

Density-Difference Estimation

Masashi Sugiyama

Tokyo Institute of Technology, Japan.

sugi@cs.titech.ac.jp

<http://sugiyama-www.cs.titech.ac.jp/~sugi>

Takafumi Kanamori

Nagoya University, Japan.

kanamori@is.nagoya-u.ac.jp

Taiji Suzuki

The University of Tokyo, Japan.

s-taiji@stat.t.u-tokyo.ac.jp

Marthinus Christoffel du Plessis

Tokyo Institute of Technology, Japan.

christo@sg.cs.titech.ac.jp

Song Liu

Tokyo Institute of Technology, Japan.

song@sg.cs.titech.ac.jp

Ichiro Takeuchi

Nagoya Institute of Technology, Japan.

takeuchi.ichiro@nitech.ac.jp

Abstract

We address the problem of estimating the *difference* between two probability densities. A naive approach is a two-step procedure of first estimating two densities separately and then computing their difference. However, such a two-step procedure does not necessarily work well because the first step is performed without regard to the second step and thus a small error incurred in the first stage can cause a big error in the second stage. In this paper, we propose a single-shot procedure for directly estimating the density difference without separately estimating two densities. We derive a non-parametric finite-sample error bound for the proposed single-shot density-difference estimator and show that it achieves the optimal convergence rate. The usefulness of the proposed method is also demonstrated experimentally.

Keywords

density difference, L^2 -distance, robustness, Kullback-Leibler divergence, kernel density estimation.

1 Introduction

When estimating a quantity consisting of two elements, a two-stage approach of first estimating the two elements separately and then approximating the target quantity based on the estimates of the two elements often performs poorly, because the first stage is carried out without regard to the second stage and thus a small error incurred in the first stage can cause a big error in the second stage. To cope with this problem, it would be more appropriate to directly estimate the target quantity in a single-shot process without separately estimating the two elements.

A seminal example that follows this general idea is pattern recognition by the *support vector machine* (Boser et al., 1992; Cortes & Vapnik, 1995; Vapnik, 1998): Instead of separately estimating two probability distributions of patterns for positive and negative classes, the support vector machine directly learns the boundary between the two classes that is sufficient for pattern recognition. More recently, a problem of estimating the ratio of two probability densities was tackled in a similar fashion (Qin, 1998; Sugiyama et al., 2008; Gretton et al., 2009; Kanamori et al., 2009; Nguyen et al., 2010; Kanamori et al., 2012; Sugiyama et al., 2012b; Sugiyama et al., 2012a): The ratio of two probability densities is directly estimated without going through separate estimation of the two probability densities.

In this paper, we further explore this line of research, and propose a method for directly estimating the *difference* between two probability densities in a single-shot process. Density differences are useful for various purposes such as class-balance estimation under class-prior change (Saerens et al., 2002; Du Plessis & Sugiyama, 2012), change-point detection in time series (Kawahara & Sugiyama, 2012; Liu et al., 2012), feature extraction (Torkkola, 2003), video-based event detection (Matsugu et al., 2011), flow cytometric data analysis (Duong et al., 2009), ultrasound image segmentation (Liu et al., 2010), non-rigid image registration (Atif et al., 2003), and image-based target recognition (Gray

& Principe, 2010).

For this density-difference estimation problem, we propose a single-shot method, called the *least-squares density-difference* (LSDD) estimator, that directly estimates the density difference without separately estimating two densities. LSDD is derived within a framework of kernel least-squares estimation, and its solution can be computed *analytically* in a computationally efficient and stable manner. Furthermore, LSDD is equipped with cross-validation, and thus all tuning parameters such as the kernel width and the regularization parameter can be systematically and objectively optimized. We derive a finite-sample error bound for the LSDD estimator in a non-parametric setup and show that it achieves the optimal convergence rate.

We also apply LSDD to L^2 -distance estimation and show that it is more accurate than the difference of KDEs, which tends to severely under-estimate the L^2 -distance (Anderson et al., 1994). Compared with the *Kullback-Leibler (KL) divergence* (Kullback & Leibler, 1951), the L^2 -distance is more robust against outliers (Basu et al., 1998; Scott, 2001; Besbeas & Morgan, 2004).

Finally, we experimentally demonstrate the usefulness of LSDD in semi-supervised class-prior estimation and unsupervised change detection.

The rest of this paper is structured as follows. In Section 2, we derive the LSDD method and investigate its theoretical properties. In Section 3, we show how the L^2 -distance can be approximated by LSDD. In Section 4, we illustrate the numerical behavior of LSDD. Finally, we conclude in Section 5.

2 Density-Difference Estimation

In this section, we propose a single-shot method for estimating the difference between two probability densities from samples, and analyze its theoretical properties.

2.1 Problem Formulation and Naive Approach

First, we formulate the problem of density-difference estimation.

Suppose that we are given two sets of independent and identically distributed samples $\mathcal{X} := \{\mathbf{x}_i\}_{i=1}^n$ and $\mathcal{X}' := \{\mathbf{x}'_{i'}\}_{i'=1}^{n'}$ drawn from probability distributions on \mathbb{R}^d with densities $p(\mathbf{x})$ and $p'(\mathbf{x})$, respectively:

$$\begin{aligned}\mathcal{X} &:= \{\mathbf{x}_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} p(\mathbf{x}), \\ \mathcal{X}' &:= \{\mathbf{x}'_{i'}\}_{i'=1}^{n'} \stackrel{\text{i.i.d.}}{\sim} p'(\mathbf{x}).\end{aligned}$$

Our goal is to estimate the difference $f(\mathbf{x})$ between $p(\mathbf{x})$ and $p'(\mathbf{x})$ from the samples \mathcal{X} and \mathcal{X}' :

$$f(\mathbf{x}) := p(\mathbf{x}) - p'(\mathbf{x}).$$

A naive approach to density-difference estimation is to use *kernel density estimators* (KDEs) (Silverman, 1986). For Gaussian kernels, the KDE-based density-difference estimator is given by

$$\tilde{f}(\mathbf{x}) := \hat{p}(\mathbf{x}) - \hat{p}'(\mathbf{x}),$$

where

$$\begin{aligned}\hat{p}(\mathbf{x}) &:= \frac{1}{n(2\pi\sigma^2)^{d/2}} \sum_{i=1}^n \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}_i\|^2}{2\sigma^2}\right), \\ \hat{p}'(\mathbf{x}) &:= \frac{1}{n'(2\pi\sigma'^2)^{d/2}} \sum_{i'=1}^{n'} \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'_{i'}\|^2}{2\sigma'^2}\right).\end{aligned}$$

The Gaussian widths σ and σ' may be determined based on cross-validation (Härdle et al.,

2004).

However, we argue that the KDE-based density-difference estimator is not the best approach because of its two-step nature: Small estimation error in each density estimate can cause a big error in the final density-difference estimate. More intuitively, good density estimators tend to be smooth and thus a density-difference estimator obtained from such smooth density estimators tends to be over-smoothed (Hall & Wand, 1988; Anderson et al., 1994, see also numerical experiments in Section 4.1.1).

To overcome this weakness, we give a single-shot procedure of directly estimating the density difference $f(\mathbf{x})$ without separately estimating the densities $p(\mathbf{x})$ and $p'(\mathbf{x})$.

2.2 Least-Squares Density-Difference Estimation

In our proposed approach, we fit a density-difference model $g(\mathbf{x})$ to the true density-difference function $f(\mathbf{x})$ under the squared loss:

$$\operatorname{argmin}_g \int (g(\mathbf{x}) - f(\mathbf{x}))^2 d\mathbf{x}. \quad (1)$$

We use the following linear-in-parameter model as $g(\mathbf{x})$:

$$g(\mathbf{x}) = \sum_{\ell=1}^b \theta_{\ell} \psi_{\ell}(\mathbf{x}) = \boldsymbol{\theta}^{\top} \boldsymbol{\psi}(\mathbf{x}), \quad (2)$$

where b denotes the number of basis functions, $\boldsymbol{\psi}(\mathbf{x}) = (\psi_1(\mathbf{x}), \dots, \psi_b(\mathbf{x}))^{\top}$ is a b -dimensional basis function vector, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_b)^{\top}$ is a b -dimensional parameter vector, and \top denotes the transpose. In practice, we use the following non-parametric Gaussian kernel model as $g(\mathbf{x})$:

$$g(\mathbf{x}) = \sum_{\ell=1}^{n+n'} \theta_{\ell} \exp\left(-\frac{\|\mathbf{x} - \mathbf{c}_{\ell}\|^2}{2\sigma^2}\right), \quad (3)$$

where $(\mathbf{c}_1, \dots, \mathbf{c}_n, \mathbf{c}_{n+1}, \dots, \mathbf{c}_{n+n'}) := (\mathbf{x}_1, \dots, \mathbf{x}_n, \mathbf{x}'_1, \dots, \mathbf{x}'_{n'})$ are Gaussian kernel centers. If $n + n'$ is large, we may use only a subset of $\mathbf{x}_1, \dots, \mathbf{x}_n, \mathbf{x}'_1, \dots, \mathbf{x}'_{n'}$ as Gaussian kernel centers.

For the model (2), the optimal parameter $\boldsymbol{\theta}^*$ is given by

$$\begin{aligned} \boldsymbol{\theta}^* &:= \operatorname{argmin}_{\boldsymbol{\theta}} \int (g(\mathbf{x}) - f(\mathbf{x}))^2 d\mathbf{x} \\ &= \operatorname{argmin}_{\boldsymbol{\theta}} \left[\int g(\mathbf{x})^2 d\mathbf{x} - 2 \int g(\mathbf{x})f(\mathbf{x})d\mathbf{x} \right] \\ &= \operatorname{argmin}_{\boldsymbol{\theta}} [\boldsymbol{\theta}^\top \mathbf{H}\boldsymbol{\theta} - 2\mathbf{h}^\top \boldsymbol{\theta}] \\ &= \mathbf{H}^{-1}\mathbf{h}, \end{aligned}$$

where \mathbf{H} is the $b \times b$ matrix and \mathbf{h} is the b -dimensional vector defined as

$$\begin{aligned} \mathbf{H} &:= \int \boldsymbol{\psi}(\mathbf{x})\boldsymbol{\psi}(\mathbf{x})^\top d\mathbf{x}, \\ \mathbf{h} &:= \int \boldsymbol{\psi}(\mathbf{x})p(\mathbf{x})d\mathbf{x} - \int \boldsymbol{\psi}(\mathbf{x}')p'(\mathbf{x}')d\mathbf{x}'. \end{aligned}$$

Note that, for the Gaussian kernel model (3), the integral in \mathbf{H} can be computed analytically as

$$\begin{aligned} H_{\ell, \ell'} &= \int \exp\left(-\frac{\|\mathbf{x} - \mathbf{c}_\ell\|^2}{2\sigma^2}\right) \exp\left(-\frac{\|\mathbf{x} - \mathbf{c}_{\ell'}\|^2}{2\sigma^2}\right) d\mathbf{x} \\ &= (\pi\sigma^2)^{d/2} \exp\left(-\frac{\|\mathbf{c}_\ell - \mathbf{c}_{\ell'}\|^2}{4\sigma^2}\right), \end{aligned}$$

where d denotes the dimensionality of \mathbf{x} .

Replacing the expectations in \mathbf{h} by empirical estimators and adding an ℓ_2 -regularizer

to the objective function, we arrive at the following optimization problem:

$$\hat{\boldsymbol{\theta}} := \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left[\boldsymbol{\theta}^\top \mathbf{H} \boldsymbol{\theta} - 2\hat{\mathbf{h}}^\top \boldsymbol{\theta} + \lambda \boldsymbol{\theta}^\top \boldsymbol{\theta} \right], \quad (4)$$

where $\lambda (\geq 0)$ is the regularization parameter and $\hat{\mathbf{h}}$ is the b -dimensional vector defined as

$$\hat{\mathbf{h}} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{\psi}(\mathbf{x}_i) - \frac{1}{n'} \sum_{i'=1}^{n'} \boldsymbol{\psi}(\mathbf{x}'_{i'}).$$

Taking the derivative of the objective function in Eq.(4) and equating it to zero, we can obtain the solution $\hat{\boldsymbol{\theta}}$ analytically as

$$\hat{\boldsymbol{\theta}} = (\mathbf{H} + \lambda \mathbf{I}_b)^{-1} \hat{\mathbf{h}},$$

where \mathbf{I}_b denotes the b -dimensional identity matrix.

Finally, a density-difference estimator $\hat{f}(\mathbf{x})$ is given as

$$\hat{f}(\mathbf{x}) = \hat{\boldsymbol{\theta}}^\top \boldsymbol{\psi}(\mathbf{x}). \quad (5)$$

We call this the *least-squares density-difference* (LSDD) estimator.

2.3 Theoretical Analysis

Here, we theoretically investigate the behavior of the LSDD estimator.

2.3.1 Parametric Convergence

First, we consider a linear parametric setup where basis functions in our density-difference model (2) are fixed.

Suppose that $n/(n+n')$ converges to $\eta \in [0, 1]$. Then the *central limit theorem* (Rao, 1965) asserts that $\sqrt{\frac{nn'}{n+n'}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ converges in law to the normal distribution with mean $\mathbf{0}$ and covariance matrix

$$\mathbf{H}^{-1}((1-\eta)\mathbf{V}_p + \eta\mathbf{V}_{p'})\mathbf{H}^{-1},$$

where \mathbf{V}_p denotes the covariance matrix of $\boldsymbol{\psi}(\mathbf{x})$ under the probability density $p(\mathbf{x})$:

$$\mathbf{V}_p := \int (\boldsymbol{\psi}(\mathbf{x}) - \boldsymbol{\psi}_p) (\boldsymbol{\psi}(\mathbf{x}) - \boldsymbol{\psi}_p)^\top p(\mathbf{x}) d\mathbf{x}, \quad (6)$$

and $\boldsymbol{\psi}_p$ denotes the expectation of $\boldsymbol{\psi}(\mathbf{x})$ under the probability density $p(\mathbf{x})$:

$$\boldsymbol{\psi}_p := \int \boldsymbol{\psi}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}.$$

This result implies that the LSDD estimator has asymptotic normality with asymptotic order $\sqrt{1/n + 1/n'}$, which is the optimal convergence rate in the parametric setup.

2.3.2 Non-Parametric Error Bound

Next, we consider a non-parametric setup where a density-difference function is learned in a Gaussian *reproducing kernel Hilbert space* (RKHS) (Aronszajn, 1950).

Let \mathcal{H}_γ be the Gaussian RKHS with width γ :

$$k_\gamma(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{\gamma^2}\right).$$

Let us consider a slightly modified LSDD estimator that is more suitable for non-

parametric error analysis: For $n' = n$,

$$\hat{f} := \arg \min_{g \in \mathcal{H}_\gamma} \left[\|g\|_{L^2}^2 - 2 \left(\frac{1}{n} \sum_{i=1}^n g(\mathbf{x}_i) + \frac{1}{n} \sum_{i'=1}^n g(\mathbf{x}'_{i'}) \right) + \lambda \|g\|_{\mathcal{H}_\gamma}^2 \right],$$

where $\|\cdot\|_{L^2}$ denotes the L^2 -norm and $\|\cdot\|_{\mathcal{H}_\gamma}$ denotes the norm in RKHS \mathcal{H}_γ .

Then we can prove that, for all $\rho, \rho' > 0$, there exists a constant K such that, for all $\tau \geq 1$ and $n \geq 1$, the non-parametric LSDD estimator with appropriate choice of λ and γ satisfies¹

$$\|\hat{f} - f\|_{L^2}^2 + \lambda \|\hat{f}\|_{\mathcal{H}_\gamma}^2 \leq K \left(n^{-\frac{2\alpha}{2\alpha+d} + \rho} + \tau n^{-1 + \rho'} \right), \quad (7)$$

with probability not less than $1 - 4e^{-\tau}$. Here, d denotes the dimensionality of input vector \mathbf{x} , and $\alpha \geq 0$ denotes the regularity of Besov space to which the true density-difference function f belongs (smaller/larger α means f is “less/more complex”; see Appendix A for its precise definition). Because $n^{-\frac{2\alpha}{2\alpha+d}}$ is the optimal learning rate in this setup (Eberts & Steinwart, 2011), the above result shows that the non-parametric LSDD estimator achieves the optimal convergence rate.

It is known that, if the naive KDE with a Gaussian kernel is used for estimating a probability density with regularity $\alpha > 2$, the optimal learning rate cannot be achieved (Farrell, 1972; Silverman, 1986). To achieve the optimal rate by KDE, we should choose a kernel specifically tailored to each regularity α (Parzen, 1962). But such a kernel is not non-negative and it is difficult to implement in practice. On the other hand, our LSDD estimator can always achieve the optimal learning rate with a Gaussian kernel without regard to regularity α .

¹Because our theoretical result is highly technical, we only describe a rough idea here. More precise statement of the result and its complete proof are provided in Appendix A, where we utilize the mathematical technique developed in Eberts and Steinwart (2011) for a regression problem.

2.4 Model Selection by Cross-Validation

The above theoretical analyses showed the superiority of LSDD. However, the practical performance of LSDD depends on the choice of models (i.e., the kernel width σ and the regularization parameter λ). Here, we show that the model can be optimized by *cross-validation* (CV).

More specifically, we first divide the samples $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$ and $\mathcal{X}' = \{\mathbf{x}'_{i'}\}_{i'=1}^{n'}$ into T disjoint subsets $\{\mathcal{X}_t\}_{t=1}^T$ and $\{\mathcal{X}'_t\}_{t=1}^T$, respectively. Then we obtain a density-difference estimate $\widehat{f}_t(\mathbf{x})$ from $\mathcal{X} \setminus \mathcal{X}_t$ and $\mathcal{X}' \setminus \mathcal{X}'_t$ (i.e., all samples without \mathcal{X}_t and \mathcal{X}'_t), and compute its hold-out error for \mathcal{X}_t and \mathcal{X}'_t as

$$\text{CV}^{(t)} := \int \widehat{f}_t(\mathbf{x})^2 d\mathbf{x} - \frac{2}{|\mathcal{X}_t|} \sum_{\mathbf{x} \in \mathcal{X}_t} \widehat{f}_t(\mathbf{x}) + \frac{2}{|\mathcal{X}'_t|} \sum_{\mathbf{x}' \in \mathcal{X}'_t} \widehat{f}_t(\mathbf{x}'),$$

where $|\mathcal{X}|$ denotes the number of elements in the set \mathcal{X} . We repeat this hold-out validation procedure for $t = 1, \dots, T$, and compute the average hold-out error as

$$\text{CV} := \frac{1}{T} \sum_{t=1}^T \text{CV}^{(t)}.$$

Finally, we choose the model that minimizes CV.

A MATLAB[®] implementation of LSDD is available from

<http://sugiyama-www.cs.titech.ac.jp/~sugi/software/LSDD/>.

(to be made public after acceptance)

3 L^2 -Distance Estimation by LSDD

In this section, we consider the problem of approximating the L^2 -distance between $p(\mathbf{x})$ and $p'(\mathbf{x})$,

$$L^2(p, p') := \int (p(\mathbf{x}) - p'(\mathbf{x}))^2 d\mathbf{x}, \quad (8)$$

from samples $\mathcal{X} := \{\mathbf{x}_i\}_{i=1}^n$ and $\mathcal{X}' := \{\mathbf{x}'_{i'}\}_{i'=1}^{n'}$ (see Section 2.1).

3.1 Basic Form

For an equivalent expression

$$L^2(p, p') = \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x} - \int f(\mathbf{x}')p'(\mathbf{x}')d\mathbf{x}',$$

if we replace $f(\mathbf{x})$ with an LSDD estimator $\widehat{f}(\mathbf{x})$ and approximate the expectations by empirical averages, the following L^2 -distance estimator can be obtained:

$$L^2(p, p') \approx \widehat{\mathbf{h}}^\top \widehat{\boldsymbol{\theta}}. \quad (9)$$

Similarly, for another expression

$$L^2(p, p') = \int f(\mathbf{x})^2 d\mathbf{x},$$

replacing $f(\mathbf{x})$ with an LSDD estimator $\widehat{f}(\mathbf{x})$ gives another L^2 -distance estimator:

$$L^2(p, p') \approx \widehat{\boldsymbol{\theta}}^\top \mathbf{H} \widehat{\boldsymbol{\theta}}. \quad (10)$$

3.2 Reduction of Bias Caused by Regularization

Eq.(9) and Eq.(10) themselves give approximations to $L^2(p, p')$. Nevertheless, we argue that the use of their combination, defined by

$$\widehat{L}^2(\mathcal{X}, \mathcal{X}') := 2\widehat{\mathbf{h}}^\top \widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}^\top \mathbf{H} \widehat{\boldsymbol{\theta}}, \quad (11)$$

is more sensible. To explain the reason, let us consider a generalized L^2 -distance estimator of the following form:

$$\beta \widehat{\mathbf{h}}^\top \widehat{\boldsymbol{\theta}} + (1 - \beta) \widehat{\boldsymbol{\theta}}^\top \mathbf{H} \widehat{\boldsymbol{\theta}}, \quad (12)$$

where β is a real scalar. If the regularization parameter λ (≥ 0) is small, then Eq.(12) can be expressed as

$$\beta \widehat{\mathbf{h}}^\top \widehat{\boldsymbol{\theta}} + (1 - \beta) \widehat{\boldsymbol{\theta}}^\top \mathbf{H} \widehat{\boldsymbol{\theta}} = \widehat{\mathbf{h}}^\top \mathbf{H}^{-1} \widehat{\mathbf{h}} - \lambda(2 - \beta) \widehat{\mathbf{h}}^\top \mathbf{H}^{-2} \widehat{\mathbf{h}} + o_p(\lambda), \quad (13)$$

where o_p denotes the probabilistic order (its derivation is given in Appendix B).

Thus, the bias introduced by regularization (i.e., the second term in the right-hand side of Eq.(13) that depends on λ) can be eliminated if $\beta = 2$, which yields Eq.(11). Note that, if no regularization is imposed (i.e., $\lambda = 0$), both Eq.(9) and Eq.(10) yield $\widehat{\mathbf{h}}^\top \mathbf{H}^{-1} \widehat{\mathbf{h}}$, the first term in the right-hand side of Eq.(13).

Eq.(11) is actually equivalent to the negative of the optimal objective value of the LSDD optimization problem without regularization (i.e., Eq.(4) with $\lambda = 0$). This can be naturally interpreted through a lower bound of $L^2(p, p')$ obtained by *Legendre-Fenchel*

convex duality (Rockafellar, 1970):

$$L^2(p, p') = \sup_g \left[2 \left(\int g(\mathbf{x})p(\mathbf{x})d\mathbf{x} - \int g(\mathbf{x})p'(\mathbf{x})d\mathbf{x} \right) - \int g(\mathbf{x})^2d\mathbf{x} \right],$$

where the supremum is attained at $g = f$. If the expectations are replaced by empirical estimators and the linear-in-parameter model (2) is used as g , the above optimization problem is reduced to the LSDD objective function without regularization (see Eq.(4)). Thus, LSDD corresponds to approximately maximizing the above lower bound and Eq.(11) is its maximum value.

Through eigenvalue decomposition of \mathbf{H} , we can show that

$$2\widehat{\mathbf{h}}^\top \widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}^\top \mathbf{H} \widehat{\boldsymbol{\theta}} \geq \widehat{\mathbf{h}}^\top \widehat{\boldsymbol{\theta}} \geq \widehat{\boldsymbol{\theta}}^\top \mathbf{H} \widehat{\boldsymbol{\theta}}.$$

Thus, our approximator (11) is not less than the plain approximators (9) and (10).

3.3 Further Bias Correction

$\widehat{\mathbf{h}}^\top \mathbf{H}^{-1} \widehat{\mathbf{h}}$, the first term in Eq.(13), is an essential part of the L^2 -distance estimator (11). However, it is actually a slightly biased estimator of the target quantity $\mathbf{h}^\top \mathbf{H}^{-1} \mathbf{h}$ ($= \boldsymbol{\theta}^{*\top} \mathbf{H} \boldsymbol{\theta}^* = \mathbf{h}^\top \boldsymbol{\theta}^*$):

$$\mathbb{E}[\widehat{\mathbf{h}}^\top \mathbf{H}^{-1} \widehat{\mathbf{h}}] = \mathbf{h}^\top \mathbf{H}^{-1} \mathbf{h} + \text{tr} \left(\mathbf{H}^{-1} \left(\frac{1}{n} \mathbf{V}_p + \frac{1}{n'} \mathbf{V}_{p'} \right) \right), \quad (14)$$

where \mathbb{E} denotes the expectation over all samples $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$ and $\mathcal{X}' = \{\mathbf{x}'_{i'}\}_{i'=1}^{n'}$, and \mathbf{V}_p and $\mathbf{V}_{p'}$ are defined by Eq.(6) (its derivation is given in Appendix C).

The second term in the right-hand side of Eq.(14) is an estimation bias that is generally non-zero. Thus, based on Eq.(14), we can construct a bias-corrected L^2 -distance estimator

as

$$\tilde{L}^2(\mathcal{X}, \mathcal{X}') := 2\hat{\mathbf{h}}^\top \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^\top \mathbf{H} \hat{\boldsymbol{\theta}} - \text{tr} \left(\mathbf{H}^{-1} \left(\frac{1}{n} \hat{\mathbf{V}}_p + \frac{1}{n'} \hat{\mathbf{V}}_{p'} \right) \right), \quad (15)$$

where $\hat{\mathbf{V}}_p$ is an empirical estimator of covariance matrix \mathbf{V}_p :

$$\hat{\mathbf{V}}_p := \frac{1}{n} \sum_{i=1}^n \left(\boldsymbol{\psi}(\mathbf{x}_i) - \hat{\boldsymbol{\psi}}_p \right) \left(\boldsymbol{\psi}(\mathbf{x}_i) - \hat{\boldsymbol{\psi}}_p \right)^\top,$$

and $\hat{\boldsymbol{\psi}}_p$ is an empirical estimator of the expectation $\boldsymbol{\psi}_p$:

$$\hat{\boldsymbol{\psi}}_p := \frac{1}{n} \sum_{i=1}^n \boldsymbol{\psi}(\mathbf{x}_i).$$

The true L^2 -distance is non-negative by definition (see Eq.(8)), but the above bias-corrected estimate can take a negative value. Following the same line as Baranchik (1964), the *positive-part* estimator may be more accurate:

$$\bar{L}^2(\mathcal{X}, \mathcal{X}') := \max \left\{ 0, \tilde{L}^2(\mathcal{X}, \mathcal{X}') \right\}.$$

However, in our preliminary experiments, $\bar{L}^2(\mathcal{X}, \mathcal{X}')$ does not always perform well particularly when \mathbf{H} is ill-conditioned. For this reason, we practically propose to use $\hat{L}^2(\mathcal{X}, \mathcal{X}')$ defined by Eq.(11).

4 Experiments

In this section, we experimentally evaluate the performance of LSDD.

4.1 Numerical Examples

First, we show numerical examples using artificial datasets.

4.1.1 LSDD vs. KDE

We experimentally compare the behavior of LSDD and the KDE-based method. Let

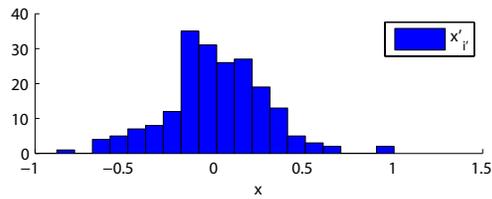
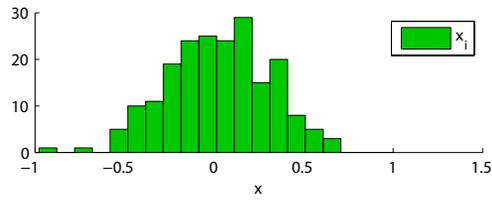
$$p(\mathbf{x}) = N(\mathbf{x}; (\mu, 0, \dots, 0)^\top, (4\pi)^{-1} \mathbf{I}_d),$$

$$p'(\mathbf{x}) = N(\mathbf{x}; (0, 0, \dots, 0)^\top, (4\pi)^{-1} \mathbf{I}_d),$$

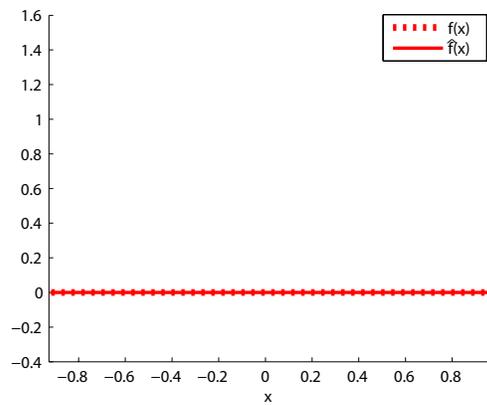
where $N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multi-dimensional normal density with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ with respect to \mathbf{x} , and \mathbf{I}_d denotes the d -dimensional identity matrix.

We first illustrate how LSDD and KDE behave under $d = 1$ and $n = n' = 200$. Figure 1 depicts the data samples, densities and density difference estimated by KDE, and density difference estimated by LSDD for $\mu = 0$ (i.e., $f(x) = p(x) - p'(x) = 0$). This shows that LSDD gives a more accurate estimate of the density difference $f(x)$ than KDE. Figure 2 depicts the results for $\mu = 0.5$ (i.e., $f(x) \neq 0$), showing again that LSDD performs well.

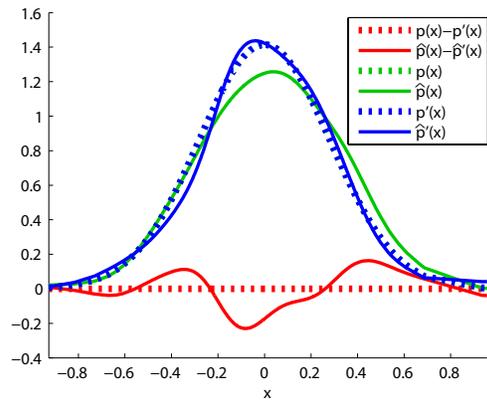
Next, we compare the L^2 -distance approximator based on LSDD and that based on KDE. For $\mu = 0, 0.2, 0.4, 0.6, 0.8$ and $d = 1, 5$, we draw $n = n' = 200$ samples from the above $p(\mathbf{x})$ and $p'(\mathbf{x})$. Figure 3 depicts the mean and standard error of estimated L^2 -distances over 100 runs as functions of mean μ . When $d = 1$, the LSDD-based L^2 -distance estimator gives accurate estimates of the true L^2 -distance, whereas the KDE-based L^2 -distance estimator slightly underestimates the true L^2 -distance. This is caused by the fact that KDE tends to provide smoother density estimates (see Figure 2(c) again). Such smoother density estimates are accurate as density estimates, but the difference of smoother density estimates yields a smaller L^2 -distance estimate (Anderson et al., 1994).



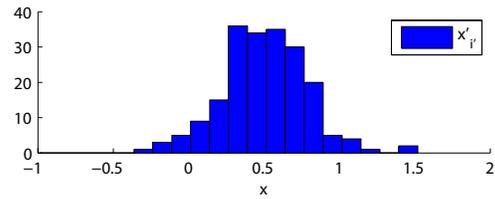
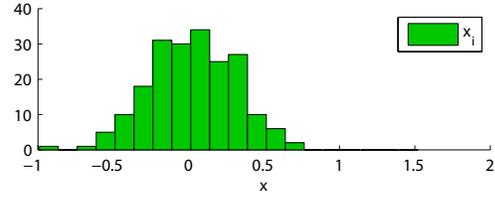
(a) Samples



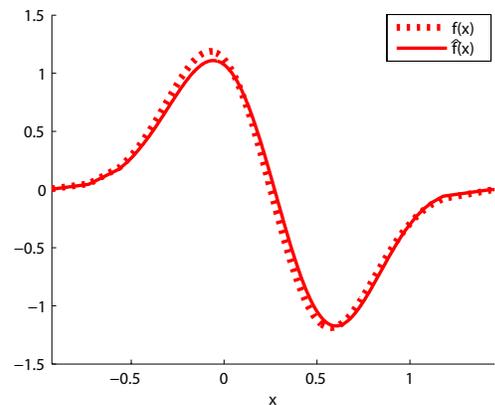
(b) LSDD



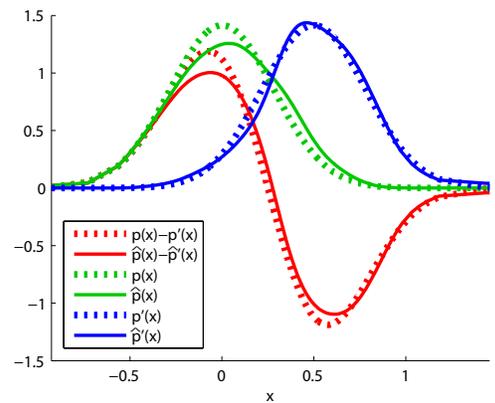
(c) KDE



(a) Samples



(b) LSDD



(c) KDE

Figure 1: Estimation of density difference when $\mu = 0$ (i.e., $f(x) = p(x) - p'(x) = 0$).

Figure 2: Estimation of density difference when $\mu = 0.5$ (i.e., $f(x) = p(x) - p'(x) \neq 0$).

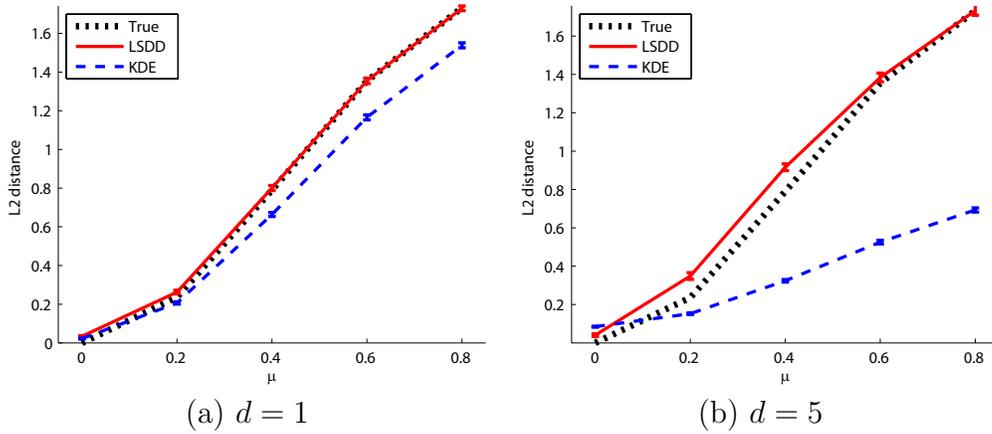


Figure 3: L^2 -distance estimation by LSDD and KDE. Means and standard errors over 100 runs are plotted.

This tendency is more significant when $d = 5$; the KDE-based L^2 -distance estimator severely underestimates the true L^2 -distance, which is a typical drawback of the two-step procedure. On the other hand, the LSDD-based L^2 -distance estimator still gives reasonably accurate estimates of the true L^2 -distance even when $d = 5$.

4.1.2 L^2 -Distance vs. KL-Divergence

The *Kullback-Leibler* (KL) divergence (Kullback & Leibler, 1951) is a popular divergence measure for comparing probability distributions. The KL-divergence from $p(\mathbf{x})$ to $p'(\mathbf{x})$ is defined as

$$\text{KL}(p||p') := \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{p'(\mathbf{x})} d\mathbf{x}.$$

First, we illustrate the difference between the L^2 -distance and the KL-divergence. For $d = 1$, let

$$\begin{aligned} p(x) &= (1 - \eta)N(x; 0, 1^2) + \eta N(x; \mu, 1/4^2), \\ p'(x) &= N(x; 0, 1^2). \end{aligned}$$

Implications of the above densities are that samples drawn from $N(x; 0, 1^2)$ are inliers, whereas samples drawn from $N(x; \mu, 1/4^2)$ are outliers. We set the outlier rate at $\eta = 0.1$ and the outlier mean at $\mu = 0, 2, 4, \dots, 10$ (see Figure 4).

Figure 5 depicts the L^2 -distance and the KL-divergence for outlier mean $\mu = 0, 2, 4, \dots, 10$. This shows that both the L^2 -distance and the KL-divergence increase as μ increases. However, the L^2 -distance is bounded from above, whereas the KL-divergence diverges to infinity as μ tends to infinity. This result implies that the L^2 -distance is less sensitive to outliers than the KL-divergence, which well agrees with the observation given in Basu et al. (1998).

Next, we draw $n = n' = 100$ samples from $p(x)$ and $p'(x)$, and estimate the L^2 -distance by LSDD and the KL-divergence by the *Kullback-Leibler importance estimation procedure*² (KLIEP) (Sugiyama et al., 2008; Nguyen et al., 2010). Figure 6 depicts the estimated L^2 -distance and KL-divergence for outlier mean $\mu = 0, 2, 4, \dots, 10$ over 100 runs. This shows that both LSDD and KLIEP reasonably capture the profiles of the true L^2 -distance and the KL-divergence, although the scale of KLIEP values is much different from the true values (see Figure 5) because the estimated normalization factor was unreliable.

Finally, based on the *permutation test* procedure (Efron & Tibshirani, 1993), we conduct hypothesis testing of the null hypothesis that densities p and p' are the same. More specifically, we first compute a distance estimator for the original datasets \mathcal{X} and \mathcal{X}' and obtain $\widehat{D}(\mathcal{X}, \mathcal{X}')$. Next, we randomly permute the $|\mathcal{X} \cup \mathcal{X}'|$ samples, and assign the first $|\mathcal{X}|$ samples to a set $\widetilde{\mathcal{X}}$ and the remaining $|\mathcal{X}'|$ samples to another set $\widetilde{\mathcal{X}'}$. Then we compute the distance estimator again using the randomly permuted datasets $\widetilde{\mathcal{X}}$ and $\widetilde{\mathcal{X}'}$ and obtain $\widetilde{D}(\widetilde{\mathcal{X}}, \widetilde{\mathcal{X}'})$. Since $\widetilde{\mathcal{X}}$ and $\widetilde{\mathcal{X}'}$ can be regarded as being drawn from the same distri-

²Estimation of the KL-divergence from data has been extensively studied recently (Wang et al., 2005; Sugiyama et al., 2008; Pérez-Cruz, 2008; Silva & Narayanan, 2010; Nguyen et al., 2010). Among them, KLIEP was shown to possess a superior convergence property and demonstrated to work well in practice. KLIEP is based on direct estimation of density ratio $p(\mathbf{x})/p'(\mathbf{x})$ without density estimation of $p(\mathbf{x})$ and $p'(\mathbf{x})$.

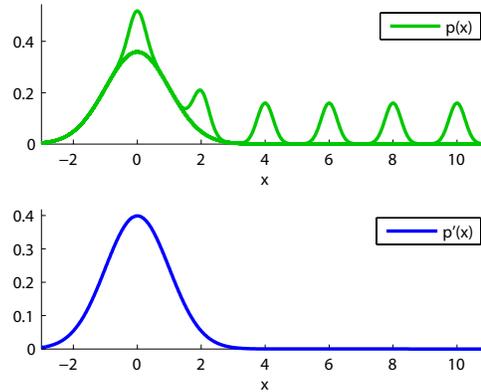


Figure 4: Comparing two densities in the presence of outliers. $p(x)$ includes outliers at $\mu = 0, 2, 4, \dots, 10$.

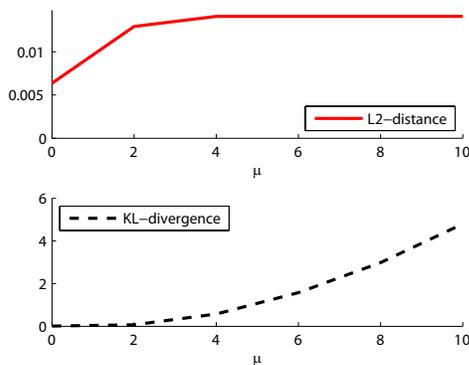


Figure 5: The true L^2 -distance and true KL-divergence as functions of outlier mean μ .

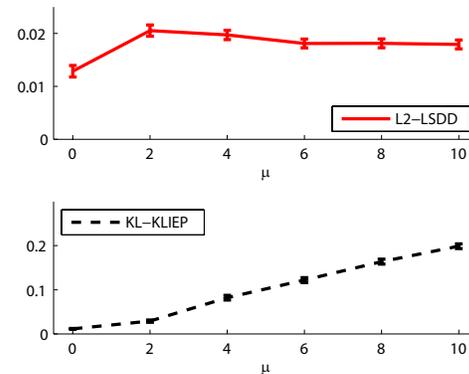


Figure 6: Means and standard errors of L^2 -distance estimation by LSDD and KL-divergence estimation by KLIEP over 100 runs.

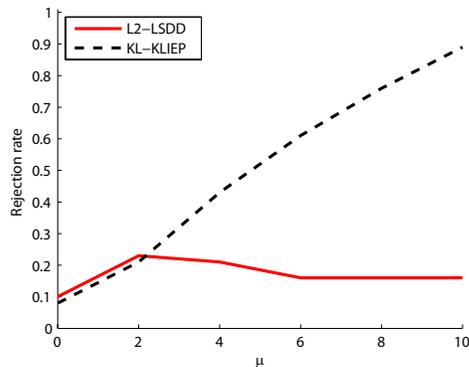


Figure 7: Two-sample test for outlier rate $\eta = 0.1$ as functions of outlier mean μ .

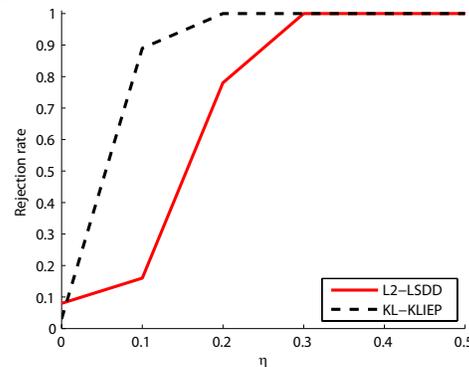


Figure 8: Two-sample test for outlier mean $\mu = 10$ as functions of outlier rate η .

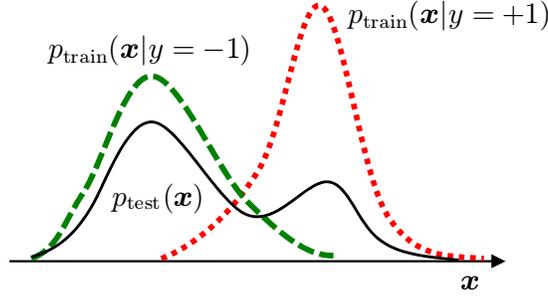
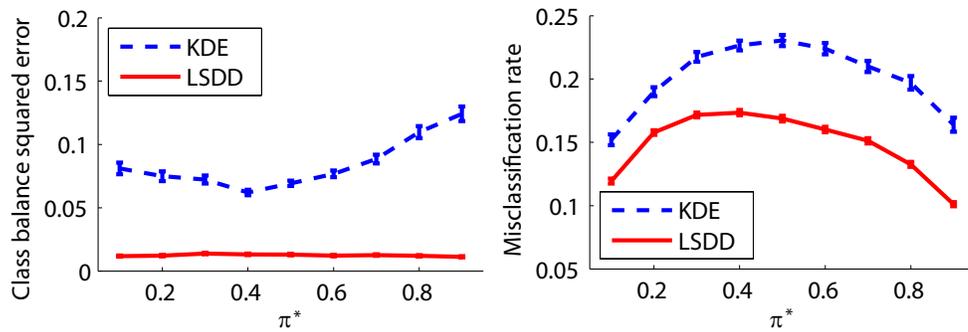


Figure 9: Schematic illustration of semi-supervised class-balance estimation.

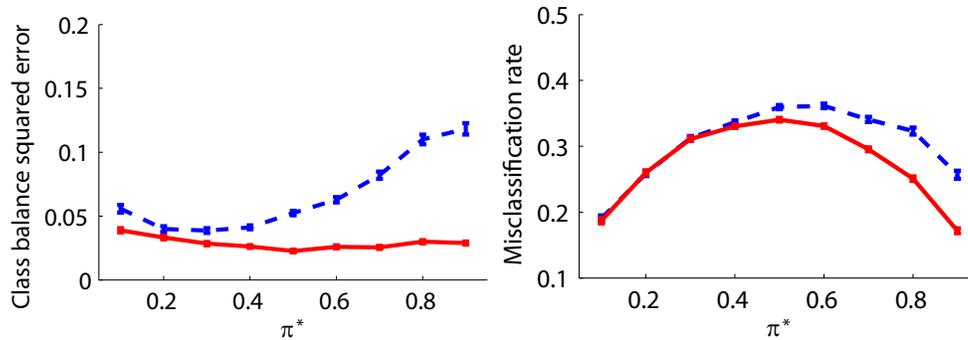
bution, $\tilde{D}(\tilde{\mathcal{X}}, \tilde{\mathcal{X}}')$ would take a value close to zero. This random permutation procedure is repeated many times, and the distribution of $\tilde{D}(\tilde{\mathcal{X}}, \tilde{\mathcal{X}}')$ under the null hypothesis (i.e., the two distributions are the same) is constructed. Finally, the p-value is approximated by evaluating the relative ranking of $\hat{D}(\mathcal{X}, \mathcal{X}')$ in the histogram of $\tilde{D}(\tilde{\mathcal{X}}, \tilde{\mathcal{X}}')$. We set the significance level at 5%.

Figure 7 depicts the rejection rate of the null hypothesis for outlier rate $\eta = 0.1$ and outlier mean $\mu = 0, 2, 4, \dots, 10$, based on the L^2 -distance estimated by LSDD and the KL-divergence estimated by KLIEP. This shows that the KLIEP-based test rejects the null hypothesis more frequently for large μ , whereas the rejection rate of the LSDD-based test is kept almost constant even when μ is changed. This result implies that the two-sample test by LSDD is more robust against outliers (i.e., two distributions tend to be regarded as the same even in the presence of outliers) than the KLIEP-based test.

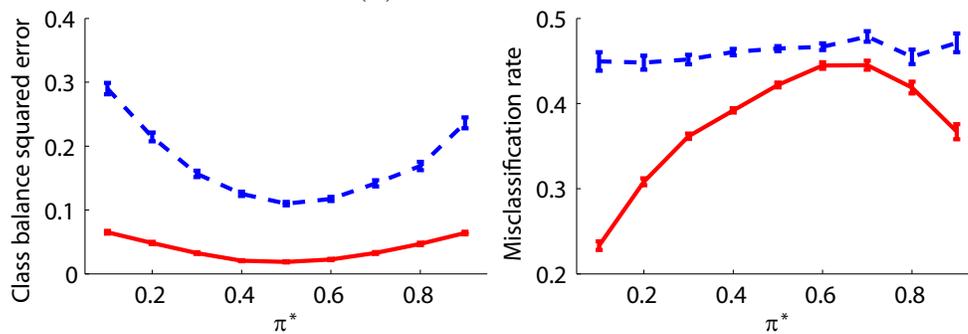
Figure 8 depicts the rejection rate of the null hypothesis for outlier mean $\mu = 10$ for outlier rate $\eta = 0, 0.05, 0.1, \dots, 0.35$. When $\eta = 0$ (i.e., no outliers), both the LSDD-based test and the KLIEP-based test accept the null hypothesis with the designated significance level approximately. When $\eta = 0.1$, the LSDD-based test still keeps a low rejection rate, whereas the KLIEP-based test tends to reject the null hypothesis. When $\eta \geq 0.3$, the LSDD-based test and the KLIEP-based test tend to reject the null hypothesis in a similar way.



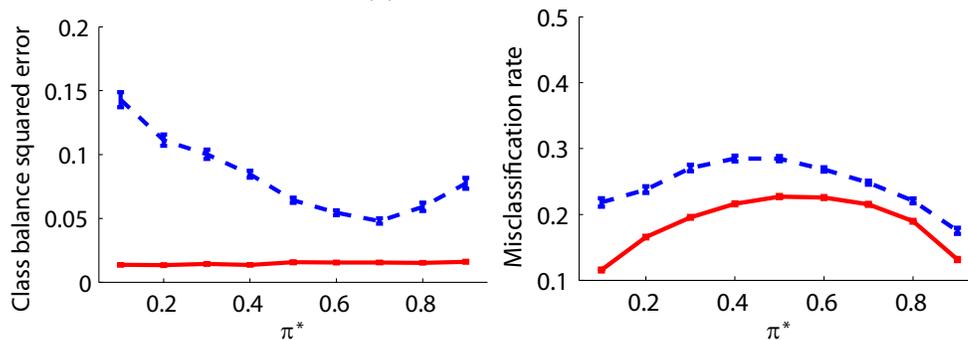
(a) Australian dataset



(b) Diabetes dataset



(c) German dataset



(d) Statlogheart dataset

Figure 10: Results of semi-supervised class-balance estimation. Left: Squared error of class balance estimation. Right: Misclassification error by a weighted regularized least-squares classifier.

4.2 Applications

Next, we apply LSDD to semi-supervised class-balance estimation under class prior change and change-point detection in time series.

4.2.1 Semi-Supervised Class-Balance Estimation

In real-world pattern recognition tasks, changes in class balance are often observed. Then significant estimation bias can be caused since the class balance in the training dataset does not reflect that of the test dataset.

Here, we consider a pattern recognition task of classifying pattern $\mathbf{x} \in \mathbb{R}^d$ to class $y \in \{+1, -1\}$. Our goal is to learn the class balance of a test dataset in a semi-supervised learning setup where unlabeled test samples are provided in addition to labeled training samples (Chapelle et al., 2006). The class balance in the test set can be estimated by matching a mixture of class-wise training input densities,

$$\pi p_{\text{train}}(\mathbf{x}|y = +1) + (1 - \pi)p_{\text{train}}(\mathbf{x}|y = -1),$$

with the test input density $p_{\text{test}}(\mathbf{x})$ (Saerens et al., 2002), where $\pi \in [0, 1]$ is a mixing coefficient to learn. See Figure 9 for schematic illustration. Here, we use the L^2 -distance estimated by LSDD and the difference of KDEs for this distribution matching.

We use four UCI benchmark datasets³, where we randomly choose 20 labeled training samples from each class and 50 unlabeled test samples following true class-prior $\pi^* = 0.1, 0.2, \dots, 0.9$. Figure 10 plots the mean and standard error of the squared difference between true and estimated class balances π and the misclassification error by a weighted regularized least-squares classifier (Rifkin et al., 2003) over 1000 runs. The results show that LSDD tends to provide better class-balance estimates, which are translated into

³<http://archive.ics.uci.edu/ml/>

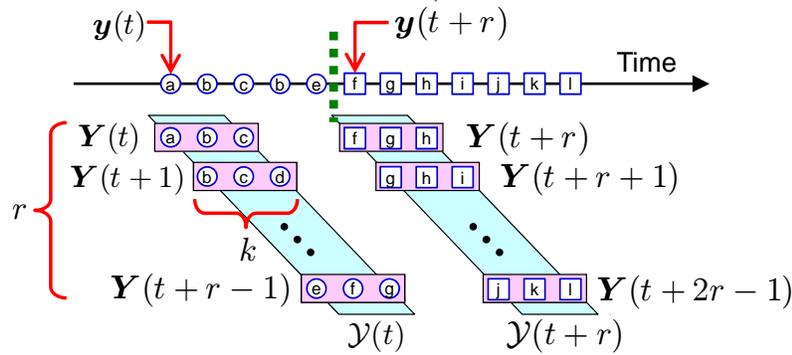


Figure 11: Schematic illustration of unsupervised change detection.

lower classification errors.

4.2.2 Unsupervised Change Detection

The objective of change detection is to discover abrupt property changes behind time-series data.

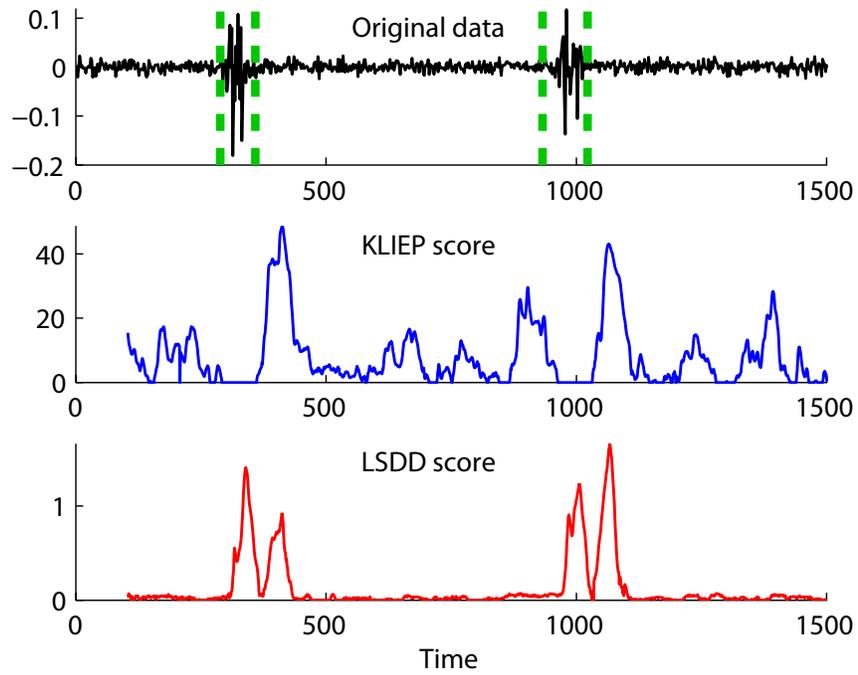
Let $\mathbf{y}(t) \in \mathbb{R}^m$ be an m -dimensional time-series sample at time t , and let

$$\mathbf{Y}(t) := [\mathbf{y}(t)^\top, \mathbf{y}(t+1)^\top, \dots, \mathbf{y}(t+k-1)^\top]^\top \in \mathbb{R}^{km}$$

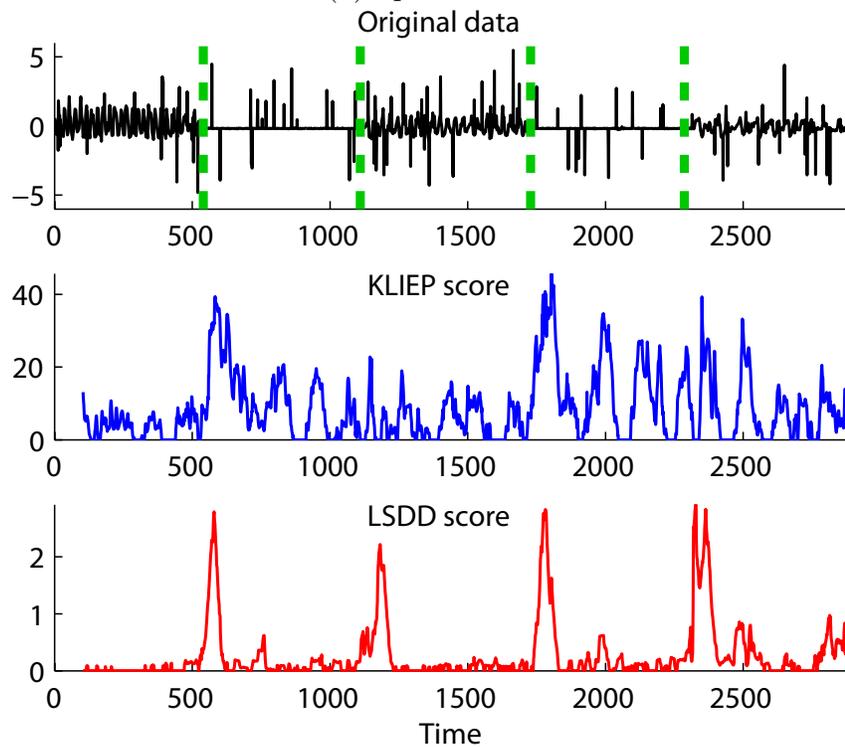
be a subsequence of time series at time t with length k . We treat the subsequence $\mathbf{Y}(t)$ as a sample, instead of a single point $\mathbf{y}(t)$, by which time-dependent information can be incorporated naturally (Kawahara & Sugiyama, 2012). Let $\mathcal{Y}(t)$ be a set of r retrospective subsequence samples starting at time t :

$$\mathcal{Y}(t) := \{\mathbf{Y}(t), \mathbf{Y}(t+1), \dots, \mathbf{Y}(t+r-1)\}.$$

Our strategy is to compute a certain dissimilarity measure between two consecutive segments $\mathcal{Y}(t)$ and $\mathcal{Y}(t+r)$, and use it as the plausibility of change points (see Figure 11). As a dissimilarity measure, we use the L^2 -distance estimated by LSDD and the KL-divergence



(a) Speech data



(b) Accelerometer data

Figure 12: Results of unsupervised change detection. Top: Original time-series. Middle: Change scores obtained by KLIEP. Bottom: Change scores obtained by LSDD.

estimated by the *KL importance estimation procedure* (KLIEP) (Sugiyama et al., 2008; Nguyen et al., 2010). We set $k = 5$ and $r = 50$.

First, we use the *IPSS SIG-SLP Corpora and Environments for Noisy Speech Recognition* (CENSREC) dataset⁴ provided by the *National Institute of Informatics, Japan*, which records human voice in a noisy environment such as a restaurant. The top graph in Figure 12(a) displays the original time-series, where true change points were manually annotated. The bottom two graphs in Figure 12(a) plot change scores obtained by KLIEP and LSDD, showing that the LSDD-based change score indicates the existence of change points more clearly than the KLIEP-based change score.

Next, we use a dataset taken from the *Human Activity Sensing Consortium (HASC) challenge 2011*⁵, which provides human activity information collected by portable three-axis accelerometers. Because the orientation of the accelerometers is not necessarily fixed, we take the ℓ_2 -norm of the 3-dimensional data. The top graph in Figure 12(b) displays the original time-series for a sequence of actions “jog”, “stay”, “stair down”, “stay”, and “stair up” (there exists 4 change points at time 540, 1110, 1728, and 2286). The bottom two graphs in Figure 12(b) depict the change scores obtained by KLIEP and LSDD, showing that the LSDD score is much more stable and interpretable than the KLIEP score.

5 Conclusions

In this paper, we proposed a method for directly estimating the difference between two probability density functions without density estimation. The proposed method, called the *least-squares density-difference* (LSDD), was derived within a framework of kernel least-squares estimation, and its solution can be computed analytically in a computation-

⁴<http://research.nii.ac.jp/src/en/CENSREC-1-C.html>

⁵<http://hasc.jp/hc2011/>

ally efficient and stable manner. Furthermore, LSDD is equipped with cross-validation, and thus all tuning parameters such as the kernel width and the regularization parameter can be systematically and objectively optimized. We showed the asymptotic normality of LSDD in a parametric setup and derived a finite-sample error bound for LSDD in a non-parametric setup. In both cases, LSDD achieves the optimal convergence rate.

We also proposed an L^2 -distance estimator based on LSDD, which nicely cancels a bias caused by regularization. The LSDD-based L^2 -distance estimator was experimentally shown to be more accurate than the difference of kernel density estimators and more robust against outliers than Kullback-Leibler divergence estimation.

Density-difference estimation is a novel research paradigm in machine learning, and we have given a simple but useful method for this emerging topic. Our future work will develop more powerful algorithms for density-difference estimation and explores a variety of applications.

Acknowledgments

The authors would like to thank Wittawat Jitkrittum for his comments. Masashi Sugiyama was supported by MEXT KAKENHI 23300069, Takafumi Kanamori was supported by MEXT KAKENHI 24500340, Taiji Suzuki was supported by MEXT KAKENHI 22700289 and the Aihara Project, the FIRST program from JSPS initiated by CSTP, Marthinus Christoffel du Plessis was supported by MEXT Scholarship, Song Liu was supported by the JST PRESTO program, and Ichiro Takeuchi was supported by MEXT KAKENHI 23700165.

References

- Anderson, N., Hall, P., & Titterington, D. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, *50*, 41–54.
- Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, *68*, 337–404.
- Atif, J., Ripoche, X., & Osorio, A. (2003). Non-rigid medical image registration by maximisation of quadratic mutual information. *IEEE 29th Annual Northeast Bioengineering Conference* (pp. 32–40).
- Baranchik, A. J. (1964). *Multiple regression and estimation of the mean of a multivariate normal distribution* (Technical Report 51). Department of Statistics, Stanford University, Stanford, CA, USA.
- Basu, A., Harris, I. R., Hjort, N. L., & Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, *85*, 549–559.
- Besbeas, P., & Morgan, B. J. T. (2004). Integrated squared error estimation of normal mixtures. *Computational Statistics & Data Analysis*, *44*, 517–526.
- Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory* (pp. 144–152). ACM Press.
- Chapelle, O., Schölkopf, B., & Zien, A. (Eds.). (2006). *Semi-supervised learning*. Cambridge, MA, USA: MIT Press.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

- Du Plessis, M. C., & Sugiyama, M. (2012). Semi-supervised learning of class balance under class-prior change by distribution matching. *Proceedings of 29th International Conference on Machine Learning (ICML2012)*. Edinburgh, Scotland.
- Duong, T., Koch, I., & Wand, M. P. (2009). Highest density difference region estimation with application to flow cytometric data. *Biometrical Journal*, *51*, 504–521.
- Eberts, M., & Steinwart, I. (2011). Optimal learning rates for least squares SVMs using Gaussian kernels. In J. Shawe-Taylor, R. S. Zemel, P. Bartlett, F. C. N. Pereira and K. Q. Weinberger (Eds.), *Advances in neural information processing systems 24*, 1539–1547.
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY, USA: Chapman & Hall/CRC.
- Farrell, R. H. (1972). On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *The Annals of Mathematical Statistics*, *43*, 170–180.
- Gray, D. M., & Principe, J. C. (2010). Quadratic mutual information for dimensionality reduction and classification. *Proceedings of SPIE* (p. 76960D).
- Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., & Schölkopf, B. (2009). Covariate shift by kernel mean matching. In J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer and N. Lawrence (Eds.), *Dataset shift in machine learning*, chapter 8, 131–160. Cambridge, MA, USA: MIT Press.
- Hall, P., & Wand, M. P. (1988). On nonparametric discrimination using density differences. *Biometrika*, *75*, 541–547.
- Härdle, W., Müller, M., Sperlich, S., & Werwatz, A. (2004). *Nonparametric and semi-parametric models*. Berlin, Germany: Springer.

- Kanamori, T., Hido, S., & Sugiyama, M. (2009). A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, *10*, 1391–1445.
- Kanamori, T., Suzuki, T., & Sugiyama, M. (2012). Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, *86*, 335–367.
- Kawahara, Y., & Sugiyama, M. (2012). Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining*, *5*, 114–127.
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*, 79–86.
- Liu, B., Cheng, H. D., Huang, J., Tian, J., Tang, X., & Liu, J. (2010). Probability density difference-based active contour for ultrasound image segmentation. *Pattern Recognition*, *43*, 2028–2042.
- Liu, S., Yamada, M., Collier, N., & Sugiyama, M. (2012). *Change-point detection in time-series data by relative density-ratio estimation* (Technical Report 1203.0453). arXiv.
- Matsugu, M., Yamanaka, M., & Sugiyama, M. (2011). Detection of activities and events without explicit categorization. *Proceedings of the 3rd International Workshop on Video Event Categorization, Tagging and Retrieval for Real-World Applications (VEC-TaR2011)* (pp. 1532–1539). Barcelona, Spain.
- 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*, 5847–5861.
- Parzen, E. (1962). On the estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, *33*, 1065–1076.

- Pérez-Cruz, F. (2008). Kullback-Leibler divergence estimation of continuous distributions. *Proceedings of IEEE International Symposium on Information Theory* (pp. 1666–1670). Nice, France.
- Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, *85*, 619–630.
- Rao, C. R. (1965). *Linear statistical inference and its applications*. New York, NY, USA: Wiley.
- Rifkin, R., Yeo, G., & Poggio, T. (2003). Regularized least-squares classification. *Advances in Learning Theory: Methods, Models and Applications* (pp. 131–154). Amsterdam, the Netherlands: IOS Press.
- Rockafellar, R. T. (1970). *Convex analysis*. Princeton, NJ, USA: Princeton University Press.
- Saerens, M., Latinne, P., & Decaestecker, C. (2002). Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, *14*, 21–41.
- Scott, D. W. (2001). Parametric statistical modeling by minimum integrated square error. *Technometrics*, *43*, 274–285.
- Silva, J., & Narayanan, S. S. (2010). Information divergence estimation based on data-dependent partitions. *Journal of Statistical Planning and Inference*, *140*, 3180–3198.
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London, UK: Chapman and Hall.
- Steinwart, I., & Christmann, A. (2008). *Support vector machines*. New York, NY, USA: Springer.

- Steinwart, I., & Scovel, C. (2004). Fast rates for support vector machines using Gaussian kernels. *The Annals of Statistics*, *35*, 575–607.
- Sugiyama, M., Suzuki, T., & Kanamori, T. (2012a). *Density ratio estimation in machine learning*. Cambridge, UK: Cambridge University Press.
- Sugiyama, M., Suzuki, T., & Kanamori, T. (2012b). Density ratio matching under the Bregman divergence: A unified framework of density ratio estimation. *Annals of the Institute of Statistical Mathematics*. to appear.
- Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Büna, P., & Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, *60*, 699–746.
- Torkkola, K. (2003). Feature extraction by non-parametric mutual information maximization. *Journal of Machine Learning Research*, *3*, 1415–1438.
- Vapnik, V. N. (1998). *Statistical learning theory*. New York, NY, USA: Wiley.
- Wang, Q., Kulkarni, S. R., & Verdú, S. (2005). Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, *51*, 3064–3074.

A Technical Details of Non-Parametric Convergence Analysis in Section 2.3.2

First, we define linear operators P_n, P, P'_n, P', Q_n, Q as

$$\begin{aligned} P_n f &:= \frac{1}{n} \sum_{i=1}^n f(\mathbf{x}_i), & P f &:= \int_{\mathbb{R}^d} f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}, \\ P'_n f &:= \frac{1}{n} \sum_{i=1}^n f(\mathbf{x}'_i), & P' f &:= \int_{\mathbb{R}^d} f(\mathbf{x}) p'(\mathbf{x}) d\mathbf{x}, \\ Q_n f &:= P_n f - P'_n f, & Q f &:= P f - P' f. \end{aligned}$$

Let \mathcal{H}_γ be an RKHS endowed with the Gaussian kernel with width γ :

$$k_\gamma(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{\gamma^2}\right).$$

A density-difference estimator \hat{f} is obtained as

$$\hat{f} := \arg \min_{f \in \mathcal{H}_\gamma} \left[\|f\|_{L^2(\mathbb{R}^d)}^2 - 2Q_n f + \lambda \|f\|_{\mathcal{H}_\gamma}^2 \right].$$

We assume the following conditions:

Assumption 1. *The densities are bounded: There exists M such that*

$$\|p\|_\infty \leq M \quad \text{and} \quad \|p'\|_\infty \leq M.$$

The density difference $f = p - p'$ is a member of Besov space with regularity α : $f \in B_{2,\infty}^\alpha$ and, for $r = \lfloor \alpha \rfloor + 1$ where $\lfloor \alpha \rfloor$ denotes the largest integer less than or equal to α ,

$$\|f\|_{B_{2,\infty}^\alpha} := \|f\|_{L_2(\mathbb{R}^d)} + \sup_{t>0} (t^{-\alpha} \omega_{r,L_2(\mathbb{R}^d)}(f, t)) < c,$$

where $B_{2,\infty}^\alpha$ is the Besov space with regularity α and $\omega_{r,L_2(\mathbb{R}^d)}$ is the r -th modulus of smoothness (see Eberts and Steinwart (2011) for the definitions).

Then we have the following theorem.

Theorem 2. *Suppose Assumption 1 is satisfied. Then, for all $\epsilon > 0$ and $p \in (0, 1)$, there exists a constant $K > 0$ depending on M, c, ϵ, p such that for all $n \geq 1$, $\tau \geq 1$, and $\lambda > 0$, the LSDD estimator \hat{f} in \mathcal{H}_γ satisfies*

$$\|\hat{f} - f\|_{L^2(\mathbb{R}^d)}^2 + \lambda \|\hat{f}\|_{\mathcal{H}_\gamma}^2 \leq K \left(\lambda \gamma^{-d} + \gamma^{2\alpha} + \frac{\gamma^{-(1-p)(1+\epsilon)d}}{\lambda^p n} + \frac{\gamma^{-\frac{2(1-p)d}{1+p}(1+\epsilon+\frac{1-p}{4})}}{\lambda^{\frac{3p-p^2}{1+p}} n^{\frac{2}{1+p}}} + \frac{\tau}{n^2 \lambda} + \frac{\tau}{n} \right),$$

with probability not less than $1 - 4e^{-\tau}$.

To prove this, we utilize the technique developed in Eberts and Steinwart (2011) for a regression problem.

Proof. First, note that

$$\|\hat{f}\|_{L^2(\mathbb{R}^d)}^2 - 2Q_n \hat{f} + \|f\|_{L^2(\mathbb{R}^d)}^2 + \lambda \|\hat{f}\|_{\mathcal{H}_\gamma}^2 \leq \|f_0\|_{L^2(\mathbb{R}^d)}^2 - 2Q_n f_0 + \|f\|_{L^2(\mathbb{R}^d)}^2 + \lambda \|f_0\|_{\mathcal{H}_\gamma}^2.$$

Therefore, we have

$$\begin{aligned} & \|\hat{f} - f\|_{L^2(\mathbb{R}^d)}^2 + \lambda \|\hat{f}\|_{\mathcal{H}_\gamma}^2 \\ &= \|\hat{f}\|_{L^2(\mathbb{R}^d)}^2 - 2Q_n \hat{f} + \|f\|_{L^2(\mathbb{R}^d)}^2 + 2(Q_n - Q)\hat{f} + \lambda \|\hat{f}\|_{\mathcal{H}_\gamma}^2 \\ &\leq \|f_0\|_{L^2(\mathbb{R}^d)}^2 - 2Q_n f_0 + \|f\|_{L^2(\mathbb{R}^d)}^2 + 2(Q_n - Q)\hat{f} + \lambda \|\hat{f}\|_{\mathcal{H}_\gamma}^2 \\ &= \|f_0\|_{L^2(\mathbb{R}^d)}^2 - 2Q f_0 + \|f\|_{L^2(\mathbb{R}^d)}^2 + 2(Q_n - Q)(\hat{f} - f_0) + \lambda \|\hat{f}\|_{\mathcal{H}_\gamma}^2 \\ &= \|f_0 - f\|_{L^2(\mathbb{R}^d)}^2 + 2(Q_n - Q)(\hat{f} - f) + 2(Q_n - Q)(f - f_0) + \lambda \|\hat{f}\|_{\mathcal{H}_\gamma}^2. \end{aligned} \quad (16)$$

Let

$$K(\mathbf{x}) := \sum_{j=1}^r \binom{r}{j} (-1)^{1-j} \frac{1}{j^d} \left(\frac{2}{\gamma\sqrt{\pi}} \right)^{\frac{d}{2}} \exp\left(-\frac{2\|\mathbf{x}\|^2}{j^2\gamma^2}\right),$$

and $\tilde{f}(\mathbf{x}) := (\gamma\sqrt{\pi})^{-\frac{d}{2}} f$. Using K and \tilde{f} , we define

$$f_0 := K * \tilde{f} := \int_{\mathbb{R}^d} \tilde{f}(y) K(x-y) dy,$$

i.e., f_0 is the convolution of K and \tilde{f} . Because of Lemma 2 in Eberts and Steinwart (2011), we have $f_0 \in \mathcal{H}_\gamma$ and

$$\begin{aligned} \|f_0\|_{\mathcal{H}_\gamma} &\leq (2^r - 1) \|\tilde{f}\|_{L^2(\mathbb{R}^d)} \quad (\because \text{Lemma 2 of Eberts and Steinwart (2011)}) \\ &\leq (2^r - 1) (\gamma\sqrt{\pi})^{-\frac{d}{2}} \|f\|_{L^2(\mathbb{R}^d)} \\ &\leq (2^r - 1) (\gamma\sqrt{\pi})^{-\frac{d}{2}} (\|p\|_{L^2(\mathbb{R}^d)} + \|p'\|_{L^2(\mathbb{R}^d)}) \\ &\leq (2^r - 1) (\gamma\sqrt{\pi})^{-\frac{d}{2}} 2\sqrt{M}. \end{aligned} \tag{17}$$

Moreover, Lemma 3 in Eberts and Steinwart (2011) gives

$$\|f_0\|_\infty \leq (2^r - 1) \|f\|_\infty \leq (2^r - 1)M, \tag{18}$$

and Lemma 1 in Eberts and Steinwart (2011) yields that there exists a constant $C_{r,2}$ such that

$$\|f_0 - f\|_{L^2(\mathbb{R}^d)}^2 \leq C_{r,2} \omega_{r,L^2(\mathbb{R}^d)}^2(f, \frac{\gamma}{2}) \leq C_{r,2} c^2 \gamma^{2\alpha}. \tag{19}$$

Now, following a similar line to Theorem 3 in Eberts and Steinwart (2011), we can show that, for all $\epsilon > 0$ and $p \in (0, 1)$, there exists a constant $C_{\epsilon,p}$ such that

$$|(P_n - P)(\hat{f} - f)| \leq \hat{f} - f.$$

To bound this, we derive the tail probability of

$$(P_n - P) \left(\frac{\widehat{f} - f}{\|\widehat{f} - f\|_{L^2(\mathbb{R}^d)}^2 + \lambda \|\widehat{f}\|_{\mathcal{H}_\gamma}^2 + r} \right),$$

where $r > 0$ is a positive real such that $r > r^*$ for

$$r^* = \min_{f \in \mathcal{H}_\gamma} \|f - f\|_{L^2(\mathbb{R}^d)}^2 + \lambda \|f\|_{\mathcal{H}_\gamma}^2.$$

Let

$$g_{f,r} = \frac{f - f}{\|f - f\|_{L^2(\mathbb{R}^d)}^2 + \lambda \|f\|_{\mathcal{H}_\gamma}^2 + r}$$

for $f \in \mathcal{H}_\gamma$ and $r > r^*$. Then we have

$$\begin{aligned} \|g_{f,r}\|_\infty &\leq \frac{\|f\|_\infty + \|f\|_\infty}{\|f - f\|_{L^2(\mathbb{R}^d)}^2 + \lambda \|f\|_{\mathcal{H}_\gamma}^2 + r} \\ &\leq \frac{\|f\|_{\mathcal{H}_\gamma} + \|f\|_\infty}{\|f - f\|_{L^2(\mathbb{R}^d)}^2 + \lambda \|f\|_{\mathcal{H}_\gamma}^2 + r} \\ &\leq \frac{1}{\lambda \|f\|_{\mathcal{H}_\gamma} + r / \|f\|_{\mathcal{H}_\gamma}} + \frac{M}{r} \leq \frac{1}{2\sqrt{r\lambda}} + \frac{M}{r}, \end{aligned}$$

and

$$Pg_{f,r}^2 = \frac{P(f - f)^2}{(\|f - f\|_{L^2(\mathbb{R}^d)}^2 + \lambda \|f\|_{\mathcal{H}_\gamma}^2 + r)^2} \leq \frac{M \|f - f\|_{L^2(\mathbb{R}^d)}^2}{(\|f - f\|_{L^2(\mathbb{R}^d)}^2 + \lambda \|f\|_{\mathcal{H}_\gamma}^2 + r)^2} \leq \frac{M}{r}.$$

Here, let

$$\mathcal{F}_r := \{f \in \mathcal{H}_\gamma \mid \|f - f\|_{L^2(\mathbb{R}^d)}^2 + \lambda \|f\|_{\mathcal{H}_\gamma}^2 \leq r\},$$

and we assume that there exists a function such that

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}_r} |(P_n - P)(f - f)| \right] \leq \varphi_n(r),$$

where \mathbb{E} denotes the expectation over all samples. Then, by the peeling device (see Theorem 7.7 in Steinwart & Christmann, 2008), we have

$$\mathbb{E} \sup_{f \in \mathcal{H}_\gamma} |(P_n - P)g_{f,r}| \leq \frac{8\varphi(r)}{r}.$$

Therefore, by Talagrand's concentration inequality, we have

$$\Pr \left[\sup_{f \in \mathcal{H}_\gamma} |(P_n - P)g_{f,r}| < \frac{10\varphi_n(r)}{r} + \sqrt{\frac{2M\tau}{nr}} + \frac{14\tau}{3n} \left(\frac{1}{2\sqrt{r\lambda}} + \frac{M}{r} \right) \right] \geq 1 - e^{-\tau}, \quad (20)$$

where $\Pr[\cdot]$ denotes the probability of an event.

From now on, we give an upper bound of φ_n . The RKHS \mathcal{H}_γ can be embedded in arbitrary Sobolev space $W^m(\mathbb{R}^d)$. Indeed, by the proof of Theorem 3.1 in Steinwart and Scovel (2004), we have

$$\|f\|_{W^m(\mathbb{R}^d)} \leq C_m \gamma^{-\frac{m}{2} + \frac{d}{4}} \|f\|_{\mathcal{H}_\gamma}$$

for all $f \in \mathcal{H}_\gamma$. Moreover, the theories of interpolation spaces give that, for all $f \in W^m(\mathbb{R}^d)$, the supremum norm of f can be bounded as

$$\|f\|_\infty \leq C'_m \|f\|_{L^2(\mathbb{R}^d)}^{1 - \frac{d}{2m}} \|f\|_{W^m(\mathbb{R}^d)}^{\frac{d}{2m}},$$

if $d < 2m$. Here we set $m = \frac{d}{2p}$. Then we have

$$\|f\|_\infty \leq C''_p \|f\|_{L^2(\mathbb{R}^d)}^{1-p} \|f\|_{\mathcal{H}_\gamma}^p \gamma^{-\frac{d(1-p)}{4}}.$$

Now, since $\mathcal{F}_r \subset (r/\lambda)^{1/2} \mathcal{B}_{\mathcal{H}_\gamma}$ and

$$P(f - f)^2 \leq M \|f - f\|_{L^2(\mathbb{R}^d)}^2 \leq Mr \quad \text{for } f \in \mathcal{F}_r$$

hold from Theorem 7.16 and Theorem 7.34 in Steinwart and Christmann (2008), we can take

$$\varphi_n(r) = \max \left\{ C_{1,p,\epsilon} \gamma^{-\frac{(1-p)(1+\epsilon)d}{2}} \left(\frac{r}{\lambda}\right)^{\frac{p}{2}} (Mr)^{\frac{1-p}{2}} n^{-1/2}, \right. \\ \left. C_{2,p,\epsilon} \gamma^{-\frac{(1-p)(1+\epsilon)d}{1+p}} \left(\frac{r}{\lambda}\right)^{\frac{p}{1+p}} \left[\left(\frac{r}{\lambda}\right)^{\frac{p}{2}} \gamma^{-\frac{d(1-p)}{4}} r^{\frac{1-p}{2}} \right]^{\frac{1-p}{1+p}} n^{-1/(1+p)} \right\},$$

where $\epsilon > 0$ and $p \in (0, 1)$ are arbitrary and $C_{1,p,\epsilon}, C_{2,p,\epsilon}$ are constants depending on p, ϵ .

In the same way, we can also obtain a bound of $\sup_{f \in \mathcal{H}_\gamma} |(P'_n - P')g_{f,r}|$.

If we set r to satisfy

$$\frac{1}{8} \geq \frac{10\varphi_n(r)}{r} + \sqrt{\frac{2M\tau}{nr}} + \frac{14\tau}{3n} \left(\frac{1}{2\sqrt{r\lambda}} + \frac{M}{r} \right), \quad (21)$$

then we have

$$|(Q_n - Q)(\hat{f} - f)| \leq \frac{1}{4} \left(r + \|\hat{f} - f\|_{L^2(\mathbb{R}^d)}^2 + \lambda \|\hat{f}\|_{\mathcal{H}_\gamma} \right), \quad (22)$$

with probability $1 - 2e^{-\tau}$. To satisfy Eq.(21), it suffices to set

$$r = C \left(\frac{\gamma^{-(1-p)(1+\epsilon)d}}{\lambda^p n} + \frac{\gamma^{-\frac{2(1-p)d}{1+p}(1+\epsilon+\frac{1-p}{4})}}{\lambda^{\frac{3p-p^2}{1+p}} n^{\frac{2}{1+p}}} + \frac{\tau}{n^2 \lambda} + \frac{\tau}{n} \right), \quad (23)$$

where C is a sufficiently large constant depending on M, ϵ, p .

Finally, we bound the term $(Q_n - Q)(f_0 - f)$. By Bernstein's inequality, we have

$$\begin{aligned} |(P_n - P)(f_0 - f)| &\leq C \left(\|f - f_0\|_{L_2(P)} \sqrt{\frac{\tau}{n}} + \frac{2^r M \tau}{n} \right) \\ &\leq C \left(\sqrt{2M} \|f - f_0\|_{L^2(\mathbb{R}^d)} \sqrt{\frac{\tau}{n}} + \frac{2^r M \tau}{n} \right) \\ &\leq C \left(\|f - f_0\|_{L^2(\mathbb{R}^d)}^2 + \frac{2M\tau}{n} + \frac{2^r M \tau}{n} \right), \end{aligned} \quad (24)$$

with probability $1 - e^{-\tau}$, where C is a universal constant. In a similar way, we can also obtain

$$|(P'_n - P')(f_0 - f)| \leq C \left(\|f - f_0\|_{L^2(\mathbb{R}^d)}^2 + \frac{2M\tau}{n} + \frac{2^r M\tau}{n} \right).$$

Combining these inequalities, we have

$$|(Q_n - Q)(f_0 - f)| \leq C \left(\|f - f_0\|_{L^2(\mathbb{R}^d)}^2 + \frac{2^r M\tau}{n} \right), \quad (25)$$

with probability $1 - 2e^{-\tau}$, where C is a universal constant.

Substituting Eqs.(22) and (25) into Eq.(16), we have

$$\begin{aligned} & \|\widehat{f} - f\|_{L^2(\mathbb{R}^d)}^2 + \lambda \|\widehat{f}\|_{\mathcal{H}_\gamma}^2 \\ & \leq 2 \left\{ \|f_0 - f\|_{L^2(\mathbb{R}^d)}^2 + C \left(\|f - f_0\|_{L^2(\mathbb{R}^d)}^2 + \frac{2^r M\tau}{n} \right) + r + \lambda \|f_0\|_{\mathcal{H}_\gamma} \right\}, \end{aligned}$$

with probability $1 - 4e^{-\tau}$. Moreover, by Eqs.(19) and (17), the right-hand side is further bounded by

$$\|\widehat{f} - f\|_{L^2(\mathbb{R}^d)}^2 + \lambda \|\widehat{f}\|_{\mathcal{H}_\gamma}^2 \leq C \left\{ \gamma^{2\alpha} + r + \lambda\gamma^{-d} + \frac{1 + \tau}{n} \right\},$$

Finally, substituting (23) into the right-hand side, we have

$$\begin{aligned} & \|\widehat{f} - f\|_{L^2(\mathbb{R}^d)}^2 + \lambda \|\widehat{f}\|_{\mathcal{H}_\gamma}^2 \\ & \leq C \left\{ \gamma^{2\alpha} + \frac{\gamma^{-(1-p)(1+\epsilon)d}}{\lambda^p n} + \frac{\gamma^{-\frac{2(1-p)d}{1+p}(1+\epsilon+\frac{1-p}{4})}}{\lambda^{\frac{3p-p^2}{1+p}} n^{\frac{2}{1+p}}} + \lambda\gamma^{-d} + \frac{\tau}{\lambda n^2} + \frac{\tau}{n} \right\}, \end{aligned}$$

with probability $1 - 4e^{-\tau}$ for $\tau \geq 1$. This gives the assertion. \square

If we set

$$\lambda = n^{-\frac{2\alpha+d}{(2\alpha+d)(1+p)+(\epsilon-p+\epsilon p)}}, \quad \gamma = n^{-\frac{1}{(2\alpha+d)(1+p)+(\epsilon-p+\epsilon p)}},$$

and take ϵ, p sufficiently small, then we immediately have the following corollary.

Corollary 1. *Suppose Assumption 1 is satisfied. Then, for all $\rho, \rho' > 0$, there exists a constant $K > 0$ depending on M, c, ρ, ρ' such that for all $n \geq 1, \tau \geq 1$, the density-difference estimator \widehat{f} with appropriate choice of γ and λ satisfies*

$$\|\widehat{f} - f\|_{L^2(\mathbb{R}^d)}^2 + \lambda \|\widehat{f}\|_{\mathcal{H}_\gamma}^2 \leq K \left(n^{-\frac{2\alpha}{2\alpha+d} + \rho} + \frac{\tau}{n^{1-\rho'}} \right), \quad (26)$$

with probability not less than $1 - 4e^{-\tau}$.

Note that $n^{-\frac{2\alpha}{2\alpha+d}}$ is the optimal learning rate to estimate a function in $B_{2,\infty}^\alpha$ (Eberts & Steinwart, 2011). Therefore, the density-difference estimator with a Gaussian kernel achieves the optimal learning rate by appropriately choosing the regularization parameter and the Gaussian width. Because the learning rate depends on α , the LSDD estimator has an adaptivity to the smoothness of the true function.

Our analysis heavily relies on the techniques developed in Eberts and Steinwart (2011) for a regression problem. The main difference is that the analysis in their paper involves a clipping procedure, which stems from the fact that the analyzed estimator requires an empirical approximation of the expectation of the square term. The Lipschitz continuity of the square function $f \mapsto f^2$ is utilized to investigate this term, and the clipping procedure is used to ensure the Lipschitz continuity. On the other hand, in the current paper, we can exactly compute $\|f\|_{L^2(\mathbb{R}^d)}^2$ so that we do not need the Lipschitz continuity.

B Derivation of Eq.(13)

When $\lambda (\geq 0)$ is small, $(\mathbf{H} + \lambda \mathbf{I}_b)^{-1}$ can be expanded as

$$(\mathbf{H} + \lambda \mathbf{I}_b)^{-1} = \mathbf{H}^{-1} - \lambda \mathbf{H}^{-2} + o_p(\lambda),$$

where o_p denotes the probabilistic order. Then Eq.(12) can be expressed as

$$\begin{aligned}
& \beta \widehat{\mathbf{h}}^\top \widehat{\boldsymbol{\theta}} + (1 - \beta) \widehat{\boldsymbol{\theta}}^\top \mathbf{H} \widehat{\boldsymbol{\theta}} \\
&= \beta \widehat{\mathbf{h}}^\top (\mathbf{H} + \lambda \mathbf{I}_b)^{-1} \widehat{\mathbf{h}} + (1 - \beta) \widehat{\mathbf{h}}^\top (\mathbf{H} + \lambda \mathbf{I}_b)^{-1} \mathbf{H} (\mathbf{H} + \lambda \mathbf{I}_b)^{-1} \widehat{\mathbf{h}} \\
&= \beta \widehat{\mathbf{h}}^\top \mathbf{H}^{-1} \widehat{\mathbf{h}} - \lambda \beta \widehat{\mathbf{h}}^\top \mathbf{H}^{-2} \widehat{\mathbf{h}} \\
&\quad + (1 - \beta) \widehat{\mathbf{h}}^\top \mathbf{H}^{-1} \widehat{\mathbf{h}} - 2\lambda(1 - \beta) \widehat{\mathbf{h}}^\top \mathbf{H}^{-2} \widehat{\mathbf{h}} + o_p(\lambda) \\
&= \widehat{\mathbf{h}}^\top \mathbf{H}^{-1} \widehat{\mathbf{h}} - \lambda(2 - \beta) \widehat{\mathbf{h}}^\top \mathbf{H}^{-2} \widehat{\mathbf{h}} + o_p(\lambda),
\end{aligned}$$

which concludes the proof.

C Derivation of Eq.(14)

Because $\mathbb{E}[\widehat{\mathbf{h}}] = \mathbf{h}$, we have

$$\begin{aligned}
\mathbb{E}[\widehat{\mathbf{h}}^\top \mathbf{H}^{-1} \widehat{\mathbf{h}} - \mathbf{h}^\top \mathbf{H}^{-1} \mathbf{h}] &= \mathbb{E}[(\widehat{\mathbf{h}} - \mathbf{h})^\top \mathbf{H}^{-1} (\widehat{\mathbf{h}} - \mathbf{h})] \\
&= \text{tr} \left(\mathbf{H}^{-1} \mathbb{E}[(\widehat{\mathbf{h}} - \mathbf{h})(\widehat{\mathbf{h}} - \mathbf{h})^\top] \right) \\
&= \text{tr} \left(\mathbf{H}^{-1} \left(\frac{1}{n} \mathbf{V}_p[\boldsymbol{\psi}] + \frac{1}{n'} \mathbf{V}_{p'}[\boldsymbol{\psi}] \right) \right),
\end{aligned}$$

which concludes the proof.