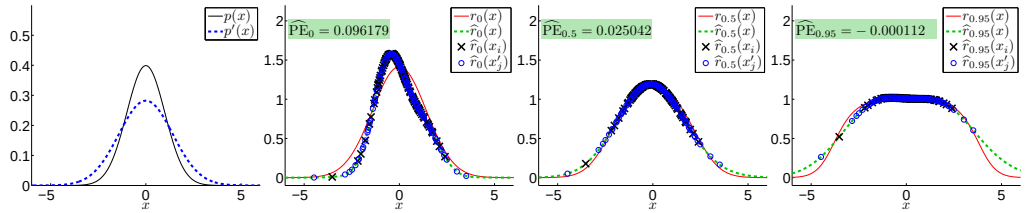
Let us represent the complexity of the function space $\mathcal{G}$ by γ ($0 < \gamma < 2$); the larger γ is, the more complex the function class $\mathcal{G}$ is (see Appendix A for its precise definition). We analyze the



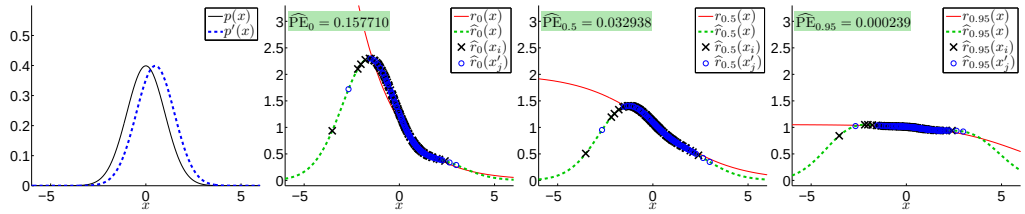
(a) $P' = N(0, 1)$: P and P' are the same.



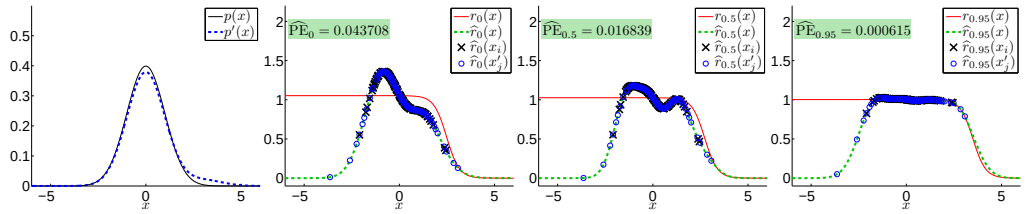
(b) $P' = N(0, 0.6)$: P' has smaller standard deviation than P.



(c) $P' = N(0, 2)$: P' has larger standard deviation than P.



(d) $P' = N(0.5, 1)$: P and P' have different means.



(e) $P' = 0.95N(0, 1) + 0.05N(3, 1)$: P' contains an additional component to P.

Figure 1: Illustrative examples of density-ratio approximation by RuLSIF. From left to right: true densities ($P = N(0, 1)$), true density-ratios, and their estimates for $\alpha = 0$, 0.5, and 0.95.

convergence rate of our PE divergence estimators as $\bar{n} := \min(n, n')$ tends to infinity for $\lambda = \lambda_{\bar{n}}$ under

$$\lambda_{\bar{n}} \to o(1) \text{ and } \lambda_{\bar{n}}^{-1} = o(\bar{n}^{2/(2+\gamma)}).$$

The first condition means that $\lambda_{\bar{n}}$ tends to zero, but the second condition means that its shrinking speed should not be too fast.

Under several technical assumptions detailed in Appendix A, we have the following asymptotic convergence results for the two PE divergence estimators $\widehat{\mathrm{PE}}_\alpha$ (7) and $\widetilde{\mathrm{PE}}_\alpha$ (8):

$$\widehat{\mathrm{PE}}_\alpha - \mathrm{PE}_\alpha = \mathcal{O}_p(\bar{n}^{-1/2} c \|r_\alpha\|_\infty + \lambda_{\bar{n}} \max(1, R(r_\alpha)^2)), \tag{10}$$

and

$$\widetilde{\mathrm{PE}}_\alpha - \mathrm{PE}_\alpha = \mathcal{O}_p\Big(\lambda_{\bar{n}}^{1/2} \|r_\alpha\|_\infty^{1/2} \max\{1, R(r_\alpha)\} + \lambda_{\bar{n}} \max\{1, \|r_\alpha\|_\infty^{(1-\gamma/2)/2}, R(r_\alpha)\|r_\alpha\|_\infty^{(1-\gamma/2)/2}, R(r_\alpha)\}\Big), \tag{11}$$

where $\mathcal{O}_p$ denotes the asymptotic order in probability,

$$c := (1+\alpha)\sqrt{\mathbb{V}_{p(\boldsymbol{x})}[r_\alpha(\boldsymbol{x})]} + (1-\alpha)\sqrt{\mathbb{V}_{p'(\boldsymbol{x})}[r_\alpha(\boldsymbol{x})]}, \tag{12}$$

and $\mathbb{V}_{p(\boldsymbol{x})}[f(\boldsymbol{x})]$ denotes the variance of $f(\boldsymbol{x})$ under $p(\boldsymbol{x})$:

$$\mathbb{V}_{p(\boldsymbol{x})}[f(\boldsymbol{x})] = \int \left(f(\boldsymbol{x}) - \int f(\boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\right)^2 p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

3.1.2 Interpretation

In both Eq.(10) and Eq.(11), the coefficients of the leading terms (i.e., the first terms) of the asymptotic convergence rates become smaller as $\|r_\alpha\|_\infty$ gets smaller. Since

$$\|r_\alpha\|_\infty = \left\| \big(\alpha + (1-\alpha)/r(\boldsymbol{x})\big)^{-1} \right\|_\infty < \tfrac{1}{\alpha} \quad \text{for } \alpha > 0,$$

larger α would be more preferable in terms of the asymptotic approximation error. Note that when $\alpha = 0$, $\|r_\alpha\|_\infty$ can tend to infinity even under a simple setting that the ratio of two Gaussian functions is considered [16, see also the numerical examples in Section 2.4 of this paper]. Thus, our proposed approach of estimating the α-relative PE divergence (with $\alpha > 0$) would be more advantageous than the naive approach of estimating the plain PE divergence (which corresponds to $\alpha = 0$) in terms of the non-parametric convergence rate.

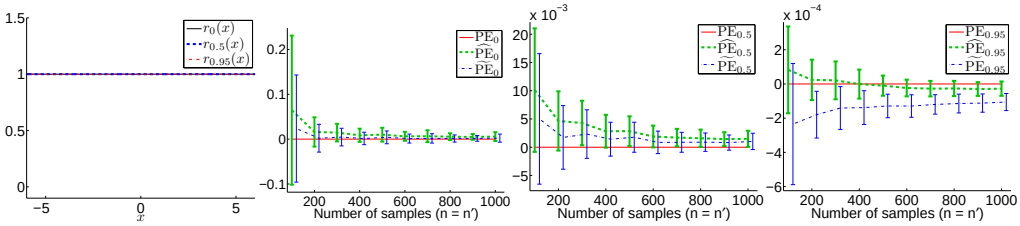
The above results also show that $\widehat{\mathrm{PE}}_\alpha$ and $\widetilde{\mathrm{PE}}_\alpha$ have different asymptotic convergence rates. The leading term in Eq.(10) is of order $\bar{n}^{-1/2}$, while the leading term in Eq.(11) is of order $\lambda_{\bar{n}}^{1/2}$, which is slightly slower (depending on the complexity γ) than $\bar{n}^{-1/2}$. Thus, $\widehat{\mathrm{PE}}_\alpha$ would be more accurate than $\widetilde{\mathrm{PE}}_\alpha$ in large sample cases. Furthermore, when $p(\boldsymbol{x}) = p'(\boldsymbol{x})$, $\mathbb{V}_{p(\boldsymbol{x})}[r_\alpha(\boldsymbol{x})] = 0$ holds and thus $c = 0$ holds (see Eq.(12)). Then the leading term in Eq.(10) vanishes and therefore $\widehat{\mathrm{PE}}_\alpha$ has the even faster convergence rate of order $\lambda_{\bar{n}}$, which is slightly slower (depending on the complexity γ) than $\bar{n}^{-1}$. Similarly, if α is close to 1, $r_\alpha(\boldsymbol{x}) \approx 1$ and thus $c \approx 0$ holds.

When $\bar{n}$ is not large enough to be able to neglect the terms of $o(\bar{n}^{-1/2})$, the terms of $O(\lambda_{\bar{n}})$ matter. If $\|r_\alpha\|_\infty$ and $R(r_\alpha)$ are large (this can happen, e.g., when α is close to 0), the coefficient of the $O(\lambda_{\bar{n}})$-term in Eq.(10) can be larger than that in Eq.(11). Then $\widetilde{\mathrm{PE}}_\alpha$ would be more favorable than $\widehat{\mathrm{PE}}_\alpha$ in terms of the approximation accuracy.

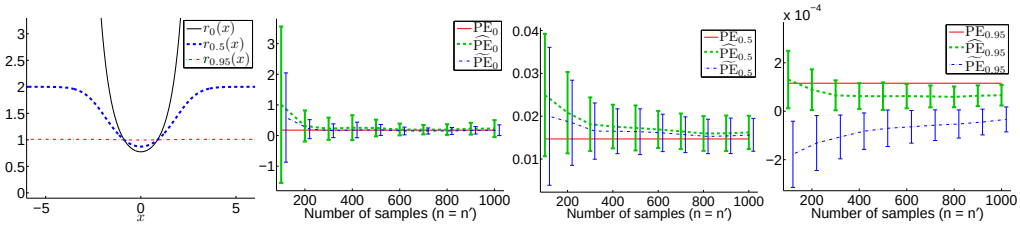
3.1.3 Numerical Illustration

Let us numerically investigate the above interpretation using the same artificial dataset as Section 2.4.

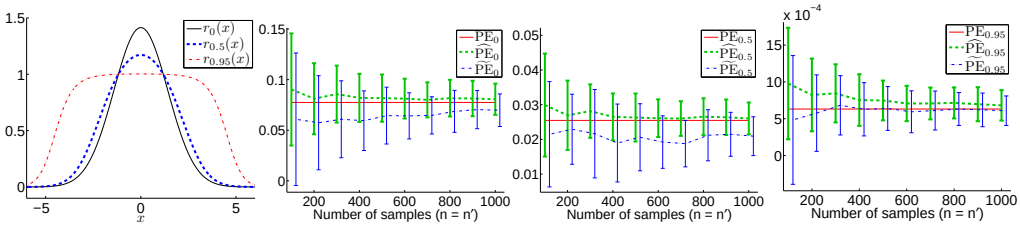
Figure 2 shows the mean and standard deviation of $\widehat{\mathrm{PE}}_\alpha$ and $\widetilde{\mathrm{PE}}_\alpha$ over 100 runs for $\alpha = 0$, 0.5, and 0.95, as functions of n ($= n'$ in this experiment). The true PE_α (which was numerically computed)



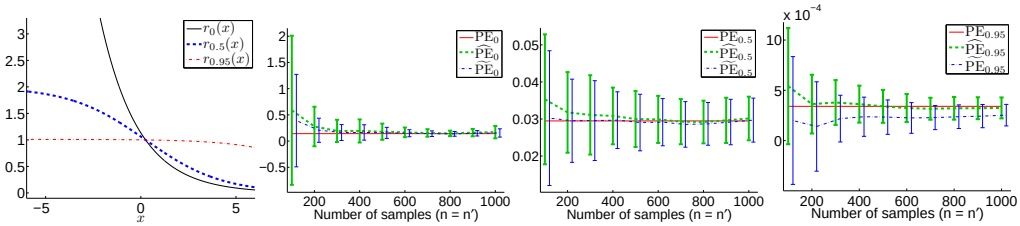
(a) $P' = N(0, 1)$: P and P' are the same.



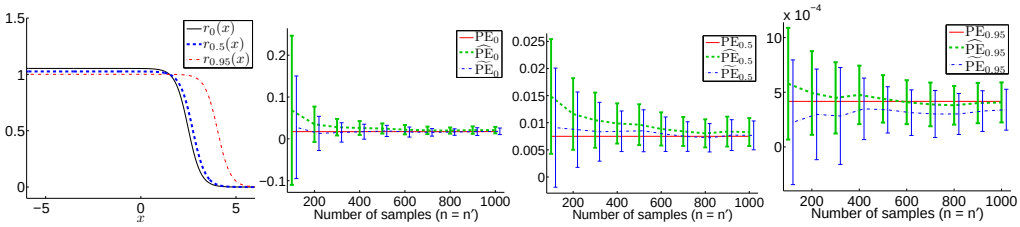
(b) $P' = N(0, 0.6)$: P' has smaller standard deviation than P.



(c) $P' = N(0, 2)$: P' has larger standard deviation than P.



(d) $P' = N(0.5, 1)$: P and P' have different means.



(e) $P' = 0.95N(0, 1) + 0.05N(3, 1)$: P' contains an additional component to P.

Figure 2: Illustrative examples of divergence estimation by RuLSIF. From left to right: true density-ratios for $\alpha = 0$, 0.5, and 0.95 ($P = N(0, 1)$), and estimation error of PE divergence for $\alpha = 0$, 0.5, and 0.95.

is also plotted in the graphs. The graphs show that both the estimators $\widehat{\mathrm{PE}}_\alpha$ and $\widetilde{\mathrm{PE}}_\alpha$ approach the true PE_α as the number of samples increases, and the approximation error tends to be smaller if α is larger.

When α is large, $\widehat{\mathrm{PE}}_\alpha$ tends to perform slightly better than $\widetilde{\mathrm{PE}}_\alpha$. On the other hand, when α is small and the number of samples is small, $\widetilde{\mathrm{PE}}_\alpha$ slightly compares favorably with $\widehat{\mathrm{PE}}_\alpha$. Overall, these numerical results well agree with our theory.

3.2 Parametric Variance Analysis

Next, we analyze the asymptotic variance of the PE divergence estimator $\widehat{\mathrm{PE}}_\alpha$ (7) under a parametric setup.

3.2.1 Theoretical Results

As the function space $\mathcal{G}$ in Eq.(9), we consider the following parametric model:

$$\mathcal{G} = \{g(\boldsymbol{x};\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^b\},$$

where b is a finite number. Here we assume that the above parametric model is *correctly specified*, i.e., it includes the true relative density-ratio function $r_\alpha(\boldsymbol{x})$: there exists $\boldsymbol{\theta}^*$ such that

$$g(\boldsymbol{x};\boldsymbol{\theta}^*) = r_\alpha(\boldsymbol{x}).$$

Here, we use RuLSIF without regularization, i.e., $\lambda = 0$ in Eq.(9).

Let us denote the variance of $\widehat{\mathrm{PE}}_\alpha$ (7) by $\mathbb{V}[\widehat{\mathrm{PE}}_\alpha]$, where randomness comes from the draw of samples $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$. Then, under a standard regularity condition for the asymptotic normality [see Section 3 of 19], $\mathbb{V}[\widehat{\mathrm{PE}}_\alpha]$ can be expressed and upper-bounded as

$$\mathbb{V}[\widehat{\mathrm{PE}}_\alpha] = \frac{1}{n}\mathbb{V}_{p(\boldsymbol{x})}\left[r_\alpha - \frac{\alpha r_\alpha(\boldsymbol{x})^2}{2}\right] + \frac{1}{n'}\mathbb{V}_{p'(\boldsymbol{x})}\left[\frac{(1-\alpha)r_\alpha(\boldsymbol{x})^2}{2}\right] + o\left(\frac{1}{n}, \frac{1}{n'}\right) \tag{13}$$

$$\leq \frac{\|r_\alpha\|_\infty^2}{n} + \frac{\alpha^2\|r_\alpha\|_\infty^4}{4n} + \frac{(1-\alpha)^2\|r_\alpha\|_\infty^4}{4n'} + o\left(\frac{1}{n}, \frac{1}{n'}\right). \tag{14}$$

Let us denote the variance of $\widetilde{\mathrm{PE}}_\alpha$ by $\mathbb{V}[\widetilde{\mathrm{PE}}_\alpha]$. Then, under a standard regularity condition for the asymptotic normality [see Section 3 of 19], the variance of $\widetilde{\mathrm{PE}}_\alpha$ is asymptotically expressed as

$$\mathbb{V}[\widetilde{\mathrm{PE}}_\alpha] = \frac{1}{n}\mathbb{V}_{p(\boldsymbol{x})}\left[\frac{r_\alpha + (1-\alpha r_\alpha)\mathbb{E}_{p(\boldsymbol{x})}[\nabla g]^\top \boldsymbol{U}_\alpha^{-1}\nabla g}{2}\right]$$
$$+ \frac{1}{n'}\mathbb{V}_{p'(\boldsymbol{x})}\left[\frac{(1-\alpha)r_\alpha\mathbb{E}_{p(\boldsymbol{x})}[\nabla g]^\top \boldsymbol{U}_\alpha^{-1}\nabla g}{2}\right] + o\left(\frac{1}{n}, \frac{1}{n'}\right), \tag{15}$$

where ∇g is the gradient vector of g with respect to $\boldsymbol{\theta}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, i.e.,

$$(\nabla g(\boldsymbol{x};\boldsymbol{\theta}^*))_j = \frac{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)}{\partial \theta_j}.$$

The matrix $\boldsymbol{U}_\alpha$ is defined by

$$\boldsymbol{U}_\alpha = \alpha\mathbb{E}_{p(\boldsymbol{x})}[\nabla g \nabla g^\top] + (1-\alpha)\mathbb{E}_{p'(\boldsymbol{x})}[\nabla g \nabla g^\top].$$

3.2.2 Interpretation

Eq.(13) shows that, up to $O\left(\frac{1}{n}, \frac{1}{n'}\right)$, the variance of $\widehat{\mathrm{PE}}_\alpha$ depends only on the true relative density-ratio $r_\alpha(\boldsymbol{x})$, not on the estimator of $r_\alpha(\boldsymbol{x})$. This means that the model complexity does not affect the asymptotic variance. Therefore, *overfitting* would hardly occur in the estimation of the relative PE divergence even when complex models are used. We note that the above superior property is applicable only to relative PE divergence estimation, not to relative density-ratio estimation. This implies

that overfitting occurs in relative density-ratio estimation, but the approximation error cancels out in relative PE divergence estimation.

On the other hand, Eq.(15) shows that the variance of $\widetilde{\mathrm{PE}}_\alpha$ is affected by the model $\mathcal{G}$, since the factor $\mathbb{E}_{p(\boldsymbol{x})}[\nabla g]^\top \boldsymbol{U}_\alpha^{-1} \nabla g$ depends on the model complexity in general. When the equality

$$\mathbb{E}_{p(\boldsymbol{x})}[\nabla g]^\top \boldsymbol{U}_\alpha^{-1} \nabla g(\boldsymbol{x}; \boldsymbol{\theta}^*) = r_\alpha(\boldsymbol{x})$$

holds, the variances of $\widetilde{\mathrm{PE}}_\alpha$ and $\widehat{\mathrm{PE}}_\alpha$ are asymptotically the same. However, in general, the use of $\widehat{\mathrm{PE}}_\alpha$ would be more recommended.

Eq.(14) shows that the variance $\mathbb{V}[\widehat{\mathrm{PE}}_\alpha]$ can be upper-bounded by the quantity depending on $\|r_\alpha\|_\infty$, which is monotonically lowered if $\|r_\alpha\|_\infty$ is reduced. Since $\|r_\alpha\|_\infty$ monotonically decreases as α increases, our proposed approach of estimating the α-relative PE divergence (with $\alpha > 0$) would be more advantageous than the naive approach of estimating the plain PE divergence (which corresponds to $\alpha = 0$) in terms of the parametric asymptotic variance.

3.2.3 Numerical Illustration

Here, we show some numerical results for illustrating the above theoretical results using the one-dimensional datasets (b) and (c) in Section 2.4. Let us define the parametric model as

$$\mathcal{G}_k = \left\{ g(x; \boldsymbol{\theta}) = \frac{r(x; \boldsymbol{\theta})}{\alpha r(x; \boldsymbol{\theta}) + 1 - \alpha} \;\middle|\; r(x; \boldsymbol{\theta}) = \exp\left(\sum_{\ell=0}^{k} \theta_\ell x^\ell \right), \boldsymbol{\theta} \in \mathbb{R}^{k+1} \right\}. \tag{16}$$

The dimension of the model $\mathcal{G}_k$ is equal to $k+1$. The α-relative density-ratio $r_\alpha(x)$ can be expressed using the ordinary density-ratio $r(x) = p(x)/p'(x)$ as

$$r_\alpha(x) = \frac{r(x)}{\alpha r(x) + 1 - \alpha}.$$

Thus, when $k > 1$, the above model $\mathcal{G}_k$ includes the true relative density-ratio $r_\alpha(x)$ of the datasets (b) and (c). We test RuLSIF with $\alpha = 0.2$ and 0.8 for the model (16) with degree $k = 1, 2, \ldots, 8$. The parameter $\boldsymbol{\theta}$ is learned so that Eq.(9) is minimized by a quasi-Newton method.

The standard deviations of $\widehat{\mathrm{PE}}_\alpha$ and $\widetilde{\mathrm{PE}}_\alpha$ for the datasets (b) and (c) are depicted in Figure 3 and Figure 4, respectively. The graphs show that the degree of models does not significantly affect the standard deviation of $\widehat{\mathrm{PE}}_\alpha$ (i.e., no overfitting), as long as the model includes the true relative density-ratio (i.e., $k > 1$). On the other hand, bigger models tend to produce larger standard deviations in $\widetilde{\mathrm{PE}}_\alpha$. Thus, the standard deviation of $\widetilde{\mathrm{PE}}_\alpha$ more strongly depends on the model complexity.

4 Experiments

In this section, we experimentally evaluate the performance of the proposed method in two-sample homogeneity test, outlier detection, and transfer learning tasks.

4.1 Two-Sample Homogeneity Test

First, we apply the proposed divergence estimator to two-sample homogeneity test.

4.1.1 Divergence-Based Two-Sample Homogeneity Test

Given two sets of samples $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P$ and $\mathcal{X}' = \{\boldsymbol{x}'_j\}_{j=1}^{n'} \overset{\text{i.i.d.}}{\sim} P'$, the goal of the two-sample homogeneity test is to test the *null hypothesis* that the probability distributions P and P' are the same against its complementary alternative (i.e., the distributions are different).

By using an estimator $\widehat{\mathrm{Div}}$ of some divergence between the two distributions P and P', homogeneity of two distributions can be tested based on the *permutation test* procedure [20] as follows:

- Obtain a divergence estimate $\widehat{\mathrm{Div}}$ using the original datasets $\mathcal{X}$ and $\mathcal{X}'$.

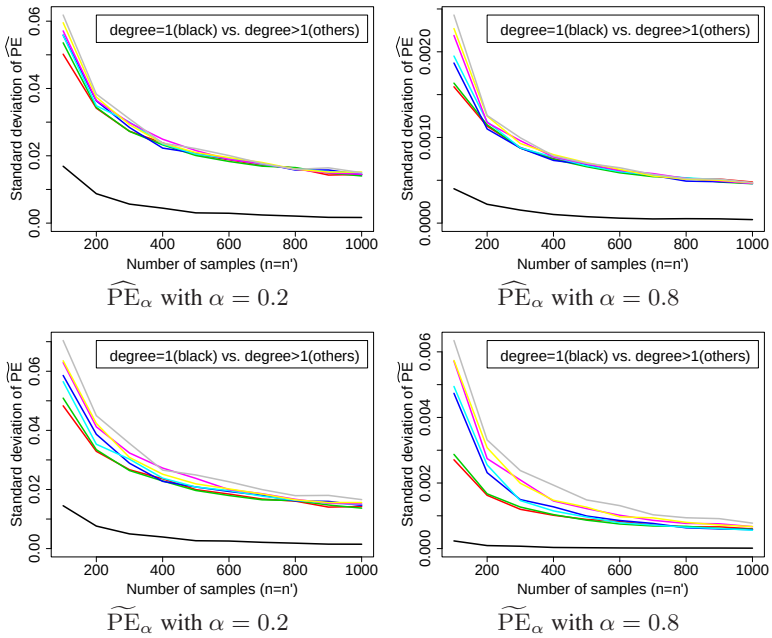


Figure 3: Standard deviations of PE estimators for dataset (b) (i.e., $P = N(0,1)$ and $P' = N(0, 0.6)$) as functions of the sample size $n = n'$.

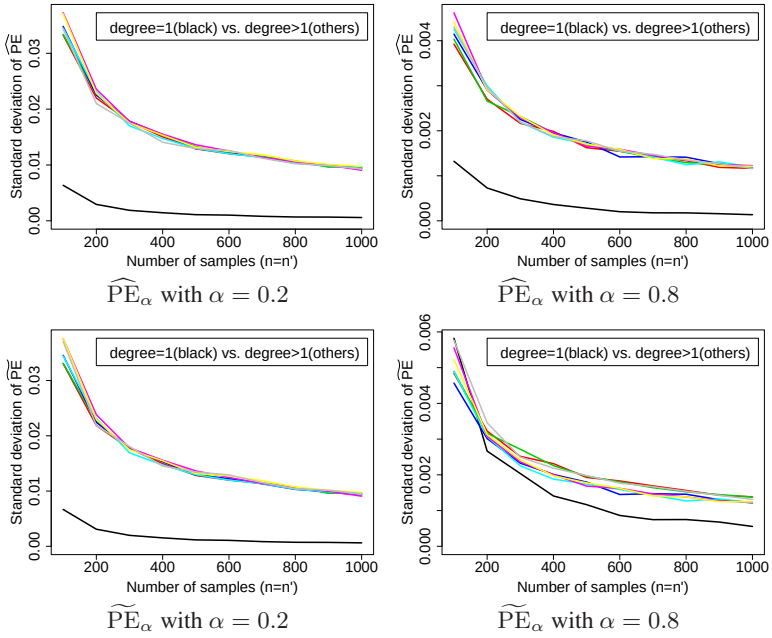


Figure 4: Standard deviations of PE estimators for dataset (c) (i.e., $P = N(0,1)$ and $P' = N(0,2)$) as functions of the sample size $n = n'$.

- 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 $\widetilde{\mathcal{X}}'$.
- Obtain a divergence estimate $\widetilde{\mathrm{Div}}$ using the randomly shuffled datasets $\widetilde{\mathcal{X}}$ and $\widetilde{\mathcal{X}}'$ (note that, since $\widetilde{\mathcal{X}}$ and $\widetilde{\mathcal{X}}'$ can be regarded as being drawn from the same distribution, $\widetilde{\mathrm{Div}}$ tends to be close to zero).
- Repeat this random shuffling procedure many times, and construct the empirical distribution of $\widetilde{\mathrm{Div}}$ under the null hypothesis that the two distributions are the same.
- Approximate the p-value by evaluating the relative ranking of the original $\widehat{\mathrm{Div}}$ in the distribution of $\widetilde{\mathrm{Div}}$.

When an asymmetric divergence such as the KL divergence [7] or the PE divergence [11] is adopted for two-sample homogeneity test, the test results depend on the choice of *directions*: a divergence from P to P' or from P' to P. **(author?)** [4] proposed to choose the direction that gives a smaller p-value—it was experimentally shown that, when the uLSIF-based PE divergence estimator is used for the two-sample homogeneity test (which is called the *least-squares two-sample homogeneity test*; LSTT), the heuristic of choosing the direction with a smaller p-value contributes to reducing 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), while the increase of 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) is kept moderate.

Below, we refer to LSTT with $p(\boldsymbol{x})/p'(\boldsymbol{x})$ as the *plain LSTT*, LSTT with $p'(\boldsymbol{x})/p(\boldsymbol{x})$ as the *reciprocal LSTT*, and LSTT with heuristically choosing the one with a smaller p-value as the *adaptive LSTT*.

4.1.2 Artificial Datasets

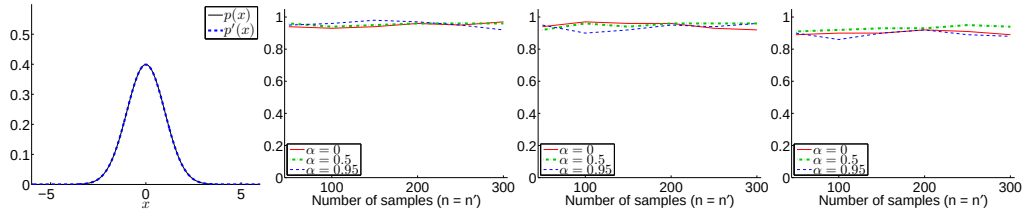
We illustrate how the proposed method behaves in two-sample homogeneity test scenarios using the artificial datasets (a)–(d) described in Section 2.4. We test the plain LSTT, reciprocal LSTT, and adaptive LSTT for $\alpha = 0$, 0.5, and 0.95, with significance level 5%.

The experimental results are shown in Figure 5. For the dataset (a) where $P = P'$ (i.e., the null hypothesis is correct), the plain LSTT and reciprocal LSTT correctly accept the null hypothesis with probability approximately 95%. This means that the type-I error is properly controlled in these methods. On the other hand, the adaptive LSTT tends to give slightly lower acceptance rates than 95% for this toy dataset, but the adaptive LSTT with $\alpha = 0.5$ still works reasonably well. This implies that the heuristic of choosing the method with a smaller p-value does not have critical influence on the type-I error.

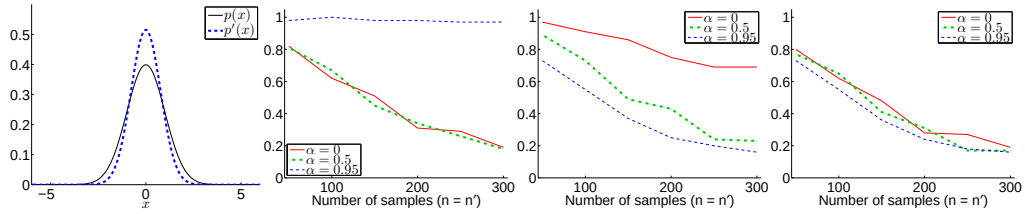
In the datasets (b), (c), and (d), P is different from P' (i.e., the null hypothesis is not correct), and thus we want to reduce the acceptance rate of the incorrect null-hypothesis as much as possible. In the plain setup for the dataset (b) and the reciprocal setup for the dataset (c), the true density-ratio functions with $\alpha = 0$ diverge to infinity, and thus larger α makes the density-ratio approximation more reliable. However, $\alpha = 0.95$ does not work well because it produces an overly-smoothed density-ratio function and thus it is hard to be distinguished from the completely constant density-ratio function (which corresponds to $P = P'$). On the other hand, in the reciprocal setup for the dataset (b) and the plain setup for the dataset (c), small α performs poorly since density-ratio functions with large α can be more accurately approximated than those with small α (see Figure 1). In the adaptive setup, large α tends to perform slightly better than small α for the datasets (b) and (c).

In the dataset (d), the true density-ratio function with $\alpha = 0$ diverges to infinity for both the plain and reciprocal setups. In this case, middle α performs the best, which well balances the trade-off between high distinguishability from the completely constant density-ratio function (which corresponds to $P = P'$) and easy approximability. The same tendency that middle α works well can also be mildly observed in the adaptive LSTT for the dataset (d).

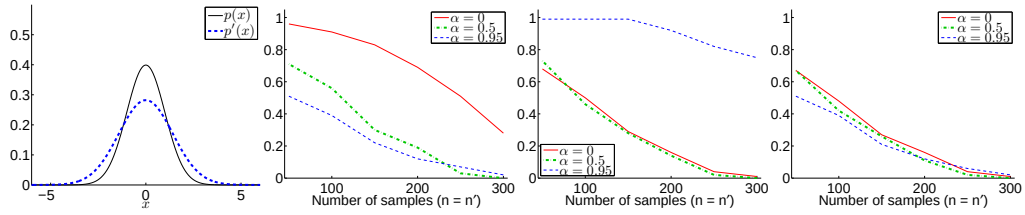
Overall, if the plain LSTT (or the reciprocal LSTT) is used, small α (or large α) sometimes works excellently. However, it performs poorly in other cases and thus the performance is unstable de-



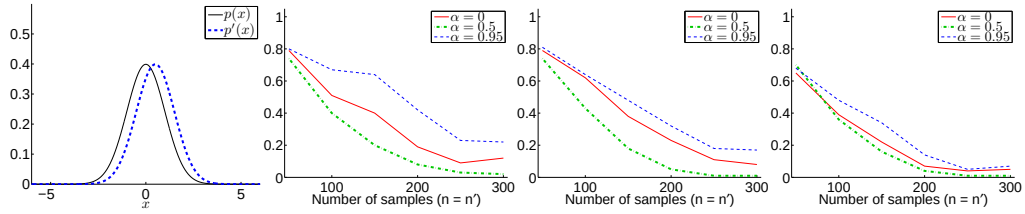
(a) $P' = N(0, 1)$: P and P' are the same.



(b) $P' = N(0, 0.6)$: P' has smaller standard deviation than P.



(c) $P' = N(0, 2)$: P' has larger standard deviation than P.



(d) $P' = N(0.5, 1)$: P and P' have different means.

Figure 5: Illustrative examples of two-sample homogeneity test based on relative divergence estimation. From left to right: true densities ($P = N(0, 1)$), the acceptance rate of the null hypothesis under the significance level 5% by plain LSTT, reciprocal LSTT, and adaptive LSTT.

pending on the true distributions. The plain LSTT (or the reciprocal LSTT) with middle α tends to perform reasonably well for all datasets. On the other hand, the adaptive LSTT was shown to nicely overcome the above instability problem when α is small or large. However, when α is set to be a middle value, the plain LSTT and the reciprocal LSTT both give similar results and thus the adaptive LSTT provides only a small amount of improvement.

Our empirical finding is that, if we have prior knowledge that one distribution has a wider support than the other distribution, assigning the distribution with a wider support to P' and setting α to be a large value seem to work well. If there is no knowledge on the true distributions or two distributions have less overlapped supports, using middle α in the adaptive setup seems to be a reasonable choice.

We will systematically investigate this issue using more complex datasets below.

4.1.3 Benchmark Datasets

Here, we apply the proposed two-sample homogeneity test to the binary classification datasets taken from the *IDA repository* [21].

We test the adaptive LSTT with the RuLSIF-based PE divergence estimator for $\alpha = 0$, 0.5, and 0.95; we also test the *maximum mean discrepancy* [MMD; 22], which is a kernel-based two-sample homogeneity test method. The performance of MMD depends on the choice of the Gaussian kernel width. Here, we adopt a version proposed by [23], which automatically optimizes the Gaussian kernel width. The p-values of MMD are computed in the same way as LSTT based on the permutation test procedure.

First, we investigate the rate of accepting the null hypothesis when the null hypothesis is correct (i.e., the two distributions are the same). We split all the positive training samples into two sets and perform two-sample homogeneity test for the two sets of samples. The experimental results are summarized in Table 1, showing that the adaptive LSTT with $\alpha = 0.5$ compares favorably with those with $\alpha = 0$ and 1 and MMD in terms of the type-I error.

Next, we consider the situation where the null hypothesis is not correct (i.e., the two distributions are different). The numerator samples are generated in the same way as above, but a half of denominator samples are replaced with negative training samples. Thus, while the numerator sample set contains only positive training samples, the denominator sample set includes both positive and negative training samples. The experimental results are summarized in Table 2, showing that the adaptive LSTT with $\alpha = 0.5$ again compares favorably with those with $\alpha = 0$ and 1. Furthermore, LSTT with $\alpha = 0.5$ tends to outperform MMD in terms of the type-II error.

Overall, LSTT with $\alpha = 0.5$ is shown to be a useful method for two-sample homogeneity test.

4.2 Inlier-Based Outlier Detection

Next, we apply the proposed method to outlier detection.

4.2.1 Density-Ratio Approach to Inlier-Based Outlier Detection

Let us consider an outlier detection problem of finding irregular samples in a dataset (called an "evaluation dataset") based on another dataset (called a "model dataset") that only contains regular samples. Defining the density ratio over the two sets of samples, we can see that the density-ratio values for regular samples are close to one, while those for outliers tend to be significantly deviated from one. Thus, density-ratio values could be used as an index of the degree of outlyingness [1, 2].

Since the evaluation dataset usually has a wider support than the model dataset, we regard the evaluation dataset as samples corresponding to the denominator density $p'(\boldsymbol{x})$, and the model dataset as samples corresponding to the numerator density $p(\boldsymbol{x})$. Then, outliers tend to have smaller density-ratio values (i.e., close to zero). As such, density-ratio approximators can be used for outlier detection.

When evaluating the performance of outlier detection methods, it is important to take into account both the *detection rate* (i.e., the amount of true outliers an outlier detection algorithm can find) and the *detection accuracy* (i.e., the amount of true inliers an outlier detection algorithm misjudges as

Table 1: Experimental results of two-sample homogeneity test for the IDA datasets. The mean (and standard deviation in the bracket) rate of accepting the null hypothesis (i.e., $P = P'$) under the significance level 5% is reported. The two sets of samples are both taken from the positive training set (i.e., the null hypothesis is correct). Methods having the mean acceptance rate 0.95 according to the one-sample t-test at the significance level 5% are specified by bold face.

Datasets	d	$n = n'$	MMD	LSTT (α = 0.0)	LSTT (α = 0.5)	LSTT (α = 0.95)
banana	2	100	**0.96(0.20)**	**0.93(0.26)**	**0.92(0.27)**	**0.92(0.27)**
thyroid	5	19	**0.96(0.20)**	**0.95(0.22)**	**0.95(0.22)**	0.88(0.33)
titanic	5	21	**0.94(0.24)**	0.86(0.35)	**0.92(0.27)**	**0.89(0.31)**
diabetes	8	85	**0.96(0.20)**	0.87(0.34)	**0.91(0.29)**	0.82(0.39)
breast-cancer	9	29	0.98(0.14)	**0.91(0.29)**	**0.94(0.24)**	**0.92(0.27)**
flare-solar	9	100	**0.93(0.26)**	**0.91(0.29)**	**0.95(0.22)**	**0.93(0.26)**
heart	13	38	1.00(0.00)	0.85(0.36)	**0.91(0.29)**	**0.93(0.26)**
german	20	100	0.99(0.10)	**0.91(0.29)**	**0.92(0.27)**	**0.89(0.31)**
ringnorm	20	100	**0.97(0.17)**	**0.93(0.26)**	**0.91(0.29)**	0.85(0.36)
waveform	21	66	0.98(0.14)	**0.92(0.27)**	**0.93(0.26)**	0.88(0.33)

Table 2: Experimental results of two-sample homogeneity test for the IDA datasets. The mean (and standard deviation in the bracket) rate of accepting the null hypothesis (i.e., $P = P'$) under the significance level 5% is reported. The set of samples corresponding to the numerator of the density ratio is taken from the positive training set and the set of samples corresponding to the denominator of the density ratio is taken from the positive training set and the negative training set (i.e., the null hypothesis is not correct). The best method having the lowest mean acceptance rate and comparable methods according to the *two-sample t-test* at the significance level 5% are specified by bold face.

Datasets	d	$n = n'$	MMD	LSTT (α = 0.0)	LSTT (α = 0.5)	LSTT (α = 0.95)
banana	2	100	0.52(0.50)	**0.10(0.30)**	**0.02(0.14)**	**0.17(0.38)**
thyroid	5	19	**0.52(0.50)**	0.81(0.39)	**0.65(0.48)**	0.80(0.40)
titanic	5	21	**0.87(0.34)**	**0.86(0.35)**	**0.87(0.34)**	**0.88(0.33)**
diabetes	8	85	**0.31(0.46)**	**0.42(0.50)**	0.47(0.50)	0.57(0.50)
breast-cancer	9	29	0.87(0.34)	**0.75(0.44)**	0.80(0.40)	0.79(0.41)
flare-solar	9	100	**0.51(0.50)**	0.81(0.39)	**0.55(0.50)**	**0.66(0.48)**
heart	13	38	0.53(0.50)	**0.28(0.45)**	**0.40(0.49)**	0.62(0.49)
german	20	100	0.56(0.50)	0.55(0.50)	**0.44(0.50)**	0.68(0.47)
ringnorm	20	100	**0.00(0.00)**	**0.00(0.00)**	**0.00(0.00)**	**0.02(0.14)**
waveform	21	66	**0.00(0.00)**	**0.00(0.00)**	**0.02(0.14)**	**0.00(0.00)**

Table 3: Mean AUC score (and the standard deviation in the bracket) over 1000 trials for the artificial outlier-detection dataset. The best method in terms of the mean AUC score and comparable methods according to the *two-sample t-test* at the significance level 5% are specified by bold face.

Input dimensionality d	RuLSIF ($\alpha = 0$)	RuLSIF ($\alpha = 0.5$)	RuLSIF ($\alpha = 0.95$)
1	**.933(.089)**	**.926(.100)**	.896 (.124)
5	**.882(.099)**	**.891(.091)**	**.894 (.086)**
10	.842(.107)	**.850(.103)**	**.859 (.092)**

outliers). Since there is a trade-off between the detection rate and the detection accuracy, we adopt the *area under the ROC curve* (AUC) as our error metric [24].

4.2.2 Artificial Datasets

First, we illustrate how the proposed method behaves in outlier detection scenarios using artificial datasets.

Let

$$P = N(0, \boldsymbol{I}_d),$$
$$P' = 0.95N(0, \boldsymbol{I}_d) + 0.05N(3d^{-1/2}\mathbf{1}_d, \boldsymbol{I}_d),$$

where d is the dimensionality of $\boldsymbol{x}$ and $\mathbf{1}_d$ is the d-dimensional vector with all one. Note that this setup is the same as the dataset (e) described in Section 2.4 when $d = 1$. Here, the samples drawn from $N(0, \boldsymbol{I}_d)$ are regarded as inliers, while the samples drawn from $N(d^{-1/2}\mathbf{1}_d, \boldsymbol{I}_d)$ are regarded as outliers. We use $n = n' = 100$ samples.

Table 3 describes the AUC values for input dimensionality $d = 1, 5$, and 10 for RuLSIF with $\alpha = 0$, 0.5, and 0.95. This shows that, as the input dimensionality d increases, the AUC values overall get smaller. Thus, outlier detection becomes more challenging in high-dimensional cases.

The result also shows that RuLSIF with small α tends to work well when the input dimensionality is low, and RuLSIF with large α works better as the input dimensionality increases. This tendency can be interpreted as follows: If α is small, the density-ratio function tends to have sharp 'hollow' for outlier points (see the leftmost graph in Figure 2(e)). Thus, as long as the true density-ratio function can be accurately estimated, small α would be preferable in outlier detection. When the data dimensionality is low, density-ratio approximation is rather easy and thus small α tends to perform well. However, as the data dimensionality increases, density-ratio approximation gets harder, and thus large α which produces a smoother density-ratio function is more favorable since such a smoother function can be more easily approximated than a 'bumpy' one produced by small α.

4.2.3 Real-World Datasets

Next, we evaluate the proposed outlier detection method using various real-world datasets:

IDA repository: The *IDA repository* [21] contains various binary classification tasks. Each dataset consists of positive/negative and training/test samples. We use positive training samples as inliers in the "model" set. In the "evaluation" set, we use at most 100 positive test samples as inliers and the first 5% of negative test samples as outliers. Thus, the positive samples are treated as inliers and the negative samples are treated as outliers.

Speech dataset: An in-house speech dataset, which contains short utterance samples recorded from 2 male subjects speaking in French with sampling rate 44.1kHz. From each utterance sample, we extracted a 50-dimensional *line spectral frequencies* vector [25]. We randomly take 200 samples from one class and assign them to the model dataset. Then we randomly take 200 samples from the same class and 10 samples from the other class.

20 Newsgroup dataset: The *20-Newsgroups* dataset[1] contains 20000 newsgroup documents, which contains the following 4 top-level categories: 'comp', 'rec', 'sci', and 'talk'. Each docu-

[1] http://people.csail.mit.edu/jrennie/20Newsgroups/

ment is expressed by a 100-dimensional bag-of-words vector of term-frequencies. We randomly take 200 samples from the 'comp' class and assign them to the model dataset. Then we randomly take 200 samples from the same class and 10 samples from one of the other classes for the evaluation dataset.

The USPS hand-written digit dataset: The *USPS* hand-written digit dataset[2] contains 9298 digit images. Each image consists of 256 (= 16×16) pixels and each pixel takes an integer value between 0 and 255 as the intensity level. We regard samples in one class as inliers and samples in other classes as outliers. We randomly take 200 samples from the inlier class and assign them to the model dataset. Then we randomly take 200 samples from the same inlier class and 10 samples from one of the other classes for the evaluation dataset.

We compare the AUC scores of RuLSIF with $\alpha = 0$, 0.5, and 0.95, and *one-class support vector machine (OSVM)* with the Gaussian kernel [26]. We used the *LIBSVM* implementation of OSVM [27]. The Gaussian width is set to the median distance between samples, which has been shown to be a useful heuristic [26]. Since there is no systematic method to determine the tuning parameter ν in OSVM, we report the results for $\nu = 0.05$ and 0.1.

The mean and standard deviation of the AUC scores over 100 runs with random sample choice are summarized in Table 4, showing that RuLSIF overall compares favorably with OSVM. Among the RuLSIF methods, small α tends to perform well for low-dimensional datasets, and large α tends to work well for high-dimensional datasets. This tendency well agrees with that for the artificial datasets (see Section 4.2.2).

4.3 Transfer Learning

Finally, we apply the proposed method to outlier detection.

4.3.1 Transductive Transfer Learning by Importance Sampling

Let us consider a problem of *semi-supervised learning* [28] from labeled training samples $\{(\boldsymbol{x}_j^{\mathrm{tr}}, y_j^{\mathrm{tr}})\}_{j=1}^{n_{\mathrm{tr}}}$ and unlabeled test samples $\{\boldsymbol{x}_i^{\mathrm{te}}\}_{i=1}^{n_{\mathrm{te}}}$. The goal is to predict a test output value y^{te} for a test input point $\boldsymbol{x}^{\mathrm{te}}$. Here, we consider the setup where the labeled training samples $\{(\boldsymbol{x}_j^{\mathrm{tr}}, y_j^{\mathrm{tr}})\}_{j=1}^{n_{\mathrm{tr}}}$ are drawn i.i.d. from $p(y|\boldsymbol{x})p_{\mathrm{tr}}(\boldsymbol{x})$, while the unlabeled test samples $\{\boldsymbol{x}_i^{\mathrm{te}}\}_{i=1}^{n_{\mathrm{te}}}$ are drawn i.i.d. from $p_{\mathrm{te}}(\boldsymbol{x})$, which is generally different from $p_{\mathrm{tr}}(\boldsymbol{x})$; the (unknown) test sample $(\boldsymbol{x}^{\mathrm{te}}, y^{\mathrm{te}})$ follows $p(y|\boldsymbol{x})p_{\mathrm{te}}(\boldsymbol{x})$. This setup means that the conditional probability $p(y|\boldsymbol{x})$ is common to training and test samples, but the marginal densities $p_{\mathrm{tr}}(\boldsymbol{x})$ and $p_{\mathrm{te}}(\boldsymbol{x})$ are generally different for training and test input points. Such a problem is called *transductive transfer learning* [29], *domain adaptation* [30], or *covariate shift* [5, 31].

Let $\mathrm{loss}(y, \widehat{y})$ be a point-wise loss function that measures a discrepancy between y and $\widehat{y}$ (at input $\boldsymbol{x}$). Then the *generalization error* which we would like to ultimately minimize is defined as

$$\mathbb{E}_{p(y|\boldsymbol{x})p_{\mathrm{te}}(\boldsymbol{x})}\left[\mathrm{loss}(y, f(\boldsymbol{x}))\right],$$

where $f(\boldsymbol{x})$ is a function model. Since the generalization error is inaccessible because the true probability $p(y|\boldsymbol{x})p_{\mathrm{te}}(\boldsymbol{x})$ is unknown, empirical-error minimization is often used in practice [8]:

$$\min_{f \in \mathcal{F}} \left[\frac{1}{n_{\mathrm{tr}}} \sum_{j=1}^{n_{\mathrm{tr}}} \mathrm{loss}(y_j^{\mathrm{tr}}, f(\boldsymbol{x}_j^{\mathrm{tr}})) \right].$$

However, under the covariate shift setup, plain empirical-error minimization is not *consistent* (i.e., it does not converge to the optimal function) if the model $\mathcal{F}$ is *misspecified* [i.e., the true function is not included in the model; see 5]. Instead, the following *importance-weighted* empirical-error minimization is consistent under covariate shift:

$$\min_{f \in \mathcal{F}} \left[\frac{1}{n_{\mathrm{tr}}} \sum_{j=1}^{n_{\mathrm{tr}}} r(\boldsymbol{x}_j^{\mathrm{tr}}) \mathrm{loss}(y_j^{\mathrm{tr}}, f(\boldsymbol{x}_j^{\mathrm{tr}})) \right],$$

[2] http://www.gaussianprocess.org/gpml/data/

Table 4: Experimental results of outlier detection for various for real-world datasets. Mean AUC score (and standard deviation in the bracket) over 100 trials is reported. The best method having the highest mean AUC score and comparable methods according to the *two-sample t-test* at the significance level 5% are specified by bold face. The datasets are sorted in the ascending order of the input dimensionality d.

Datasets	d	OSVM ($\nu = 0.05$)	OSVM ($\nu = 0.1$)	RuLSIF ($\alpha = 0$)	RuLSIF ($\alpha = 0.5$)	RuLSIF ($\alpha = 0.95$)
IDA:banana	2	**.668 (.105)**	**.676(.120)**	.597(.097)	.619(.101)	.623 (.115)
IDA:thyroid	5	.760 (.148)	**.782(.165)**	**.804(.148)**	**.796(.178)**	.722 (.153)
IDA:titanic	5	**.757 (.205)**	**.752(.191)**	**.750(.182)**	.701(.184)	.712 (.185)
IDA:diabetes	8	**.636 (.099)**	.610(.090)	.594(.105)	.575(.105)	**.663 (.112)**
IDA:b-cancer	9	**.741 (.160)**	.691(.147)	**.707(.148)**	**.737(.159)**	**.733 (.160)**
IDA:f-solar	9	.594 (.087)	.590(.083)	**.626(.102)**	**.612(.100)**	.584 (.114)
IDA:heart	13	.714 (.140)	.694(.148)	**.748(.149)**	**.769(.134)**	.726 (.127)
IDA:german	20	**.612 (.069)**	**.604(.084)**	**.605(.092)**	**.597(.101)**	**.605 (.095)**
IDA:ringnorm	20	**.991 (.012)**	**.993(.007)**	.944(.091)	.971(.062)	**.992 (.010)**
IDA:waveform	21	.812 (.107)	.843(.123)	**.879(.122)**	**.875(.117)**	**.885 (.102)**
Speech	50	.788 (.068)	**.830(.060)**	.804(.101)	**.821(.076)**	**.836 (.083)**
20News ('rec')	100	.598 (.063)	.593(.061)	.628(.105)	.614(.093)	**.767 (.100)**
20News ('sci')	100	.592 (.069)	.589(.071)	.620(.094)	.609(.087)	**.704 (.093)**
20News ('talk')	100	.661 (.084)	.658(.084)	.672(.117)	.670(.102)	**.823 (.078)**
USPS (1 vs. 2)	256	.889 (.052)	**.926(.037)**	.848(.081)	.878(.088)	.898 (.051)
USPS (2 vs. 3)	256	.823 (.053)	.835(.050)	.803(.093)	.818(.085)	**.879 (.074)**
USPS (3 vs. 4)	256	.901 (.044)	.939(.031)	.950(.056)	.961(.041)	**.984 (.016)**
USPS (4 vs. 5)	256	.871 (.041)	.890(.036)	.857(.099)	.874(.082)	**.941 (.031)**
USPS (5 vs. 6)	256	.825 (.058)	.859(.052)	.863(.078)	.867(.068)	**.901 (.049)**
USPS (6 vs. 7)	256	.910 (.034)	.950(.025)	.972(.038)	.984(.018)	**.994 (.010)**
USPS (7 vs. 8)	256	.938 (.030)	.967(.021)	.941(.053)	.951(.039)	**.980 (.015)**
USPS (8 vs. 9)	256	.721 (.072)	.728(.073)	.721(.084)	.728(.083)	**.761 (.096)**
USPS (9 vs. 0)	256	.920 (.037)	.966(.023)	.982(.048)	.989(.022)	**.994 (.011)**

where $r(\boldsymbol{x})$ is called the *importance* [32] in the context of covariate shift adaptation:

$$r(\boldsymbol{x}) := \frac{p_{\mathrm{te}}(\boldsymbol{x})}{p_{\mathrm{tr}}(\boldsymbol{x})}.$$

However, since importance-weighted learning is not *statistically efficient* (i.e., it tends to have larger variance), slightly *flattening* the importance weights is practically useful for stabilizing the estimator. **(author?)** [5] proposed to use the *exponentially-flattened importance weights* as

$$\min_{f \in \mathcal{F}} \left[\frac{1}{n_{\mathrm{tr}}} \sum_{j=1}^{n_{\mathrm{tr}}} r(\boldsymbol{x}_j^{\mathrm{tr}})^{\tau} \mathrm{loss}(y_j^{\mathrm{tr}}, f(\boldsymbol{x}_j^{\mathrm{tr}})) \right],$$

where $0 \leq \tau \leq 1$ is called the *exponential flattening parameter*. $\tau = 0$ corresponds to plain empirical-error minimization, while $\tau = 1$ corresponds to importance-weighted empirical-error minimization; $0 < \tau < 1$ will give an intermediate estimator that balances the trade-off between statistical efficiency and consistency. The exponential flattening parameter τ can be optimized by model selection criteria such as the *importance-weighted Akaike information criterion* for regular models [5], the *importance-weighted subspace information criterion* for linear models [33], and *importance-weighted cross-validation* for arbitrary models [6].

One of the potential drawbacks of the above exponential flattering approach is that estimation of $r(\boldsymbol{x})$ (i.e., $\tau = 1$) is rather hard, as shown in this paper. Thus, when $r(\boldsymbol{x})$ is estimated poorly, all flattened weights $r(\boldsymbol{x})^{\tau}$ are also unreliable and then covariate shift adaptation does not work well in practice. To cope with this problem, we propose to use *relative importance weights* alternatively:

$$\min_{f \in \mathcal{F}} \left[\frac{1}{n_{\mathrm{tr}}} \sum_{j=1}^{n_{\mathrm{tr}}} r_{\alpha}(\boldsymbol{x}_j^{\mathrm{tr}}) \mathrm{loss}(y_j^{\mathrm{tr}}, f(\boldsymbol{x}_j^{\mathrm{tr}})) \right],$$

where $r_{\alpha}(\boldsymbol{x})$ $(0 \leq \alpha \leq 1)$ is the α-relative importance weight defined by

$$r_{\alpha}(\boldsymbol{x}) := \frac{p_{\mathrm{te}}(\boldsymbol{x})}{(1-\alpha) p_{\mathrm{te}}(\boldsymbol{x}) + \alpha p_{\mathrm{tr}}(\boldsymbol{x})}.$$

Note that, compared with the definition of the α-relative density-ratio (1), α and $(1-\alpha)$ are swapped in order to be consistent with exponential flattening. Indeed, the relative importance weights play a similar role to exponentially-flattened importance weights; $\alpha = 0$ corresponds to plain empirical-error minimization, while $\alpha = 1$ corresponds to importance-weighted empirical-error minimization; $0 < \alpha < 1$ will give an intermediate estimator that balances the trade-off between efficiency and consistency. We note that the relative importance weights and exponentially flattened importance weights agree only when $\alpha = \tau = 0$ and $\alpha = \tau = 1$; for $0 < \alpha = \tau < 1$, they are generally different.

A possible advantage of the above relative importance weights is that its estimation for $0 < \alpha < 1$ does not depend on that for $\alpha = 1$, unlike exponentially-flattened importance weights. Since α-relative importance weights for $0 < \alpha < 1$ can be reliably estimated by RuLSIF proposed in this paper, the performance of covariate shift adaptation is expected to be improved. Below, we experimentally investigate this effect.

4.3.2 Artificial Datasets

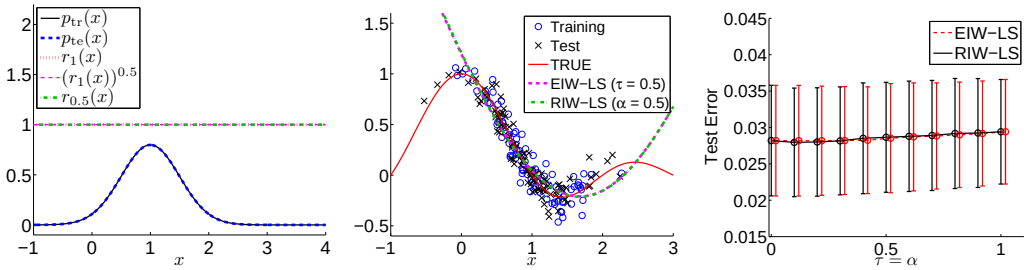
First, we illustrate how the proposed method behaves in covariate shift adaptation using one-dimensional artificial datasets.

In this experiment, we employ the following kernel regression model:

$$f(x; \boldsymbol{\beta}) = \sum_{i=1}^{n_{\mathrm{te}}} \beta_i \exp\left(-\frac{(x - x_i^{\mathrm{te}})^2}{2\rho^2} \right),$$

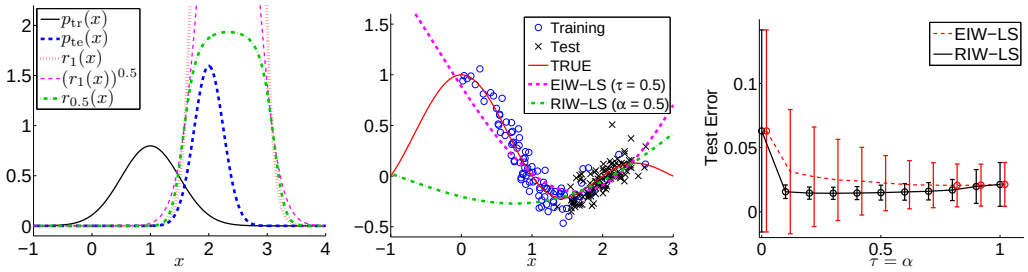
where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{n_{\mathrm{te}}})^{\top}$ is the parameter to be learned and ρ is the Gaussian width. The parameter $\boldsymbol{\beta}$ is learned by *relative importance-weighted least-squares* (RIW-LS):

$$\widehat{\boldsymbol{\beta}}_{\mathrm{RIW-LS}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \left[\frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \widehat{r}_{\alpha}(x_j^{\mathrm{tr}}) \left(f(x_j^{\mathrm{tr}}; \boldsymbol{\beta}) - y_j^{\mathrm{tr}} \right)^2 \right],$$



(a) Densities and ratios (b) Learned functions (c) Test error

Figure 6: Illustrative example of transfer learning under no distribution change.



(a) Densities and ratios (b) Learned functions (c) Test error

Figure 7: Illustrative example of transfer learning under covariate shift.

or *exponentially-flattened importance-weighted least-squares* (EIW-LS):

$$\widehat{\boldsymbol{\beta}}_{\rm EIW-LS} = \mathop{\rm argmin}_{\boldsymbol{\beta}} \left[\frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \widehat{r}(x_j^{\rm tr})^{\tau} \left(f(x_j^{\rm tr}; \boldsymbol{\beta}) - y_j^{\rm tr} \right)^2 \right].$$

The relative importance weight $\widehat{r}_{\alpha}(x_j^{\rm tr})$ is estimated by RuLSIF, and the exponentially-flattened importance weight $\widehat{r}(x_j^{\rm tr})^{\tau}$ is estimated by uLSIF (i.e., RuLSIF with $\alpha = 1$). The Gaussian width ρ is chosen by 5-fold *importance-weighted cross-validation* [6].

First, we consider the case where input distributions do not change:

$$P_{\rm tr} = P_{\rm te} = N(1, 0.25).$$

The densities and their ratios are plotted in Figure 6(a). The training output samples $\{y_j^{\rm tr}\}_{j=1}^{n_{\rm tr}}$ are generated as

$$y_j^{\rm tr} = {\rm sinc}(x_j^{\rm tr}) + \epsilon_j^{\rm tr},$$

where $\{\epsilon_j^{\rm tr}\}_{j=1}^{n_{\rm tr}}$ is additive noise following $N(0, 0.01)$. We set $n_{\rm tr} = 100$ and $n_{\rm te} = 200$. Figure 6(b) shows a realization of training and test samples as well as learned functions obtained by RIW-LS with $\alpha = 0.5$ and EIW-LS with $\tau = 0.5$. This shows that RIW-LS with $\alpha = 0.5$ and EIW-LS with $\tau = 0.5$ give almost the same functions, and both functions fit the true function well in the test region. Figure 6(c) shows the mean and standard deviation of the test error under the squared loss over 200 runs, as functions of the relative flattening parameter α in RIW-LS and the exponential flattening parameter τ in EIW-LS. The method having a lower mean test error and another method that is comparable according to the *two-sample t-test* at the significance level 5% are specified by '$\circ$'. As can be observed, the proposed RIW-LS compares favorably with EIW-LS.

Next, we consider the situation where input distribution changes (Figure 7(a)):

$$P_{\rm tr} = N(1, 0.25),$$
$$P_{\rm te} = N(2, 0.1).$$

The output values are created in the same way as the previous case. Figure 7(b) shows a realization of training and test samples as well as learned functions obtained by RIW-LS with $\alpha = 0.5$ and

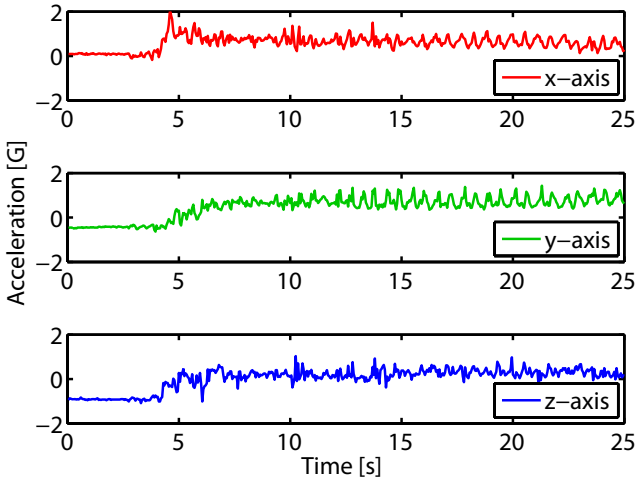


Figure 8: An example of three-axis accelerometer data for "walking" collected by *iPod touch*.

EIW-LS with $\tau = 0.5$. This shows that RIW-LS with $\alpha = 0.5$ fits the true function slightly better than EIW-LS with $\tau = 0.5$ in the test region. Figure 7(c) shows that the proposed RIW-LS tends to outperform EIW-LS, and the standard deviation of the test error for RIW-LS is much smaller than EIW-LS. This is because EIW-LS with $0 < \tau < 1$ is based on an importance estimate with $\tau = 1$, which tends to have high fluctuation. Overall, the stabilization effect of relative importance estimation was shown to improve the test accuracy.

4.3.3 Real-World Datasets

Finally, we evaluate the proposed transfer learning method on a real-world transfer learning task.

We consider the problem of human activity recognition from accelerometer data collected by *iPod touch*[3]. In the data collection procedure, subjects were asked to perform a specific action such as walking, running, and bicycle riding. The duration of each task was arbitrary and the sampling rate was 20Hz with small variations. An example of three-axis accelerometer data for "walking" is plotted in Figure 8.

To extract features from the accelerometer data, each data stream was segmented in a sliding window manner with window width 5 seconds and sliding step 1 second. Depending on subjects, the position and orientation of *iPod touch* was arbitrary—held by hand or kept in a pocket or a bag. For this reason, we decided to take the ℓ_2-norm of the 3-dimensional acceleration vector at each time step, and computed the following 5 orientation-invariant features from each window: *mean*, *standard deviation*, *fluctuation of amplitude*, *average energy*, and *frequency-domain entropy* [34, 35].

Let us consider a situation where a new user wants to use the activity recognition system. However, since the new user is not willing to label his/her accelerometer data due to troublesomeness, no labeled sample is available for the new user. On the other hand, unlabeled samples for the new user and labeled data obtained from existing users are available. Let labeled training data $\{(\boldsymbol{x}_j^{\mathrm{tr}}, y_j^{\mathrm{tr}})\}_{j=1}^{n_{\mathrm{tr}}}$ be the set of labeled accelerometer data for 20 existing users. Each user has at most 100 labeled samples for each action. Let unlabeled test data $\{\boldsymbol{x}_i^{\mathrm{te}}\}_{i=1}^{n_{\mathrm{te}}}$ be unlabeled accelerometer data obtained from the new user.

We use *kernel logistic regression* (KLR) for activity recognition. We compare the following four methods:

- Plain KLR without importance weights (i.e., $\alpha = 0$ or $\tau = 0$).
- KLR with relative importance weights for $\alpha = 0.5$.
- KLR with exponentially-flattened importance weights for $\tau = 0.5$.

[3] `http://alkan.mns.kyutech.ac.jp/web/data.html`

Table 5: Experimental results of transfer learning in real-world human activity recognition. Mean classification accuracy (and the standard deviation in the bracket) over 100 runs for activity recognition of a new user is reported. The method having the lowest mean classification accuracy and comparable methods according to the *two-sample t-test* at the significance level 5% are specified by bold face.

Task	KLR ($\alpha = 0, \tau = 0$)	RIW-KLR ($\alpha = 0.5$)	EIW-KLR ($\tau = 0.5$)	IW-KLR ($\alpha = 1, \tau = 1$)
Walks vs. run	0.803 (0.082)	**0.889(0.035)**	**0.882(0.039)**	**0.882 (0.035)**
Walks vs. bicycle	0.880 (0.025)	**0.892(0.035)**	0.867(0.054)	0.854 (0.070)
Walks vs. train	0.985 (0.017)	**0.992(0.008)**	0.989(0.011)	0.983 (0.021)

- KLR with plain importance weights (i.e., $\alpha = 1$ or $\tau = 1$).

The experiments are repeated 100 times with different sample choice for $n_{\mathrm{tr}} = 500$ and $n_{\mathrm{te}} = 200$. Table 5 depicts the classification accuracy for three binary-classification tasks: *walk vs. run*, *walk vs. riding a bicycle*, and *walk vs. taking a train*. The classification accuracy is evaluated for 800 samples from the new user that are not used for classifier training (i.e., the 800 test samples are different from 200 unlabeled samples). The table shows that KLR with relative importance weights for $\alpha = 0.5$ compares favorably with other methods in terms of the classification accuracy. KLR with plain importance weights and KLR with exponentially-flattened importance weights for $\tau = 0.5$ are outperformed by KLR without importance weights in the *walk vs. riding a bicycle* task due to the instability of importance weight estimation for $\alpha = 1$ or $\tau = 1$.

Overall, the proposed relative density-ratio estimation method was shown to be useful also in transfer learning under covariate shift.

5 Conclusion

In this paper, we proposed to use a relative divergence for robust distribution comparison. We gave a computationally efficient method for estimating the relative Pearson divergence based on direct relative density-ratio approximation. We theoretically elucidated the convergence rate of the proposed divergence estimator under non-parametric setup, which showed that the proposed approach of estimating the relative Pearson divergence is more preferable than the existing approach of estimating the plain Pearson divergence. Furthermore, we proved that the asymptotic variance of the proposed divergence estimator is independent of the model complexity under a correctly-specified parametric setup. Thus, the proposed divergence estimator hardly overfits even with complex models. Experimentally, we demonstrated the practical usefulness of the proposed divergence estimator in two-sample homogeneity test, inlier-based outlier detection, and transductive transfer learning under covariate shift.

In addition to two-sample homogeneity test, outlier detection, and transfer learning, density ratios were shown to be useful for tackling various machine learning problems, including multi-task learning [36, 37], independence test [38], feature selection [39], causal inference [40], independent component analysis [41], dimensionality reduction [15], unpaired data matching [42], clustering [43], conditional density estimation [44], and probabilistic classification [45]. Thus, it would be promising to explore more applications of the proposed relative density-ratio approximator beyond two-sample homogeneity test, outlier detection, and transfer learning tasks.

Acknowledgments

MY was supported by the JST PRESTO program, TS was partially supported by MEXT KAKENHI 22700289 and Aihara Project, the FIRST program from JSPS, initiated by CSTP, TK was partially supported by Grant-in-Aid for Young Scientists (20700251), HH was supported by the FIRST program, and MS was partially supported by SCAT, AOARD, and the FIRST program.

A Technical Details of Non-Parametric Convergence Analysis

Here, we give the technical details of the non-parametric convergence analysis described in Section 3.1.

A.1 Results

For notational simplicity, we define linear operators $P, P_n, P', P'_{n'}$ as

$$Pf := \mathrm{E}_p f, \quad P_n f := \frac{\sum_{j=1}^{n} f(\boldsymbol{x}_j)}{n},$$
$$P'f := \mathrm{E}_q f, \quad P'_{n'} f := \frac{\sum_{i=1}^{n'} f(\boldsymbol{x}'_i)}{n'}.$$

For $\alpha \in [0, 1]$, we define $S_{n,n'}$ and S as

$$S_{n,n'} = \alpha P_n + (1-\alpha) P'_{n'}, \quad S = \alpha P + (1-\alpha) P'.$$

We estimate the Pearson divergence between p and $\alpha p + (1-\alpha) q$ through estimating the density ratio

$$g^* := \frac{p}{\alpha p + (1-\alpha) p'}.$$

Let us consider the following density ratio estimator:

$$\begin{aligned}\widehat{g} &:= \operatorname*{argmin}_{g \in \mathcal{G}} \left[\frac{1}{2} \left(\alpha P_n + (1-\alpha) P'_{n'} \right) g^2 - P_n g + \frac{\lambda_{\bar{n}}}{2} R(g)^2 \right] \\ &= \operatorname*{argmin}_{g \in \mathcal{G}} \left(\frac{1}{2} S_{n,n'} g^2 - P_n g + \frac{\lambda_{\bar{n}}}{2} R(g)^2 \right).\end{aligned}$$

where $\bar{n} = \min(n, n')$ and $R(g)$ is a non-negative regularization functional such that

$$\sup_{\boldsymbol{x}} [|g(\boldsymbol{x})|] \leq R(g). \tag{17}$$

A possible estimator of the Pearson (PE) divergence $\widehat{\mathrm{PE}}_\alpha$ is

$$\widehat{\mathrm{PE}}_\alpha := P_n \widehat{g} - \frac{1}{2} S_{n,n'} \widehat{g}^2 - \frac{1}{2}.$$

Another possibility is

$$\widetilde{\mathrm{PE}}_\alpha := \frac{1}{2} P_n \widehat{g} - \frac{1}{2}.$$

A useful example is to use a *reproducing kernel Hilbert space* [RKHS; 18] as $\mathcal{G}$ and the RKHS norm as $R(g)$. Suppose $\mathcal{G}$ is an RKHS associated with bounded kernel $k(\cdot, \cdot)$:

$$\sup_{\boldsymbol{x}} [k(\boldsymbol{x}, \boldsymbol{x})] \leq C.$$

Let $\| \cdot \|_{\mathcal{G}}$ denote the norm in the RKHS $\mathcal{G}$. Then $R(g) = \sqrt{C} \|g\|_{\mathcal{G}}$ satisfies Eq.(17):

$$g(\boldsymbol{x}) = \langle k(\boldsymbol{x}, \cdot), g(\cdot) \rangle \leq \sqrt{k(\boldsymbol{x}, \boldsymbol{x})} \|g\|_{\mathcal{G}} \leq \sqrt{C} \|g\|_{\mathcal{G}},$$

where we used the reproducing property of the kernel and Schwartz's inequality. Note that the Gaussian kernel satisfies this with $C = 1$. It is known that the Gaussian kernel RKHS spans a dense subset in the set of continuous functions. Another example of RKHSs is Sobolev space. The canonical norm for this space is the integral of the squared derivatives of functions. Thus the regularization term $R(g) = \|g\|_{\mathcal{G}}$ imposes the solution to be smooth. The RKHS technique in Sobolev space has been well exploited in the context of spline models [46]. We intend that the regularization term $R(g)$ is a generalization of the RKHS norm. Roughly speaking, $R(g)$ is like a "norm" of the function space $\mathcal{G}$.

We assume that the true density-ratio function $g^*(\boldsymbol{x})$ is contained in the model $\mathcal{G}$ and is bounded from above:

$$g^*(\boldsymbol{x}) \leq M_0 \quad \text{for all} \quad \boldsymbol{x} \in \mathcal{D}_{\mathrm{X}}.$$

Let $\mathcal{G}_M$ be a *ball* of $\mathcal{G}$ with radius $M > 0$:

$$\mathcal{G}_M := \{g \in \mathcal{G} \mid R(g) \leq M\}.$$

To derive the convergence rate of our estimator, we utilize the *bracketing entropy* that is a complexity measure of a function class [see p. 83 of 47].

Definition 1. *Given two functions l and u, the bracket $[l, u]$ is the set of all functions f with $l(\boldsymbol{x}) \leq f(\boldsymbol{x}) \leq u(\boldsymbol{x})$ for all $\boldsymbol{x}$. An ϵ-bracket with respect to $L_2(\tilde{p})$ is a bracket $[l, u]$ with $\|l - u\|_{L_2(\tilde{p})} < \epsilon$. The bracketing entropy $\mathcal{H}_{[]}(\mathcal{F}, \epsilon, L_2(\tilde{p}))$ is the logarithm of the minimum number of ϵ-brackets with respect to $L_2(\tilde{p})$ needed to cover a function set $\mathcal{F}$.*

We assume that there exists γ $(0 < \gamma < 2)$ such that, for all $M > 0$,

$$\mathcal{H}_{[]}(\mathcal{G}_M, \epsilon, L_2(p)) = O\left(\left(\frac{M}{\epsilon}\right)^{\gamma}\right), \ \mathcal{H}_{[]}(\mathcal{G}_M, \epsilon, L_2(p')) = O\left(\left(\frac{M}{\epsilon}\right)^{\gamma}\right). \tag{18}$$

This quantity represents a complexity of function class $\mathcal{G}$—the larger γ is, the more complex the function class $\mathcal{G}$ is because, for larger γ, more brackets are needed to cover the function class. The Gaussian RKHS satisfies this condition for arbitrarily small γ [48]. Note that when $R(g)$ is the RKHS norm, the condition (18) holds for all $M > 0$ if that holds for $M = 1$.

Then we have the following theorem.

Theorem 1. *Let $\bar{n} = \min(n, n')$, $M_0 = \|g^*\|_\infty$, and $c = (1+\alpha)\sqrt{P(g^* - Pg^*)^2} + (1-\alpha)\sqrt{P'(g^* - P'g^*)^2}$. Under the above setting, if $\lambda_{\bar{n}} \to 0$ and $\lambda_{\bar{n}}^{-1} = o(\bar{n}^{2/(2+\gamma)})$, then we have*

$$\widehat{\mathrm{PE}}_\alpha - \mathrm{PE}_\alpha = \mathcal{O}_p(\lambda_{\bar{n}} \max(1, R(g^*)^2) + \bar{n}^{-1/2} c M_0),$$

and

$$\widetilde{\mathrm{PE}}_\alpha - \mathrm{PE}_\alpha = \mathcal{O}_p(\lambda_{\bar{n}} \max\{1, M_0^{\frac{1}{2}(1-\frac{\gamma}{2})}, R(g^*) M_0^{\frac{1}{2}(1-\frac{\gamma}{2})}, R(g^*)\} + \lambda_{\bar{n}}^{\frac{1}{2}} \max\{M_0^{\frac{1}{2}}, M_0^{\frac{1}{2}} R(g^*)\}),$$

where $\mathcal{O}_p$ denotes the asymptotic order in probability.

In the proof of Theorem 1, we use the following auxiliary lemma.

Lemma 1. *Under the setting of Theorem 1, if $\lambda_{\bar{n}} \to 0$ and $\lambda_{\bar{n}}^{-1} = o(\bar{n}^{2/(2+\gamma)})$, then we have*

$$\|\widehat{g} - g^*\|_{L_2(S)} = \mathcal{O}_p(\lambda_{\bar{n}}^{1/2} \max\{1, R(g^*)\}), \quad R(\widehat{g}) = \mathcal{O}_p(\max\{1, R(g^*)\}),$$

where $\|\cdot\|_{L_2(S)}$ denotes the $L_2(\alpha p + (1-\alpha)q)$-norm.

A.2 Proof of Lemma 1

First, we prove Lemma 1.

From the definition, we obtain

$$\begin{aligned}
&\frac{1}{2} S_{n,n'} \widehat{g}^2 - P_n \widehat{g} + \lambda_{\bar{n}} R(\widehat{g})^2 \leq \frac{1}{2} S_{n,n'} g^{*2} - P_n g^* + \lambda_{\bar{n}} R(g^*)^2 \\
\Rightarrow \quad &\frac{1}{2} S_{n,n'} (\widehat{g} - g^*)^2 - S_{n,n'}(g^*(g^* - \widehat{g})) - P_n(\widehat{g} - g^*) + \lambda_{\bar{n}}(R(\widehat{g})^2 - R(g^*)^2) \leq 0.
\end{aligned}$$

On the other hand, $S(g^*(g^* - \widehat{g})) = P(g^* - \widehat{g})$ indicates

$$\begin{aligned}
&\frac{1}{2}(S - S_{n,n'})(\widehat{g} - g^*)^2 - (S - S_{n,n'})(g^*(g^* - \widehat{g})) - (P - P_n)(\widehat{g} - g^*) - \lambda_{\bar{n}}(R(\widehat{g})^2 - R(g^*)^2) \\
&\geq \frac{1}{2} S(\widehat{g} - g^*)^2.
\end{aligned}$$

Therefore, to bound $\|\widehat{g} - g^*\|_{L_2(S)}$, it suffices to bound the left-hand side of the above inequality.

Define $\mathcal{F}_M$ and $\mathcal{F}_M^2$ as

$$\mathcal{F}_M := \{g - g^* \mid g \in \mathcal{G}_M\} \quad \text{and} \quad \mathcal{F}_M^2 := \{f^2 \mid f \in \mathcal{F}_M\}.$$

To bound $|(S - S_{n,n'})(\widehat{g} - g^*)^2|$, we need to bound the bracketing entropies of $\mathcal{F}_M^2$. We show that

$$\mathcal{H}_{[]}(\mathcal{F}_M^2, \delta, L_2(p)) = O\left(\left(\frac{(M+M_0)^2}{\delta}\right)^\gamma\right),$$
$$\mathcal{H}_{[]}(\mathcal{F}_M^2, \delta, L_2(q)) = O\left(\left(\frac{(M+M_0)^2}{\delta}\right)^\gamma\right).$$

This can be shown as follows. Let f_L and f_U be a δ-bracket for $\mathcal{G}_M$ with respect to $L_2(p)$; $f_L(x) \leq f_U(x)$ and $\|f_L - f_U\|_{L_2(p)} \leq \delta$. Without loss of generality, we can assume that $\|f_L\|_{L_\infty}, \|f_U\|_{L_\infty} \leq M + M_0$. Then f'_U and f'_L defined as

$$f'_U(x) := \max\{f_L^2(x), f_U^2(x)\},$$
$$f'_L(x) := \begin{cases} \min\{f_L^2(x), f_U^2(x)\} & (\text{sign}(f_L(x)) = \text{sign}(f_U(x))), \\ 0 & (\text{otherwise}) \end{cases},$$

are also a bracket such that $f'_L \leq g^2 \leq f'_U$ for all $g \in \mathcal{G}_M$ s.t. $f_L \leq g \leq f_U$ and $\|f'_L - f'_U\|_{L_2(p)} \leq 2\delta(M + M_0)$ because $\|f_L - f_U\|_{L_2(p)} \leq \delta$ and the following relation is met:

$$\begin{aligned}
(f'_L(x) - f'_U(x))^2 &\leq \begin{cases} (f_L^2(x) - f_U^2(x))^2 & (\text{sign}(f_L(x)) = \text{sign}(f_U(x))), \\ \max\{f_L^4(x), f_U^4(x)\} & (\text{otherwise}) \end{cases} \\
&\leq \begin{cases} (f_L(x) - f_U(x))^2 (f_L(x) + f_U(x))^2 & (\text{sign}(f_L(x)) = \text{sign}(f_U(x))), \\ \max\{f_L^4(x), f_U^4(x)\} & (\text{otherwise}) \end{cases} \\
&\leq \begin{cases} (f_L(x) - f_U(x))^2 (f_L(x) + f_U(x))^2 & (\text{sign}(f_L(x)) = \text{sign}(f_U(x))), \\ (f_L(x) - f_U(x))^2 (|f_L(x)| + |f_U(x)|)^2 & (\text{otherwise}) \end{cases} \\
&\leq 4(f_L(x) - f_U(x))^2 (M + M_0)^2.
\end{aligned}$$

Therefore the condition for the bracketing entropies (18) gives $\mathcal{H}_{[]}(\mathcal{F}_M^2, \delta, L_2(p)) = O\left(\left(\frac{(M+M_0)^2}{\delta}\right)^\gamma\right)$. We can also show that $\mathcal{H}_{[]}(\mathcal{F}_M^2, \delta, L_2(q)) = O\left(\left(\frac{(M+M_0)^2}{\delta}\right)^\gamma\right)$ in the same fashion.

Let $f := \widehat{g} - g^*$. Then, as in Lemma 5.14 and Theorem 10.6 in **(author?)** [49], we obtain

$$\begin{aligned}
&|(S_{n,n'} - S)(f^2)| \leq \alpha|(P_n - P)(f^2)| + (1-\alpha)|(P'_{n'} - P')(f^2)| \\
&= \alpha \mathcal{O}_p\left(\frac{1}{\sqrt{\bar{n}}} \|f^2\|_{L_2(P)}^{1-\frac{\gamma}{2}} (1 + R(\widehat{g})^2 + M_0^2)^{\frac{\gamma}{2}} \vee \bar{n}^{-\frac{2}{2+\gamma}} (1 + R(\widehat{g})^2 + M_0^2)\right) \\
&\quad + (1-\alpha)\mathcal{O}_p\left(\frac{1}{\sqrt{\bar{n}}} \|f^2\|_{L_2(P')}^{1-\frac{\gamma}{2}} (1 + R(\widehat{g})^2 + M_0^2)^{\frac{\gamma}{2}} \vee \bar{n}^{-\frac{2}{2+\gamma}} (1 + R(\widehat{g})^2 + M_0^2)\right) \\
&\leq \mathcal{O}_p\left(\frac{1}{\sqrt{\bar{n}}} \|f^2\|_{L_2(S)}^{1-\frac{\gamma}{2}} (1 + R(\widehat{g})^2 + M_0^2)^{\frac{\gamma}{2}} \vee \bar{n}^{-\frac{2}{2+\gamma}} (1 + R(\widehat{g})^2 + M_0^2)\right), \qquad (19)
\end{aligned}$$

where $a \vee b = \max(a, b)$ and we used

$$\alpha\|f^2\|_{L_2(P)}^{1-\frac{\gamma}{2}} + (1-\alpha)\|f^2\|_{L_2(P')}^{1-\frac{\gamma}{2}} \leq \left(\int f^4 \mathrm{d}(\alpha P + (1-\alpha)P')\right)^{\frac{1}{2}(1-\frac{\gamma}{2})} = \|f^2\|_{L_2(S)}^{1-\frac{\gamma}{2}}$$

by Jensen's inequality for a concave function. Since

$$\|f^2\|_{L_2(S)} \leq \|f\|_{L_2(S)} \sqrt{2(1 + R(\widehat{g})^2 + M_0^2)},$$

the right-hand side of Eq.(19) is further bounded by

$$\begin{aligned}
&|(S_{n,n'} - S)(f^2)| \\
&= \mathcal{O}_p\left(\frac{1}{\sqrt{\bar{n}}} \|f\|_{L_2(S)}^{1-\frac{\gamma}{2}} (1 + R(\widehat{g})^2 + M_0^2)^{\frac{1}{2}+\frac{\gamma}{4}} \vee \bar{n}^{-\frac{2}{2+\gamma}} (1 + R(\widehat{g})^2 + M_0^2)\right). \qquad (20)
\end{aligned}$$

Similarly, we can show that

$$
\begin{aligned}
&|(S_{n,n'} - S)(g^*(g^* - \widehat{g}))| \\
&\quad = \mathcal{O}_p\left(\frac{1}{\sqrt{\bar{n}}}\|f\|_{L_2(S)}^{1-\frac{\gamma}{2}}(1 + R(\widehat{g})M_0 + M_0^2)^{\frac{\gamma}{2}} \vee \bar{n}^{-\frac{2}{2+\gamma}}(1 + R(\widehat{g})M_0 + M_0^2)\right), \qquad (21)
\end{aligned}
$$

and

$$
\begin{aligned}
&|(P_n - P)(g^* - \widehat{g})| = \mathcal{O}_p\left(\frac{1}{\sqrt{\bar{n}}}\|f\|_{L_2(P)}^{1-\frac{\gamma}{2}}(1 + R(\widehat{g}) + M_0)^{\frac{\gamma}{2}} \vee \bar{n}^{-\frac{2}{2+\gamma}}(1 + R(\widehat{g}) + M_0)\right) \\
&\leq \mathcal{O}_p\left(\frac{1}{\sqrt{\bar{n}}}\|f\|_{L_2(S)}^{1-\frac{\gamma}{2}}(1 + R(\widehat{g}) + M_0)^{\frac{\gamma}{2}} M_0^{\frac{1}{2}(1-\frac{\gamma}{2})} \vee \bar{n}^{-\frac{2}{2+\gamma}}(1 + R(\widehat{g}) + M_0)\right), \qquad (22)
\end{aligned}
$$

where we used

$$
\|f\|_{L_2(P)} = \sqrt{\int f^2 \mathrm{d}P} = \sqrt{\int f^2 g^* \mathrm{d}S} \leq M_0^{\frac{1}{2}} \sqrt{\int f^2 \mathrm{d}S}
$$

in the last inequality. Combining Eqs.(20), (21), and (22), we can bound the $L_2(S)$-norm of f as

$$
\begin{aligned}
&\frac{1}{2}\|f\|_{L_2(S)}^2 + \lambda_{\bar{n}} R(\widehat{g})^2 \\
&\leq \lambda_{\bar{n}} R(g^*)^2 + \mathcal{O}_p\left(\frac{1}{\sqrt{\bar{n}}}\|f\|_{L_2(S)}^{1-\frac{\gamma}{2}}(1 + R(\widehat{g})^2 + M_0^2)^{\frac{1}{2}+\frac{\gamma}{4}} \vee \bar{n}^{-\frac{2}{2+\gamma}}(1 + R(\widehat{g})^2 + M_0^2)\right). \qquad (23)
\end{aligned}
$$

The following is similar to the argument in Theorem 10.6 in **(author?)** [49], but we give a simpler proof.

By Young's inequality, we have $a^{\frac{1}{2}-\frac{\gamma}{4}} b^{\frac{1}{2}+\frac{\gamma}{4}} \leq (\frac{1}{2} - \frac{\gamma}{4})a + (\frac{1}{2} + \frac{\gamma}{4})b \leq a + b$ for all $a, b > 0$. Applying this relation to Eq.(23), we obtain

$$
\begin{aligned}
&\frac{1}{2}\|f\|_{L_2(S)}^2 + \lambda_{\bar{n}} R(\widehat{g})^2 \\
&\leq \lambda_{\bar{n}} R(g^*)^2 + \mathcal{O}_p\left(\|f\|_{L_2(S)}^{2(\frac{1}{2}-\frac{\gamma}{4})}\left\{\bar{n}^{-\frac{2}{2+\gamma}}(1 + R(\widehat{g})^2 + M_0^2)\right\}^{\frac{1}{2}+\frac{\gamma}{4}} \vee \bar{n}^{-\frac{2}{2+\gamma}}(1 + R(\widehat{g})^2 + M_0^2)\right) \\
&\overset{\text{Young}}{\leq} \lambda_{\bar{n}} R(g^*)^2 + \frac{1}{4}\|f\|_{L_2(S)}^2 + \mathcal{O}_p\left(\bar{n}^{-\frac{2}{2+\gamma}}(1 + R(\widehat{g})^2 + M_0^2) + \bar{n}^{-\frac{2}{2+\gamma}}(1 + R(\widehat{g})^2 + M_0^2)\right) \\
&= \lambda_{\bar{n}} R(g^*)^2 + \frac{1}{4}\|f\|_{L_2(S)}^2 + \mathcal{O}_p\left(\bar{n}^{-\frac{2}{2+\gamma}}(1 + R(\widehat{g})^2 + M_0^2)\right),
\end{aligned}
$$

which indicates

$$
\frac{1}{4}\|f\|_{L_2(S)}^2 + \lambda_{\bar{n}} R(\widehat{g})^2 \leq \lambda_{\bar{n}} R(g^*)^2 + o_p\left(\lambda_{\bar{n}}(1 + R(\widehat{g})^2 + M_0^2)\right).
$$

Therefore, by moving $o_p(\lambda_{\bar{n}} R(\widehat{g})^2)$ to the left hind side, we obtain

$$
\begin{aligned}
\frac{1}{4}\|f\|_{L_2(S)}^2 + \lambda_{\bar{n}}(1 - o_p(1)) R(\widehat{g})^2 &\leq \mathcal{O}_p\left(\lambda_{\bar{n}}(1 + R(g^*)^2 + M_0^2)\right) \\
&\leq \mathcal{O}_p\left(\lambda_{\bar{n}}(1 + R(g^*)^2)\right).
\end{aligned}
$$

This gives

$$
\begin{aligned}
&\|f\|_{L_2(S)} = \mathcal{O}_p(\lambda_{\bar{n}}^{\frac{1}{2}} \max\{1, R(g^*)\}), \\
&R(\widehat{g}) = \mathcal{O}_p(\sqrt{1 + R(g^*)^2}) = \mathcal{O}_p(\max\{1, R(g^*)\}).
\end{aligned}
$$

Consequently, the proof of Lemma 1 was completed.

A.3 Proof of Theorem 1

Based on Lemma 1, we prove Theorem 1.

As in the proof of Lemma 1, let $f := \widehat{g} - g^*$. Since $(\alpha P + (1-\alpha)P')(fg^*) = S(fg^*) = Pf$, we have

$$\begin{aligned}
\widehat{\mathrm{PE}}_\alpha - \mathrm{PE}_\alpha &= \frac{1}{2}S_{n,n'}\widehat{g}^2 - P_n\widehat{g} - (\frac{1}{2}Sg^{*2} - Pg^*) \\
&= \frac{1}{2}S_{n,n'}(f+g^*)^2 - P_n(f+g^*) - \left(\frac{1}{2}Sg^{*2} - Pg^*\right) \\
&= \frac{1}{2}Sf^2 + \frac{1}{2}(S_{n,n'} - S)f^2 + (S_{n,n'} - S)(g^*f) - (P_n - P)f \\
&\quad + \frac{1}{2}(S_{n,n'} - S)g^{*2} - (P_ng^* - Pg^*). \qquad (24)
\end{aligned}$$

Below, we show that each term of the right-hand side of the above equation is $\mathcal{O}_p(\lambda_{\bar{n}})$. By the central limit theorem, we have

$$\begin{aligned}
&\frac{1}{2}(S_{n,n'} - S)g^{*2} - (P_ng^* - Pg^*) \\
&\quad = \mathcal{O}_p\left(\bar{n}^{-1/2}M_0\left((1+\alpha)\sqrt{P(g^* - Pg^*)^2} + (1-\alpha)\sqrt{P'(g^* - P'g^*)^2}\right)\right).
\end{aligned}$$

Since Lemma 1 gives $\|f\|_2 = \mathcal{O}_p(\lambda_{\bar{n}}^{\frac{1}{2}}\max(1, R(g^*)))$ and $R(\widehat{g}) = \mathcal{O}_p(\max(1, R(g^*)))$, Eqs.(20), (21), and (22) in the proof of Lemma 1 imply

$$\begin{aligned}
|(S_{n,n'} - S)f^2| &= \mathcal{O}_p\left(\frac{1}{\sqrt{\bar{n}}}\|f\|_{L_2(S)}^{1-\frac{\gamma}{2}}(1+R(g^*))^{1+\frac{\gamma}{2}} \vee \bar{n}^{-\frac{2}{2+\gamma}}R(g^*)^2\right) \\
&\le \mathcal{O}_p(\lambda_{\bar{n}}\max(1, R(g^*)^2)), \\
|(S_{n,n'} - S)(g^*f)| &= \mathcal{O}_p\left(\frac{1}{\sqrt{\bar{n}}}\|f\|_{L_2(S)}^{1-\frac{\gamma}{2}}(1+R(\widehat{g})M_0 + M_0^2)^{\frac{\gamma}{2}} \vee \bar{n}^{-\frac{2}{2+\gamma}}(1+R(\widehat{g})M_0 + M_0^2)\right) \\
&\le \mathcal{O}_p(\lambda_{\bar{n}}\max(1, R(g^*)M_0^{\frac{\gamma}{2}}, M_0^{\gamma}R(g^*)^{1-\frac{\gamma}{2}}, M_0R(g^*), M_0^2)) \\
&\le \mathcal{O}_p(\lambda_{\bar{n}}\max(1, R(g^*)M_0^{\frac{\gamma}{2}}, M_0R(g^*))), \\
&\le \mathcal{O}_p(\lambda_{\bar{n}}\max(1, R(g^*)^2)), \\
|(P_n - P)f| &\le \mathcal{O}_p\left(\frac{1}{\sqrt{\bar{n}}}\|f\|_{L_2(S)}^{1-\frac{\gamma}{2}}(1+R(\widehat{g}) + M_0)^{\frac{\gamma}{2}}M_0^{\frac{1}{2}(1-\frac{\gamma}{2})} \vee \bar{n}^{-\frac{2}{2+\gamma}}(1+R(\widehat{g}) + M_0)\right) \\
&= \mathcal{O}_p(\lambda_{\bar{n}}\max(1, M_0^{\frac{1}{2}(1-\frac{\gamma}{2})}, R(g^*)M_0^{\frac{1}{2}(1-\frac{\gamma}{2})}, R(g^*))) \qquad (25) \\
&\le \mathcal{O}_p(\lambda_{\bar{n}}\max(1, R(g^*)^2)),
\end{aligned}$$

where we used $\lambda_{\bar{n}}^{-1} = o(\bar{n}^{2/(2+\gamma)})$ and $M_0 \le R(g^*)$. Lemma 1 also implies

$$Sf^2 = \|f\|_2^2 = \mathcal{O}_p(\lambda_{\bar{n}}\max(1, R(g^*)^2)).$$

Combining these inequalities with Eq.(24) implies

$$\widehat{\mathrm{PE}}_\alpha - \mathrm{PE}_\alpha = \mathcal{O}_p(\lambda_{\bar{n}}\max(1, R(g^*)^2) + n^{-1/2}cM_0),$$

where we again used $M_0 \le R(g^*)$.

On the other hand, we have

$$\begin{aligned}
\widetilde{\mathrm{PE}}_\alpha - \mathrm{PE}_\alpha &= \frac{1}{2}P_n\widehat{g} - \frac{1}{2}Pg^* \\
&= \frac{1}{2}\left[(P_n - P)(\widehat{g} - g^*) + P(\widehat{g} - g^*) + (P_n - P)g^*\right]. \qquad (26)
\end{aligned}$$

Eq.(25) gives

$$(P_n - P)(\widehat{g} - g^*) = \mathcal{O}_p(\lambda_{\bar{n}}\max(1, M_0^{\frac{1}{2}(1-\frac{\gamma}{2})}, R(g^*)M_0^{\frac{1}{2}(1-\frac{\gamma}{2})}, R(g^*))).$$

We also have

$$P(\widehat{g}-g^*) \le \|\widehat{g}-g^*\|_{L_2(P)} \le \|\widehat{g}-g^*\|_{L_2(S)} M_0^{\frac{1}{2}} = \mathcal{O}_p(\lambda_{\bar{n}}^{\frac{1}{2}} \max(M_0^{\frac{1}{2}}, M_0^{\frac{1}{2}} R(g^*))),$$

and

$$(P_n - P)g^* = O_p(\bar{n}^{-\frac{1}{2}} \sqrt{P(g^* - Pg^*)^2}) \le O_p(\bar{n}^{-\frac{1}{2}} M_0) \le \mathcal{O}_p(\lambda_{\bar{n}}^{\frac{1}{2}} \max(M_0^{\frac{1}{2}}, M_0^{\frac{1}{2}} R(g^*))),$$

Therefore by substituting these bounds into the relation (26), one observes that

$$\begin{aligned}&\widetilde{\mathrm{PE}}_\alpha - \mathrm{PE}_\alpha \\ =&\mathcal{O}_p(\lambda_{\bar{n}}^{\frac{1}{2}} \max(M_0^{\frac{1}{2}}, M_0^{\frac{1}{2}} R(g^*)) + \lambda_{\bar{n}} \max(1, M_0^{\frac{1}{2}(1-\frac{\gamma}{2})}, R(g^*) M_0^{\frac{1}{2}(1-\frac{\gamma}{2})}, R(g^*))). \end{aligned} \tag{27}$$

This completes the proof. ■

B Technical Details of Parametric Variance Analysis

Here, we give the technical details of the parametric variance analysis described in Section 3.2.

B.1 Results

For the estimation of the α-relative density-ratio (1), the statistical model

$$\mathcal{G} = \{g(\boldsymbol{x};\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^b\}$$

is used where b is a finite number. Let us consider the following estimator of α-relative density-ratio,

$$\widehat{g} = \operatorname*{argmin}_{g\in\mathcal{G}} \frac{1}{2}\left\{\frac{\alpha}{n}\sum_{i=1}^{n}(g(\boldsymbol{x}_i))^2 + \frac{1-\alpha}{n'}\sum_{j=1}^{n'}(g(\boldsymbol{x}'_j))^2\right\} - \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{x}_i).$$

Suppose that the model is correctly specified, i.e., there exists $\boldsymbol{\theta}^*$ such that

$$g(\boldsymbol{x};\boldsymbol{\theta}^*) = r_\alpha(\boldsymbol{x}).$$

Then, under a mild assumption [see Theorem 5.23 of 19], the estimator $\widehat{g}$ is consistent and the estimated parameter $\widehat{\boldsymbol{\theta}}$ satisfies the asymptotic normality in the large sample limit. Then, a possible estimator of the α-relative Pearson divergence PE_α is

$$\widehat{\mathrm{PE}}_\alpha = \frac{1}{n}\sum_{i=1}^{n}\widehat{g}(\boldsymbol{x}_i) - \frac{1}{2}\left\{\frac{\alpha}{n}\sum_{i=1}^{n}(\widehat{g}(\boldsymbol{x}_i))^2 + \frac{1-\alpha}{n'}\sum_{j=1}^{n'}(\widehat{g}(\boldsymbol{x}'_j))^2\right\} - \frac{1}{2}.$$

Note that there are other possible estimators for PE_α such as

$$\widetilde{\mathrm{PE}}_\alpha = \frac{1}{2n}\sum_{i=1}^{n}\widehat{g}(\boldsymbol{x}_i) - \frac{1}{2}.$$

We study the asymptotic properties of $\widehat{\mathrm{PE}}_\alpha$. The expectation under the probability p (p') is denoted as $\mathbb{E}_{p(\boldsymbol{x})}[\cdot]$ ($\mathbb{E}_{p'(\boldsymbol{x})}[\cdot]$). Likewise, the variance is denoted as $\mathbb{V}_{p(\boldsymbol{x})}[\cdot]$ ($\mathbb{V}_{p'(\boldsymbol{x})}[\cdot]$). Then, we have the following theorem.

Theorem 2. *Let $\|r\|_\infty$ be the sup-norm of the standard density ratio $r(\boldsymbol{x})$, and $\|r_\alpha\|_\infty$ be the sup-norm of the α-relative density ratio, i.e.,*

$$\|r_\alpha\|_\infty = \frac{\|r\|_\infty}{\alpha\|r\|_\infty + 1 - \alpha}.$$

The variance of $\widehat{\mathrm{PE}}_\alpha$ is denoted as $\mathbb{V}[\widehat{\mathrm{PE}}_\alpha]$. Then, under the regularity condition for the asymptotic normality, we have the following upper bound of $\mathbb{V}[\widehat{\mathrm{PE}}_\alpha]$:

$$\begin{aligned}\mathbb{V}[\widehat{\mathrm{PE}}_\alpha] &= \frac{1}{n}\mathbb{V}_{p(\boldsymbol{x})}\left[r_\alpha - \frac{\alpha r_\alpha^2}{2}\right] + \frac{1}{n'}\mathbb{V}_{p'(\boldsymbol{x})}\left[\frac{(1-\alpha)r_\alpha^2}{2}\right] + o\left(\frac{1}{n}, \frac{1}{n'}\right) \\ &\le \frac{\|r_\alpha\|_\infty^2}{n} + \frac{\alpha^2\|r_\alpha\|_\infty^4}{4n} + \frac{(1-\alpha)^2\|r_\alpha\|_\infty^4}{4n'} + o\left(\frac{1}{n}, \frac{1}{n'}\right).\end{aligned}$$

Theorem 3. *The variance of $\widetilde{\mathrm{PE}}_\alpha$ is denoted as $\mathbb{V}[\widetilde{\mathrm{PE}}_\alpha]$. Let ∇g be the gradient vector of g with respect to $\boldsymbol{\theta}$ at $\boldsymbol{\theta}=\boldsymbol{\theta}^*$, i.e., $(\nabla g(\boldsymbol{x};\boldsymbol{\theta}^*))_j = \frac{\partial g(\boldsymbol{x};\boldsymbol{\theta}^*)}{\partial \theta_j}$. The matrix $\boldsymbol{U}_\alpha$ is defined by*

$$\boldsymbol{U}_\alpha = \alpha\mathbb{E}_{p(\boldsymbol{x})}[\nabla g \nabla g^\top] + (1-\alpha)\mathbb{E}_{p'(\boldsymbol{x})}[\nabla g \nabla g^\top].$$

Then, under the regularity condition, the variance of $\widetilde{\mathrm{PE}}_\alpha$ is asymptotically given as

$$\begin{aligned}\mathbb{V}[\widetilde{\mathrm{PE}}_\alpha] = \frac{1}{n}\mathbb{V}_{p(\boldsymbol{x})}\left[\frac{r_\alpha + (1-\alpha r_\alpha)\mathbb{E}_{p(\boldsymbol{x})}[\nabla g]^\top \boldsymbol{U}_\alpha^{-1}\nabla g}{2}\right] \\ + \frac{1}{n'}\mathbb{V}_{p'(\boldsymbol{x})}\left[\frac{(1-\alpha)r_\alpha \mathbb{E}_{p(\boldsymbol{x})}[\nabla g]^\top \boldsymbol{U}_\alpha^{-1}\nabla g}{2}\right] + o\left(\frac{1}{n},\frac{1}{n'}\right).\end{aligned}$$

B.2 Proof of Theorem 2

Let $\widehat{\boldsymbol{\theta}}$ be the estimated parameter, i.e., $\widehat{g}(\boldsymbol{x}) = g(\boldsymbol{x};\widehat{\boldsymbol{\theta}})$. Suppose that $r_\alpha(\boldsymbol{x}) = g(\boldsymbol{x};\boldsymbol{\theta}^*) \in \mathcal{G}$ holds. Let $\delta\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$, then the asymptotic expansion of $\widehat{\mathrm{PE}}_\alpha$ is given as

$$\begin{aligned}\widehat{\mathrm{PE}}_\alpha &= \frac{1}{n}\sum_{i=1}^n g(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}}) - \frac{1}{2}\left\{\frac{\alpha}{n}\sum_{i=1}^n g(\boldsymbol{x}_i;\widehat{\boldsymbol{\theta}})^2 + \frac{1-\alpha}{n'}\sum_{j=1}^{n'} g(\boldsymbol{x}'_j;\widehat{\boldsymbol{\theta}})^2\right\} - \frac{1}{2} \\ &= \mathrm{PE}_\alpha + \frac{1}{n}\sum_{i=1}^n (r_\alpha(\boldsymbol{x}_i) - \mathbb{E}_{p(\boldsymbol{x})}[r_\alpha]) + \frac{1}{n}\sum_{i=1}^n \nabla g(\boldsymbol{x}_i;\boldsymbol{\theta}^*)^\top \delta\boldsymbol{\theta} \\ &\quad - \frac{1}{2}\left\{\frac{\alpha}{n}\sum_{i=1}^n (r_\alpha(\boldsymbol{x}_i)^2 - \mathbb{E}_{p(\boldsymbol{x})}[r_\alpha^2]) + \frac{1-\alpha}{n'}\sum_{j=1}^{n'} (r_\alpha(\boldsymbol{x}'_j)^2 - \mathbb{E}_{p'(\boldsymbol{x})}[r_\alpha^2])\right\} \\ &\quad - \left\{\frac{\alpha}{n}\sum_{i=1}^n r_\alpha(\boldsymbol{x}_i)\nabla g(\boldsymbol{x}_i;\boldsymbol{\theta}^*) + \frac{1-\alpha}{n'}\sum_{j=1}^{n'} r_\alpha(\boldsymbol{x}'_j)\nabla g(\boldsymbol{x}'_j;\boldsymbol{\theta}^*)\right\}^\top \delta\boldsymbol{\theta} + o_p\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n'}}\right).\end{aligned}$$

Let us define the linear operator G as

$$Gf = \frac{1}{\sqrt{n}}\sum_{i=1}^n (f(\boldsymbol{x}_i) - \mathbb{E}_{p(\boldsymbol{x})}[f]).$$

Likewise, the operator G' is defined for the samples from p'. Then, we have

$$\begin{aligned}&\widehat{\mathrm{PE}}_\alpha - \mathrm{PE}_\alpha \\ &= \frac{1}{\sqrt{n}}G\big(r_\alpha - \frac{\alpha}{2}r_\alpha^2\big) - \frac{1}{\sqrt{n'}}G'\big(\frac{1-\alpha}{2}r_\alpha^2\big) \\ &\quad + \left\{\mathbb{E}_{p(\boldsymbol{x})}[\nabla g] - \alpha\mathbb{E}_{p(\boldsymbol{x})}[r_\alpha\nabla g] - (1-\alpha)\mathbb{E}_{p'(\boldsymbol{x})}[r_\alpha\nabla g]\right\}^\top \delta\boldsymbol{\theta} + o_p\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n'}}\right) \\ &= \frac{1}{\sqrt{n}}G\big(r_\alpha - \frac{\alpha}{2}r_\alpha^2\big) - \frac{1}{\sqrt{n'}}G'\big(\frac{1-\alpha}{2}r_\alpha^2\big) + o_p\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n'}}\right).\end{aligned}$$

The second equality follows from

$$\mathbb{E}_{p(\boldsymbol{x})}[\nabla g] - \alpha\mathbb{E}_{p(\boldsymbol{x})}[r_\alpha\nabla g] - (1-\alpha)\mathbb{E}_{p'(\boldsymbol{x})}[r_\alpha\nabla g] = 0.$$

Then, the asymptotic variance is given as

$$\mathbb{V}[\widehat{\mathrm{PE}}_\alpha] = \frac{1}{n}\mathbb{V}_{p(\boldsymbol{x})}\left[r_\alpha - \frac{\alpha}{2}r_\alpha^2\right] + \frac{1}{n'}\mathbb{V}_{p'(\boldsymbol{x})}\left[\frac{1-\alpha}{2}r_\alpha^2\right] + o\left(\frac{1}{n},\frac{1}{n'}\right). \tag{28}$$

We confirm that both $r_\alpha - \frac{\alpha}{2}r_\alpha^2$ and $\frac{1-\alpha}{2}r_\alpha^2$ are non-negative and increasing functions with respect to r for any $\alpha \in [0,1]$. Since the result is trivial for $\alpha = 1$, we suppose $0 \le \alpha < 1$. The function $r_\alpha - \frac{\alpha}{2}r_\alpha^2$ is represented as

$$r_\alpha - \frac{\alpha}{2}r_\alpha^2 = \frac{r(\alpha r + 2 - 2\alpha)}{2(\alpha r + 1 - \alpha)^2},$$

and thus, we have $r_\alpha - \frac{\alpha}{2} r_\alpha^2 = 0$ for $r = 0$. In addition, the derivative is equal to

$$\frac{\partial}{\partial r} \frac{r(\alpha r + 2 - 2\alpha)}{2(\alpha r + 1 - \alpha)^2} = \frac{(1-\alpha)^2}{(\alpha r + 1 - \alpha)^3},$$

which is positive for $r \geq 0$ and $\alpha \in [0, 1)$. Hence, the function $r_\alpha - \frac{\alpha}{2} r_\alpha^2$ is non-negative and increasing with respect to r. Following the same line, we see that $\frac{1-\alpha}{2} r_\alpha^2$ is non-negative and increasing with respect to r. Thus, we have the following inequalities,

$$0 \leq r_\alpha(\boldsymbol{x}) - \frac{\alpha}{2} r_\alpha(\boldsymbol{x})^2 \leq \|r_\alpha\|_\infty - \frac{\alpha}{2} \|r_\alpha\|_\infty^2,$$
$$0 \leq \frac{1-\alpha}{2} r_\alpha(\boldsymbol{x})^2 \leq \frac{1-\alpha}{2} \|r_\alpha\|_\infty^2.$$

As a result, upper bounds of the variances in Eq.(28) are given as

$$\mathbb{V}_{p(\boldsymbol{x})}\left[r_\alpha - \frac{\alpha}{2} r_\alpha^2 \right] \leq \left(\|r_\alpha\|_\infty - \frac{\alpha}{2} \|r_\alpha\|_\infty^2 \right)^2,$$
$$\mathbb{V}_{p'(\boldsymbol{x})}\left[\frac{1-\alpha}{2} r_\alpha^2 \right] \leq \frac{(1-\alpha)^2}{4} \|r_\alpha\|_\infty^4.$$

Therefore, the following inequality holds,

$$\begin{aligned}
\mathbb{V}[\widehat{\mathrm{PE}}_\alpha] &\leq \frac{1}{n} \left(\|r_\alpha\|_\infty - \frac{\alpha \|r_\alpha\|_\infty^2}{2} \right)^2 + \frac{1}{n'} \cdot \frac{(1-\alpha)^2 \|r_\alpha\|_\infty^4}{4} + o\left(\frac{1}{n}, \frac{1}{n'} \right) \\
&\leq \frac{\|r_\alpha\|_\infty^2}{n} + \frac{\alpha^2 \|r_\alpha\|_\infty^4}{4n} + \frac{(1-\alpha)^2 \|r_\alpha\|_\infty^4}{4n'} + o\left(\frac{1}{n}, \frac{1}{n'} \right),
\end{aligned}$$

which completes the proof.

B.3 Proof of Theorem 3

The estimator $\widehat{\boldsymbol{\theta}}$ is the optimal solution of the following problem:

$$\min_{\theta \in \Theta} \left[\frac{1}{2n} \sum_{i=1}^{n} \alpha g(x_i; \boldsymbol{\theta})^2 + \frac{1}{2n'} \sum_{j=1}^{n'} (1-\alpha) g(x'_j; \boldsymbol{\theta})^2 - \frac{1}{n} \sum_{i=1}^{n} g(x_i; \boldsymbol{\theta}) \right].$$

Then, the extremal condition yields the equation,

$$\frac{\alpha}{n} \sum_{i=1}^{n} g(x_i; \widehat{\boldsymbol{\theta}}) \nabla g(x_i; \widehat{\boldsymbol{\theta}}) + \frac{1-\alpha}{n'} \sum_{j=1}^{n'} g(x'_j; \widehat{\boldsymbol{\theta}}) \nabla g(x'_j; \widehat{\boldsymbol{\theta}}) - \frac{1}{n} \sum_{i=1}^{n} \nabla g(x_i; \widehat{\boldsymbol{\theta}}) = 0.$$

Let $\delta \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$. The asymptotic expansion of the above equation around $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ leads to

$$\frac{1}{n} \sum_{i=1}^{n} (\alpha r_\alpha(x_i) - 1) \nabla g(x_i; \boldsymbol{\theta}^*) + \frac{1-\alpha}{n'} \sum_{j=1}^{n'} r_\alpha(x'_j) \nabla g(x'_j; \boldsymbol{\theta}^*) + \boldsymbol{U}_\alpha \delta \boldsymbol{\theta} + o_p\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n'}} \right) = \mathbf{0}.$$

Therefore, we obtain

$$\delta \boldsymbol{\theta} = \frac{1}{\sqrt{n}} G((1 - \alpha r_\alpha) \boldsymbol{U}_\alpha^{-1} \nabla g) - \frac{1}{\sqrt{n'}} G'((1-\alpha) r_\alpha \boldsymbol{U}_\alpha^{-1} \nabla g) + o_p\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n'}} \right).$$

Next, we compute the asymptotic expansion of $\widetilde{\mathrm{PE}}_\alpha$:

$$\begin{aligned}
\widetilde{\mathrm{PE}}_\alpha &= \frac{1}{2} \mathbb{E}_{p(\boldsymbol{x})}[r_\alpha] + \frac{1}{2n} \sum_{i=1}^{n} (r_\alpha(x_i) - \mathbb{E}_{p(\boldsymbol{x})}[r_\alpha]) \\
&\quad + \frac{1}{2n} \sum_{i=1}^{n} \nabla g(x_i; \boldsymbol{\theta}^*)^\top \delta \boldsymbol{\theta} - \frac{1}{2} + o_p\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n'}} \right) \\
&= \mathrm{PE}_\alpha + \frac{1}{2\sqrt{n}} G(r_\alpha) + \frac{1}{2} \mathbb{E}_{p(\boldsymbol{x})}[\nabla g]^\top \delta \boldsymbol{\theta} + o_p\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n'}} \right).
\end{aligned}$$

Substituting $\delta\boldsymbol{\theta}$ into the above expansion, we have

$$\widetilde{\mathrm{PE}}_\alpha - \mathrm{PE}_\alpha = \frac{1}{2\sqrt{n}} G(r_\alpha + (1-\alpha r_\alpha)\mathbb{E}_{p(\boldsymbol{x})}[\nabla g]^\top \boldsymbol{U}_\alpha^{-1}\nabla g) \\ - \frac{1}{2\sqrt{n'}} G'((1-\alpha) r_\alpha \mathbb{E}_{p(\boldsymbol{x})}[\nabla g]^\top \boldsymbol{U}_\alpha^{-1}\nabla g) + o_p\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n'}}\right).$$

As a result, we have

$$\mathbb{V}[\widetilde{\mathrm{PE}}_\alpha] = \frac{1}{n}\mathbb{V}_{p(\boldsymbol{x})}\left[\frac{r_\alpha + (1-\alpha r_\alpha)\mathbb{E}_{p(\boldsymbol{x})}[\nabla g]^\top \boldsymbol{U}_\alpha^{-1}\nabla g}{2}\right] \\ + \frac{1}{n'}\mathbb{V}_{p'(\boldsymbol{x})}\left[\frac{(1-\alpha) r_\alpha \mathbb{E}_{p(\boldsymbol{x})}[\nabla g]^\top \boldsymbol{U}_\alpha^{-1}\nabla g}{2}\right] + o\left(\frac{1}{n}, \frac{1}{n'}\right),$$

which completes the proof.

References

[1] A. Smola, L. Song, and C. H. Teo. Relative novelty detection. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS2009)*, pages 536–543, 2009.

[2] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori. Statistical outlier detection using direct density ratio estimation. *Knowledge and Information Systems*, 26(2):309–336, 2011.

[3] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, Cambridge, MA, 2007.

[4] M. Sugiyama, T. Suzuki, Y. Itoh, T. Kanamori, and M. Kimura. Least-squares two-sample test. *Neural Networks*, 2011. to appear.

[5] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[6] M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, May 2007.

[7] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[8] V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, 1998.

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

[10] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[11] K. Pearson. 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, 1900.

[12] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 28:131–142, 1966.

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

[14] T. Kanamori, S. Hido, and M. Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

[15] T. Suzuki and M. Sugiyama. Sufficient dimension reduction via squared-loss mutual information estimation. In Y. W. Teh and M. Tiggerington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS2010)*, volume 9 of *JMLR Workshop and Conference Proceedings*, pages 804–811, Sardinia, Italy, May 13-15 2010.

[16] C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In J. Lafferty, C. K. I. Williams, R. Zemel, J. Shawe-Taylor, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 442–450. 2010.

[17] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, USA, 1970.

[18] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[19] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

[20] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, NY, 1993.

[21] G. Rätsch, T. Onoda, and K.-R. Müller. Soft margins for adaboost. *Machine Learning*, 42(3):287–320, 2001.

[22] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[23] B. Sriperumbudur, K. Fukumizu, A. Gretton, G. Lanckriet, and B. Schölkopf. Kernel choice and classifiability for RKHS embeddings of probability distributions. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1750–1758. MIT Press, Cambridge, MA, 2009.

[24] A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159, 1997.

[25] A. Kain and M. W. Macon. Spectral voice conversion for text-to-speech synthesis. In *Proceedings of 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP1998)*, pages 285–288, 1998.

[26] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

[27] C.-C. Chang and C.h-J. Lin. *LIBSVM: A Library for Support Vector Machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[28] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, 2006.

[29] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[30] J. Jiang and C. Zhai. Instance weighting for domain adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 264–271, 2007.

[31] M. Sugiyama and M. Kawanabe. *Covariate Shift Adaptation: Toward Machine Learning in Non-Stationary Environments*. MIT Press, Cambridge, MA, USA, 2011. to appear.

[32] G. S. Fishman. *Monte Carlo: Concepts, Algorithms, and Applications*. Springer-Verlag, Berlin, 1996.

[33] M. Sugiyama and K.-R. Müller. Input-dependent estimation of generalization error under covariate shift. *Statistics & Decisions*, 23(4):249–279, 2005.

[34] L. Bao and S. S. Intille. Activity recognition from user-annotated acceleration data. In *Proceedings of the 2nd IEEE International Conference on Pervasive Computing*, pages 1–17, 2004.

[35] N. B. Bharatula, M. Stager, P. Lukowicz, and G Troster. Empirical study of design choices in multi-sensor context recognition systems. In *Proceedings of the 2nd International Forum on Applied Wearable Computing*, pages 79–93, 2005.

[36] S. Bickel, J. Bogojeska, T. Lengauer, and T. Scheffer. Multi-task learning for HIV therapy screening. In A. McCallum and S. Roweis, editors, *Proceedings of 25th Annual International Conference on Machine Learning (ICML2008)*, pages 56–63, Jul. 5–9 2008.

[37] J. Simm, M. Sugiyama, and T. Kato. Computationally efficient multi-task learning with least-squares probabilistic classifiers. *IPSJ Transactions on Computer Vision and Applications*, 3:1–8, 2011.

[38] M. Sugiyama and T. Suzuki. Least-squares independence test. *IEICE Transactions on Information and Systems*, E94-D(6), 2011.

[39] T. Suzuki, M. Sugiyama, T. Kanamori, and J. Sese. Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinformatics*, 10(1):S52, 2009.

[40] M. Yamada and M. Sugiyama. Dependence minimizing regression with model selection for non-linear causal inference under non-Gaussian noise. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI2010)*, pages 643–648, Atlanta, Georgia, USA, Jul. 11–15 2010. The AAAI Press.

[41] T. Suzuki and M. Sugiyama. Least-squares independent component analysis. *Neural Computation*, 23(1):284–301, 2011.

[42] M. Yamada and M. Sugiyama. Cross-domain object matching with model selection. In G. Gordon, D. Dunson, and M. Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS2011)*, Fort Lauderdale, Florida, USA, Apr. 11-13 2011.

[43] M. Kimura and M. Sugiyama. Dependence-maximization clustering with least-squares mutual information. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 2011.

[44] M. Sugiyama, I. Takeuchi, T. Suzuki, T. Kanamori, H. Hachiya, and D. Okanohara. Least-squares conditional density estimation. *IEICE Transactions on Information and Systems*, E93-D(3):583–594, 2010.

[45] M. Sugiyama. Superfast-trainable multi-class probabilistic classifier by least-squares posterior fitting. *IEICE Transactions on Information and Systems*, E93-D(10):2690–2701, 2010.

[46] G. Wahba. *Spline Model for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia and Pennsylvania, 1990.

[47] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, 1996.

[48] I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *The Annals of Statistics*, 35(2):575–607, 2007.

[49] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.