Relative Density-Ratio Estimation for Robust Distribution Comparison

Makoto Yamada  
yamada.makoto@lab.ntt.co.jp  
NTT Communication Science Laboratories, NTT Corporation, Seika-cho, Kyoto, 619-0237, Japan

Taiji Suzuki  
s-taiji@stat.t.u-tokyo.ac.jp  
Department of Mathematical Informatics, University of Tokyo, Bunkyo-ku, Tokyo, 113-8656, Japan

Takafumi Kanamori  
kanamori@is.nagoya-u.ac.jp  
Department of Computer Science and Mathematical Informatics, Nagoya University, Chikusa-ku, Nagoya, Aichi, 464-8601, Japan

Hirotaka Hachiya  
hacchan@gmail.com

Masashi Sugiyama  
sugi@cs.titech.ac.jp  
Department of Computer Science, Tokyo Institute of Technology, Meguro-ku, Tokyo, 152-8552, Japan

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 a challenging task in practice. In this letter, we 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 nonparametric 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 outlier detection (Smola, Song, & Teo, 2009; Hido, Tsuboi, Kashima, Sugiyama, & Kanamori, 2011), two-sample homogeneity test (Gretton, Borgwardt, Rasch, Schölkopf, & Smola, 2007; Sugiyama, Suzuki, Itoh, Kanamori, & Kimura, 2011), and transfer learning (Shimodaira, 2000; Sugiyama, Krauledat, & Müller, 2007), for example.

A standard approach to comparing probability densities $p(x)$ and $p'(x)$ would be to estimate a divergence from $p(x)$ to $p'(x)$, such as the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951):

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

A naive way to estimate the KL divergence is to separately approximate the densities $p(x)$ and $p'(x)$ from data and plug the estimated densities into the above definition. However, since density estimation is known to be a hard task (Vapnik, 1998), this approach does not work well unless a good parametric model is available. Recently a divergence estimation approach has been proposed that directly approximates the density ratio,

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

without going through separate approximation of densities $p(x)$ and $p'(x)$ (Sugiyama et al., 2008; Nguyen, Wainwright, & Jordan, 2010). Such density-ratio approximation methods were proved to achieve the optimal nonparametric convergence rate in the mini-max sense.

However, the KL divergence estimation via density-ratio approximation is computationally rather expensive due to the nonlinearity introduced by the log term. To cope with this problem, another divergence, called the Pearson (PE) divergence (Pearson, 1900), is useful. The PE divergence from $p(x)$ to $p'(x)$ is defined as

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

The PE divergence is a squared-loss variant of the KL divergence, and both belong to the class of the Ali-Silvey-Csiszár divergences (which is also known as the $f$-divergences; see Ali & Silvey, 1966; Csiszár, 1967). Thus, the PE and KL divergences share similar properties; they are nonnegative and vanish if and only if $p(x) = p'(x)$.
Similar to the KL divergence estimation, the PE divergence can also be accurately estimated based on density-ratio approximation (Kanamori, Hido, & Sugiyama, 2009): the density-ratio approximator called unconstrained least-squares importance fitting (uLSIF) gives the PE divergence estimator analytically, which can be computed 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 (Hido et al., 2011), two-sample homogeneity test (Sugiyama et al., 2011), and dimensionality reduction (Suzuki & Sugiyama, 2010).

In this letter, we first establish the nonparametric convergence rate of the uLSIF-based PE divergence estimator that 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: \( \max_x r(x) \). This implies that in the region where the denominator density \( p'(x) \) takes small values, the density ratio \( r(x) = p(x)/p'(x) \) tends to take large values, and therefore the overall convergence speed becomes slow. More critically, density ratios can diverge to infinity under a rather simple setting, as when the ratio of two gaussian functions is considered (Cortes, Mansour, & Mohri, 2010). 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 \( \alpha \)-relative divergence estimation. In the proposed approach, we estimate the quantity called \( \alpha \)-relative divergence, which is the divergence from \( p(x) \) to the \( \alpha \)-mixture density \( \alpha p(x) + (1 - \alpha) p'(x) \) for \( 0 \leq \alpha < 1 \). For example, the \( \alpha \)-relative PE divergence is given by

\[
\text{PE}_{\alpha}[p(x), p'(x)] := \text{PE}[p(x), \alpha p(x) + (1 - \alpha) p'(x)] \\
= \frac{1}{2} \int \left( \frac{p(x)}{\alpha p(x) + (1 - \alpha) p'(x)} - 1 \right)^2 (\alpha p(x) + (1 - \alpha) p'(x)) \, dx.
\]

We estimate the \( \alpha \)-relative divergence by direct approximation of the \( \alpha \)-relative density-ratio:

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

A notable advantage of this approach is that the \( \alpha \)-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 \( \alpha \)-relative PE divergence estimator based on \( \alpha \)-relative density-ratio
approximation is more favorable than the ordinary density-ratio approach in terms of nonparametric convergence speed.

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

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

The rest of this letter is structured as follows. In section 2, our proposed relative PE divergence estimator is described. In section 3, we provide nonparametric 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. In section 5, we conclude 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 $\{x_i\}_{i=1}^n$ from a $d$-dimensional distribution $P$ with density $p(x)$ and i.i.d. samples $\{x'_j\}_{j=1}^{n'}$ from another $d$-dimensional distribution $P'$ with density $p'(x)$:

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

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

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

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

Let $r_\alpha(x)$ be the $\alpha$-relative density ratio of $p(x)$ and $p'(x)$:

$$r_\alpha(x) := \frac{p(x)}{\alpha p(x) + (1 - \alpha) p'(x)} = \frac{p(x)}{q_\alpha(x)}. \quad (2.1)$$
We define the $\alpha$-relative PE divergence from $p(x)$ to $p'(x)$ as
\[
\text{PE}_\alpha := \frac{1}{2} \mathbb{E}_{q_{\alpha}(x)} \left[ (r_{\alpha}(x) - 1)^2 \right],
\] (2.2)
where $\mathbb{E}_{p(x)}[f(x)]$ denotes the expectation of $f(x)$ under $p(x)$:
\[
\mathbb{E}_{p(x)}[f(x)] = \int f(x)p(x)dx.
\]
When $\alpha = 0$, $\text{PE}_\alpha$ is reduced to the ordinary PE divergence. Thus, the $\alpha$-relative PE divergence can be regarded as a smoothed extension of the ordinary PE divergence.

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

2.2 Direct Approximation of $\alpha$-Relative Density Ratios. Here, we describe a method for approximating the $\alpha$-relative density ratio, equation 2.1.

We model the $\alpha$-relative density ratio $r_{\alpha}(x)$ by the following kernel model:
\[
g(x; \theta) := \sum_{\ell=1}^{n} \theta_{\ell} K(x, x_{\ell}),
\]
where $\theta := (\theta_{1}, \ldots, \theta_{n})^T$ are parameters to be learned from data samples, $^T$ denotes the transpose of a matrix or a vector, and $K(x, x')$ is a kernel basis function. In the experiments, we use the gaussian kernel,
\[
K(x, x') = \exp \left( - \frac{||x - x'||^2}{2\sigma^2} \right),
\] (2.3)
where $\sigma (> 0)$ is the kernel width.

The parameters $\theta$ in the model $g(x; \theta)$ are determined so that the following expected squared-error $J$ is minimized:
\[
J(\theta) := \frac{1}{2} \mathbb{E}_{q_{\alpha}(x)} [(g(x; \theta) - r_{\alpha}(x))^2]
\]
\[
= \frac{\alpha}{2} \mathbb{E}_{p(x)}[g(x; \theta)^2] + \frac{(1 - \alpha)}{2} \mathbb{E}_{p'(x)}[g(x; \theta)^2]
\]
\[-\mathbb{E}_{p(x)}[g(x; \theta)] + \text{Const.},
\]
where we used $r_\alpha(x)q_\alpha(x) = p(x)$ in the third term. Approximating the expectations by empirical averages, we obtain the following optimization problem:

$$\hat{\theta} := \arg\min_{\theta \in \mathbb{R}^n} \left[ \frac{1}{2} \theta^\top \hat{H} \theta - \hat{h}^\top \theta + \frac{\lambda}{2} \theta^\top \theta \right],$$  

(2.4)

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

$$\hat{H}_{\ell, \ell'} := \frac{\alpha}{n} \sum_{i=1}^{n} K(x_i, x_{\ell})K(x_i, x_{\ell'}) + \frac{(1-\alpha)}{n'} \sum_{j=1}^{n'} K(x'_j, x_{\ell})K(x'_j, x_{\ell'}).$$  

(2.5)

$\hat{h}$ is the $n$-dimensional vector with the $\ell$th element,

$$\hat{h}_\ell := \frac{1}{n} \sum_{i=1}^{n} K(x_i, x_{\ell}).$$

It is easy to confirm that the solution of equation 2.4 can be analytically obtained as

$$\hat{\theta} = (\hat{H} + \lambda I_n)^{-1} \hat{h},$$

where $I_n$ denotes the $n$-dimensional identity matrix. Finally, a density-ratio estimator is given as

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

(2.6)

When $\alpha = 0$, the above method is reduced to a direct density-ratio estimator called unconstrained least-squares importance fitting (uLSIF; Kanamori et al., 2009). Thus, the method can be regarded as an extension of uLSIF to the $\alpha$-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 $\sigma$ in the case of the gaussian kernel) and the regularization parameter $\lambda$. 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 (Kanamori et al., 2009).
2.3 $\alpha$-Relative PE Divergence Estimation Based on RuLSIF. Using an estimator of the $\alpha$-relative density ratio $r_\alpha(x)$, we can construct estimators of the $\alpha$-relative PE divergence, equation 2.2. After a few lines of calculation, we can show that the $\alpha$-relative PE divergence, equation 2.2, is equivalently expressed as

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

$$= \frac{1}{2} \mathbb{E}_{p(x)} \left[ r_\alpha(x) \right] - \frac{1}{2}.$$  

Note that the first line can also be obtained by Legendre-Fenchel convex duality of the divergence functional (Rockafellar, 1970).

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

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

$$\widetilde{PE}_\alpha := \frac{1}{2n} \sum_{i=1}^{n} \widehat{r}(x_i) - \frac{1}{2}.$$  

We note that the $\alpha$-relative PE divergence, equation 2.2, 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 $\widehat{PE}_\alpha$ has superior theoretical properties (see section 3), and the second one, $\widetilde{PE}_\alpha$, is simple to compute.

2.4 Illustrative Examples. Here, we numerically illustrate the behavior of RuLSIF, equation 2.6, using toy data sets. Let the numerator distribution be $P = N(0, 1)$, where $N(\mu, \sigma^2)$ denotes the normal distribution with mean $\mu$ and variance $\sigma^2$. 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 $\alpha = 0, 0.5, \text{and } 0.95$.

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 $\alpha$-relative density ratios get smoother as $\alpha$ increases. In particular, in data sets b and d,
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 $\alpha = 0, 0.5, \text{and} 0.95$.

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 $\alpha$ gets large, the estimation quality of RuLSIF tends to be improved since the complexity of the true density-ratio functions is reduced.
Note that in dataset \( a \) where, \( p(x) = p'(x) \), the true density ratio \( r_\alpha(x) \) does not depend on \( \alpha \) since \( r_\alpha(x) = 1 \) for any \( \alpha \). However, the estimated density ratios still depend on \( \alpha \) through the matrix \( \hat{H} \) (see equation 2.5).

### 3 Theoretical Analysis

In this section, we analyze the theoretical properties of the proposed PE divergence estimators. More specifically, we provide nonparametric 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 the practical insights we can gain from the theoretical results here. We describe all the mathematical details of the nonparametric 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

\[
\arg\min_{g \in \mathcal{G}} \left[ \frac{\alpha}{2n} \sum_{i=1}^{n} g(x_i)^2 + \frac{(1-\alpha)}{2n'} \sum_{j=1}^{n'} g(x'_j)^2 - \frac{1}{n} \sum_{i=1}^{n} g(x_i) + \frac{\lambda}{2} R(g)^2 \right],
\]

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

#### 3.1 Nonparametric Convergence Analysis

First, we elucidate the nonparametric 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; Aronszajn, 1950) such as the gaussian kernel space and \( R(\cdot) \) as the associated RKHS norm.

**3.1.1 Theoretical Results.** We represent the complexity of the function space \( \mathcal{G} \) by \( \gamma (0 < \gamma < 2) \); the larger \( \gamma \) is, the more complex the function class \( \mathcal{G} \) is (see appendix A 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 appendix A, we have the following asymptotic convergence results for the two PE divergence estimators $\hat{\text{PE}}_\alpha$, equation 2.7, and $\tilde{\text{PE}}_\alpha$, equation 2.8:

$$\hat{\text{PE}}_\alpha - \text{PE}_\alpha = O_p(\bar{n}^{-1/2}c\|r\|_\infty + \lambda \max(1, R(r)^2))$$  \hspace{1cm} (3.2)

and

$$\tilde{\text{PE}}_\alpha - \text{PE}_\alpha = O_p(\lambda^{1/2} \|r\|_\infty^{1/2} \max(1, R(r))) + \lambda \max(1, \|r\|_\infty^{(1-\gamma/2)/2}, R(r)\|r\|_\infty^{(1-\gamma/2)/2}, R(r)))$$ \hspace{1cm} (3.3)

where $O_p$ denotes the asymptotic order in probability,

$$c := (1 + \alpha)\sqrt{\nabla p(x)[r(x)]} + (1 - \alpha)\sqrt{\nabla p'(x)[r(x)]},$$ \hspace{1cm} (3.4)

and $\nabla p(x)[f(x)]$ denotes the variance of $f(x)$ under $p(x)$:

$$\nabla p(x)[f(x)] = \int \left( f(x) - \int f(x)p(x)dx \right)^2 p(x)dx.$$

3.1.2 Interpretation. In both equations 3.2 and 3.3, the coefficients of the leading terms (i.e., the first terms) of the asymptotic convergence rates become smaller as $\|r\|_\infty$ gets smaller. Since

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

larger $\alpha$ would be more preferable in terms of the asymptotic approximation error. Note that when $\alpha = 0$, $\|r\|_\infty$ can tend to infinity even under a simple setting that the ratio of two gaussian functions is considered (Cortes et al., 2010; see also the numerical examples in section 2.4 of this letter). Thus, our proposed approach of estimating the $\alpha$-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 nonparametric convergence rate.

The above results also show that $\hat{\text{PE}}_\alpha$ and $\tilde{\text{PE}}_\alpha$ have different asymptotic convergence rates. The leading term in equation 3.2 is of order $\bar{n}^{-1/2}$, while the leading term in equation 3.3 is of order $\lambda^{1/2}$, which is slightly slower (depending on the complexity $\gamma$) than $\bar{n}^{-1/2}$. Thus, $\hat{\text{PE}}_\alpha$ would be more accurate than $\tilde{\text{PE}}_\alpha$ in large sample cases. Furthermore, when $p(x) = p'(x)$, $\nabla p(x)[r(x)] = 0$ holds, and thus $c = 0$ holds (see equation 3.4). Then the
leading term in equation 3.2 vanishes, and therefore \( \hat{\text{PE}}_\alpha \) has an even faster convergence rate of order \( \lambda \bar{n} \), which is slightly slower (depending on the complexity \( \gamma \)) than \( \bar{n}^{-1} \). Similarly, if \( \alpha \) is close to 1, \( r_\alpha(x) \approx 1 \) 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 \( \alpha \) is close to 0), the coefficient of the \( O(\lambda \bar{n}) \)-term in equation 3.2 can be larger than that in equation 3.3. Then \( \hat{\text{PE}}_\alpha \) would be more favorable than \( \tilde{\text{PE}}_\alpha \) in terms of approximation accuracy.

3.1.3 Numerical Illustration. Let us numerically investigate the above interpretation using the same artificial data set as in section 2.4.

Figure 2 shows the mean and standard deviation of \( \hat{\text{PE}}_\alpha \) and \( \tilde{\text{PE}}_\alpha \) over 100 runs for \( \alpha = 0, 0.5, \) and \( 0.95 \), as functions of \( n (= n' \) in this experiment). The true \( \text{PE}_\alpha \) (which was numerically computed) is also plotted in the graphs. The graphs show that both estimators, \( \hat{\text{PE}}_\alpha \) and \( \tilde{\text{PE}}_\alpha \), approach the true \( \text{PE}_\alpha \) as the number of samples increases, and the approximation error tends to be smaller if \( \alpha \) is larger.

When \( \alpha \) is large, \( \hat{\text{PE}}_\alpha \) tends to perform slightly better than \( \tilde{\text{PE}}_\alpha \). When \( \alpha \) is small and the number of samples is small, \( \tilde{\text{PE}}_\alpha \) slightly compares favorably with \( \hat{\text{PE}}_\alpha \). Overall, these numerical results agree well with our theory.

3.2 Parametric Variance Analysis. Next, we analyze the asymptotic variance of the PE divergence estimator \( \hat{\text{PE}}_\alpha \), equation 2.7, under a parametric setup.

3.2.1 Theoretical Results. As the function space \( \mathcal{G} \) in equation 3.1, we consider the following parametric model:

\[
\mathcal{G} = \{g(x; \theta) | \theta \in \Theta \subset \mathbb{R}^b\},
\]

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

\[
g(x; \theta^*) = r_\alpha(x).
\]

Here, we use RuLSIF without regularization, that is, \( \lambda = 0 \) in equation 3.1.

Let us denote the variance of \( \hat{\text{PE}}_\alpha \), equation 2.7 by \( \text{Var}[\hat{\text{PE}}_\alpha] \), where randomness comes from the draw of samples \( \{x_i\}_{i=1}^n \) and \( \{x'_j\}_{j=1}^n \). Then, under a standard regularity condition for the asymptotic normality (see section 3
Relative Density-Ratio Estimation for Robust Distribution Comparison 1335

(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 $0.95$. 
of van der Vaart, 2000), \(\mathbb{V}[\hat{PE}_\alpha]\) can be expressed and upper-bounded as

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

(3.5)

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

(3.6)

Let us denote the variance of \(\hat{PE}_\alpha\) by \(\mathbb{V}[\hat{PE}_\alpha]\). Then, under a standard regularity condition for the asymptotic normality (see section 3 of van der Vaart, 2000), the variance of \(\hat{PE}_\alpha\) is asymptotically expressed as

\[
\mathbb{V}[\hat{PE}_\alpha] = \frac{1}{n} \mathbb{V}_{p(x)} \left[ r_\alpha + \frac{(1 - \alpha) r_\alpha(x)^2}{2} \right]
\]

\[
+ \frac{1}{n'} \mathbb{V}_{p'(x)} \left[ \frac{(1 - \alpha) r_\alpha E_p(x) [\nabla g] U_\alpha^{-1} \nabla g]}{2} \right] + o\left( \frac{1}{n}, \frac{1}{n'} \right),
\]  

(3.7)

where \(\nabla g\) is the gradient vector of \(g\) with respect to \(\theta\) at \(\theta = \theta^*\), that is,

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

The matrix \(U_\alpha\) is defined by

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

3.2.2 Interpretation. Equation 3.5 shows that up to \(O\left( \frac{1}{n}, \frac{1}{n'} \right)\), the variance of \(\hat{PE}_\alpha\) depends on only the true relative density ratio \(r_\alpha(x)\), not on the estimator of \(r_\alpha(x)\). 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, equation 3.7 shows that the variance of \(\hat{PE}_\alpha\) is affected by the model \(\mathcal{G}\), since the factor \(\mathbb{E}_{p(x)}[\nabla g] U_\alpha^{-1} \nabla g\) depends on the model complexity in general. When the equality

\[
\mathbb{E}_{p(x)}[\nabla g] U_\alpha^{-1} \nabla g(x; \theta^*) = r_\alpha(x)
\]
holds, the variances of $\hat{\text{PE}}_\alpha$ and $\tilde{\text{PE}}_\alpha$ are asymptotically the same. However, in general, the use of $\hat{\text{PE}}_\alpha$ would be more recommended.

Equation 3.6 shows that the variance $\mathbb{V}[\hat{\text{PE}}_\alpha]$ 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 $\alpha$ increases, our proposed approach of estimating the $\alpha$-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 parametric asymptotic variance.

3.2.3 Numerical Illustration. Here, we show some numerical results for illustrating the above theoretical results using the one-dimensional data sets b and c in section 2.4. Let us define the parametric model as

$$G_k = \left\{ g(x; \theta) = \frac{r(x; \theta)}{\alpha r(x; \theta) + 1 - \alpha} \right\} \text{ where } r(x; \theta) = \exp \left( \sum_{\ell=0}^{k} \theta_{\ell} x^{\ell} \right), \theta \in \mathbb{R}^{k+1} \right\}. \quad (3.8)$$

The dimension of model $G_k$ is equal to $k + 1$. The $\alpha$-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 $G_k$ includes the true relative density ratio $r_\alpha(x)$ of data sets b and c. We test RuLSIF with $\alpha = 0.2$ and $0.8$ for model 3.8 with degree $k = 1, 2, \ldots, 8$. The parameter $\theta$ is learned so that equation 3.1 is minimized by a quasi-Newton method.

The standard deviations of $\hat{\text{PE}}_\alpha$ and $\tilde{\text{PE}}_\alpha$ for data sets b and c are depicted in Figures 3 and 4, respectively. The graphs show that the degree of models does not significantly affect the standard deviation of $\hat{\text{PE}}_\alpha$ (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 $\tilde{\text{PE}}_\alpha$. Thus, the standard deviation of $\hat{\text{PE}}_\alpha$ more strongly depends on model complexity.

4 Experiments

In this section, we experimentally evaluate the performance of the proposed method in a 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 $\mathcal{X} = \{x_i\}_{i=1}^n \text{i.i.d. } P$ and $\mathcal{X'} = \{x'_j\}_{j=1}^{n'} \text{i.i.d. } P'$, 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 $\widehat{\text{Div}}$ of some divergence between the two distributions $P$ and $P'$, the homogeneity of two distributions can be tested based on the permutation test procedure (Efron & Tibshirani, 1993) as follows:

- Obtain a divergence estimate $\widehat{\text{Div}}$ using the original data sets $\mathcal{X}$ and $\mathcal{X'}$.
- Randomly permute the $|\mathcal{X} \cup \mathcal{X'}|$ samples, and assign the first $|\mathcal{X}|$ samples to a set $\tilde{\mathcal{X}}$ and the remaining $|\mathcal{X'}|$ samples to another set $\tilde{\mathcal{X'}}$.
Relative Density-Ratio Estimation for Robust Distribution Comparison

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

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

When an asymmetric divergence such as the KL divergence (Kullback & Leibler, 1951) or PE divergence (Pearson, 1900) 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 $P$. Sugiyama et al. (2011) proposed choosing 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 type 2 errors (the probability of accepting incorrect null hypotheses: two distributions are judged to be the same when they are actually different), while the increase of type 1 errors (the probability of rejecting correct null hypotheses: two distributions are judged to be different when they are actually the same) is kept moderate.

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

4.1.2 Artificial Data Sets. We illustrate how the proposed method behaves in two-sample homogeneity test scenarios using artificial data sets a to d described in section 2.4. We test the plain LSTT, reciprocal LSTT, and adaptive LSTT for $\alpha = 0, 0.5, 0.95$, with a significance level of 5%.

The experimental results are shown in Figure 5. For data set 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 type 1 error is properly controlled in these methods. The adaptive LSTT tends to give slightly lower acceptance rates than 95% for this toy data set, 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 a critical influence on type 1 errors.

In data sets b, c, and d, $P$ is different from $P'$ (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 data set b and the reciprocal setup for data set c, the true density-ratio functions with $\alpha = 0$ diverge to infinity, and thus larger $\alpha$ 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 distinguish from the completely constant density-ratio function (which corresponds to $P = P'$). In the reciprocal setup for data set b and the plain setup for data set e, small $\alpha$ performs poorly since density-ratio functions with large $\alpha$ can more accurately be approximated than those with small $\alpha$ (see Figure 1). In the adaptive setup, large $\alpha$ tends to perform slightly better than small $\alpha$ for data sets b and c.

In data set d, the true density-ratio function with $\alpha = 0$ diverges to infinity for both the plain and reciprocal setups. In this case, middle $\alpha$ 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 $\alpha$ works well can also be mildly observed in the adaptive LSTT for data set d.

Overall, if the plain LSTT (or the reciprocal LSTT) is used, small $\alpha$ (or large $\alpha$) sometimes works excellently. However, it performs poorly in other cases, and thus the performance is unstable depending on the true distributions. The plain LSTT (or the reciprocal LSTT) with middle $\alpha$ tends to perform
Figure 5: Illustrative examples of the 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.

reasonably well for all data sets. On the other hand, the adaptive LSTT was shown to nicely overcome the instability problem when \(\alpha\) is small or large. However, when \(\alpha\) is set to be a middle value, both the plain LSTT and the reciprocal LSTT give similar results, and thus the adaptive LSTT provides only small improvement.

Our empirical finding is that if we have prior knowledge that one distribution has wider support than the other distribution, assigning the distribution with a wider support to \(P'\) and setting \(\alpha\) to be a large value seem to work well. If there is no knowledge on the true distributions or two
Table 1: Experimental Results of Two-Sample Homogeneity Test for the IDA Data Sets.

| Data Sets     | d | n = n' | MMD   | LSTT ($\alpha = 0.0$) | LSTT ($\alpha = 0.5$) | LSTT ($\alpha = 0.95$) |
|---------------|---|--------|-------|-----------------------|-----------------------|-----------------------|
| 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) |

Notes: The mean (and standard deviations in parentheses) rate of accepting the null hypothesis (i.e., $P = P'$) under significance level 5% is reported. The two sets of samples are 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 in boldface.

distributions have fewer overlapped supports, using middle $\alpha$ in the adaptive setup seems to be a reasonable choice.

We will systematically investigate this issue using more complex data sets.

4.1.3 Benchmark Data Sets. Here, we apply the proposed two-sample homogeneity test to the binary classification data sets taken from the IDA repository (Rätsch, Onoda, & Müller, 2001).

We test the adaptive LSTT with the RuLSIF-based PE divergence estimator for $\alpha = 0, 0.5, 0.95$; we also test the maximum mean discrepancy (MMD; Borgwardt et al., 2006), 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 Sriperumbudur, Fukumizu, Gretton, Lanckriet, and Schölkopf (2009), 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 the positive training samples into two sets and perform a 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 type 1 error.
Table 2: Experimental Results of Two-Sample Homogeneity Test for the IDA Data Sets.

| Data Sets   | $d$ | $n = n'$ | MMD $\alpha = 0.0$ | LSTT $\alpha = 0.5$ | LSTT $\alpha = 0.95$ |
|-------------|-----|----------|---------------------|---------------------|---------------------|
| 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)         |

Notes: The mean (and standard deviations in parentheses) rate of accepting the null hypothesis (i.e., $P = P'$) under 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 significance level 5% are in boldface.

Next, we consider the situation where the null hypothesis is not correct (the two distributions are different). The numerator samples are generated in the same way as above, but half of the 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 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 type 2 error.

Overall, LSTT with $\alpha = 0.5$ is shown to be a useful method for a 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 data set (called an evaluation data set) based on another data set (called a model dataset) that contains only 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 deviate...
Table 3: Mean AUC Score (and Standard Deviations in Parentheses) over 1000 Trials for the Artificial Outlier-Detection Data Set.

| 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)               |

Note: 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 in boldface.

significantly from one. Thus, density-ratio values could be used as an index of the degree of outlyingness (Smola et al., 2009; Hido et al., 2011).

Since the evaluation data set usually has wider support than the model data set, we regard the evaluation data set as samples corresponding to the denominator density $p'(x)$ and the model data set as samples corresponding to the numerator density $p(x)$. 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 number of true outliers an outlier detection algorithm can find) and detection accuracy (i.e., the number of true inliers an outlier detection algorithm misjudges as outliers). Since there is a trade-off between detection rate and detection accuracy, we adopt the area under the ROC curve (AUC) as our error metric (Bradley, 1997).

4.2.2 Artificial Data Sets. First, we illustrate how the proposed method behaves in outlier detection scenarios using artificial data sets.

Let

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

where $d$ is the dimensionality of $x$ and $1_d$ is the $d$-dimensional vector with all one. Note that this setup is the same as data set e described in section 2.4 when $d = 1$. Here, the samples drawn from $N(0, I_d)$ are regarded as inliers, while the samples drawn from $N(d^{-1/2}1_d, I_d)$ 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 $\alpha$ tends to work well when the input dimensionality is low, and RuLSIF with large $\alpha$ works better as the input dimensionality increases. This tendency can be interpreted as follows: if $\alpha$ is small, the density-ratio function tends to have a sharp hollow for outlier points (see the left-most graph in Figure 2e). Thus, as long as the true density-ratio function can be accurately estimated, small $\alpha$ would be preferable in the outlier detection. When data dimensionality is low, density-ratio approximation is rather easy, and thus small $\alpha$ tends to perform well. However, as the data dimensionality increases, density-ratio approximation gets harder, and thus large $\alpha$, which produces a smoother density-ratio function, is more favorable, since a smoother function can be more easily approximated than a bumpy one produced by small $\alpha$.

4.2.3 Real-World Data Sets. Next, we evaluate the proposed outlier detection method using various real-world data sets:

- **IDA repository**: The IDA repository (Rätsch et al., 2001) contains various binary classification tasks. Each data set consists of positive or negative and training or 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 data set**: An in-house speech data set, which contains short utterance samples recorded from two male subjects speaking in French with a sampling rate of 44.1 kHz. From each utterance sample, we extracted a 50-dimensional line spectral frequencies vector (Kain & Macon, 1998). We randomly take 200 samples from one class and assign them to the model data set. Then we randomly take 200 samples from the same class and 10 samples from the other class.

- **20 Newsgroup data set**: The 20-Newsgroups data set (http://people.csail.mit.edu/jrennie/20Newsgroups/) contains 20,000 newsgroup documents, which contains the following four top-level categories: comp, rec, sci, and talk. Each document 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 data set. Then we randomly take 200 samples from the same class and 10 samples from one of the other classes for the evaluation data set.

- **USPS handwritten digit data set**: The USPS handwritten digit data set (http://www.gaussianprocess.org/gpml/data/) contains 9298 digit images. Each image consists of 256 ($=16 \times 16$) 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 data set. Then we randomly take 200 samples from the same inlier class and 10 samples from one of the other classes for the evaluation data set.

We compare the AUC scores of RuLSIF with $\alpha = 0, 0.5, \text{ and } 0.95$ and one-class support vector machine (OSVM) with the gaussian kernel (Schölkopf, Platt, Shawe-Taylor, Smola, & Williamson, 2001). We used the LIBSVM implementation of OSVM (Chang & Lin, 2001). The gaussian width is set to the median distance between samples, which has been shown to be a useful heuristic (Schölkopf et al., 2001). Since there is no systematic method to determine the tuning parameter $\nu$ in OSVM, we report the results for $\nu = 0.05 \text{ and } 0.1$.

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 $\alpha$ tends to perform well for low-dimensional data sets, and large $\alpha$ tends to work well for high-dimensional data sets. This tendency agrees well with that for the artificial data sets (see section 4.2.2).

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

4.3.1 Transductive Transfer Learning by Importance Sampling. Let us consider a problem of semisupervised learning (Chapelle, Schölkopf, & Zien, 2006) from labeled training samples $\{(x_{tr}^j, y_{tr}^j)\}_{j=1}^{n_{tr}}$ and unlabeled test samples $\{x_{te}^i\}_{i=1}^{n_{te}}$. The goal is to predict a test output value $y_{te}$ for a test input point $x_{te}$. Here, we consider the setup where the labeled training samples $\{(x_{tr}^j, y_{tr}^j)\}_{j=1}^{n_{tr}}$ are drawn i.i.d. from $p(y|x)p_{tr}(x)$, while the unlabeled test samples $\{x_{te}^i\}_{i=1}^{n_{te}}$ are drawn i.i.d. from $p_{te}(x)$, which is generally different from $p_{tr}(x)$; the (unknown) test sample $(x_{te}^i, y_{te}^i)$ follows $p(y|x)p_{te}(x)$. This setup means that the conditional probability $p(y|x)$ is common to training and test samples, but the marginal densities $p_{tr}(x)$ and $p_{te}(x)$ are generally different for training and test input points. Such a problem is called transductive transfer learning (Pan & Yang, 2010), domain adaptation (Jiang & Zhai, 2007), or covariate shift (Shimodaira, 2000; Sugiyama & Kawanabe, 2012).

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

$$E_{p(y|x)p_{te}(x)}[\text{loss}(y, f(x))],$$
Table 4: Experimental Results of Outlier Detection for Various Real-World Data Sets.

| Data Sets | 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) | .849(.149) | .769(.134) | .726(.127) |
| 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) |

Notes: Mean AUC score (and standard deviations in parentheses) 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 in boldface. The data sets are sorted in ascending order of the input dimensionality $d$.

where $f(x)$ is a function model. Since the generalization error is inaccessible because the true probability $p(y|x)p_{te}(x)$ is unknown, empirical error minimization is often used in practice (Vapnik, 1998):

$$
\min_{f \in \mathcal{F}} \left[ \frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \text{loss}(y_{tr}^j, f(x_{tr}^j)) \right].
$$

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 Shimodaira, 2000). Instead, the following importance-weighted
empirical error minimization is consistent under covariate shift:

$$\min_{f \in \mathcal{F}} \left[ \frac{1}{n_{\text{tr}}} \sum_{j=1}^{n_{\text{u}}} r(x_{\text{tr}}^j) \text{loss}(y_{\text{tr}}^j, f(x_{\text{tr}}^j)) \right],$$

where $r(x)$ is called the importance (Fishman, 1996) in the context of covariate shift adaptation:

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

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. Shimodaira (2000) proposed using the exponentially flattened importance weights as

$$\min_{f \in \mathcal{F}} \left[ \frac{1}{n_{\text{tr}}} \sum_{j=1}^{n_{\text{u}}} r(x_{\text{tr}}^j)^\tau \text{loss}(y_{\text{tr}}^j, f(x_{\text{tr}}^j)) \right],$$

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 $\tau$ can be optimized by model selection criteria such as the importance-weighted Akaike information criterion for regular models (Shimodaira, 2000), the importance-weighted subspace information criterion for linear models (Sugiyama & Müller, 2005), and importance-weighted cross-validation for arbitrary models (Sugiyama et al., 2007).

One of the potential drawbacks of the above exponential flattening approach is that estimation of $r(x)$ (i.e., $\tau = 1$) is rather hard, as shown in this letter. Thus, when $r(x)$ is estimated poorly, all flattened weights $r(x)^\tau$ 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,

$$\min_{f \in \mathcal{F}} \left[ \frac{1}{n_{\text{tr}}} \sum_{j=1}^{n_{\text{u}}} r_\alpha(x_{\text{tr}}^j) \text{loss}(y_{\text{tr}}^j, f(x_{\text{tr}}^j)) \right],$$

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

$$r_\alpha(x) := \frac{p_{\text{te}}(x)}{(1 - \alpha) p_{\text{te}}(x) + \alpha p_{\text{tr}}(x)}.$$
Note that compared with the definition of the $\alpha$-relative density-ratio (1), $\alpha$ 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 $\alpha$-relative importance weights for $0 < \alpha < 1$ can be reliably estimated by RuLSIF proposed in this letter, the performance of covariate shift adaptation is expected to be improved. Below, we experimentally investigate this effect.

4.3.2 Artificial Data Sets. First, we illustrate how the proposed method behaves in covariate shift adaptation using one-dimensional artificial data sets.

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

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

where $\beta = (\beta_1, \ldots, \beta_n)^T$ is the parameter to be learned and $\rho$ is the gaussian width. The parameter $\beta$ is learned by relative importance-weighted least squares (RIW-LS):

$$\hat{\beta}_{RIW-LS} = \arg\min_{\beta} \left[ \frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \hat{r}_\alpha (x^{tr}_j) \left( f(x^{tr}_j; \beta) - y^{tr}_j \right)^2 \right],$$

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

$$\hat{\beta}_{EIW-LS} = \arg\min_{\beta} \left[ \frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \hat{r}(x^{tr}_j) \left( f(x^{tr}_j; \beta) - y^{tr}_j \right)^2 \right].$$

The relative importance weight $\hat{r}_\alpha (x^{tr}_j)$ is estimated by RuLSIF, and the exponentially flattened importance weight $\hat{r}(x^{tr}_j)$ is estimated by uLSIF (i.e., RuLSIF with $\alpha = 1$). The gaussian width $\rho$ is chosen by five-fold importance-weighted cross-validation (Sugiyama et al., 2007).
Figure 6: Illustrative example of transfer learning under no distribution change.

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

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

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

$$y_{tr}^{(i)} = \text{sinc}(x_{tr}^{(i)}) + \epsilon_{tr}^{(i)},$$

where \(\{\epsilon_{tr}^{(i)}\}_{i=1}^{n_{tr}}\) is additive noise following \(N(0, 0.01)\). We set \(n_{tr} = 100\) and \(n_{te} = 200\). Figure 6b 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 6c shows the mean and standard deviation of the test error under the squared loss over 200 runs as functions of the relative flattening parameter \(\alpha\) in RIW-LS and the exponential flattening parameter \(\tau\) 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 (see Figure 7a):

$$P_{tr} = N(1, 0.25),$$

$$P_{te} = N(2, 0.1).$$

The output values are created in the same way as the previous case. Figure 7b 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 \(\alpha = 0.5\) fits the true function slightly better than
EIW-LS with $\tau = 0.5$ in the test region. Figure 7c 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 Data Sets. 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 (http://alkan.mns.kyutech.ac.jp). In the data collection procedure, subjects were asked to perform a specific action such as walking, running, or bicycle riding. The duration of each task was arbitrary, and the sampling rate was 20 Hz with small variations. An example of three-axis accelerometer data for “walking” is plotted in Figure 8.
Table 5: Experimental Results of Transfer Learning in Real-World Human Activity Recognition.

| 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)                     |

Notes: Mean classification accuracy (and standard deviations in parentheses) 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 in bold.

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 $\ell_2$-norm of the three-dimensional acceleration vector at each time step and computed the following five orientation-invariant features from each window: mean, standard deviation, fluctuation of amplitude, average energy, and frequency-domain entropy (Bao & Intille, 2004; Bharatula, Stager, Lukowicz, & Troster, 2005).

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 go to the trouble of labeling his or her accelerometer data, no labeled sample is available for the new user. However, unlabeled samples for the new user and labeled data obtained from existing users are available. Let labeled training data $\{(x_{tr}^j, y_{tr}^j)\}_{j=1}^{n_tr}$ 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 $\{x_{te}^i\}_{i=1}^{n_te}$ 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 $\tau = 0.5$
- KLR with plain importance weights (i.e., $\alpha = 1$ or $\tau = 1$)

The experiments are repeated 100 times with different sample choice for $n_{tr} = 500$ and $n_{te} = 200$. Table 5 depicts the classification accuracy for three binary-classification tasks: walk versus run, walk versus riding a bicycle, and walk versus 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 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 versus 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 in transfer learning under covariate shift.

5 Conclusion

In this letter, we proposed using 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 a nonparametric setup, which showed that the proposed approach of estimating the relative Pearson divergence is preferable to the existing approach of estimating the plain Pearson divergence. Furthermore, we proved that the asymptotic variance of the proposed divergence estimator is independent of 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 a two-sample homogeneity test, inlier-based outlier detection, and transductive transfer learning under covariate shift.

In addition to the two-sample homogeneity test, outlier detection, and transfer learning, density ratios were shown to be useful for tackling various machine learning problems, including multitask learning (Bickel, Bogojeska, Lengauer, & Scheffer, 2008; Simm, Sugiyama, & Kato, 2011), independence test (Sugiyama & Suzuki, 2011), feature selection (Suzuki, Sugiyama, Kanamori, & Sese, 2009), causal inference (Yamada & Sugiyama, 2010), independent component analysis (Suzuki & Sugiyama, 2011), dimensionality reduction (Suzuki & Sugiyama, 2010), unpaired data matching (Yamada & Sugiyama, 2011), clustering (Kimura & Sugiyama, 2011), conditional density estimation (Sugiyama et al., 2010), and probabilistic classification (Sugiyama, 2010). 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.

Appendix A: Technical Details of Nonparametric Convergence Analysis

Here, we give the technical details of the nonparametric convergence analysis described in section 3.1.
A.1 Results. For notational simplicity, we define linear operators $P, P_n, P', P'_n$ as

$$P f := \mathbb{E} p f, \quad P_n f := \frac{\sum_{j=1}^n f(x_j)}{n},$$

$$P' f := \mathbb{E} q f, \quad P'_n f := \frac{\sum_{i=1}^{n'} f(x'_i)}{n'}.$$  

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:

$$\hat{g} := \arg\min_{g \in \mathcal{G}} \left[ \frac{1}{2} \left( \alpha P_n + (1 - \alpha)P'_n \right) g^2 - P_n g + \frac{\lambda_{\bar{n}}}{2} R(g)^2 \right],$$

$$= \arg\min_{g \in \mathcal{G}} \left( \frac{1}{2} S_{n,n'} \hat{g}^2 - P_n \hat{g} + \frac{\lambda_{\bar{n}}}{2} R(g)^2 \right),$$

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

$$\sup_x |g(x)| \leq R(g). \tag{A.1}$$

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

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

Another possibility is

$$\tilde{PE}_\alpha := \frac{1}{2} P_n \hat{g} - \frac{1}{2}.$$

A useful example is to use a reproducing kernel Hilbert space (RKHS; Aronszajn, 1950) as $\mathcal{G}$ and the RKHS norm as $R(g)$. Suppose $\mathcal{G}$ is an RKHS
associated with bounded kernel \( k(\cdot, \cdot) \):

\[
\sup_x [k(x, x)] \leq C.
\]

Let \( \| \cdot \|_G \) denote the norm in the RKHS \( G \). Then \( R(g) = \sqrt{C} \| g \|_G \) satisfies equation A.1:

\[
g(x) = \langle k(x, \cdot), g(\cdot) \rangle \leq \sqrt{k(x, x)} \| g \|_G \leq \sqrt{C} \| g \|_G,
\]

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 \( R(g) = \| g \|_G \) imposes the solution to be smooth. The RKHS technique in Sobolev space has been well exploited in the context of spline models (Wahba, 1990). 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 \( G \).

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

\[
g^*(x) \leq M_0 \text{ for all } x \in \mathcal{D}_X.
\]

Let \( G_M \) be a ball of \( G \) with radius \( M > 0 \):

\[
G_M := \{ g \in G \mid R(g) \leq M \}.
\]

To derive the convergence rate of our estimator, we utilize the bracketing entropy that is a complexity measure of a function class (van der Vaart & Wellner, 1996).

**Definition 1.** Given two functions \( l \) and \( u \), the bracket \( [l, u] \) is the set of all functions \( f \) with \( l(x) \leq f(x) \leq u(x) \) for all \( x \). An \( \epsilon \)-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 \( \epsilon \)-brackets with respect to \( L_2(\tilde{p}) \) needed to cover a function set \( \mathcal{F} \).

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

\[
\mathcal{H}_{[]}([G_M, \epsilon, L_2(p)]) = O \left( \left( \frac{M}{\epsilon} \right)^\gamma \right), \mathcal{H}_{[]}([G_M, \epsilon, L_2(p^\prime)]) = O \left( \left( \frac{M}{\epsilon} \right)^\gamma \right).
\]

(A.2)
This quantity represents a complexity of function class \( \mathcal{G} \)—the larger \( \gamma \) is, the more complex the function class \( \mathcal{G} \) is because, for larger \( \gamma \), more brackets are needed to cover the function class. The gaussian RKHS satisfies this condition for arbitrarily small \( \gamma \) (Steinwart & Scovel, 2007). Note that when \( R(g) \) is the RKHS norm, condition A.2 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 \( c = (1 + \alpha)\sqrt{P(g^* - P g^*)^2} + (1 - \alpha)\sqrt{P'(g^* - P' g^*)^2} \). Under the above setting, if \( \lambda_n \to 0 \) and \( \lambda_n^{-1} = o(\bar{n}^{2/(2+\gamma)}) \), then we have

\[
\widehat{P}_E \alpha - P E_\alpha = O_p(\lambda_n \max(1, R(g^*)) + \bar{n}^{-1/2}cM_0)
\]

and

\[
\widehat{P}_E \alpha - P E_\alpha = O_p(\lambda_n \max\{1, M_0^{\frac{1}{2}}(1-\frac{\gamma}{2}), R(g^*)M_0^{\frac{1}{2}}(1-\frac{\gamma}{2}), R(g^*)\} + \lambda_n \frac{1}{2} \max\{M_0^{\frac{1}{2}}, \bar{M} g^*(R(g^*))\}).
\]

where \( O_p \) 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 \( \lambda_n \to 0 \) and \( \lambda_n^{-1} = o(\bar{n}^{2/(2+\gamma)}) \), then we have

\[
\| \hat{g} - g^* \|_{L_2(S)} = O_p(\lambda_n^{1/2} \max(1, R(g^*))), \quad R(\hat{g}) = O_p(\max(1, R(g^*))).
\]

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

\[
\frac{1}{2} S_{n,n'}(\hat{g} - g^*)^2 - P_{n,h} S_{n,n'}(\hat{g} - g^*)^2 + \lambda_n R(\hat{g})^2 \leq \frac{1}{2} S_{n,n'}(g^* - P_{n,h} g^*)^2 + \lambda_n R(g^*)^2
\]

\[
\Rightarrow \frac{1}{2} S_{n,n'}(\hat{g} - g^*)^2 - S_{n,n'}(g^* - P_{n} (\hat{g} - g^*)) - P_{n} (\hat{g} - g^*) + \lambda_n (R(\hat{g})^2 - R(g^*)^2) \leq 0.
\]

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

\[
\frac{1}{2} (S - S_{n,n'}) (\hat{g} - g^*)^2 - (S - S_{n,n'}) (g^* (g^* - \hat{g})) - (P - P_{n}) (\hat{g} - g^*)
\]

\[
- \lambda_n (R(\hat{g})^2 - R(g^*)^2) \geq \frac{1}{2} S(\hat{g} - g^*)^2.
\]
Therefore, to bound $\|\hat{g} - g^*\|_{L^2(S)}$, it suffices to bound the left-hand side of the above inequality.

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

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

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

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

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

This can be shown as follows. Let $f_L$ and $f_U$ be a $\delta$-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

$$f'_U(x) := \max\{f^2_L(x), f^2_U(x)\},$$

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

are also a bracket such that $f'_L \leq \hat{g} \leq f'_U$ 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:

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

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

$$\leq \begin{cases} (f_L(x) - f_U(x))^2(f_L(x) + f_U(x))^2, & \text{sign}(f_L(x)) = \text{sign}(f_U(x)), \\ (f_L(x) - f_U(x))^2(|f_L(x)| + |f_U(x)|)^2, & \text{(otherwise)} \end{cases},$$

$$\leq 4(f'_L(x) - f'_U(x))^2(M + M_0)^2.$$
Therefore, the condition for the bracketing entropies, equation A.2, gives \( \mathcal{H}(\mathcal{F}_M, \delta, \mathcal{L}_2(p)) = O((\frac{(M+M_0)^2}{\delta})^{\tilde{\nu}}) \). We can also show that \( \mathcal{H}(\mathcal{F}_M, \delta, \mathcal{L}_2(q)) = O((\frac{(M+M_0)^2}{\delta})^{\tilde{\nu}}) \) in the same fashion.

Let \( f := \hat{g} - \hat{g}^* \). Then, as in lemma 5.14 and theorem 10.6 in van de Geer (2000), we obtain

\[
| (S_{n,n'} - S)(f^2) | \leq \alpha | (P_n - P)(f^2) | + (1 - \alpha) | (P_{n'} - P')(f^2) | \\
= \alpha O_p \left( \frac{1}{\sqrt{n}} \| f^2 \|_{L_2(P)}^{1 - \frac{\gamma}{2}} (1 + R(\hat{g})^2 + M_0^2) \right) \land \bar{n}^{-\frac{2}{4\tilde{\nu}}} (1 + R(\hat{g})^2 + M_0^2) \\
+ (1 - \alpha) O_p \left( \frac{1}{\sqrt{n}} \| f^2 \|_{L_2(P')}^{1 - \frac{\gamma}{2}} (1 + R(\hat{g})^2 + M_0^2) \right) \land \bar{n}^{-\frac{2}{4\tilde{\nu}}} (1 + R(\hat{g})^2 + M_0^2) \\
\leq O_p \left( \frac{1}{\sqrt{n}} \| f^2 \|_{L_2(S)}^{1 - \frac{\gamma}{2}} (1 + R(\hat{g})^2 + M_0^2) \right) \land \bar{n}^{-\frac{2}{4\tilde{\nu}}} (1 + R(\hat{g})^2 + M_0^2) \right).
\]

(A.3)

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

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

by Jensen’s inequality for a concave function. Since

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

the right-hand side of equation A.3 is further bounded by

\[
| (S_{n,n'} - S)(f^2) | \\
= O_p \left( \frac{1}{\sqrt{n}} \| f \|_{L_2(S)}^{1 - \frac{\gamma}{2}} (1 + R(\hat{g})^2 + M_0^2) \right) \land \bar{n}^{-\frac{2}{4\tilde{\nu}}} (1 + R(\hat{g})^2 + M_0^2) \right).
\]

(A.4)

Similarly, we can show that

\[
| (S_{n,n'} - S)(g^*(\hat{g} - \hat{g}^*)) | \\
= O_p \left( \frac{1}{\sqrt{n}} \| f \|_{L_2(S)}^{1 - \frac{\gamma}{2}} (1 + R(\hat{g})M_0 + M_0^2) \right) \land \bar{n}^{-\frac{2}{4\tilde{\nu}}} (1 + R(\hat{g})M_0 + M_0^2) \right).
\]

(A.5)
Relative Density-Ratio Estimation for Robust Distribution Comparison

and

\[
|(P_n - P)(g^* - \hat{g})| = O_p \left( \frac{1}{\sqrt{n}} \|f\|_{L^2(P)}^{1 - \frac{z}{2}} (1 + R(\hat{g}) + M_0)^{\frac{z}{2}} \sqrt{n}^{-\frac{z}{2}} (1 + R(\bar{g}) + M_0) \right)
\]

\[
\leq O_p \left( \frac{1}{\sqrt{n}} \|f\|_{L^2(S)}^{1 - \frac{z}{2}} (1 + R(\bar{g}) + M_0)^{\frac{z}{2}} M_0^{1 - \frac{z}{2}} \sqrt{n}^{-\frac{z}{2}} (1 + R(\bar{g}) + M_0) \right).
\]

(A.6)

where we used

\[
\|f\|_{L^2(P)} = \sqrt{\int f^2 dP} = \sqrt{\int f^2 g^* dS} \leq M_0^{\frac{1}{2}} \sqrt{\int f^2 dS}
\]

in the last inequality. Combining equations A.4 to A.6, we can bound the \(L^2(S)\)-norm of \(f\) as

\[
\frac{1}{2} \|f\|_{L^2(S)}^2 + \lambda_\theta R(\bar{g})^2
\]

\[
\leq \lambda_\theta R(g^*)^2 + O_p \left( \frac{1}{\sqrt{n}} \|f\|_{L^2(S)}^{1 - \frac{z}{2}} (1 + R(\bar{g})^2 + M_0^2) \right)^{1 - \frac{z}{2}} \sqrt{n}^{-\frac{z}{2}} (1 + R(\bar{g}) + M_0).
\]

(A.7)

The following is similar to the argument in theorem 10.6 in van de Geer (2000), but we give a simpler proof.

By Young's inequality, we have \(a^{\frac{1}{2} - \frac{z}{4}} b^{\frac{1}{4} + \frac{z}{4}} \leq (\frac{1}{2} - \frac{z}{4})a + (\frac{1}{4} + \frac{z}{4})b \leq a + b\) for all \(a, b > 0\). Applying this relation to equation A.7, we obtain

\[
\frac{1}{2} \|f\|_{L^2(S)}^2 + \lambda_\theta R(\bar{g})^2
\]

\[
\leq \lambda_\theta R(g^*)^2 + O_p \left( \|f\|_{L^2(S)}^{\frac{1}{2} - \frac{z}{4}} (\tilde{n}^{-\frac{z}{4}} (1 + R(\bar{g})^2 + M_0^2))^{1 - \frac{z}{4}} \sqrt{n}^{-\frac{z}{4}} (1 + R(\bar{g})^2 + M_0) \right)^{1 - \frac{z}{4}} \sqrt{n}^{-\frac{z}{4}} (1 + R(\bar{g}) + M_0).
\]

Young

\[
\leq \lambda_\theta R(g^*)^2 + \frac{1}{4} \|f\|_{L^2(S)}^2 + O_p \left( \tilde{n}^{-\frac{z}{2}} (1 + R(\bar{g})^2 + M_0) \right) + \tilde{n}^{-\frac{z}{2}} (1 + R(\bar{g}) + M_0)
\]

\[
= \lambda_\theta R(g^*)^2 + \frac{1}{4} \|f\|_{L^2(S)}^2 + O_p (\tilde{n}^{-\frac{z}{2}} (1 + R(\bar{g})^2 + M_0))
\]
which indicates
\[
\frac{1}{4} \| f \|_{L_2(S)}^2 + \lambda \tilde{R}(\tilde{g})^2 \leq \lambda \tilde{R}(\tilde{g}^*)^2 + \sigma_p (\lambda \tilde{R}(1 + R(\tilde{g})^2 + M_0^2)).
\]

Therefore, by moving \( \sigma_p (\lambda \tilde{R}(\tilde{g})^2) \) to the left-hand side, we obtain
\[
\frac{1}{4} \| f \|_{L_2(S)}^2 + \lambda \tilde{R}(1 - \sigma_p(1))R(\tilde{g})^2 \leq \sigma_p (\lambda \tilde{R}(1 + R(\tilde{g})^2) + M_0^2)
\leq \sigma_p (\lambda \tilde{R}(1 + R(\tilde{g})^2)).
\]

This gives
\[
\| f \|_{L_2(S)} = \sigma_p (\lambda \tilde{R} \max\{1, R(\tilde{g})\}),
\]
\[
R(\tilde{g}) = \sigma_p (\sqrt{1 + R(\tilde{g})^2}) = \sigma_p (\max\{1, R(\tilde{g})\}).
\]

Consequently, the proof of lemma 1 is completed.

**A.3 Proof of Theorem 1.** Based on lemma 1, we prove theorem 1.

As in the proof of lemma 1, let \( f := \hat{g} - g^* \). Since \( (\alpha P + (1 - \alpha)P') (fg^*) = S(fg^*) = Pf \), we have
\[
\hat{P}E_a - PE_a = \frac{1}{2} S_{n,n'} \hat{g}^2 - P_n \hat{g} - \left( \frac{1}{2} Sg^* - Pg^* \right)
= \frac{1}{2} S_{n,n'} (f + g^*)^2 - P_n (f + g^*) - \left( \frac{1}{2} Sg^* - Pg^* \right)
= \frac{1}{2} Sf^2 + \frac{1}{2} (S_{n,n'} - S) f^2 + (S_{n,n'} - S)(g^* f) - (P_n - P) f
+ \frac{1}{2} (S_{n,n'} - S) g^*^2 - (P_n g^* - P g^*).
\]

Below, we show that each term of the right-hand side of the above equation is \( \sigma_p (\lambda \tilde{R}) \). By the central limit theorem, we have
\[
\frac{1}{2} (S_{n,n'} - S) g^*^2 - (P_n g^* - P g^*)
= \sigma_p \left( \bar{n}^{-1/2} M_0 \left( (1 + \alpha) \sqrt{P(g^* - Pg^*)^2} + (1 - \alpha) \sqrt{P'(g^* - Pg^*)^2} \right) \right).
\]
Since lemma 1 gives \( \|f\|_2 = O_p(\lambda^{-1}_n \max(1, R(g^*)) ) \) and \( R(\hat{\gamma}) = O_p(\max(1, R(g^*)) ) \), equations A.4 to A.6 in the proof of lemma 1 imply

\[
|\langle s_n, n - s_n \rangle f^2| = O_p \left( \frac{1}{\sqrt{n}} \|f\|_{L_2, \hat{n}}(1 + R(g^*))^{1 + 2^{2/7}} \vee \hat{n}^{-\frac{2}{27}} R(g^*)^2 \right)
\leq O_p(\lambda^{-1}_n \max(1, R(g^*)^2)) ,
\]

\[
|\langle s_n, f \rangle - S f^2| = O_p \left( \frac{1}{\sqrt{n}} \|f\|_{L_2, \hat{n}}(1 + R(\hat{\gamma}))M_0 + M_0^2 \right)^{2/7} \vee \hat{n}^{-\frac{2}{27}} (1 + R(\hat{\gamma})M_0 + M_0^2)
\leq O_p(\lambda^{-1}_n \max(1, R(g^*)M_0^{2/7}M_0^0R(\gamma^*/M_0, M_0))
\leq O_p(\lambda^{-1}_n \max(1, R(g^*)M_0^0))
\leq O_p(\lambda^{-1}_n \max(1, R(g^*)^2)) ,
\]

\[
|\langle P_n - P \rangle f| \leq O_p \left( \frac{1}{\sqrt{n}} \|f\|_{L_2, \hat{n}}(1 + R(\hat{\gamma}) + M_0)^{2/7}M_0^1(1 - 2^{2/7}) \vee \hat{n}^{-\frac{2}{27}} (1 + R(\hat{\gamma}) + M_0)
\right)
= O_p(\lambda^{-1}_n \max(1, M_0^{1 - 2^{2/7}}R(g^*)M_0^{1 - 2^{2/7}}R(g^*)))
\leq O_p(\lambda^{-1}_n \max(1, R(g^*)^2)) ,
\]

(A.9)

where we used \( \lambda^{-1}_n = o(\hat{n}^{-2/(2+\gamma)}) \) and \( M_0 \leq R(g^*) \). Lemma 1 also implies

\[
Sf^2 = \|f\|_2 = O_p(\lambda^{-1}_n \max(1, R(g^*)^2)) .
\]

Combining these inequalities with equation A.8 implies

\[
\hat{P}_E - P_E = O_p(\lambda^{-1}_n \max(1, R(g^*)^2) + n^{-1/2}cM_0) ,
\]

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

On the other hand, we have

\[
\hat{P}_E - P_E = \frac{1}{2}P_n\hat{\gamma} - \frac{1}{2}P\gamma^* = \frac{1}{2}[(P_n - P)(\hat{\gamma} - \gamma^*) + P(\hat{\gamma} - \gamma^*) + (P_n - P)\gamma^*].
\]

(A.10)
Equation A.9 gives

\[(P_n - P)(\hat{g} - g^*) = O_p(\lambda_{\bar{n}} \max(1, M_0^{\frac{1}{2}(1 - \frac{\gamma}{2})}, R(g^*)M_0^{\frac{1}{2}(1 - \frac{\gamma}{2})}, R(g^*))).\]

We also have

\[P(\hat{g} - g^*) \leq \|\hat{g} - g^*\|_{L_2(P)} \leq \|\hat{g} - g^*\|_{L_2(S)}M_0^{\frac{1}{2}} = O_p(\lambda_{\bar{n}} \max(M_0^{\frac{1}{2}}, M_0^{\frac{1}{2}}R(g^*))))\]

and

\[(P_n - P)g^* = O_p(\bar{\bar{n}}^{-\frac{1}{2}} \sqrt{P(g^* - Pg^*)^2}) \leq O_p(\bar{\bar{n}}^{-\frac{1}{2}}M_0) \leq O_p(\lambda_{\bar{n}} \max(M_0^{\frac{1}{2}}, M_0^{\frac{1}{2}}R(g^*))).\]

Therefore by substituting these bounds into the relation A.10, we observe that

\[\hat{\theta}_{E_a} - \theta = O_p(\lambda_{\bar{n}}^{\frac{1}{2}} \max(M_0^{\frac{1}{2}}, M_0^{\frac{1}{2}}R(g^*)))\]

\[+\lambda_{\bar{n}} \max(1, M_0^{\frac{1}{2}(1 - \frac{\gamma}{2})}, R(g^*)M_0^{\frac{1}{2}(1 - \frac{\gamma}{2})}, R(g^*)).\]  \hspace{1cm} (A.11)

This completes the proof.

Appendix 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 \(\alpha\)-relative density ratio, equation 2.1, the statistical model

\[\mathcal{G} = \{g(x; \theta) | \theta \in \Theta \subset \mathbb{R}^b\}\]

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

\[\hat{g} = \arg\min_{g \in \mathcal{G}} \left\{ \frac{1}{2} \left\{ \frac{\alpha}{n} \sum_{i=1}^{n} (g(x_i))^2 + \frac{1 - \alpha}{n'} \sum_{j=1}^{n'} (g(x'_j))^2 \right\} - \frac{1}{n} \sum_{i=1}^{n} g(x_i) \right\}.\]
Suppose that the model is correctly specified; there exists $\theta^*$ such that
\[ g(x; \theta^*) = r_\alpha(x). \]

Then, under a mild assumption (see theorem 5.23 of van der Vaart, 2000), the estimator $\hat{g}$ is consistent and the estimated parameter $\hat{\theta}$ satisfies the asymptotic normality in the large sample limit. Then a possible estimator of the $\alpha$-relative Pearson divergence $PE_\alpha$ is
\[
\hat{PE}_\alpha = \frac{1}{n} \sum_{i=1}^{n} \hat{g}(x_i) - \frac{1}{2} \left\{ \frac{\alpha}{n} \sum_{i=1}^{n} \hat{g}(x_i)^2 + \frac{1-\alpha}{n'} \sum_{j=1}^{n'} \hat{g}(x'_j)^2 \right\} - \frac{1}{2}.
\]

Note that there are other possible estimators for $PE_\alpha$ such as
\[
\tilde{PE}_\alpha = \frac{1}{2n} \sum_{i=1}^{n} \hat{g}(x_i) - \frac{1}{2}.
\]

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

**Theorem 2.** Let $\|r\|_\infty$ be the sup-norm of the standard density ratio $r(x)$ and $\|r_\alpha\|_\infty$ be the sup-norm of the $\alpha$-relative density ratio,
\[
\|r_\alpha\|_\infty = \frac{\|r\|_\infty}{\alpha\|r\|_\infty + 1 - \alpha}.
\]

The variance of $\hat{PE}_\alpha$ is denoted as $\nabla[\hat{PE}_\alpha]$. Then, under the regularity condition for the asymptotic normality, we have the following upper bound of $\nabla[\hat{PE}_\alpha]$:
\[
\nabla[\hat{PE}_\alpha] = \frac{1}{n} \nabla_{p(x)} \left[ r_\alpha - \frac{\alpha r^2_\alpha}{2} \right] + \frac{1}{n'} \nabla_{p'(x)} \left[ \frac{(1-\alpha)r^2_\alpha}{2} \right] + o \left( \frac{1}{n}, \frac{1}{n'} \right)
\leq \frac{\|r_\alpha\|^2}{n} + \frac{\alpha^2\|r_\alpha\|^4}{4n} + \frac{(1-\alpha)^2\|r_\alpha\|^4}{4n'} + o \left( \frac{1}{n}, \frac{1}{n'} \right).
\]

**Theorem 3.** The variance of $\tilde{PE}_\alpha$ is denoted as $\nabla[\tilde{PE}_\alpha]$. Let $\nabla g$ be the gradient vector of $g$ with respect to $\theta$ at $\theta = \theta^*$, that is, $(\nabla g(x; \theta^*))_j = \frac{\partial g(x; \theta^*)}{\partial \theta_j}$. The matrix $U_\alpha$ is defined by
\[
U_\alpha = \alpha E_{p(x)}[\nabla g \nabla g^\top] + (1-\alpha) E_{p'(x)}[\nabla g \nabla g^\top].
\]
Then, under the regularity condition, the variance of \( \hat{\mathbb{P}} \hat{E}_a \) is asymptotically given as

\[
\mathbb{V}[\hat{\mathbb{P}} \hat{E}_a] = \frac{1}{n} \mathbb{V}_{p(\cdot)} \left[ \frac{r_a + (1 - \alpha r_a) E_{p(\cdot)} [\nabla g] \mathbb{U}_a^{-1} \nabla g}{2} \right] + \frac{1}{n'} \mathbb{V}_{p'(\cdot)} \left[ \frac{(1 - \alpha) r_a E_{p(\cdot)} [\nabla g] \mathbb{U}_a^{-1} \nabla g}{2} \right] + o \left( \frac{1}{n}, \frac{1}{n'} \right).
\]

**B.2 Proof of Theorem 2.** Let \( \hat{\theta} \) be the estimated parameter, \( \hat{g}(x) = g(x; \hat{\theta}) \). Suppose that \( r_a(x) = g(x; \theta^*) \in \mathcal{G} \) holds. Let \( \delta \theta = \theta - \theta^* \). Then the asymptotic expansion of \( \hat{\mathbb{P}} \hat{E}_a \) is given as

\[
\hat{\mathbb{P}} \hat{E}_a = \frac{1}{n} \sum_{i=1}^{n} g(x_i; \hat{\theta}) - \frac{1}{2} \left\{ \frac{\alpha}{n} \sum_{i=1}^{n} g(x_i; \hat{\theta})^2 + \frac{1 - \alpha}{n'} \sum_{j=1}^{n'} g(x'_j; \hat{\theta})^2 \right\} - \frac{1}{2}
\]

\[
= \mathbb{P} \hat{E}_a + \frac{1}{n} \sum_{i=1}^{n} (r_a(x_i) - E_{p(\cdot)}[r_a]) + \frac{1}{n} \sum_{i=1}^{n} \nabla g(x_i; \theta^*) \mathbb{U}_a^{-1} \nabla g - \frac{1}{2} \left\{ \frac{\alpha}{n} \sum_{i=1}^{n} (r_a(x_i)^2 - E_{p(\cdot)}[r_a^2]) + \frac{1 - \alpha}{n'} \sum_{j=1}^{n'} (r_a(x'_j)^2 - E_{p'(\cdot)}[r_a^2]) \right\}
\]

\[
- \left\{ \frac{\alpha}{n} \sum_{i=1}^{n} r_a(x_i) \nabla g(x_i; \theta^*) + \frac{1 - \alpha}{n'} \sum_{j=1}^{n'} r_a(x'_j) \nabla g(x'_j; \theta^*) \right\}^\top \times \delta \theta + o_p \left( \frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n'}} \right).
\]

Let us define the linear operator \( G \) as

\[
G f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(x_i) - E_{p(\cdot)}[f]).
\]

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

\[
\hat{\mathbb{P}} \hat{E}_a - \mathbb{P} \hat{E}_a = \frac{1}{\sqrt{n}} G \left( \alpha \frac{1}{2} r_a^2 \right) - \frac{1}{\sqrt{n'}} G' \left( \frac{1 - \alpha}{2} r_a^2 \right) + [E_{p(\cdot)}[\nabla g]] - \alpha E_{p(\cdot)}[\nabla g] - (1 - \alpha) E_{p'(\cdot)}[\nabla g] \mathbb{U}_a^{-1} \nabla g + o_p \left( \frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n'}} \right)
\]

\[
= \frac{1}{\sqrt{n}} G \left( \alpha \frac{1}{2} r_a^2 \right) - \frac{1}{\sqrt{n'}} G' \left( \frac{1 - \alpha}{2} r_a^2 \right) + o_p \left( \frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n'}} \right).
\]
The second equality follows from
\[ \mathbb{E}_{p(x)}[\nabla g] - \alpha \mathbb{E}_{p(x)}[r_\alpha \nabla g] - (1 - \alpha) \mathbb{E}_{p'(x)}[r_\alpha \nabla g] = 0. \]

Then the asymptotic variance is given as
\[
\mathbb{V}[\hat{\mathbb{P}} E_{\alpha}] = \frac{1}{n} \mathbb{V}_p(x) \left[ r_\alpha - \frac{\alpha^2 r_\alpha^2}{2} \right] + \frac{1}{n'} \mathbb{V}_{p'}(x) \left[ \frac{1 - \alpha}{2} r_\alpha^2 \right] + o\left( \frac{1}{n}, \frac{1}{n'} \right). \tag{B.1}
\]

We confirm that both \( r_\alpha - \frac{\alpha r_\alpha^2}{2} \) and \( \frac{1 - \alpha}{2} r_\alpha^2 \) are nonnegative and increasing functions with respect to \( r \) for any \( \alpha \in [0, 1] \). Since the result is trivial for \( \alpha = 1 \), we suppose \( 0 \leq \alpha < 1 \). The function \( r_\alpha - \frac{\alpha r_\alpha^2}{2} \) is represented as
\[
r_\alpha - \frac{\alpha r_\alpha^2}{2} = r(\alpha r + 2 - 2\alpha) \frac{1}{2(\alpha r + 1 - \alpha)^2},
\]
and thus we have \( r_\alpha - \frac{\alpha r_\alpha^2}{2} = 0 \) 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 \( r_\alpha - \frac{\alpha r_\alpha^2}{2} \) is nonnegative and increasing with respect to \( r \). Following the same line, we see that \( \frac{1 - \alpha}{2} r_\alpha^2 \) is nonnegative and increasing with respect to \( r \). Thus we have the following inequalities,
\[
0 \leq r_\alpha(x) - \frac{\alpha}{2} r_\alpha(x)^2 \leq \|r_\alpha\|_\infty - \frac{\alpha}{2} \|r_\alpha\|_\infty^2,
\]
\[
0 \leq \frac{1 - \alpha}{2} r_\alpha(x)^2 \leq \frac{1 - \alpha}{2} \|r_\alpha\|_\infty^2.
\]
As a result, upper bounds of the variances in equation B.1 are given as
\[
\mathbb{V}_p(x) \left[ r_\alpha - \frac{\alpha r_\alpha^2}{2} \right] \leq \left( \|r_\alpha\|_\infty - \frac{\alpha}{2} \|r_\alpha\|_\infty^2 \right)^2,
\]
\[
\mathbb{V}_{p'}(x) \left[ \frac{1 - \alpha}{2} r_\alpha^2 \right] \leq \frac{(1 - \alpha)^2}{4} \|r_\alpha\|_\infty^4.
\]
Therefore, the following inequality holds,

$$V[\hat{PE}_\alpha] \leq \frac{1}{n} \left( \| r_\alpha \|_\infty - \frac{\alpha \| r_\alpha \|^2}{2} \right)^2 + \frac{1}{n'} \left( (1 - \alpha)^2 \| r_\alpha \|^4 \right) + o \left( \frac{1}{n}, \frac{1}{n'} \right)$$

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

which completes the proof.

**B.3 Proof of Theorem 3.** The estimator $\hat{\theta}$ is the optimal solution of the following problem:

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

Then the extremal condition yields the equation,

$$\frac{\alpha}{n} \sum_{i=1}^{n} g(x^p_i; \hat{\theta}) \nabla g(x^p_i; \hat{\theta}) + \frac{1 - \alpha}{n'} \sum_{j=1}^{n'} g(x^q_j; \hat{\theta}) \nabla g(x^q_j; \hat{\theta})$$

$$- \frac{1}{n} \sum_{i=1}^{n} \nabla g(x^p_i; \hat{\theta}) = 0.$$

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

$$\frac{1}{n} \sum_{i=1}^{n} (\alpha r_\alpha(x^p_i) - 1) \nabla g(x^p_i; \theta^*) + \frac{1 - \alpha}{n'} \sum_{j=1}^{n'} r_\alpha(x^q_j) \nabla g(x^q_j; \theta^*)$$

$$+ U_\alpha \delta \theta + o_p \left( \frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n'}} \right) = 0.$$

Therefore, we obtain

$$\delta \theta = \frac{1}{\sqrt{n}} G((1 - \alpha r_\alpha)U_\alpha^{-1} \nabla g) - \frac{1}{\sqrt{n'}} G'((1 - \alpha) r_\alpha U_\alpha^{-1} \nabla g)$$

$$+ o_p \left( \frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n'}} \right).$$
Next, we compute the asymptotic expansion of $\tilde{PE}_\alpha$:

\[
\tilde{PE}_\alpha = \frac{1}{2} E_p(x)[r_\alpha] + \frac{1}{2n} \sum_{i=1}^{n} (r_\alpha(x_i^p) - E_p(x)[r_\alpha]) \\
+ \frac{1}{2n} \sum_{i=1}^{n} \nabla g(x_i^p; \theta^*)^\top \delta \theta - \frac{1}{2} o_p \left( \frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n'}} \right)
\]

\[
= PE_\alpha + \frac{1}{2\sqrt{n}} G(r_\alpha) + \frac{1}{2} E_{p(x)}[\nabla g] U^{-1} \nabla g + o_p \left( \frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n'}} \right).
\]

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

\[
\tilde{PE}_\alpha - PE_\alpha = \frac{1}{2\sqrt{n}} G(r_\alpha + (1 - \alpha r_\alpha) E_{p(x)}[\nabla g] U^{-1} \nabla g) \\
- \frac{1}{2\sqrt{n'}} G'(1 - \alpha) r_\alpha E_{p(x)}[\nabla g] U^{-1} \nabla g + o_p \left( \frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n'}} \right).
\]

As a result, we have

\[
\nabla[\tilde{PE}_\alpha] = \frac{1}{n} \nabla_{p(x)} \left[ \frac{r_\alpha + (1 - \alpha r_\alpha) E_{p(x)}[\nabla g] U^{-1} \nabla g}{2} \right] \\
+ \frac{1}{n'} \nabla_{p'(x)} \left[ \frac{(1 - \alpha) r_\alpha E_{p(x)}[\nabla g] U^{-1} \nabla g}{2} \right] + o \left( \frac{1}{n}, \frac{1}{n'} \right),
\]

which completes the proof.

**Acknowledgments**

M.Y. was supported by the JST PRESTO program. T.S. was partially supported by MEXT KAKENHI 22700289 and Aihara Project, the FIRST program from JSPS, initiated by CSTP. T.K. was partially supported by a Grant-in-Aid for Young Scientists (20700251), H.H. was supported by the FIRST program, and M.S. was partially supported by SCAT, AOARD, and the FIRST program.

**References**

Ali, S. M., & Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B, 28*, 131–142.
Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68, 337–404.

Bao, L., & Intille, S. S. (2004). Activity recognition from user-annotated acceleration data. In *Proceedings of the 2nd IEEE International Conference on Pervasive Computing* (pp. 1–17). Piscataway, NJ: IEEE.

Bharatula, N. B., Stager, M., Lukowicz, P., & Troster, G. (2005). Empirical study of design choices in multi-sensor context recognition systems. In *Proceedings of the 2nd International Forum on Applied Wearable Computing* (pp. 79–93). Berlin: Springer-Verlag.

Bickel, S., Bogojeska, J., Lengauer, T., & Scheffer, T. (2008). Multi-task learning for HIV therapy screening. In A. McCallum & S. Roweis (Eds.), *Proceedings of 25th Annual International Conference on Machine Learning (ICML2008)* (pp. 56–63). Madison, WI: Omnipress.

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

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

Chang, C.-C., & Lin, C.-J. (2001). *LIBSVM: A library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chapelle, O., Schölkopf, B., & Zien, A. (Eds.). (2006). *Semi-supervised learning*. Cambridge, MA: MIT Press.

Cortes, C., Mansour, Y., & Mohri, M. (2010). Learning bounds for importance weighting. In J. Lafferty, C.K.I. Williams, R. Zemel, J. Shawe-Taylor, & A. Culotta (Eds.), *Advances in neural information processing systems 23* (pp. 442–450). Red Hook, NY: Curran.

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

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

Fishman, G. S. (1996). *Monte Carlo: Concepts, algorithms, and applications*. Berlin: Springer-Verlag.

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

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

Jiang, J., & Zhai, C. (2007). Instance weighting for domain adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics* (pp. 264–271). Stroudsburg, PA: Association for Computational Linguistics.

Kain, A., & Macon, M. W. (1998). Spectral voice conversion for text-to-speech synthesis. In *Proceedings of 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 285–288). Piscataway, NJ: IEEE.
Kanamori, T., Hido, S., & Sugiyama, M. (2009). A least-squares approach to direct importance estimation. *Journal of Machine Learning Research, 10*, 1391–1445.

Kimura, M., & Sugiyama, M. (2011). Dependence-maximization clustering with least-squares mutual information. *Journal of Advanced Computational Intelligence and Intelligent Informatics, 15*, 800–805.

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

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

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

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

Rätsch, G., Onoda, T., & Müller, K.-R. (2001). Soft margins for Adaboost. *Machine Learning, 42*(3), 287–320.

Rockafellar, R. T. (1970). *Convex analysis*. Princeton, NJ: Princeton University Press.

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

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

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

Smola, A., Song, L., & Teo, C. H. (2009). Relative novelty detection. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS2009)* (pp. 536–543). Brookline, MA: Microtome.

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

Steinwart, I., & Scovel, C. (2007). Fast rates for support vector machines using gaussian kernels. *Annals of Statistics, 35*(2), 575–607.

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

Sugiyama, M., & Kawanabe, M. (2012). *Covariate shift adaptation: Toward machine learning in non-stationary environments*. Cambridge, MA: MIT Press.

Sugiyama, M., Krauledat, M., & Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research, 8*, 985–1005.
Sugiyama, M., & Müller, K.-R. (2005). Input-dependent estimation of generalization error under covariate shift. *Statistics and Decisions*, *23*(4), 249–279.

Sugiyama, M., & Suzuki, T. (2011). Least-squares independence test. *IEICE Transactions on Information and Systems*, *E94-D*(6), 1333–1336.

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

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

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

Suzuki, T., & Sugiyama, M. (2010). Sufficient dimension reduction via squared-loss mutual information estimation. In Y. W. Teh, & M. Tiggerington (Eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS2010)* (pp. 804–811). London: BioMed Central.

Suzuki, T., & Sugiyama, M. (2011). Least-squares independent component analysis. *Neural Computation*, *23*(1), 284–301.

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

van de Geer, S. (2000). *Empirical processes in M-estimation*. Cambridge: Cambridge University Press.

van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge: Cambridge University Press.

van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes: With applications to statistics*. New York: Springer.

Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley.

Wahba, G. (1990). *Spline model for observational data*. Philadelphia: Society for Industrial and Applied Mathematics.

Yamada, M., & Sugiyama, M. (2010). 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)* (pp. 643–648). Palo Alto, CA: AAAI Press.

Yamada, M., & Sugiyama, M. (2011). Cross-domain object matching with model selection. In G. Gordon, D. Dunson, & M. Dudík (Eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Brookline, MA: Microtome.

Received July 12, 2012; accepted December 20, 2012.