Statistical Analysis of Kernel-Based Least-Squares Density-Ratio Estimation

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

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

Masashi Sugiyama
Tokyo Institute of Technology, Tokyo, Japan
sugi@cs.titech.ac.jp
http://sugiyama-www.cs.titech.ac.jp/~sugi/

Abstract

The ratio of two probability densities can be used for solving various machine learning tasks such as covariate shift adaptation (importance sampling), outlier detection (likelihood-ratio test), feature selection (mutual information), and conditional probability estimation. Several methods of directly estimating the density ratio have recently been developed, e.g., moment matching estimation, maximum-likelihood density-ratio estimation, and least-squares density-ratio fitting. In this paper, we propose a kernelized variant of the least-squares method for density-ratio estimation, which is called kernel unconstrained least-squares importance fitting (KuLSIF). We investigate its fundamental statistical properties including a non-parametric convergence rate, an analytic-form solution, and a leave-one-out cross-validation score. We further study its relation to other kernel-based density-ratio estimators. In experiments, we numerically compare various kernel-based density-ratio estimation methods, and show that KuLSIF compares favorably with other approaches.

1 Introduction

The problem of estimating the ratio of two probability densities is attracting a great deal of attention these days, since the density ratio can be used for various purposes (Sugiyama et al., 2009; Sugiyama et al., 2012), such as covariate shift adaptation (Shimodaira, 2000; Zadrozny, 2004; Sugiyama & Müller, 2005; Huang et al., 2007; Sugiyama et al., 2007; Bickel et al., 2009; Quiñonero-Candela et al., 2009; Sugiyama & Kawanabe, 2011), outlier detection (Hido et al., 2008; Smola et al., 2009; Kawahara & Sugiyama, 2011; Hido et al., 2011), divergence estimation (Nguyen et al., 2010; Suzuki et al., 2008; Suzuki et al., 2009), and conditional probability estimation (Sugiyama et al., 2010; Sugiyama, 2010).

A naive approach to density-ratio estimation is to first separately estimate two probability densities (corresponding to the numerator and the denominator of the ratio), and then take the ratio of the estimated densities. However, density estimation is known to be a hard problem particularly in high-dimensional cases unless we have simple and good parametric density models (Vapnik, 1998; Härdle et al., 2004; Kanamori et al., 2010), which may not be the case in practice.

For reliable statistical inference, it is important to develop methods of directly estimating the density ratio without going through density estimation. In the context of case-control studies, Qin (1998) has proposed a direct method of estimating the density ratio by matching moments of the two distributions. Another density-ratio estimation approach uses the M-estimator (Nguyen et al., 2010) based on non-asymptotic variational characterization of the f-divergence (Ali & Silvey, 1966; Csiszár, 1967). See also Sugiyama et al. (2008a) for a similar algorithm using the Kullback-Leibler divergence. Kanamori et al. (2009) have developed a squared-loss version of the M-estimator for linear density-ratio models called *unconstrained Least-Squares Importance Fitting* (uLSIF), and have shown that uLSIF possesses superior computational properties. That is, a closed-form solution is available and the leave-one-out cross-validation score can be analytically computed. As another approach, one can use logistic regression for the inference of density ratios, since the ratio of two probability densities is directly connected to the posterior probability of labels in classification problems. Using the Bayes formula, the estimated posterior probability can be transformed to an estimator of density ratios (Bickel et al., 2007).

Various kernel-based approaches are also available for density-ratio estimation. The *kernel mean matching* (KMM) method (Gretton et al., 2009) directly gives estimates of the density ratio by matching the two distributions using universal reproducing kernel Hilbert spaces (Steinwart, 2001). KMM can be regarded as a kernelized variant of Qin's moment matching estimator (Qin, 1998). Nguyen's approach based on the M-estimator (Nguyen et al., 2010) also has a kernelized variant. Non-parametric convergence properties of the M-estimator in reproducing kernel Hilbert spaces have been elucidated under the Kullback-Leibler divergence (Nguyen et al., 2010; Sugiyama et al., 2008b). For the density-ratio estimation, one can also apply kernel logistic regression (Wahba et al., 1993; Zhu & Hastie, 2001), instead of conventional linear logistic models for the inference of the posterior distribution in classification problems.

In this paper, we first propose a kernelized variant of uLSIF (called KuLSIF), and show that the solution of KuLSIF as well as its leave-out-out cross-validation score can be computed analytically, as the original uLSIF for linear models. We then elucidate the statistical consistency and convergence rate of KuLSIF based on the argument on non-parametric bounds (van de Geer, 2000; Nguyen et al., 2010). We further study the relation between KuLSIF and other kernel-based density-ratio estimators. Finally, statistical performance of KuLSIF is numerically compared with other kernel-based density-ratio estimators in experiments.

The rest of this paper is organized as follows. In Section 2, we formulate the problem of density-ratio estimation and briefly review the existing least-squares method. In Section 3, we describe the kernelized variant of uLSIF, and show its statistical properties such as the convergence rate and availability of the analytic-form solution and the analytic-form leave-one-out cross-validation score. In Section 4, we investigate the relation between KuLSIF and other kernel-based density-ratio estimators. In Section 5, we experimentally investigate computational efficiency and statistical performance of KuLSIF. Finally, in Section 6, we conclude by summarizing our contributions and showing possible future directions. Detailed proofs and calculations are deferred to Appendix. In a companion paper (Kanamori et al., 2011), computational properties of the KuLSIF method are further investigated from the viewpoint of condition numbers.

2 Estimation of Density Ratios

In this section, we formulate the problem of density-ratio estimation and briefly review the least-squares density-ratio estimator.

2.1 Formulation and Notations

Consider two probability distributions P and Q on a probability space $\mathcal{Z}$. Assume that the distributions P and Q have the probability densities p and q, respectively. We assume $p(x) > 0$ for all $x \in \mathcal{Z}$. Suppose that we are given two sets of independent and identically distributed (i.i.d.) samples,

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} P, \qquad Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} Q. \tag{1}$$

Our goal is to estimate the density ratio

$$w_0(x) = \frac{q(x)}{p(x)} \ (\geq 0)$$

based on the observed samples.

We summarize some notations to be used throughout the paper. For two integers n and m, $n \wedge m$ denotes $\min\{m, n\}$. For a vector a in the Euclidean space, $\|a\|$ denotes the Euclidean norm. Given a probability distribution P and a random variable $h(X)$, we

denote the expectation of $h(X)$ under P by $\int h dP$ or $\int h(x)P(dx)$. Let $\|\cdot\|_\infty$ be the infinity norm, and $\|\cdot\|_P$ be the L_2-norm under the probability P, i.e., $\|h\|_P^2 = \int |h|^2 dP$. For a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ (Aronszajn, 1950), the inner product and the norm on $\mathcal{H}$ are denoted as $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{H}}$, respectively.

Below we review the least-squares approach to density-ratio estimation proposed by Kanamori et al. (2009).

2.2 Least-Squares Approach

The linear model

$$\widehat{w}(x) \;=\; \sum_{i=1}^{B} \alpha_i h_i(x) \tag{2}$$

is assumed for the estimation of the density ratio w_0, where the coefficients $\alpha_1, \ldots, \alpha_B$ are the parameters of the model. The basis functions $h_i,\ i = 1, \ldots, B$ are chosen so that the non-negativity condition $h_i(x) \geq 0$ is satisfied. A practical choice would be the Gaussian kernel function $h_i(x) = e^{-\|x-c_i\|^2/2\sigma^2}$ with appropriate kernel center $c_i \in \mathcal{Z}$ and kernel width σ (Sugiyama et al., 2008a).

The *unconstrained least-squares importance fitting* (uLSIF) (Kanamori et al., 2009) estimates the parameter α based on the squared error:

$$\frac{1}{2}\int (\widehat{w} - w_0)^2 dP \;=\; \frac{1}{2}\int \widehat{w}^2 dP - \int \widehat{w} dQ + \frac{1}{2}\int w_0^2 dP.$$

The last term in the above expression is a constant and can be safely ignored when minimizing the squared error of the estimator $\widehat{w}$. Therefore, the solution of the following minimization problem over the linear model (2),

$$\min_{w} \;\frac{1}{2n}\sum_{i=1}^{n}(w(X_i))^2 - \frac{1}{m}\sum_{j=1}^{m} w(Y_j) + \lambda \cdot \mathrm{Reg}(\alpha), \tag{3}$$

is expected to approximate the true density-ratio w_0, where the regularization term $\mathrm{Reg}(\alpha)$ with the regularization parameter λ is introduced to avoid overfitting. Let $\widehat{\alpha}$ be the optimal solution of (3) under the linear model (2). Then the estimator of w_0 is given as $\widehat{w}(x) = \sum_{i=1}^{B} \widehat{\alpha}_i h_i(x)$. There are several ways to impose the non-negativity condition $\widehat{w}(x) \geq 0$ (Kanamori et al., 2009). Here, truncation of $\widehat{w}$ defined as

$$\widehat{w}_+(x) = \max\{\widehat{w}(x),\ 0\}$$

is used to ensure the non-negativity of the estimator.

It is worthwhile to point out that uLSIF can be regarded as an example of the M-estimator (Nguyen et al., 2010) with the quadratic loss function, i.e., $\phi^*(f) = f^2/2$ in Nguyen's notation. Nguyen's M-estimator is constructed based on the f-divergence from Q and P. Due to the asymmetry of f-divergence, the estimation error evaluated with

respect to P rather than Q is obtained in uLSIF. uLSIF has an advantage in computation over other M-estimators: When $\mathrm{Reg}(\alpha) = \|\alpha\|^2/2$, the estimator $\widehat{\alpha}$ can be obtained in an analytic form. As a result, the leave-one-out cross-validation (LOOCV) score can also be computed in a closed form (Kanamori et al., 2009), which allows us to compute the LOOCV score very efficiently. LOOCV is an (almost) unbiased estimator of the prediction error and can be used for determining hyper-parameters such as the regularization parameter and the Gaussian kernel width. In addition, for the L_1-regularization $\mathrm{Reg}(\alpha) = \sum_{i=1}^{B} |\alpha_i|$, Kanamori et al. (2009) applied the path-following algorithm to regularization parameter estimation, which highly contributes to reducing the computational cost in the model selection phase.

3 Kernel uLSIF

The purpose of this paper is to show that a kernelized variant of uLSIF (which we refer to as *kernel uLSIF*; KuLSIF) has good theoretical properties and thus useful. In this section, we formalize the KuLSIF algorithm and show its fundamental statistical properties.

3.1 uLSIF on RKHS

We assume that the model for the density ratio is an RKHS $\mathcal{H}$ endowed with a kernel function k on $\mathcal{Z} \times \mathcal{Z}$, and we consider the optimization problem (3) on $\mathcal{H}$. Then, the estimator $\widehat{w}$ is obtained as an optimal solution of

$$\min_{w} \ \frac{1}{2n}\sum_{i=1}^{n}(w(X_i))^2 - \frac{1}{m}\sum_{j=1}^{m} w(Y_j) + \frac{\lambda}{2}\|w\|_{\mathcal{H}}^2, \quad \text{s.t.} \quad w \in \mathcal{H}, \tag{4}$$

where the regularization term $\frac{\lambda}{2}\|w\|_{\mathcal{H}}^2$ with the regularization parameter λ (≥ 0) is introduced to avoid overfitting. We may also consider the truncated estimator $\widehat{w}_+ = \max\{\widehat{w}, 0\}$. The estimator based on the loss function (4) is called KuLSIF.

The computation of KuLSIF is efficiently conducted. For infinite-dimensional $\mathcal{H}$, the problem (4) is an infinite-dimensional optimization problem. The representer theorem (Kimeldorf & Wahba, 1971), however, is applicable to RKHSs, which allows us to transform the infinite-dimensional optimization problem to a finite-dimensional one. Let K_{11}, K_{12}, and K_{21} be the sub-matrices of the Gram matrix:

$$(K_{11})_{ii'} \ = \ k(X_i, X_{i'}), \quad (K_{12})_{ij} \ = \ k(X_i, Y_j), \quad K_{21} \ = \ K_{12}^{\top},$$

where $i, i' = 1, \ldots, n,\ j, j' = 1, \ldots, m$. Then, detailed analysis leads us to the specific form of the solution as follows.

Theorem 1 (Analytic Solution of KuLSIF)**.** *Suppose $\lambda > 0$. Then, the KuLSIF estimator is given as*

$$\widehat{w}(z) \ = \ \sum_{i=1}^{n} \bar{\alpha}_i k(z, X_i) + \frac{1}{m\lambda}\sum_{j=1}^{m} k(z, Y_j).$$

The coefficients $\bar{\alpha} = (\bar{\alpha}_1, \ldots, \bar{\alpha}_n)^\top$ *are given by the solution of the linear equation*

$$\left(\frac{1}{n}K_{11} + \lambda I_n\right)\alpha = -\frac{1}{nm\lambda}K_{12}\mathbf{1}_m, \tag{5}$$

where I_n *is the* n *by* n *identity matrix and* $\mathbf{1}_m$ *is the column vector defined as* $\mathbf{1}_m = (1, \ldots, 1)^\top \in \Re^m$.

The proof is deferred to Appendix A. Theorem 1 implies that it is sufficient to find n variables $\bar{\alpha}_1, \ldots, \bar{\alpha}_n$ to obtain the estimator $\widehat{w}$ and that the estimator has the analytic-form solution.

Theorem 1 also guarantees that the parameters in the KuLSIF estimator are obtained by the solution of the following optimization problem:

$$\min_{\alpha} \ \frac{1}{2}\alpha^\top\left(\frac{1}{n}K_{11} + \lambda I_n\right)\alpha + \frac{1}{nm\lambda}\mathbf{1}_m^\top K_{21}\alpha, \quad \alpha \in \Re^n, \tag{6}$$

where we used the fact that the solution of $Ax = b$ is given as the minimizer of $\frac{1}{2}x^\top Ax - b^\top x$, when A is positive-semidefinite. When the sample size of n is large, numerically optimizing the quadratic function in (6) can be computationally more efficient than directly solving the linear equation (5). In Section 5, numerical experiments are carried out to investigate the computational efficiency of KuLSIF.

3.2 Leave-One-Out Cross-Validation

The leave-one-out cross-validation (LOOCV) score for the KuLSIF estimator can also be obtained analytically as well as the coefficient parameters of the kernel model. Let us measure the accuracy of the KuLSIF estimator, $\widehat{w}_+ = \max\{\widehat{w}, 0\}$, by

$$\frac{1}{2}\int \widehat{w}_+^2 dP - \int \widehat{w}_+ dQ,$$

which is equal to the squared error of $\widehat{w}_+$ up to a constant term. Then the LOOCV score of $\widehat{w}_+$ under the squared error is defined as

$$\text{LOOCV} \ = \ \frac{1}{n \wedge m}\sum_{\ell=1}^{n\wedge m}\left\{\frac{1}{2}(\widehat{w}_+^{(\ell)}(x_\ell))^2 - \widehat{w}_+^{(\ell)}(y_\ell)\right\}, \tag{7}$$

where $\widehat{w}_+^{(\ell)} = \max\{\widehat{w}^{(\ell)}, 0\}$ is the estimator based on training samples except[1] x_ℓ and y_ℓ. The hyper-parameters achieving the minimum value of LOOCV are chosen.

Thanks to the analytic-form solution shown in Theorem 1, the leave-one-out solution $\widehat{w}^{(\ell)}$ can be computed efficiently from $\widehat{w}$ by the use of the Sherman-Woodbury-Morrison formula (Golub & Loan, 1996). Details of the analytic LOOCV expression are presented in Appendix B.

[1]The index of removed samples can be different for x and y, i.e., x_{ℓ_1} and y_{ℓ_2} $(\ell_1 \neq \ell_2)$ can be removed. For the sake of simplicity, however, we suppose that the samples x_ℓ and y_ℓ are removed in the computation of LOOCV.

3.3 Statistical Consistency of KuLSIF

The following theorem reveals the convergence rate of the KuLSIF estimator.

Theorem 2 (Convergence Rate of KuLSIF)**.** *Let $\mathcal{Z}$ be a probability space, and $\mathcal{H}$ be the RKHS endowed with the kernel function k defined on $\mathcal{Z}\times\mathcal{Z}$. Suppose that $\sup_{x\in\mathcal{Z}} k(x,x) < \infty$, and that the bracketing entropy $H_B(\delta, \mathcal{H}_M, P)$ is bounded above by $O(M/\delta)^\gamma$, where γ is a constant satisfying $0 < \gamma < 2$ (see Appendix C for the detailed definition). Set the regularization parameter $\lambda = \lambda_{n,m}$ so that*[2]

$$\lim_{n,m\to\infty} \lambda_{n,m} = 0, \qquad \lambda_{n,m}^{-1} = O((n \wedge m)^{1-\delta}), \quad (n, m \to \infty),$$

where $n \wedge m = \min\{n, m\}$ and δ is an arbitrary number satisfying $1 - 2/(2+\gamma) < \delta < 1$. Then, for $q/p = w_0 \in \mathcal{H}$, we have

$$\|\widehat{w}_+ - w_0\|_P \;\le\; \|\widehat{w} - w_0\|_P \;=\; O_p(\lambda_{n,m}^{1/2}),$$

where $\|\cdot\|_P$ is the L_2-norm under the probability P.

The proof is available in Appendix C. See Nguyen et al. (2010) and Sugiyama et al. (2008b) for similar convergence analysis for the logarithmic loss function. The condition $\lim_{n,m\to\infty} \lambda_{n,m} = 0$ means that the regularization parameter $\lambda_{n,m}$ should vanish asymptotically, but the condition $\lambda_{n,m}^{-1} = O((n\wedge m)^{1-\delta})$ means that the regularization parameter $\lambda_{n,m}$ should not vanish too fast. As shown in the proof of the theorem, the assumption $w_0 \in \mathcal{H}$ imposes $\sup_{x\in\mathcal{Z}} w_0(x) \le \|w_0\|_{\mathcal{H}} \sup_{x\in\mathcal{Z}} \sqrt{k(x,x)} < \infty$. For example, the ratio of two Gaussian distributions with different means or variances does not satisfy this condition (see Yamada et al., 2011 for how to handle such a situation).

Remark 1. *Suppose that $\mathcal{Z}$ is a compact set and k is the Gaussian kernel. Then, for any small $\gamma > 0$, the condition*

$$H_B(\delta, \mathcal{H}_M, P) \;=\; O\left(\frac{M}{\delta}\right)^\gamma, \quad (M/\delta \to \infty) \tag{8}$$

holds (Cucker & Smale, 2002, Theorem D in Chap III, Section 5). More precisely, Cucker and Smale (2002) proved that the entropy number with the supremum norm is bounded above by $c(M/\delta)^\gamma$ for M, $\delta > 0$, where c is a positive constant. In addition, the bracketing entropy $H_B(\delta, \mathcal{H}_M, P)$ is bounded above by the entropy number with the supremum norm due to the second inequality of Lemma 2.1 in van de Geer (2000). As a result, the convergence rate in Theorem 2 is given as $O_p(\lambda_{n,m}^{1/2}) = O_p(1/(n\wedge m)^{(1-\delta)/2})$ for $0 < \delta < 1$. By choosing small $\delta > 0$ (i.e., $\lambda_{n,m}$ vanishes fast), the convergence rate will get close to that for parametric models, i.e., $O_p(1/\sqrt{n\wedge m})$.

[2]The multivariate big-O notation $f(n,m) = O(g(n,m)), (n,m\to\infty)$ implies that there exist $C > 0$, $n_0 > 0$, and $m_0 > 0$ such that the inequality $|f(n,m)| \le C|g(n,m)|$ holds for all $n > n_0$ and all $m > m_0$.

Table 1: Summary of density ratio estimators. "#parameters" is the number of parameters other than regularization parameters.

Estimator	loss function	#parameters	estimates of density ratio	model selection
KuLSIF	quadratic loss	n	$\max\{\widehat{w}, 0\}, \quad \widehat{w} \in \mathcal{H}$	possible
KMM	quadratic loss	n	$\max\{\widehat{w}, 0\}, \quad \widehat{w} \in \mathcal{H}$	not possible
KL-div	conjugate of log loss	m	$\max\{\widehat{w}, 0\}, \quad \widehat{w} \in \mathcal{H}$	possible
KLR	log-likelihood loss	$n+m$	$\frac{n}{m}e^{-\widehat{f}}, \quad \widehat{f} \in \mathcal{H}$	possible
RKDE	log-likelihood loss	0	ratio of KDEs	possible

Remark 2. *In the KuLSIF estimator, we do not need the assumption that the target density-ratio w_0 is bounded below by a positive constant. In the M-estimator with the Kullback-Leibler divergence proposed by Nguyen et al. (2010), the function defined by*

$$f(w) = \log \frac{w + w_0}{w_0}$$

is considered. The boundedness condition of w_0 such that $w_0 \geq c > 0$ with some constant c leads to the fact that $f(w)$ is Lipschitz in w, which is a crucial property in the proof of Nguyen et al. (2010). In our proof in Appendix C, we use different transformation of w, and thus, the boundedness condition is not needed.

4 Relation to Existing Kernel-Based Estimators

In this section, we discuss the relation between KuLSIF and other kernel-based density-ratio estimators. Properties of density-ratio estimation methods are summarized in Table 1.

4.1 Kernel Mean Matching (KMM)

The *kernel mean matching* (KMM) method allows us to directly obtain an estimate of $w_0(x)$ at $X_1, \ldots, X_n$ without going through density estimation (Gretton et al., 2009).

The basic idea of KMM is to find $w_0(x)$ such that the mean discrepancy between non-linearly transformed samples drawn from P and Q is minimized in a *universal reproducing kernel Hilbert space* (Steinwart, 2001). We introduce the definition of universal kernels below.

Definition 1 (Definition 4.52 in Steinwart, 2001)**.** *A continuous kernel k on a compact metric space $\mathcal{Z}$ is called universal if the RKHS $\mathcal{H}$ of k is dense in the set of all continuous functions on $\mathcal{Z}$, that is, for every continuous function g on $\mathcal{Z}$ and all $\varepsilon > 0$, there exists an $f \in \mathcal{H}$ such that $\|f - g\|_\infty < \varepsilon$. The corresponding RKHS is called a universal RKHS.*

The Gaussian kernel on a compact set $\mathcal{Z}$ is an example of universal kernels. Let $\mathcal{H}$ be a universal RKHS endowed with a universal kernel function $k : \mathcal{Z} \times \mathcal{Z} \to \Re$. Then, one can infer the density ratio w_0 by solving the following minimization problem:

$$\begin{aligned} \min_{w} \quad & \frac{1}{2} \left\| \int w(x)k(\cdot, x)P(dx) - \int k(\cdot, y)Q(dy) \right\|_{\mathcal{H}}^2, \\ \text{s.t.} \quad & \int w dP = 1 \quad \text{and} \quad w \geq 0. \end{aligned} \tag{9}$$

Huang et al. (2007) proved that the solution of (9) is given as $w = w_0$, when Q is absolutely continuous with respect to P.

An empirical version of the above problem is reduced to the following convex quadratic program:

$$\begin{aligned} \min_{w_1,\ldots,w_n} \quad & \frac{1}{2n} \sum_{i,j=1}^{n} w_i w_j k(X_i, X_j) - \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{n} w_i k(X_i, Y_j), \\ \text{s.t.} \quad & \left| \frac{1}{n} \sum_{i=1}^{n} w_i - 1 \right| \leq \epsilon \quad \text{and} \quad 0 \leq w_1, w_2, \ldots, w_n \leq B. \end{aligned} \tag{10}$$

The tuning parameters, $B \geq 0$ and $\epsilon \geq 0$, control the regularization effects. The optimal solution $(\widehat{w}_1, \ldots, \widehat{w}_n)$ is an estimate of the density ratio at the samples from P, i.e., $w_0(X_1), \ldots, w_0(X_n)$. KMM does not estimate the function w_0 on $\mathcal{Z}$ but the values on sample points, while the assumption that $w_0 \in \mathcal{H}$ is not required.

We study the relation between KuLSIF and KMM. Below, we assume that the true density-ratio $w_0 = q/p$ is included in the RKHS $\mathcal{H}$. Let $\Phi(w)$ be

$$\Phi(w) = \int k(\cdot, x)w(x)P(dx) - \int k(\cdot, y)Q(dy). \tag{11}$$

Then the loss function of KMM on $\mathcal{H}$ under the population distribution is written as

$$L_{\mathrm{KMM}}(w) = \frac{1}{2}\|\Phi(w)\|_{\mathcal{H}}^2.$$

In the estimation phase, an empirical approximation of L_{KMM} is optimized in the KMM algorithm. On the other hand, the (unregularized) loss function of KuLSIF is given by

$$L_{\mathrm{KuLSIF}}(w) \;=\; \frac{1}{2} \int w^2 dP - \int w dQ.$$

Both L_{KMM} and L_{KuLSIF} are minimized at the true density-ratio $w_0 \in \mathcal{H}$. Although some linear constraints may be introduced in the optimization phase, we study the optimization problems of L_{KMM} and L_{KuLSIF} without constraints. This is because when the sample size tends to infinity, the optimal solutions of L_{KMM} and L_{KuLSIF} without constraints automatically satisfy the required constraints such as $\int w dP = 1$ and $w \geq 0$.

We consider the extremal condition of $L_{\mathrm{KuLSIF}}(w)$ at w_0. Substituting

$$w = w_0 + \delta \cdot v, \quad (\delta \in \Re,\, v \in \mathcal{H})$$

into $L_{\mathrm{KuLSIF}}(w)$, we have

$$L_{\mathrm{KuLSIF}}(w_0 + \delta v) - L_{\mathrm{KuLSIF}}(w_0) \;=\; \delta \left\{ \int w_0 v dP - \int v dQ \right\} + \frac{\delta^2}{2} \int v^2 dP.$$

Since $L_{\mathrm{KuLSIF}}(w_0 + \delta v)$ is minimized at $\delta = 0$, the derivative of $L_{\mathrm{KuLSIF}}(w_0 + \delta v)$ at $\delta = 0$ vanishes, i.e.,

$$\int w_0 v dP - \int v dQ = 0. \tag{12}$$

The equality (12) holds for arbitrary $v \in \mathcal{H}$. Using the reproducing property of the kernel function k, we can derive another expression of (12) as follows

$$\begin{aligned} \int w_0 v dP - \int v dQ \;&=\; \int w_0(x) \langle k(\cdot, x), v \rangle_{\mathcal{H}} P(dx) - \int \langle k(\cdot, y), v \rangle_{\mathcal{H}} Q(dy) \\ &=\; \left\langle \int k(\cdot, x) w_0(x) P(dx) - \int k(\cdot, y) Q(dy),\; v \right\rangle_{\mathcal{H}} \\ &=\; \big\langle \Phi(w_0),\; v \big\rangle_{\mathcal{H}} = 0, \quad \forall v \in \mathcal{H}. \end{aligned} \tag{13}$$

Rigorous proof of the above formula is shown in Appendix D. As a result, we obtain $\Phi(w_0) = 0$. The above expression implies that $\Phi(w)$ is the Gâteaux derivative (Zeidler, 1986, Section 4.2) of L_{KuLSIF} at $w \in \mathcal{H}$, that is,

$$\frac{d}{d\delta} L_{\mathrm{KuLSIF}}(w + \delta \cdot v)\Big|_{\delta=0} = \langle \Phi(w),\, v \rangle_{\mathcal{H}} \tag{14}$$

holds for all $v \in \mathcal{H}$. Let DL_{KuLSIF} be the Gâteaux derivative of L_{KuLSIF} over the RKHS $\mathcal{H}$. Then we have $DL_{\mathrm{KuLSIF}} = \Phi$, and the equality

$$L_{\mathrm{KMM}}(w) \;=\; \frac{1}{2} \| DL_{\mathrm{KuLSIF}}(w) \|_{\mathcal{H}}^2 \tag{15}$$

holds. Tsuboi et al. (2008) have pointed out a similar relation for the M-estimator based on the Kullback-Leibler divergence.

Now we give an interpretation of (15) through an analogous optimization example in the Euclidean space. Let $f : \Re^d \to \Re$ be a differentiable function, and consider the optimization problem $\min_x f(x)$. At an optimal solution x_0, the extremal condition $\nabla f(x_0) = 0$ should hold, where ∇f is the gradient of f with respect to x. Thus, instead of minimizing f, minimization of $\|\nabla f(x)\|^2$ also provides the minimizer of f. This corresponds to the relation between KuLSIF and KMM:

$$\begin{aligned} \text{KuLSIF} \;&\Longleftrightarrow\; \min_x f(x), \\ \text{KMM} \;&\Longleftrightarrow\; \min_x \frac{1}{2} \|\nabla f(x)\|^2. \end{aligned}$$

In other words, in order to find the solution of the equation

$$\Phi(w) = 0, \tag{16}$$

KMM tries to minimize the norm of $\Phi(w)$. The "dual" expression of (16) is given as

$$\langle \Phi(w), v \rangle_{\mathcal{H}} = 0, \quad \forall v \in \mathcal{H}. \tag{17}$$

By "integrating" $\langle \Phi(w), v \rangle_{\mathcal{H}}$, we obtain the loss function L_{KuLSIF}.

Remark 3. *Gretton et al. (2006) proposed the maximum mean discrepancy (MMD) criterion to measure the discrepancy between two probability distributions P and Q. When the constant function 1 is included in the RKHS $\mathcal{H}$, MMD between P and Q is equal to $2 \times L_{\mathrm{KMM}}(1)$. Due to the equality (15), we find that MMD is also expressed as $\|DL_{\mathrm{KuLSIF}}(1)\|_{\mathcal{H}}^2$, that is, the squared norm of the derivative of L_{KuLSIF} at $1 \in \mathcal{H}$. This quantity will be related to the discrepancy between the constant function 1 and the true density-ratio $w_0 = q/p$.*

In the original KMM method, the density-ratio values on training samples $X_1, \ldots, X_n$ are estimated (Gretton et al., 2009). Here, we consider its inductive variant, i.e., estimating the function w_0 on $\mathcal{Z}$ using the loss function of KMM. Given samples (1), the empirical loss function of inductive KMM is defined as

$$\min_{w} \ \frac{1}{2}\big\|\widehat{\Phi}(w) + \lambda w\big\|_{\mathcal{H}}^2, \quad w \in \mathcal{H}, \tag{18}$$

where $\widehat{\Phi}(w)$ is defined as

$$\widehat{\Phi}(w) = \frac{1}{n}\sum_{i=1}^{n} k(\cdot\,, X_i) w(X_i) - \frac{1}{m}\sum_{j=1}^{m} k(\cdot\,, Y_j).$$

Note that $\widehat{\Phi}(w) + \lambda w$ in (18) is the Gâteaux derivative of the empirical loss function of KuLSIF in (4) including the regularization term. The optimal solution of (18) is the same as that of KuLSIF, and hence, the same results as Theorem 1 and Theorem 2 hold for the inductive version of the KMM estimator. The computational efficiency, however, could be different. We show numerical examples of the computational cost in Section 5.

In a companion paper (Kanamori et al., 2011), we further investigate the computational properties of the KuLSIF method from the viewpoint of condition numbers (see Section 8.7 of Luenberger & Ye, 2008), and reveal that KuLSIF is computationally more efficient than KMM.

Another difference between KuLSIF and the inductive variant of KMM lies in model selection. As shown in Section 3.2, KuLSIF is equipped with cross-validation, and thus model selection can be performed systematically. On the other hand, the KMM objective function (9) is defined in terms of the RKHS norm. This implies that once kernel parameters (such as the Gaussian kernel width) are changed, the definition of the objective function is also changed and therefore naively performing cross-validation may not be valid in KMM. The regularization parameter in KMM may be optimized by cross-validation for a fixed RKHS.

4.2 M-Estimator with the Kullback-Leibler Divergence (KL-div)

The M-estimator based on the Kullback-Leibler (KL) divergence (Nguyen et al., 2010) also directly gives an estimate of the density ratio without going through density estimation. The KL divergence $I(Q, P)$ is defined as

$$\begin{aligned} I(Q,P) &= -\int \log \frac{p(z)}{q(z)} dQ(z) \\ &= -\inf_w \left[-\int \log(w(z)) dQ(z) + \int w(z) dP(z) - 1 \right], \end{aligned} \tag{19}$$

where the second equality follows from the conjugate dual function of the logarithmic function and the infimum is taken over all measurable functions. Detailed derivation is shown in Nguyen et al. (2010). The optimal solution of (19) is given as $w(z) = q(z)/p(z)$, and thus, the empirical approximation of (19) leads to the loss function for the estimation of density ratios.

The kernel-based estimator $\widehat{w}(z)$ is defined as an optimal solution of

$$\min_w -\frac{1}{m}\sum_{j=1}^{m} \log(w(Y_j)) + \frac{1}{n}\sum_{i=1}^{n} w(X_i) + \frac{\lambda}{2}\|w\|_{\mathcal{H}}^2, \quad w \in \mathcal{H},$$

where $\mathcal{H}$ is an RKHS. We may also use the truncated one $\widehat{w} = \max\{w, 0\}$ as the estimator of the density ratio. Nguyen et al. (2010) proved that the RKHS $\mathcal{H}$ endowed with the Gaussian kernel and regularization parameter $\lambda = (m \wedge m)^{\delta-1}$, $(0 < \delta < 1)$ leads to a consistent estimator under a boundedness assumption on $w_0 = q/p$. Due to the representer theorem (Kimeldorf & Wahba, 1971), we see that the above infinite-dimensional optimization problem is reduced to a finite-dimensional one.

Furthermore, the optimal solution of KL-div has a similar form to that shown in Theorem 1, and one needs to estimate only m parameters when samples (1) are observed (Nguyen et al., 2010). Actually, this property holds for general M-estimators with all f-divergences (Ali & Silvey, 1966; Csiszár, 1967); see Kanamori et al. (2011) for details.

Note that model selection of the KL-div method can be systematically carried out based on cross-validation in terms of the KL-divergence (Sugiyama et al., 2008b).

4.3 Kernel Logistic Regression (KLR)

Another approach to directly estimating the density ratio is to use a probabilistic classifier. Let b be a binary random variable. For the conditional probability $p(z|b)$, we assume that

$$\begin{aligned} p(z) &= p(z|b = +1), \\ q(z) &= p(z|b = -1), \end{aligned}$$

hold. That is, b plays a role as a 'class label' for discriminating 'numerator' and 'denominator'. An application of the Bayes theorem yields that the density ratio can be expressed in terms of the class label b as

$$w_0(z) = \frac{q(z)}{p(z)} = \frac{p(b=+1)}{p(b=-1)} \frac{p(b=-1|z)}{p(b=+1|z)}.$$

The ratio of class-prior probabilities, $p(b=+1)/p(b=-1)$, can be simply estimated from the numbers of samples from P and Q, and the class-posterior probability $p(b|z)$ can be estimated by discrimination methods such as logistic regression. Below we briefly explain the kernel logistic regression method (Wahba et al., 1993; Zhu & Hastie, 2001).

The kernel logistic regression method employs a model of the following form for expressing the class-posterior probability $p(b|z)$:

$$p(b|z) = \frac{1}{1+\exp\left(-bf(z)\right)}, \qquad f \in \mathcal{H},$$

where $\mathcal{H}$ is an RKHS on $\mathcal{Z}$. The function $f \in \mathcal{H}$ is learned so that the negative regularized log-likelihood based on training samples (1) is minimized:

$$\min_f \ \frac{1}{n+m}\left[\sum_{i=1}^{n} \log(1+e^{-f(X_i)}) + \sum_{j=1}^{m} \log(1+e^{f(Y_j)})\right] + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2, \qquad f \in \mathcal{H},$$

where λ is the regularization parameter. Let $\widehat{f}$ be an optimal solution. Then the density ratio can be estimated by

$$\widehat{w}(z) = \frac{n}{m} e^{-\widehat{f}(z)}.$$

Note that we do not need to truncate the negative part of $\widehat{w}(z)$, since the estimator $\widehat{w}$ takes only positive values by construction.

Model selection of the KLR-based density-ratio estimator is performed by cross-validation in terms of the classification accuracy measured by the log-likelihood of the logistic model.

4.4 Ratio of Kernel Density Estimators (RKDE)

The *kernel density estimator* (KDE) is a non-parametric technique to estimate a probability density function $p(x)$ from its i.i.d. samples $\{x_k\}_{k=1}^n$. For the Gaussian kernel $k_\sigma(x,x') = \exp\{-\|x-x'\|^2/(2\sigma^2)\}$, KDE is expressed as

$$\widehat{p}(x) = \frac{1}{n(2\pi\sigma^2)^{d/2}} \sum_{k=1}^{n} k_\sigma(x, x_k).$$

The accuracy of KDE heavily depends on the choice of the kernel width σ, which can be optimized by cross-validation in terms of the log-likelihood. See Härdle et al. (2004) for details.

KDE can be used for density-ratio estimation by first obtaining density estimators $\widehat{p}(x)$ and $\widehat{q}(y)$ separately from $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, and then estimating the density ratio by $\widehat{q}(z)/\widehat{p}(z)$. This estimator is referred to as the ratio of kernel density estimators (RKDE). A potential limitation of RKDE is that division by an estimated density $\widehat{p}(z)$ is involved, which tends to magnify the estimation error of $q(z)$. This is critical when the number of available samples is limited. Therefore, the KDE-based approach may not be reliable in high-dimensional problems.

5 Simulation Studies

In this section, we numerically compare the computational cost and the statistical performance of proposed and existing density-ratio estimators.

5.1 Computational Costs

First, we experimentally investigate the computational cost of KuLSIF and KMM. The *IDA data sets* (Rätsch et al., 2001) are used, which are binary classification data sets consisting of positive/negative and training/test samples (see Table 3). We use large data sets in IDA: `titanic`, `waveform`, `banana`, `ringnorm`, and `twonorm`, and compare the computation time of KuLSIF (6) with that of the inductive KMM (18). The solutions are numerically computed by minimizing the objective functions using the BFGS quasi-Newton method implemented in the `optim` function in the R environment (R Development Core Team, 2009). For KuLSIF, we also investigate the computation time for directly solving the linear equation (5) by the function `solve` in R. Note that theoretically all methods share the same solution (see Section 4.1).

In the first experiments, the data set corresponding to the distribution P consists of all positive test samples, and all negative test samples are assigned to the other data set corresponding to Q. Therefore, the target density-ratio may be far from the constant function $w_0(x) = 1$. Table 2(a) shows the average computation time over 20 runs. In the table, 'KuLSIF(numerical)', 'KuLSIF(direct)', and 'KMM' denote KuLSIF numerically minimizing the loss function, KuLSIF directly solving the linear equation, and the inductive variant of KMM (numerically minimizing the loss function), respectively. In the second experiments, samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are both randomly taken from all (i.e., both positive and negative) test samples. Hence, the target density-ratio is almost equal to the constant function $w_0(x) = 1$. Table 2(b) shows the average computation time over 20 runs.

The results show that, for large data sets, KuLSIF(numerical) is computationally more efficient than KuLSIF(direct). Experimentally, the computational cost of KuLSIF(numerical) is approximately proportional to n^2, while that of KuLSIF(direct) takes the order of n^3. Thus, for large data sets, computing the solution by numerically minimizing the quadratic loss function will be more advantageous than directly solving the linear equation. KMM is computationally highly demanding for all cases.

Table 2: The averaged computation time (sec.) of KuLSIF(numerical), KuLSIF(direct), and KMM are presented. (a) The data set from P is randomly taken from positive test samples, and that from Q is randomly taken from negative test samples. (b) Two data sets $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are both randomly taken from all (i.e., both positive and negative) test samples. The data sets are arranged in ascending order of the sample size n. Results of the method having the lowest mean are described by bold face.

(a) The true density-ratio is far from a constant

data set	n	m	KuLSIF (numerical)	KuLSIF (direct)	KMM
titanic	1327	2775	6.11	**1.45**	57.96
waveform	3032	6168	52.74	**16.96**	1713.71
banana	4383	5417	97.64	**52.97**	1539.65
ringnorm	6933	7067	**145.37**	177.96	4346.32
twonorm	7002	6998	**145.61**	226.20	1944.79

(b) The true density-ratio is close to a constant

data set	n	m	KuLSIF (numerical)	KuLSIF (direct)	KMM
titanic	2052	2050	10.20	**5.13**	91.97
waveform	4600	4600	63.55	**58.55**	3078.64
banana	4900	4900	112.21	**78.08**	1408.91
ringnorm	7000	7000	**135.70**	258.03	3201.78
twonorm	7000	7000	**133.44**	243.46	3584.25

5.2 Stability of Estimators

By using synthetic data, we study how the dimension of the data affects the estimation accuracy. The probabilities P and Q are defined as the Gaussian distribution with increasing dimension ranges from 1 to 10, and the sample size is set to $m = n = 500$. The covariance matrix of both distributions is given as the identity matrix. The mean vector of P is the null vector, and that of Q is equal to $\mu \boldsymbol{e}_1$, where $\boldsymbol{e}_1$ is the standard unit vector with only the first component being 1. Then, the density ratio is equal to $w_0(x) = \exp\{x_1\mu - \mu^2/2\}$ for $x = (x_1, \ldots, x_d)$.

For each case of $\mu = 0$ and $\mu = 1$, we compare four kernel-based estimators: KuLSIF, the M-estimator with the Kullback-Leibler divergence (KL-div), kernel logistic regression (KLR), and the ratio of kernel density estimators (RKDE). See Section 4 for details of each estimator. The `solve` function in R is used for computing the KuLSIF solution, and the `optim` function in R is used for computing the KL-div solution. For computing the KLR solution, we use the `myKLR` package (Rüping, 2003), which is a C++ implementation of the algorithm proposed by Keerthi et al. (2005) to solve the dual problem. For computing the RKDE, we use our own implementation in R.

In all estimators, the Gaussian kernel is used. Except RKDE, the kernel width σ is

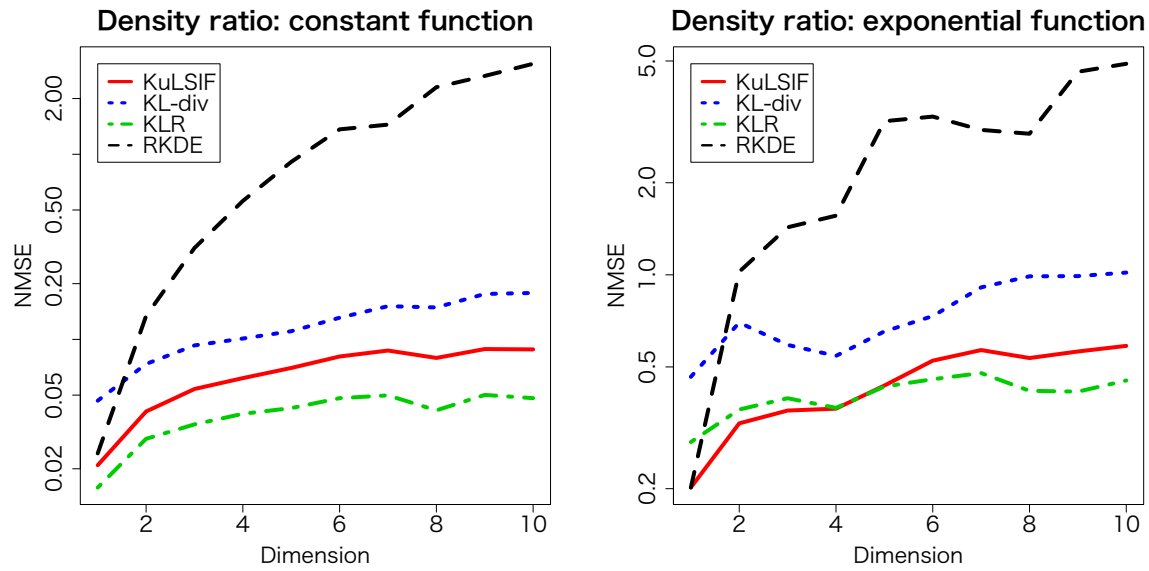


Figure 1: The median of NMSEs is depicted as functions of the input dimensionality. Left panel: the case of $w_0(x) = 1$ (i.e., $\mu = 0$). Right panel: the case of $w_0(x) = \exp\{x_1 - 1/2\}$ (i.e., $\mu = 1$).

set to the median of $\|z - z'\|$ among all pairs of distinct training points, z and z'. This is a standard heuristic for the choice of the Gaussian kernel width (Schölkopf & Smola, 2002). For RKDE, the kernel width is chosen by using CV among 20 candidates around the value determined by the above median heuristics. The regularization parameter λ is set to $\lambda = 1/(n \wedge m)^{0.9}$. The estimation accuracy of the density-ratio estimator $\widehat{w}(z)$ is evaluated by the normalized mean squared error (NMSE) over the test points $\widetilde{z}_1, \ldots, \widetilde{z}_N$:

$$\mathrm{NMSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\widehat{w}(\widetilde{z}_i)}{\frac{1}{N}\sum_{k=1}^{N}\widehat{w}(\widetilde{z}_k)} - \frac{w_0(\widetilde{z}_i)}{\frac{1}{N}\sum_{k=1}^{N}w_0(\widetilde{z}_k)}\right)^2. \tag{20}$$

In many applications of density ratios, only the relative size of the density ratio is required and the normalization factor is not essential (Sugiyama et al., 2009; Sugiyama et al., 2012). On the other hand, the target of the current experiment is density ratio estimation itself. Thus, evaluating the error under normalization would be reasonable.

Figure 1 presents the median NMSE for each estimator as functions of the input dimensionality. Since the average NMSE of the RKDE took extremely large values high-dimensional data, we decided to use the median NMSE for evaluation. We see that the RKDE immediately gets unstable for multi-dimensional data, whereas the other three estimators provide stable prediction for all data sets.

5.3 Statistical Performance

Finally, we experimentally compare the statistical performance of four kernel-based estimators: KuLSIF, the M-estimator with the Kullback-Leibler divergence (KL-div), kernel

logistic regression (KLR), and the ratio of kernel density estimators (RKDE).

We again use the IDA data sets (see Table 3 for details). Bayes error in Table 3 denotes the test error of the best classifier reported in Rätsch et al. (2001). First, we explain how to prepare the training data set using the IDA data sets. Given the training data for binary classification, $(z_1, b_1), \ldots, (z_t, b_t) \in \mathcal{Z} \times \{+1, -1\}$, the posterior probability of binary labels is estimated by the support vector machine with the Gaussian kernel, where Platt's approach (Platt, 2000) is used.[3] The estimated class-posterior probability is denoted as $\widehat{P}(b|z)$, and let $\widehat{P}_\eta(b|z)$ be

$$\widehat{P}_\eta(b|z) = (1-\eta)\widehat{P}(b|z) + \frac{\eta}{2}, \qquad 0 \leq \eta \leq 1,$$

for $z \in \mathcal{Z}$ and $b \in \{+1, -1\}$. The probability $\widehat{P}_0(b|z)$ will lead to the Bayes error close to the values described in Table 3, whereas $\widehat{P}_1(b|z)$ indicates the uniform probability on the binary labels. Hence, the Bayes error of $\widehat{P}_1(b|z)$ is equal to 0.5. Then, the label of the training input z_i is *reassigned* according to the conditional probability $\widehat{P}_\eta(b|z_i)$. The input points with the reassigned label $+1$ are regarded as samples from the probability distribution P, and those with the label -1 are regarded as samples from the probability distribution Q. As such, we have the training data set (1), and the density ratio is estimated based on these training samples. Thus, the true density-ratio is approximately given by

$$\widetilde{w}_0(z) = \frac{n}{m} \cdot \frac{\widehat{P}_\eta(b=-1|z)}{\widehat{P}_\eta(b=+1|z)}.$$

Note that $\widetilde{w}_0$ is close to the constant function 1, when η is close to 1. The estimation accuracy of the estimator $\widehat{w}(z)$ is evaluated by the NMSE (20). Here the test points in the NMSE are uniformly sampled from all the test input vectors of the classification data set. Hence, the distribution of $\widetilde{z}_i$ is not the same as the probability distribution P, unless $\eta = 1$.

Some examples of estimated density-ratios are depicted in Figure 2 and Figure 3. The training samples are generated from the data set `banana` or `german` with the mixing parameter $\eta = 0.01$. In these figures, the index of test samples $\widetilde{z}_i$ is arranged in the ascending order of the density-ratio values $\widetilde{w}_0(\widetilde{z}_i)$. The solid increasing line denotes $\widetilde{w}_0(\widetilde{z}_i)$ for each test point, and $\circ$'s in the plots are estimated values. When the input dimension is low (see Figure 2), all the methods including RKDE perform reasonably well. However, for the high-dimensional data (see Figure 3), RKDE severely overfits due to division by an estimated density. As illustrated in Section 5.2, this leads to the instability of estimation by RKDE, and the prediction ability becomes poor. The other three direct density-ratio estimators provide reasonably stable prediction even when the dimension is high.

[3] In the Platt's approach, the conditional probability is estimated by the model $\widehat{P}(b|z) = 1/\{1 + \exp(-b(\alpha\widehat{h}(z) + \beta))\}, \alpha, \beta \in \Re$, where $\widehat{h} : \mathcal{Z} \to \Re$ is the decision function estimated by support vector machine. Maximum likelihood estimation is used to estimate the parameter α, β.

Table 3: The dimension of the data domain and the Bayes error are shown. The Bayes error denotes the lowest test error reported in the original paper (Rätsch et al., 2001). The data sets are arranged in the ascending order of the Bayes error. "#samples" is the total sample size, i.e., $n + m$. "Iterations" denotes the number of trials of estimation to compute the average performance of estimators.

data set	dimension	Bayes error (%)	#samples ($n + m$)	iterations
ringnorm	20	1.5	7000	20
twonorm	20	2.6	7000	20
image	18	2.7	1010	20
thyroid	5	4.2	75	50
splice	60	9.5	2175	20
waveform	21	9.8	4600	20
banana	2	10.7	4900	20
heart	13	16.0	100	50
titanic	3	22.4	2051	20
diabetes	8	23.2	300	50
german	20	23.6	300	50
breast-cancer	9	24.8	77	50
flare-solar	9	32.4	400	50

Next, we compare the following methods: KuLSIF, KuLSIF with leave-one-out cross-validation (LOOCV), KL-div, KL-div with 5-fold cross-validation (CV), KLR, KLR with CV, and RKDE with CV. For KuLSIF, KL-div, and KLR, LOOCV or CV is used to choose the regularization parameter λ from $2^k/(n \wedge m)^{0.9}, k = -5, -4, \ldots, 4, 5$. We also test a fixed value $\lambda = (m \wedge n)^{-0.9}$ for KuLSIF, KL-div, and KLR. As shown in Section 3, the regularization parameter $\lambda = (m \wedge n)^{-0.9}$ guarantees the statistical consistency of KuLSIF and KL-div under mild assumptions. Statistical properties of KLR have been studied by Bartlett et al. (2006), Bartlett and Tewari (2007), Steinwart (2005), and Park (2009). Especially, Steinwart (2005) has proved that, under mild assumptions, KLR with $\lambda = (m + n)^{-0.9}$ has the statistical consistency. When the training samples are balanced, i.e., the ratio of sample size m/n converges to a positive constant, the regularization parameter $\lambda = (m \wedge n)^{-0.9}$ guarantees the statistical consistency of KLR. In all estimators, the Gaussian kernel is used. Except RKDE, the kernel width σ is set by using the standard heuristic introduced in Section 5.2. For RKDE, the kernel width is chosen by using CV among 20 candidates around the value determined by the above median heuristics.

For each data set, training samples generated by setting $\eta = 0.01, 0.1, 0.5$ or 1 in $P_\eta(b|z)$ are respectively prepared. For each training set, the NMSE of each estimator is computed. By using the NMSE over the uniformly distributed test samples, the estimation accuracy on the whole data domain is evaluated, while Theorem 2 does not guarantee the statistical consistency for that test distribution. To compute the average performance, the above experiments are repeated multiple times as described in Table 3. The numerical results are presented in Tables 4–7 and Figure 4. In the tables, data sets are arranged in the

ascending order of the Bayes error shown in Table 3. In Table 4, the NMSEs under $\eta = 1$ are presented. In this case, the class-posterior probability satisfies $\widehat{P}_1(b|z) = 0.5$, and hence, the density ratio is close to the constant function. Then, estimators with strong regularization will provide good results. Indeed, methods using LOOCV or CV such as KuLSIF(LOOCV), KL-div(CV), and KLR(CV) achieve the lowest NMSEs. Especially, KLR(CV) is significantly better than the others. For small η, other estimators except RKDE also present good statistical performance (see Tables 5–7).

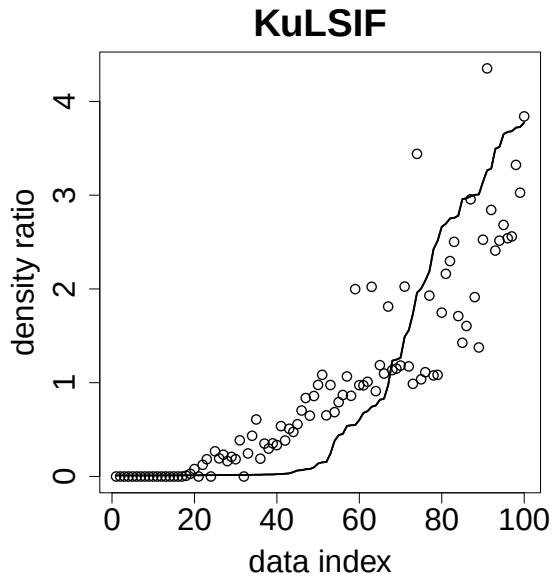
At the bottom of each table, the relative computational costs are also described. The computation time depends on parameters included in the optimization algorithm. In order to reduce the computation time of KL-div and KL-div(CV), we used the `optim` function with the stopping criterion `reltol`$=0.5 \times 10^{-3}$, instead of the default value `reltol`$=10^{-8}$. On the other hand, KuLSIF using the `solve` function is numerically accurate. From the experimental results, we see that KuLSIF dominates the other methods in terms of the computational efficiency. KuLSIF(LOOCV) with the analytic-form expression of the LOOCV score also has a computational advantage over KL-div(CV), KLR(CV), and RKDE(CV). Note that the relative computational cost of KL-div is large for small η. This phenomenon is theoretically studied in a companion paper (Kanamori et al., 2011).

In Figure 4, the NMSEs of estimators described in these tables are plotted as functions of η. The NMSEs of RKDE are not shown since they are much larger than the others. We see that KLR and KLR(CV) are sensitive to η for the data set with low Bayes error such as `ringnorm`, `twonorm`, `image`, `thyroid`, `splice`, `waveform`, and `banana`. On the other hand, KuLSIF, KuLSIF(LOOCV), and KL-div(CV) present moderate NMSEs for a wide range of η. See Tables 4–7 for more details.

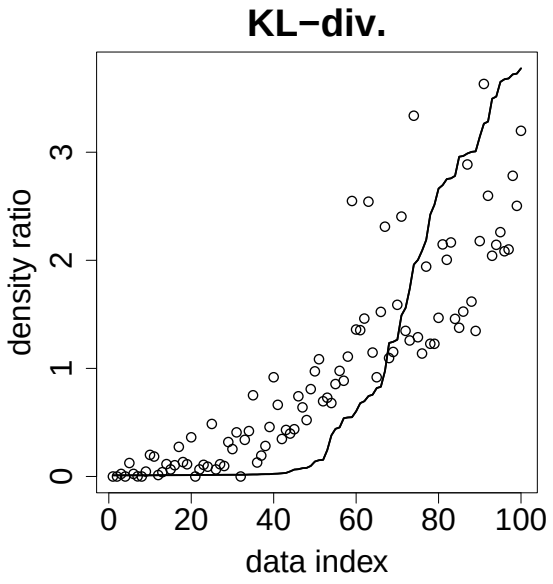
6 Conclusions

In this paper, we addressed the problem of estimating the ratio of two probability densities. We proposed a kernel-based least-squares density-ratio estimator called KuLSIF, and investigated its statistical properties such as consistency and the rate of convergence. We also showed that, not only the estimator, but also the leave-one-out cross-validation score can be analytically obtained for KuLSIF. This highly contributes to reducing the computational cost. Then we pointed out that KuLSIF and an inductive variant of kernel mean matching (KMM) actually share the same solution. Hence, the statistical properties of KuLSIF are inherited to KMM. However, we showed through numerical experiments that KuLSIF is computationally much more efficient than KMM. We further experimentally showed that KuLSIF overall compares favorably with other density-ratio estimators such as the M-estimator with the Kullback-Leibler divergence, kernel logistic regression, and the ratio of kernel density estimators.

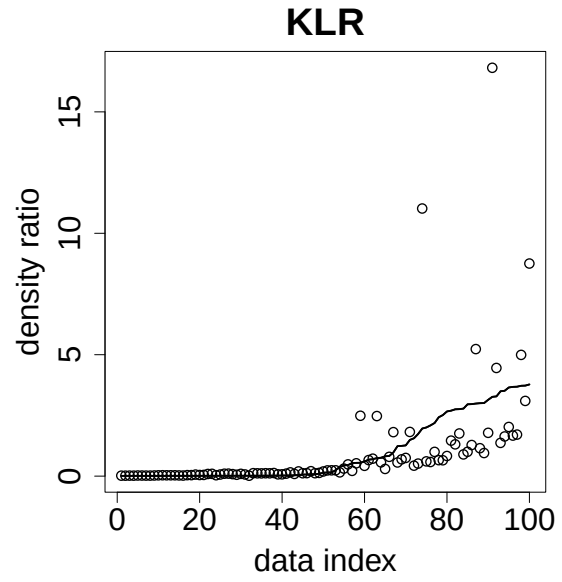
Our definition of KuLSIF (see Eq.(4)) does not contain a non-negativity constraint on the learned density-ratio function. We may add a non-negativity constraint $w \geq 0$ to (4) as Kanamori et al. (2009) did. However, by the additional constraint, we can no longer obtain the solution analytically. When the sample size is large, the estimator $\widehat{w}$



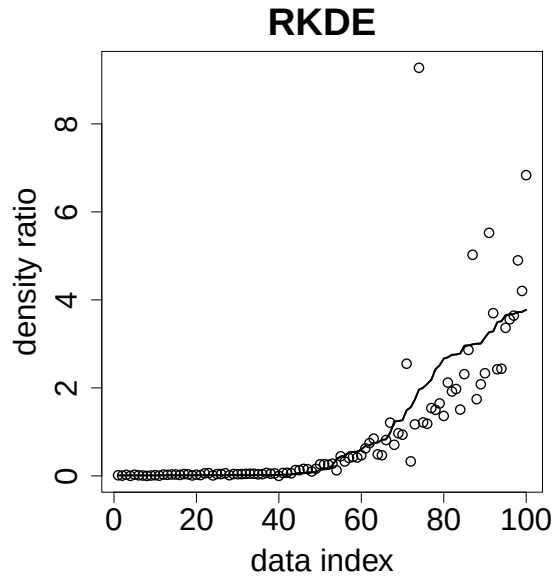
(a) KuLSIF: NMSE= 0.634



(b) KL-div: NMSE= 0.749

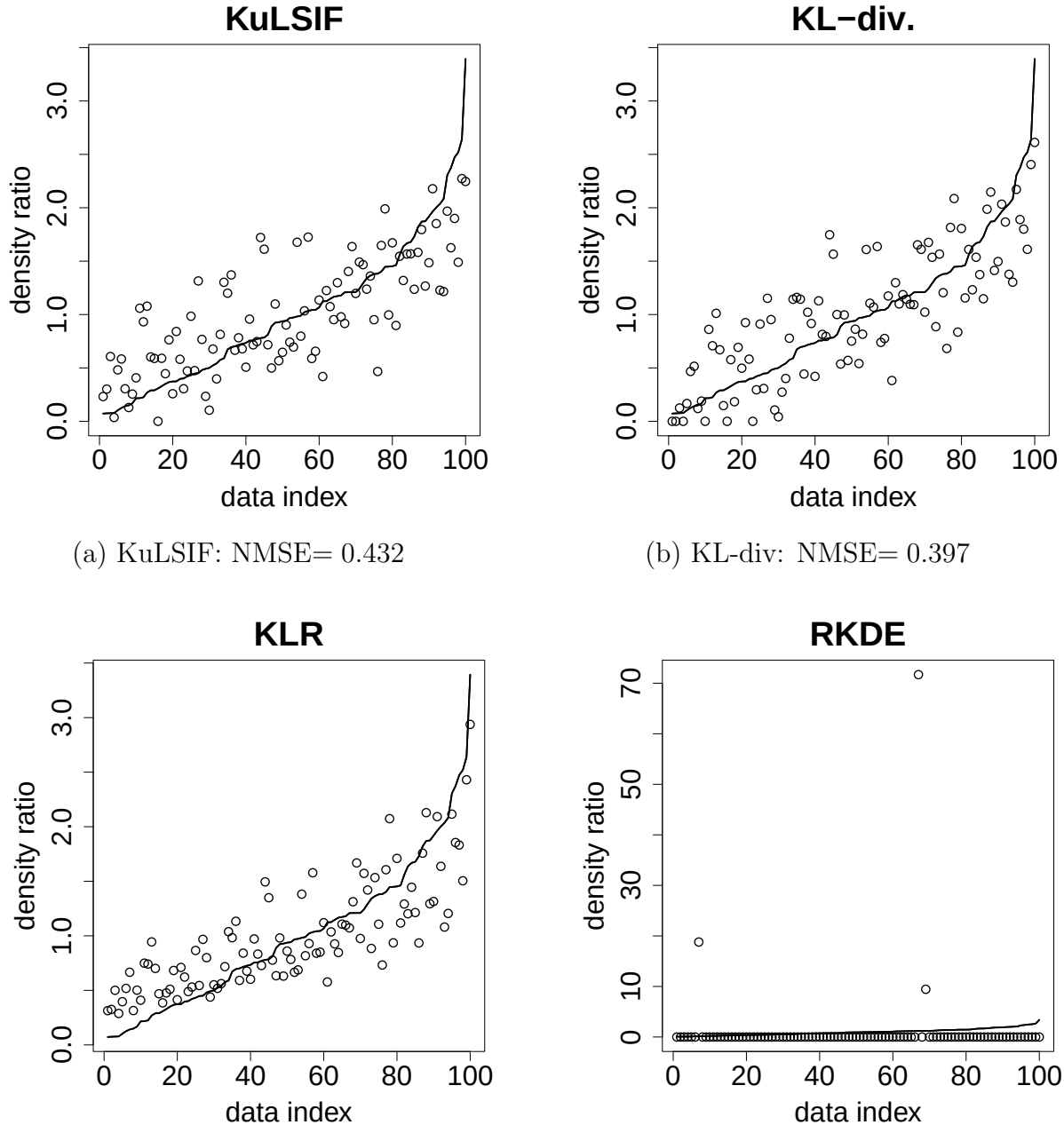


(c) KLR: NMSE= 1.922



(d) RKDE: NMSE= 0.960

Figure 2: Estimation of density ratios for the data set `banana` is shown. The dimension of the data is 2. The solid line is the true density-ratio $\widetilde{w}_0$ with $\eta = 0.01$, and $\circ$'s are predicted values of the density ratio. The data index is arranged in the ascending order of the density ratio, and thus the solid line is an increasing function.



(a) KuLSIF: NMSE= 0.432

(b) KL-div: NMSE= 0.397

(c) KLR: NMSE= 0.372

(d) RKDE: NMSE= 7.435

Figure 3: Estimation of density ratios for the data set `german` is shown. The dimension of the data is 20. The solid line is the true density-ratio $\widetilde{w}_0$ with $\eta = 0.01$, and $\circ$'s are predicted values of the density ratio. The data index is arranged in the ascending order of the density ratio, and thus the solid line is an increasing function.

Table 4: Average NMSEs of each estimator for training data with $\eta = 1$. The lowest NMSEs are in bold, and those with the second lowest are also in bold if they are not significantly different from the lowest NMSEs under the Wilcoxon paired rank-sum test with the significance level 5%. "—" in RKDE denotes the fact that numerically 0/0 occurred frequently and thus valid results were not obtained.

data set ($\eta = 1$)	Bayes err. (%)	KuLSIF ($\lambda = (m \wedge n)^{-0.9}$)	KuLSIF (LOOCV)	KL-div ($\lambda = (m \wedge n)^{-0.9}$)	KL-div (CV)	KLR ($\lambda = (m \wedge n)^{-0.9}$)	KLR (CV)	RKDE (CV)
ringnorm	50	0.23±0.01	0.18±0.01	0.35±0.02	0.25±0.01	0.17±0.01	**0.03±0.01**	41.7±9.31
twonorm	50	0.23±0.01	0.15±0.01	0.33±0.03	0.16±0.02	0.17±0.01	**0.03±0.01**	30.8±12.4
image	50	0.26±0.03	0.22±0.03	0.45±0.06	0.23±0.04	0.19±0.02	**0.05±0.03**	—
thyroid	50	0.49±0.11	0.53±0.16	0.57±0.13	0.53±0.14	0.25±0.09	**0.10±0.22**	4.39±2.23
splice	50	0.31±0.02	0.14±0.01	0.39±0.05	0.15±0.02	0.22±0.02	**0.03±0.03**	19.9±6.63
waveform	50	0.23±0.01	0.13±0.01	0.37±0.05	0.14±0.03	0.17±0.01	**0.04±0.02**	17.8±5.49
banana	50	0.10±0.02	0.09±0.02	0.52±0.14	0.15±0.07	0.09±0.02	**0.05±0.02**	23.9±23.9
heart	50	0.36±0.06	0.26±0.07	0.54±0.10	0.31±0.18	0.23±0.05	**0.13±0.23**	5.35±1.27
titanic	50	0.11±0.03	0.13±0.10	0.57±0.20	0.14±0.04	0.10±0.03	**0.03±0.02**	6.07±12.3
diabetes	50	0.36±0.05	0.30±0.05	0.46±0.06	0.30±0.07	0.25±0.06	**0.07±0.09**	9.72±2.98
german	50	0.35±0.04	0.25±0.03	0.48±0.06	0.26±0.04	0.23±0.04	**0.03±0.04**	10.9±1.88
breast-cancer	50	0.37±0.09	0.30±0.13	0.51±0.13	0.34±0.21	0.23±0.07	**0.11±0.21**	4.09±1.28
flare-solar	50	0.21±0.06	0.17±0.06	0.31±0.08	0.17±0.08	0.21±0.05	**0.05±0.06**	12.8±2.55
relative computational cost		1.0	38.0	2.0	132.1	5.0	207.1	45.0

Table 5: Average NMSEs of each estimator for training data with $\eta = 0.5$. The lowest NMSEs are in bold, and those with the second lowest are also in bold if they are not significantly different from the lowest NMSEs under the Wilcoxon paired rank-sum test with the significance level 5%. "—" in RKDE denotes the fact that numerically 0/0 occurred frequently and thus valid results were not obtained.

data set ($\eta = 0.5$)	Bayes err. (%)	KuLSIF ($\lambda = (m \wedge n)^{-0.9}$)	KuLSIF (LOOCV)	KL-div ($\lambda = (m \wedge n)^{-0.9}$)	KL-div (CV)	KLR ($\lambda = (m \wedge n)^{-0.9}$)	KLR (CV)	RKDE (CV)
ringnorm	25.7	**0.29±0.01**	**0.29±0.01**	0.41±0.01	0.34±0.02	0.33±0.01	0.31±0.01	42.8±12.6
twonorm	26.3	0.34±0.02	**0.31±0.02**	0.44±0.03	**0.33±0.05**	0.44±0.03	0.43±0.03	30.8±9.89
image	26.3	**0.45±0.07**	**0.46±0.07**	0.49±0.08	0.47±0.05	0.50±0.07	0.55±0.10	—
thyroid	27.1	0.34±0.11	**0.31±0.12**	0.35±0.11	0.39±0.19	0.32±0.08	0.43±0.15	4.60±2.08
splice	29.8	**0.42±0.01**	**0.42±0.01**	0.51±0.02	0.47±0.02	0.45±0.01	0.57±0.06	23.1±5.92
waveform	29.9	0.29±0.02	**0.25±0.02**	0.30±0.02	0.26±0.01	0.29±0.02	0.31±0.03	19.8±7.50
banana	30.4	0.28±0.01	**0.26±0.02**	0.58±0.06	0.37±0.04	0.41±0.04	0.37±0.05	19.91±21.4
heart	33.0	0.38±0.08	0.37±0.07	0.47±0.11	0.46±0.17	**0.36±0.07**	0.45±0.18	5.10±1.41
titanic	36.2	0.14±0.02	**0.13±0.03**	0.58±0.25	0.18±0.04	0.15±0.02	**0.12±0.02**	1.71±7.14
diabetes	36.6	0.32±0.05	0.30±0.05	0.37±0.06	0.31±0.06	**0.27±0.04**	0.29±0.06	8.86±3.00
german	36.8	0.34±0.05	0.32±0.05	0.41±0.06	0.33±0.08	**0.28±0.04**	0.31±0.07	10.8±1.98
breast-cancer	37.4	0.36±0.08	0.30±0.12	0.44±0.10	0.38±0.20	**0.24±0.07**	0.33±0.21	3.79±1.49
flare-solar	41.2	0.24±0.05	0.25±0.06	0.29±0.07	0.25±0.05	0.23±0.05	**0.22±0.05**	12.0±2.96
relative computational cost		1.0	38.9	3.8	179.3	5.3	218.4	47.7

Table 6: Average NMSEs of each estimator for training data with $\eta = 0.1$. The lowest NMSEs are in bold, and those with the second lowest are also in bold if they are not significantly different from the lowest NMSEs under the Wilcoxon paired rank-sum test with the significance level 5%. "—" in RKDE denotes the fact that numerically 0/0 occurred frequently and thus valid results were not obtained.

data set ($\eta = 0.1$)	Bayes err. (%)	KuLSIF ($\lambda = (m \wedge n)^{-0.9}$)	KuLSIF (LOOCV)	KL-div ($\lambda = (m \wedge n)^{-0.9}$)	KL-div (CV)	KLR ($\lambda = (m \wedge n)^{-0.9}$)	KLR (CV)	RKDE (CV)
ringnorm	6.3	0.34±0.01	**0.32±0.02**	0.61±0.02	0.50±0.04	0.82±0.02	0.98±0.03	31.7±8.78
twonorm	7.3	**0.41±0.02**	**0.41±0.02**	0.65±0.03	0.55±0.13	1.34±0.07	1.31±0.10	36.4±10.3
image	7.4	0.79±0.14	0.89±0.37	0.76±0.12	**0.67±0.05**	0.95±0.12	1.66±0.43	—
thyroid	8.8	0.36±0.10	0.37±0.11	**0.31±0.11**	0.44±0.19	0.39±0.09	0.61±0.26	2.41±2.21
splice	13.6	0.54±0.02	**0.49±0.01**	0.85±0.04	0.67±0.08	0.76±0.05	2.91±0.56	21.7±6.93
waveform	13.8	0.36±0.04	0.35±0.04	0.39±0.04	**0.32±0.02**	0.53±0.04	0.74±0.11	26.1±9.02
banana	14.7	0.49±0.02	**0.46±0.04**	0.93±0.06	0.83±0.07	1.28±0.08	1.74±0.35	37.8±16.4
heart	19.4	0.58±0.13	0.58±0.13	**0.51±0.11**	0.66±0.15	0.57±0.13	0.79±0.32	5.20±1.31
titanic	25.2	0.18±0.02	**0.15±0.04**	0.60±0.18	0.23±0.04	0.22±0.02	**0.15±0.03**	0.16±0.05
diabetes	25.9	0.39±0.07	0.40±0.07	0.40±0.08	0.41±0.08	**0.37±0.06**	0.47±0.13	10.1±2.31
german	26.2	0.45±0.07	0.46±0.07	**0.45±0.06**	0.46±0.08	0.50±0.08	0.53±0.13	11.2±1.66
breast-cancer	27.3	0.37±0.08	0.35±0.10	0.41±0.08	0.43±0.16	**0.31±0.10**	0.40±0.23	3.85±1.36
flare-solar	34.2	0.30±0.07	0.30±0.07	0.35±0.07	0.33±0.08	**0.25±0.06**	0.28±0.09	11.6±3.43
relative computational cost		1.0	38.9	4.6	237.7	5.6	226.4	49.9

Table 7: Average NMSEs of each estimator for training data with $\eta = 0.01$. The lowest NMSEs are in bold, and those with the second lowest are also in bold if they are not significantly different from the lowest NMSEs under the Wilcoxon paired rank-sum test with the significance level 5%. "—" in RKDE denotes the fact that numerically 0/0 occurred frequently and thus valid results were not obtained.

data set ($\eta = 0.01$)	Bayes err. (%)	KuLSIF ($\lambda = (m \wedge n)^{-0.9}$)	KuLSIF (LOOCV)	KL-div ($\lambda = (m \wedge n)^{-0.9}$)	KL-div (CV)	KLR ($\lambda = (m \wedge n)^{-0.9}$)	KLR (CV)	RKDE (CV)
ringnorm	2.0	0.49±0.01	**0.44±0.01**	0.72±0.02	0.62±0.03	1.10±0.03	2.28±0.17	21.9±15.4
twonorm	3.1	**0.38±0.01**	0.48±0.04	0.71±0.05	0.66±0.09	2.11±0.10	3.72±0.73	43.9±10.7
image	3.1	1.26±0.13	1.44±0.31	1.14±0.16	**1.03±0.11**	1.24±0.12	2.04±0.66	—
thyroid	4.7	0.64±0.15	0.64±0.15	**0.61±0.14**	**0.60±0.15**	0.68±0.14	0.67±0.20	2.66±2.23
splice	9.9	1.10±0.04	0.98±0.05	1.36±0.06	1.21±0.07	**0.75±0.03**	5.70±1.94	27.1±5.62
waveform	10.2	**0.47±0.04**	0.49±0.05	0.59±0.05	0.50±0.04	0.66±0.04	1.11±0.10	38.8±8.54
banana	11.1	1.00±0.04	**0.78±0.04**	1.59±0.09	1.50±0.08	1.61±0.08	4.30±0.77	44.3±9.82
heart	16.3	0.79±0.17	0.77±0.16	**0.69±0.15**	0.78±0.20	0.76±0.16	0.87±0.38	5.10±1.48
titanic	22.7	0.19±0.02	**0.12±0.04**	0.70±0.31	0.25±0.07	0.23±0.02	0.14±0.03	**0.12±0.04**
diabetes	23.5	0.49±0.11	0.50±0.11	0.50±0.10	0.52±0.12	**0.48±0.09**	0.54±0.19	9.91±2.67
german	23.9	0.62±0.14	0.62±0.14	**0.60±0.13**	0.61±0.13	0.68±0.16	0.63±0.18	11.1±1.68
breast-cancer	25.0	0.37±0.13	**0.35±0.16**	0.40±0.12	0.42±0.18	**0.32±0.17**	0.40±0.23	3.62±1.34
flare-solar	32.6	0.34±0.06	0.34±0.13	0.37±0.07	0.37±0.07	**0.27±0.06**	0.30±0.09	12.1±3.02
relative computational cost		1.0	39.4	5.2	254.3	5.7	230.9	50.4

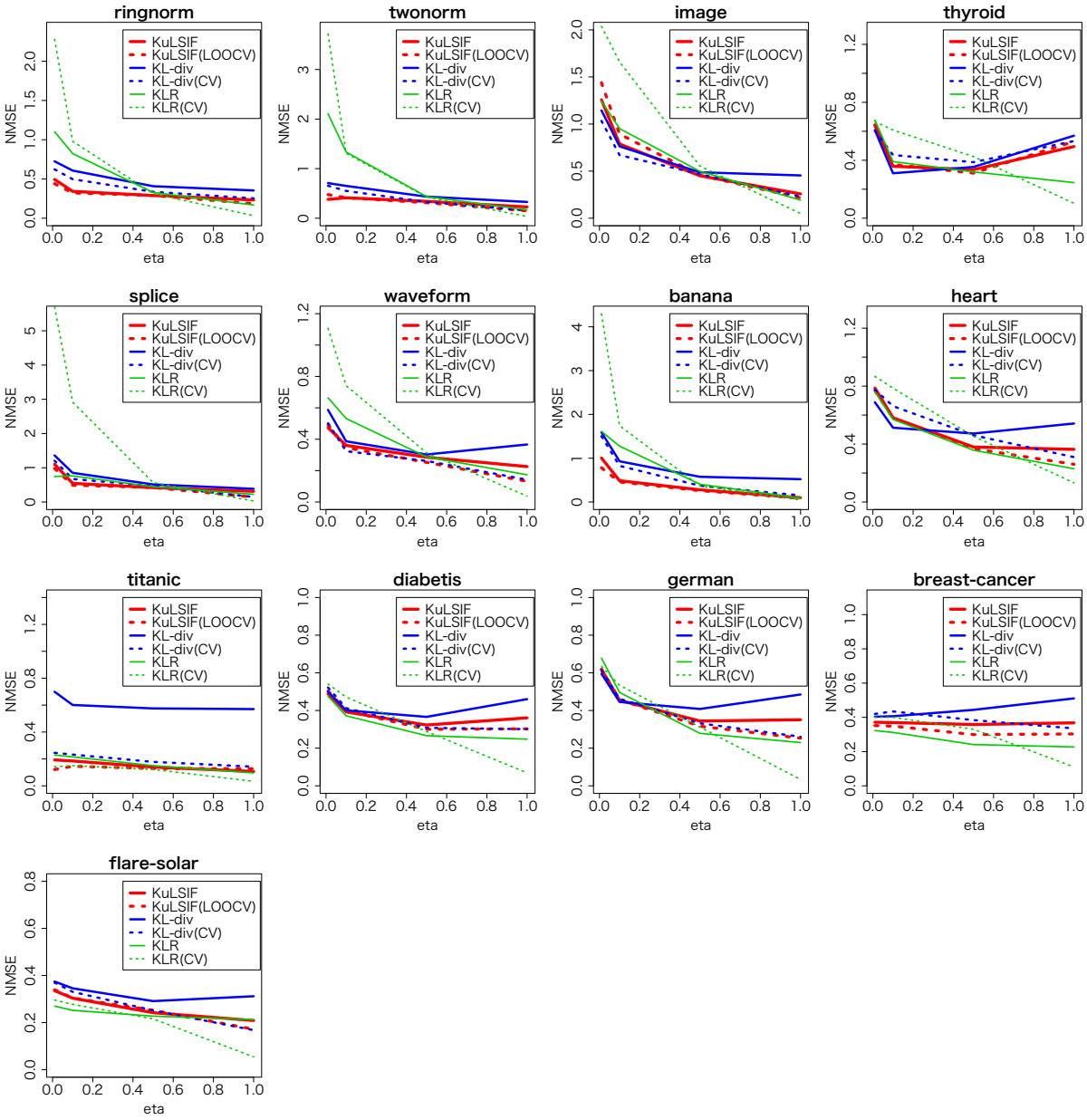


Figure 4: The NMSEs are depicted as functions of η.

obtained by (4) will be a non-negative function without additional constraints. Thus, the estimator $\widehat{w}$ (and its cut-off version $\widehat{w}_+$) will be asymptotically the same as the one obtained by imposing the nonnegative constraint on (4). For the small sample size, however, the estimator $\widehat{w}$ can take negative values, and the cut-off estimator $\widehat{w}_+$ may have statistical bias. Thus, we need careful treatment to obtain a good estimator in practice. In nonparametric density estimation, it was shown that nonnegative estimators cannot be unbiased (Rosenblatt, 1956). We conjecture that a similar result also holds in the inference of density ratios, which needs to be investigated in our future work.

Acknowledgment

The authors are grateful to the anonymous reviewers for their helpful comments. The work of T. Kanamori was partially supported by Grant-in-Aid for Young Scientists (20700251). T. Suzuki was supported in part by Global COE Program "The research and training center for new development in mathematics", MEXT, Japan. M. Sugiyama was supported by SCAT, AOARD, and the JST PRESTO program.

A Proof of Theorem 1

Proof. Applying the representer theorem (Kimeldorf & Wahba, 1971), we see that an optimal solution of (4) has the form of

$$w = \sum_{j=1}^{n} \alpha_j k(\cdot, X_j) + \sum_{\ell=1}^{m} \beta_\ell k(\cdot, Y_\ell). \tag{21}$$

Let K_{11}, K_{12}, K_{21}, and K_{22} be the sub-matrices of the Gram matrix:

$$(K_{11})_{ii'} = k(X_i, X_{i'}), \quad (K_{12})_{ij} = k(X_i, Y_j), \quad K_{21} = K_{12}^\top, \quad (K_{22})_{jj'} = k(Y_j, Y_{j'}),$$

where $i, i' = 1, \ldots, n,\ j, j' = 1, \ldots, m$. Then, the extremal condition of (4) with respect to parameters $\alpha = (\alpha_1, \ldots, \alpha_n)^\top$ and $\beta = (\beta_1, \ldots, \beta_n)^\top$ is given as

$$\begin{aligned}
&\frac{1}{n} K_{11}(K_{11}\alpha + K_{12}\beta) - \frac{1}{m} K_{12}\mathbf{1}_m + \lambda K_{11}\alpha + \lambda K_{12}\beta = 0, \quad \text{and} \\
&\frac{1}{n} K_{21}(K_{11}\alpha + K_{12}\beta) - \frac{1}{m} K_{22}\mathbf{1}_m + \lambda K_{22}\beta + \lambda K_{21}\alpha = 0.
\end{aligned}$$

An easy computation shows that the above extremal condition is satisfied at the parameter α which is defined as the solution of the linear equation (5) and $\beta = \frac{1}{m\lambda}(1, \ldots, 1)^\top$. □

B Leave-One-Out Cross-Validation of KuLSIF

The procedure to compute the leave-one-out cross-validation score of KuLSIF is presented here. Let $K_{11}^{(\ell)} \in \Re^{(n-1)\times(n-1)}$ and $K_{12}^{(\ell)} = K_{21}^{(\ell)\top} \in \Re^{(n-1)\times(m-1)}$ be the Gram matrices of

samples except x_ℓ and y_ℓ, respectively. According to Theorem 1, the estimated parameters $\widetilde{\alpha}^{(\ell)}$ and $\widetilde{\beta}^{(\ell)}$ of

$$\widehat{w}^{(\ell)}(z) = \sum_{i \neq \ell} \alpha_i k(z, X_i) + \sum_{j \neq \ell} \beta_j k(z, Y_j)$$

is equal to

$$\widetilde{\alpha}^{(\ell)} \;=\; -\frac{1}{(m-1)\lambda}(K_{11}^{(\ell)} + (n-1)\lambda I_{n-1})^{-1} K_{12}^{(\ell)} \mathbf{1}_{m-1}, \quad \widetilde{\beta}^{(\ell)} \;=\; \frac{1}{(m-1)\lambda}\mathbf{1}_{m-1},$$

where I_{n-1} denotes the $(n-1)$ by $(n-1)$ identity matrix. Hence, the parameter $\widetilde{\alpha}^{(\ell)}$ is the solution of the following convex quadratic problem,

$$\min_{\alpha} \; \frac{1}{2}\alpha^\top (K_{11}^{(\ell)} + (n-1)\lambda I_{n-1})\alpha + \frac{1}{(m-1)\lambda}\mathbf{1}_{m-1}^\top K_{21}^{(\ell)}\alpha, \quad \alpha \in \Re^{n-1}.$$

The same solution can be obtained by solving

$$\begin{aligned} \min_{\alpha} \; & \frac{1}{2}\alpha^\top (K_{11} + (n-1)\lambda I_n)\alpha + \frac{1}{(m-1)\lambda}(\mathbf{1}_m - \boldsymbol{e}_{m,\ell})^\top K_{21}\alpha, \\ & \text{s.t.} \;\; \alpha \in \Re^n, \; \alpha_\ell = 0, \end{aligned} \tag{22}$$

where $\boldsymbol{e}_{m,\ell} \in \Re^m$ is the standard unit vector with only the ℓ-th component being 1. The optimal solution of (22) denoted by $\alpha^{(\ell)}$ is equal to

$$\alpha^{(\ell)} = (K_{11} + (n-1)\lambda I_n)^{-1} \left(-\frac{1}{(m-1)\lambda} K_{12}(\mathbf{1}_m - \boldsymbol{e}_{m,\ell}) - c_\ell \boldsymbol{e}_{n,\ell} \right),$$

where c_ℓ is determined so that $\alpha_\ell^{(\ell)} = 0$. The estimator $\widetilde{\alpha}^{(\ell)} \in \Re^{n-1}$ is equal to the $(n-1)$-dimensional vector consisting of $\alpha^{(\ell)}$ except the ℓ-th component, i.e., $\widetilde{\alpha}^{(\ell)} = (\alpha_1^{(\ell)}, \ldots, \alpha_{\ell-1}^{(\ell)}, \alpha_{\ell+1}^{(\ell)}, \ldots, \alpha_n^{(\ell)})^\top$. Let $\widehat{\beta}^{(\ell)}$ be

$$\widehat{\beta}^{(\ell)} = \frac{1}{(m-1)\lambda}(\mathbf{1}_m - \boldsymbol{e}_{m,\ell}),$$

then we have

$$\widehat{w}^{(\ell)}(z) = \sum_{i=1}^{n} \alpha_i^{(\ell)} k(z, X_i) + \sum_{j=1}^{m} \widehat{\beta}_j^{(\ell)} k(z, Y_j).$$

We consider an analytic expression of the leave-one-out score. Let the matrices A and B be the parameters of the leave-one-out estimator,

$$A = (\alpha^{(1)}, \ldots, \alpha^{(n \wedge m)}) \in \Re^{n \times (n \wedge m)}, \quad B = (\beta^{(1)}, \ldots, \beta^{(n \wedge m)}) \in \Re^{m \times (n \wedge m)},$$

the matrix $G \in \Re^{n \times n}$ be $G = (K_{11} + (n-1)\lambda I_n)^{-1}$, and $E \in \Re^{m \times (n \wedge m)}$ be the matrix defined as

$$E_{ij} = \begin{cases} 1 & i \neq j, \\ 0 & i = j. \end{cases}$$

Let $S \in \Re^{n\times(n\wedge m)}$ be

$$S = -\frac{1}{(m-1)\lambda} K_{12} E,$$

and $T \in \Re^{n\times(n\wedge m)}$ be

$$T_{ij} = \begin{cases} \dfrac{(GS)_{ii}}{G_{ii}} & i = j, \\ 0 & i \neq j. \end{cases}$$

Then, we obtain

$$A = G(S - T), \quad B = \frac{1}{(m-1)\lambda} E.$$

Let $K_X \in \Re^{(n\wedge m)\times(n+m)}$ be the sub-matrix of $(K_{11}K_{12})$ formed by the first $n \wedge m$ rows and all columns. Similarly, let $K_Y \in \Re^{(n\wedge m)\times(n+m)}$ be the sub-matrix of $(K_{21}K_{22})$ formed by the first $n \wedge m$ rows and all columns. Let the product $U * U'$ be the element-wise multiplication of matrices U and U' of the same size, i.e., the (i,j) element is given by $U_{ij}U'_{ij}$. Then, we have

$$\begin{aligned}
\widehat{w}_X &= (\widehat{w}^{(1)}(X_1), \ldots, \widehat{w}^{(n\wedge m)}(X_{n\wedge m}))^\top = (K_X * (A^\top \; B^\top))\mathbf{1}_{n+m}, \\
\widehat{w}_Y &= (\widehat{w}^{(1)}(Y_1), \ldots, \widehat{w}^{(n\wedge m)}(Y_{n\wedge m}))^\top = (K_Y * (A^\top \; B^\top))\mathbf{1}_{n+m}, \\
\widehat{w}_{X+} &= (\widehat{w}_+^{(1)}(X_1), \ldots, \widehat{w}_+^{(n\wedge m)}(X_{n\wedge m}))^\top = \max\{\widehat{w}_X, 0\}, \\
\widehat{w}_{Y+} &= (\widehat{w}_+^{(1)}(Y_1), \ldots, \widehat{w}_+^{(n\wedge m)}(Y_{n\wedge m}))^\top = \max\{\widehat{w}_Y, 0\},
\end{aligned}$$

where the max operation for a vector is applied in the element-wise manner. As a result, LOOCV (7) is equal to

$$\text{LOOCV} \;=\; \frac{1}{n\wedge m}\left\{\frac{1}{2}\widehat{w}_{X+}^\top \widehat{w}_{X+} - \mathbf{1}_{n\wedge m}^\top \widehat{w}_{Y+}\right\}.$$

C Proof of Theorem 2

We summarize some notations to be used in the proof. Given a probability distribution P and a random variable $h(X)$, we denote the expectation of $h(X)$ under P by $\int h dP$. Given samples $X_1, \ldots, X_n$ from P, the empirical distribution is denoted by P_n. The expectation $\int h dP_n$ denotes the empirical means of $h(X)$, that is, $\frac{1}{n}\sum_{i=1}^n h(X_i)$. We also use the notation $\int h\, d(P - P_n)$ to represent $\int h dP - \frac{1}{n}\sum_{i=1}^n h(X_i)$. Let $\mathcal{H}$ be the RKHS endowed with the kernel k. The norm and inner product on $\mathcal{H}$ are denoted by $\|\cdot\|_{\mathcal{H}}$ and $\langle\cdot,\cdot\rangle_{\mathcal{H}}$, respectively. Let $\|\cdot\|_\infty$ be the infinity norm, and for distribution function P, define the L_2 norm by

$$\|g\|_P \;=\; \left(\int |g|^2 dP\right)^{1/2},$$

and let $L_2(P)$ be the metric space defined by this distance.

Since $\sup_{x\in\mathcal{Z}} k(x,x)$ is assumed to be bounded above, without loss of generality we assume $\sup_{x\in\mathcal{Z}} k(x,x) \leq 1$. The constant factor of the kernel function does not affect the following proof.

We now define the bracketing entropy of the set of functions. For any fixed $\delta > 0$, a covering for function class $\mathcal{F}$ using the metric $L_2(P)$ is a collection of functions which allows $\mathcal{F}$ to be covered using $L_2(P)$ balls of radius δ centered at these functions. Let $N_B(\delta, \mathcal{F}, P)$ be the smallest number of N for which there exist pairs of functions $\{(g_j^L, g_j^U) \in L_2(P) \times L_2(P) \mid j = 1, \ldots, N\}$ such that $\|g_j^L - g_j^U\|_P \leq \delta$, and such that for each $f \in \mathcal{F}$, there exists j satisfying $g_j^L \leq f \leq g_j^U$. Then, $H_B(\delta, \mathcal{F}, P) = \log N_B(\delta, \mathcal{F}, P)$ is called the *bracketing entropy* of $\mathcal{F}$ (van de Geer, 2000, Definition 2.2).

For $w \in \mathcal{H}$, we have $\|w\|_P \leq \|w\|_\infty \leq \|w\|_{\mathcal{H}}$, because for any $x \in \mathcal{Z}$, the inequalities

$$|w(x)| = |\langle w, k(\cdot, x)\rangle_{\mathcal{H}}| \leq \|w\|_{\mathcal{H}} \sup_x \sqrt{k(x,x)} \leq \|w\|_{\mathcal{H}}$$

hold. Let $\mathcal{G} = \{v^2 \mid v \in \mathcal{H}\}$ and we define a measure of complexity $J : \mathcal{G} \to \Re$ by

$$J(g) = \inf\{\, \|v\|_{\mathcal{H}}^2 \mid g = v^2,\, v \in \mathcal{H}\,\}.$$

Let $\mathcal{H}_M$ and $\mathcal{G}_M$ be

$$\begin{aligned} \mathcal{H}_M &= \{v \in \mathcal{H} \mid \|v\|_{\mathcal{H}} < M\}, \\ \mathcal{G}_M &= \{v^2 \mid v \in \mathcal{H}_{\sqrt{M}}\} = \{g \in \mathcal{G} \mid J(g) < M\}. \end{aligned} \tag{23}$$

It is straightforward to verify the second equality of (23).

The following proposition is crucial to prove the convergence property of KuLSIF.

Proposition 1 (Lemma 5.14 in van de Geer (2000))**.** *Let $\mathcal{F} \subset L_2(P)$ be a function class, and the map $I(f)$ be a measure of complexity of $f \in \mathcal{F}$, where I is a non-negative functional on $\mathcal{F}$ and $I(f_0) < \infty$ for a fixed $f_0 \in \mathcal{F}$. We now define $\mathcal{F}_M = \{f \in \mathcal{F} \mid I(f) < M\}$ satisfying $\mathcal{F} = \cup_{M\geq 1}\mathcal{F}_M$. Suppose that there exist $c_0 > 0$ and $0 < \gamma < 2$ such that*

$$\sup_{f\in\mathcal{F}_M} \|f - f_0\|_P \leq c_0 M, \qquad \sup_{\substack{f\in\mathcal{F}_M \\ \|f-f_0\|_P\leq\delta}} \|f - f_0\|_\infty \leq c_0 M, \quad \text{for all } \delta > 0,$$

and that $H_B(\delta, \mathcal{F}_M, P) = O\left(M/\delta\right)^\gamma$. Then, we have

$$\sup_{f\in\mathcal{F}} \frac{\left|\int (f - f_0) d(P - P_n)\right|}{D(f)} = O_p(1), \quad (n \to \infty),$$

where $D(f)$ is defined by

$$D(f) = \frac{\|f - f_0\|_P^{1-\gamma/2} I(f)^{\gamma/2}}{\sqrt{n}} \vee \frac{I(f)}{n^{2/(2+\gamma)}}$$

and $a \vee b$ denotes $\max\{a, b\}$.

In van de Geer (2000), the probabilistic order is evaluated for each case of $\|f - f_0\| \leq n^{-1/(2+\gamma)} I(f)$ and $\|f - f_0\| > n^{-1/(2+\gamma)} I(f)$, respectively. When the supremum is taken over $\{f \in \mathcal{F} \mid \|f - f_0\| \leq n^{-1/(2+\gamma)} I(f)\}$, $D(f)$ is equal to $I(f)/n^{2/(2+\gamma)}$, and the probabilistic order above is obtained from the first formula of Lemma 5.14 (van de Geer, 2000). In the same way, we obtain the probabilistic order for $\|f - f_0\| > n^{-1/(2+\gamma)} I(f)$. The sum of the probabilistic upper bounds for these two cases provides the result in the above proposition.

We use Proposition 1 to derive an upper bound of

$$\int (\widehat{w} - w_0) d(Q - Q_m), \quad \text{and} \quad \int (\widehat{w}^2 - w_0^2) d(P - P_n).$$

Lemma 1. *The bracketing entropy of $\mathcal{G}_M$ is bounded above by*

$$H_B(\delta, \mathcal{G}_M, P) \;=\; O\left(\frac{M}{\delta}\right)^{\gamma}.$$

Proof. Let $v_1^L, v_1^U, v_2^L, v_2^U, \ldots, v_N^L, v_N^U \in L_2(P)$ be coverings of $\mathcal{H}_{\sqrt{M}}$ in the sense of bracketing, such that $\|v_i^L - v_i^U\|_P \leq \delta$ holds for $i = 1, \ldots, N$. Then, for any $v \in \mathcal{H}_{\sqrt{M}}$ there exists i such that $v_i^L \leq v \leq v_i^U$ holds. We can choose these functions such that $\|v_i^{L(U)}\|_\infty \leq \sqrt{M}$ is satisfied for all $i = 1, \ldots, N$, since for any $v \in \mathcal{H}_{\sqrt{M}}$, the inequality $\|v\|_\infty \leq \|v\|_{\mathcal{H}} < \sqrt{M}$ holds. For example, replace $v_i^{L(U)}$ with $\min\{\sqrt{M}, \max\{-\sqrt{M}, v_i^{L(U)}\}\} \in L_2(P)$. Let $\bar{v}_i^L$ and $\bar{v}_i^U$ be

$$\bar{v}_i^L(x) = \begin{cases} (v_i^L(x))^2, & v_i^L(x) \geq 0, \\ (v_i^U(x))^2, & v_i^U(x) \leq 0, \\ 0, & v_i^L(x) < 0 < v_i^U(x), \end{cases}$$
$$\bar{v}_i^U = \max\{(v_i^L)^2, (v_i^U)^2\},$$

for $i = 1, \ldots, N$. Then, $\bar{v}_i^L \leq \bar{v}_i^U$ holds. Moreover, for any $v \in \mathcal{H}_{\sqrt{M}}$ satisfying $v_i^L \leq v \leq v_i^U$, we have $\bar{v}_i^L \leq v^2 \leq \bar{v}_i^U$. By definition, we also have

$$\begin{aligned} 0 \;\leq\; \bar{v}_i^U(x) - \bar{v}_i^L(x) \;&\leq\; \max\{|v_i^U(x)^2 - v_i^L(x)^2|,\ |v_i^U(x) - v_i^L(x)|^2\} \\ &\leq\; (|v_i^U(x)| + |v_i^L(x)|) \cdot |v_i^U(x) - v_i^L(x)| \;\leq\; 2\sqrt{M} |v_i^U(x) - v_i^L(x)|, \end{aligned}$$

and thus, $\|\bar{v}_i^U - \bar{v}_i^L\|_P \leq 2\sqrt{M} \|v_i^U - v_i^L\|_P$ holds. Due to (8), we obtain

$$H_B(2\sqrt{M}\delta, \mathcal{G}_M, P) \leq H_B(\delta, \mathcal{H}_{\sqrt{M}}, P) = O\left(\frac{\sqrt{M}}{\delta}\right)^{\gamma}.$$

Hence, $H_B(\delta, \mathcal{G}_M, P) = O\left(M/\delta\right)^{\gamma}$ holds. □

Lemma 2. *Assume the condition of Theorem 2. Then, for the KuLSIF estimator $\widehat{w}$, we have*

$$\left|\int(\widehat{w}-w_0)d(Q-Q_m)\right| \;=\; O_p\left(\frac{\|w_0-\widehat{w}\|_P^{1-\gamma/2}\|\widehat{w}\|_{\mathcal{H}}^{\gamma/2}}{\sqrt{m}}\vee\frac{\|\widehat{w}\|_{\mathcal{H}}}{m^{2/(2+\gamma)}}\right),$$
$$\left|\int(\widehat{w}^2-w_0^2)d(P-P_n)\right| \;=\; O_p\left(\frac{\|\widehat{w}-w_0\|_P^{1-\gamma/2}(1+\|\widehat{w}\|_{\mathcal{H}})^{1+\gamma/2}}{\sqrt{n}}\vee\frac{\|\widehat{w}\|_{\mathcal{H}}^2}{n^{2/(2+\gamma)}}\right).$$

Proof. There exists $c_0>0$ such that

$$\sup_{w\in\mathcal{H}_M}\|w-w_0\|_P \;\le\; c_0M,\qquad \sup_{\substack{w\in\mathcal{H}_M\\ \|w-w_0\|_P\le\delta}}\|w-w_0\|_\infty \;\le\; c_0M, \tag{24}$$
$$\sup_{g\in\mathcal{G}_M}\|g-w_0^2\|_P \;\le\; c_0M,\qquad \sup_{\substack{g\in\mathcal{G}_M\\ \|g-w_0^2\|_P\le\delta}}\|g-w_0^2\|_\infty \;\le\; c_0M. \tag{25}$$

The inequalities in (25) are derived as follows. For $g\in\mathcal{G}_M$, there exists $v\in\mathcal{H}$ such that $v^2=g$ and $\|v\|_{\mathcal{H}}^2<M$, and then, we have

$$\begin{aligned}\|g-w_0^2\|_P \;&\le\; \|g-w_0^2\|_\infty \;\le\; \|v\|_\infty^2+\|w_0\|_\infty^2\\ &\le\; \|v\|_{\mathcal{H}}^2+\|w_0\|_\infty^2 \;\le\; M+\|w_0\|_\infty^2 \;\le\; c_0M,\quad (M\ge 1).\end{aligned}$$

In the same way, (24) also holds.

Set $\mathcal{F}$ be $\mathcal{H}$ and $I(w)=\|w\|_{\mathcal{H}}$ in Proposition 1. Taking (24) into account, we have

$$\sup_{w\in\mathcal{H}}\frac{\left|\int(w_0-w)d(Q-Q_m)\right|}{D(w)} \;=\; O_p(1),$$

where $D(w)$ is defined as

$$D(w) \;=\; \frac{\|w_0-w\|_P^{1-\gamma/2}\|w\|_{\mathcal{H}}^{\gamma/2}}{\sqrt{m}}\vee\frac{\|w\|_{\mathcal{H}}}{m^{2/(2+\gamma)}}.$$

In the same way, by setting $\mathcal{F}$ be $\mathcal{G}$ and $I(g)=J(g)$ in Proposition 1, we have

$$\sup_{w\in\mathcal{H}}\frac{\left|\int(w^2-w_0^2)d(P-P_n)\right|}{E(w)} \;=\; O_p(1),$$

where $E(w)$ is defined as

$$E(w) \;=\; \frac{\|w^2-w_0^2\|_P^{1-\gamma/2}J(w^2)^{\gamma/2}}{\sqrt{n}}\vee\frac{J(w^2)}{n^{2/(2+\gamma)}}.$$

Note that $\|w^2 - w_0^2\|_P \leq (\|w_0\|_\infty + \|w\|_{\mathcal{H}})\|w - w_0\|_P = O((1 + \|w\|_{\mathcal{H}})\|w - w_0\|_P)$ and $J(w^2) \leq \|w\|_{\mathcal{H}}^2$. Then, we obtain

$$E(w) \leq \frac{\|w - w_0\|_P^{1-\gamma/2}(1 + \|w\|_{\mathcal{H}})^{1+\gamma/2}}{\sqrt{n}} \vee \frac{\|w\|_{\mathcal{H}}^2}{n^{2/(2+\gamma)}}.$$

□

Now we show the proof of Theorem 2.

Proof. The estimator $\widehat{w}$ satisfies the inequality

$$\frac{1}{2}\int \widehat{w}^2 dP_n - \int \widehat{w} dQ_m + \frac{\lambda}{2}\|\widehat{w}\|_{\mathcal{H}}^2 \leq \frac{1}{2}\int w_0^2 dP_n - \int w_0 dQ_m + \frac{\lambda}{2}\|w_0\|_{\mathcal{H}}^2.$$

Then, we have

$$\begin{aligned}
\frac{1}{2}\|\widehat{w} - w_0\|_P^2 &= \int (w_0 - \widehat{w}) dQ + \frac{1}{2}\int (\widehat{w}^2 - w_0^2) dP \\
&\leq \int (w_0 - \widehat{w}) dQ + \frac{1}{2}\int (\widehat{w}^2 - w_0^2) dP \\
&\quad + \int (\widehat{w} - w_0) dQ_m + \frac{1}{2}\int (w_0^2 - \widehat{w}^2) dP_n + \frac{\lambda}{2}\|w_0\|_{\mathcal{H}}^2 - \frac{\lambda}{2}\|\widehat{w}\|_{\mathcal{H}}^2.
\end{aligned}$$

As a result, we have

$$\begin{aligned}
&\frac{1}{2}\|\widehat{w} - w_0\|_P^2 + \frac{\lambda}{2}\|\widehat{w}\|_{\mathcal{H}}^2 \\
&\leq \left| \int (\widehat{w} - w_0) d(Q - Q_m) \right| + \frac{1}{2}\left| \int (\widehat{w}^2 - w_0^2) d(P - P_n) \right| + \frac{\lambda}{2}\|w_0\|_{\mathcal{H}}^2 \\
&\leq \frac{\lambda}{2}\|w_0\|_{\mathcal{H}}^2 + O_p\left(\frac{\|w_0 - \widehat{w}\|_P^{1-\gamma/2}(1 + \|\widehat{w}\|_{\mathcal{H}})^{1+\gamma/2}}{\sqrt{n \wedge m}} \vee \frac{(1 + \|\widehat{w}\|_{\mathcal{H}})^2}{(n \wedge m)^{2/(2+\gamma)}} \right),
\end{aligned}$$

where Lemma 2 is used.

We need to study three possibilities:

$$\frac{1}{2}\|w_0 - \widehat{w}\|_P^2 + \frac{\lambda}{2}\|\widehat{w}\|_{\mathcal{H}}^2 \leq O_p(\lambda), \tag{26}$$

$$\frac{1}{2}\|w_0 - \widehat{w}\|_P^2 + \frac{\lambda}{2}\|\widehat{w}\|_{\mathcal{H}}^2 \leq O_p\left(\frac{\|w_0 - \widehat{w}\|_P^{1-\gamma/2}(1 + \|\widehat{w}\|_{\mathcal{H}})^{1+\gamma/2}}{\sqrt{n \wedge m}} \right), \tag{27}$$

$$\frac{1}{2}\|w_0 - \widehat{w}\|_P^2 + \frac{\lambda}{2}\|\widehat{w}\|_{\mathcal{H}}^2 \leq O_p\left(\frac{(1 + \|\widehat{w}\|_{\mathcal{H}})^2}{(n \wedge m)^{2/(2+\gamma)}} \right). \tag{28}$$

One of the above inequalities should be satisfied. We study each inequality below.

Case (26): we have

$$\frac{1}{2}\|w_0 - \widehat{w}\|_P^2 \leq O_p(\lambda), \quad \frac{\lambda}{2}\|\widehat{w}\|_{\mathcal{H}}^2 \;\leq\; O_p(\lambda),$$

and hence the inequalities $\|w_0 - \widehat{w}\|_P \leq O_p(\lambda^{1/2})$ and $\|\widehat{w}\|_{\mathcal{H}} \leq O_p(1)$ hold.
Case (27): we have

$$\begin{aligned}
\|w_0 - \widehat{w}\|_P^2 \;&\leq\; O_p\left(\frac{\|w_0 - \widehat{w}\|_P^{1-\gamma/2}(1+\|\widehat{w}\|_{\mathcal{H}})^{1+\gamma/2}}{(n\wedge m)^{1/2}}\right), \\
\lambda\|\widehat{w}\|_{\mathcal{H}}^2 \;&\leq\; O_p\left(\frac{\|w_0 - \widehat{w}\|_P^{1-\gamma/2}(1+\|\widehat{w}\|_{\mathcal{H}})^{1+\gamma/2}}{(n\wedge m)^{1/2}}\right).
\end{aligned}$$

The first inequality provides

$$\|w_0 - \widehat{w}\|_P \;\leq\; O_p\left(\frac{1+\|\widehat{w}\|_{\mathcal{H}}}{(n\wedge m)^{1/(2+\gamma)}}\right).$$

Thus, the second inequality leads to

$$\begin{aligned}
\lambda\|\widehat{w}\|_{\mathcal{H}}^2 \;&\leq\; O_p\left(\frac{\|w_0 - \widehat{w}\|_P^{1-\gamma/2}(1+\|\widehat{w}\|_{\mathcal{H}})^{1+\gamma/2}}{(n\wedge m)^{1/2}}\right) \\
&\leq\; O_p\left(\left(\frac{1+\|\widehat{w}\|_{\mathcal{H}}}{(n\wedge m)^{1/(2+\gamma)}}\right)^{1-\gamma/2}\frac{(1+\|\widehat{w}\|_{\mathcal{H}})^{1+\gamma/2}}{(n\wedge m)^{1/2}}\right) \\
&=\; O_p\left(\frac{(1+\|\widehat{w}\|_{\mathcal{H}})^2}{(n\wedge m)^{2/(2+\gamma)}}\right).
\end{aligned}$$

Hence, we have

$$\|\widehat{w}\|_{\mathcal{H}} \;\leq\; O_p\left(\frac{1}{\lambda^{1/2}(n\wedge m)^{1/(2+\gamma)}}\right) \;=\; o_p(1).$$

Then, we obtain

$$\|w_0 - \widehat{w}\|_P \;\leq\; O_p\left(\frac{1}{(n\wedge m)^{1/(2+\gamma)}}\right) \;\leq\; O_p(\lambda^{1/2}).$$

Case (28): we have

$$\|w_0 - \widehat{w}\|_P^2 \;\leq\; O_p\left(\frac{(1+\|\widehat{w}\|_{\mathcal{H}})^2}{(n\wedge m)^{2/(2+\gamma)}}\right), \qquad \lambda\|\widehat{w}\|_{\mathcal{H}}^2 \;\leq\; O_p\left(\frac{(1+\|\widehat{w}\|_{\mathcal{H}})^2}{(n\wedge m)^{2/(2+\gamma)}}\right).$$

Then, as shown in the case (27), we have $\|\widehat{w}\|_{\mathcal{H}} = o_p(1)$. Hence, we obtain

$$\|w_0 - \widehat{w}\|_P \;\leq\; O_p\left(\frac{1}{(n\wedge m)^{1/(2+\gamma)}}\right) \;\leq\; O_p(\lambda^{1/2}).$$

□

D Proof of (13)

Theorem 3. *Let $\mathcal{H}$ be the reproducing kernel Hilbert space endowed with the kernel function k on $\mathcal{Z} \times \mathcal{Z}$, and suppose $\sup_{x \in \mathcal{Z}} k(x,x) < \infty$. Then, for $w, v \in \mathcal{H}$, the equality*

$$\int wv dP - \int v dQ = \langle \Phi(w), v \rangle_{\mathcal{H}}$$

holds, where $\Phi(w)$ is defined by (11).

Proof. For all $w \in \mathcal{H}$, $\sup_{x \in \mathcal{Z}} |w(x)|$ is bounded. This is because $|w(x)| = |\langle w, k(\cdot, x) \rangle_{\mathcal{H}}| \leq \|w\|_{\mathcal{H}} \sqrt{k(x,x)} < \infty$. For a fixed $w \in \mathcal{H}$, the function $\int wv dP - \int v dQ$ is linear and bounded as the function of $v \in \mathcal{H}$. Indeed, the linearity is clear, and the boundedness is shown by

$$\begin{aligned}
\left| \int wv dP - \int v dQ \right| &\leq \int |v(x)||w(x)| P(dx) + \int |v(y)| Q(dy) \\
&= \int |\langle v, k(\cdot, x) \rangle_{\mathcal{H}}||w(x)| P(dx) + \int |\langle v, k(\cdot, y) \rangle_{\mathcal{H}}| Q(dy) \\
&\leq \sup_{x \in \mathcal{Z}} \sqrt{k(x,x)} \left(\int |w(x)| P(dx) + 1 \right) \|v\|_{\mathcal{H}}.
\end{aligned}$$

Since $|w(x)|$ is bounded, the integral above is finite. Then, by the Riesz representation theorem (Reed & Simon, 1972, Theorem II.4), there exists $\Psi : \mathcal{H} \to \mathcal{H}$ such that

$$\int wv dP - \int v dQ = \langle \Psi(w), v \rangle_{\mathcal{H}}$$

holds for all $w, v \in \mathcal{H}$. For $v = k(\cdot, x_0) \in \mathcal{H}$, we have

$$(\Psi(w))(x_0) = \int k(x_0, x) w(x) P(dx) - \int k(x_0, y) Q(dy),$$

where we used the symmetry of the kernel function. We see that the function Ψ is the same as the function Φ defined by (11). □

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.

Bartlett, P. L., Jordan, M. I., & McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, *101*, 138–156.

Bartlett, P. L., & Tewari, A. (2007). Sparseness vs estimating conditional probabilities: Some asymptotic results. *Journal of Machine Learning Research*, *8*, 775–790.

Bickel, S., Brückner, M., & Scheffer, T. (2007). Discriminative learning for differing training and test distributions. *Proceedings of the 24th International Conference on Machine Learning* (pp. 81–88).

Bickel, S., Brückner, M., & Scheffer, T. (2009). Discriminative learning under covariate shift. *Journal of Machine Learning Research*, *10*, 2137–2155.

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

Cucker, F., & Smale, S. (2002). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, *39*, 1–49.

Golub, G. H., & Loan, C. F. V. (1996). *Matrix computations*. Baltimore, MD: Johns Hopkins University Press.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. J. (2006). A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems 19* (pp. 513–520).

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

Härdle, W., Müller, M., Sperlich, S., & Werwatz, A. (2004). *Nonparametric and semiparametric models*. Springer Series in Statistics. Berlin: Springer.

Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M., & Kanamori, T. (2008). Inlier-based outlier detection via direct density ratio estimation. *Proceedings of IEEE International Conference on Data Mining (ICDM2008)* (pp. 223–232). Pisa, Italy.

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

Huang, J., Smola, A., Gretton, A., Borgwardt, K. M., & Schölkopf, B. (2007). Correcting sample selection bias by unlabeled data. *Advances in Neural Information Processing Systems 19* (pp. 601–608). Cambridge, MA: MIT Press.

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

Kanamori, T., Suzuki, T., & Sugiyama, M. (2011). Kernel-based density ratio estimation: Part II, condition number analysis. *Machine Learning*. submitted.

Kawahara, Y., & Sugiyama, M. (2011). Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining*. to appear.

Keerthi, S. S., Duan, K., Shevade, S. K., & Poo, A. N. (2005). A fast dual algorithm for kernel logistic regression. *Machine Learning*, *61*, 151–165.

Kimeldorf, G. S., & Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, *33*, 82–95.

Luenberger, D., & Ye, Y. (2008). *Linear and nonlinear programming*. Springer.

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.

Park, C. (2009). Convergence rates of generalization errors for margin-based classification. *Journal of Statistical Planning and Inference*, *139*, 2543–2551.

Platt, J. C. (2000). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. *Advances in Large Margin Classifiers*, 61–74.

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

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. (Eds.). (2009). *Dataset shift in machine learning*. Cambridge, MA: MIT Press.

R Development Core Team (2009). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

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

Reed, M., & Simon, B. (1972). *Functional analysis*. New York: Academic Press.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, *27*, 832–837.

Rüping, S. (2003). myklr - kernel logistic regression. University of Dortmund, Department of Computer Science.

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

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

Smola, A., Song, L., & Teo, C. H. (2009). Relative novelty detection. *Twelfth International Conference on Artificial Intelligence and Statistics* (pp. 536–543).

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

Steinwart, I. (2005). Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, *51*, 128–142.

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

Sugiyama, M., Kanamori, T., Suzuki, T., Hido, S., Sese, J., Takeuchi, I., & Wang, L. (2009). A density-ratio framework for statistical data processing. *IPSJ Transactions on Computer Vision and Applications*, *1*, 183–208.

Sugiyama, M., & Kawanabe, M. (2011). *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. Cambridge, MA, USA: MIT Press. to appear.

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 & Decisions*, *23*, 249–279.

Sugiyama, M., Nakajima, S., Kashima, H., von Bünau, P., & Kawanabe, M. (2008a). Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in Neural Information Processing Systems 20* (pp. 1433–1440). Cambridge, MA: MIT Press.

Sugiyama, M., Suzuki, T., & Kanamori, T. (2012). *Density ratio estimation in machine learning*. Cambridge, UK: Cambridge University Press. to appear.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., & Kawanabe, M. (2008b). 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*, 583–594.

Suzuki, T., Sugiyama, M., Sese, J., & Kanamori, T. (2008). Approximating mutual information by maximum likelihood density ratio estimation. *JMLR Workshop and Conference Proceedings* (pp. 5–20).

Suzuki, T., Sugiyama, M., & Tanaka, T. (2009). Mutual information approximation via maximum likelihood estimation of density ratio. *Proceedings of 2009 IEEE International Symposium on Information Theory (ISIT2009)* (pp. 463–467). Seoul, Korea.

Tsuboi, Y., Kashima, H., Hido, S., Bickel, S., & Sugiyama, M. (2008). Direct density ratio estimation for large-scale covariate shift adaptation. *SDM* (pp. 443–454).

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

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

Wahba, G., Gu, C., & Y. (1993). Soft classification, a.k.a. risk estimation, via penalized log likelihood and smoothing spline analysis of variance. *The Mathematics of Generalization*. Addison-Wesley.

Yamada, M., Suzuki, T., Kanamori, T., Hachiya, H., & Sugiyama, M. (2011). Relative density-ratio estimation for robust distribution comparison. *Advances in Neural Information Processing Systems 24*. to appear.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. *Proceedings of the Twenty-First International Conference on Machine Learning*. New York, NY: ACM Press.

Zeidler, E. (1986). *Nonlinear functional analysis and its applications, I: Fixed-point theorems*. Springer-Verlag.

Zhu, J., & Hastie, T. (2001). Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics* (pp. 1081–1088). MIT Press.