Computational Complexity of Kernel-Based Density-Ratio Estimation: A Condition Number Analysis

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

In this study, the computational properties of a kernel-based least-squares density-ratio estimator are investigated from the viewpoint of *condition numbers*. The condition number of the Hessian matrix of the loss function is closely related to the convergence rate of optimization and the numerical stability. We use smoothed analysis techniques and theoretically demonstrate that the kernel least-squares method has a smaller condition number than other M-estimators. This implies that the kernel least-squares method has desirable computational properties. In addition, an alternate formulation of the kernel least-squares estimator that possesses an even smaller condition number is presented. The validity of the theoretical analysis is verified through numerical experiments.

1 Introduction

In this section, we introduce background materials of our target problem addressed in this study.

1.1 Density-Ratio Estimation

Recently, methods of directly estimating the ratio of two probability densities without going through density estimation have been developed. These methods can be used to solve various machine learning tasks such as importance sampling, divergence estimation, mutual information estimation, and conditional probability estimation (Sugiyama et al., 2009; Sugiyama et al., 2012).

The *kernel mean matching* (KMM) method (Gretton et al., 2009) directly yields density ratio estimates by efficiently matching the two distributions using a special property of the universal reproducing kernel Hilbert spaces (RKHSs) (Steinwart, 2001). Another approach is the M-estimator (Nguyen et al., 2010), which is based on the non-asymptotic variational characterization of the φ-divergence (Ali & Silvey, 1966; Csiszár, 1967). See Sugiyama et al. (2008a) for a similar algorithm that uses the Kullback-Leibler divergence. Non-parametric convergence properties of the M-estimator in RKHSs have been elucidated under the Kullback-Leibler divergence (Nguyen et al., 2010; Sugiyama et al., 2008b). A squared-loss version of the M-estimator for linear density-ratio models called *unconstrained Least-Squares Importance Fitting* (uLSIF) has also been developed (Kanamori et al., 2009). The squared-loss version was also shown to possess useful computational properties, e.g., a closed-form solution is available, and the leave-one-out cross-validation score can be computed analytically. A kernelized variant of uLSIF was recently proposed, and its statistical consistency was studied (Kanamori et al., 2012).

In this paper, we study loss functions of M-estimators. In Nguyen et al. (2010), a general framework of the density-ratio estimation has been established (also see Sugiyama et al., 2011b). However, when we estimate the density ratio for real-world data analysis, it becomes necessary to choose an M-estimator from infinitely many candidates. Hence it is important to study which M-estimator should be chosen in practice. The suitability of the estimator depends on the chosen criterion. In learning problems, there are mainly two criteria for choosing the estimator: 1) the estimation accuracy and 2) the computational cost. Kanamori et al. (2012) studied the choice of loss functions in density-ratio estimation from the viewpoint of the estimation accuracy. In the present paper, we focus on the computational cost associated with density-ratio estimators.

1.2 Condition Numbers

In numerical analysis, the computational cost is closely related to the so-called *condition number* (von Neumann & Goldstine, 1947; Turing, 1948; Eckart & Young, 1936). Indeed, the condition number appears as a parameter in complexity bounds for a variety of efficient iterative algorithms in linear algebra, linear and convex optimization, and homotopy

methods for solving systems of polynomial equations (Luenberger & Ye, 2008; Nocedal & Wright, 1999; Renegar, 1995; Renegar, 1987; Smale, 1981; Demmel, 1997).

The definition of the condition number depends on the problem. In computational tasks involving matrix manipulations, a typical definition of the condition number is the ratio of the maximum and minimum singular values of the matrix given as the input of the problem under consideration. For example, consider solving the linear equation $Ax = b$. The input of the problem is the matrix A, and the computational cost to find the solution can be evaluated by the condition number of A, denoted hereafter as $\kappa(A)$. hereafter. Specifically, when an iterative algorithm is applied to solving $Ax = b$, the number of iterations required to converge to a solution is evaluated using $\kappa(A)$. In general, a problem with a larger condition number results in a higher computational cost. Since the condition number is independent of the algorithm, it is expected to represent the essential difficulty of the problem.

To evaluate the efficiency of numerical algorithms, a two-stage approach is frequently used: In the first stage, the relation between the computational cost $c(A)$ of an algorithm with input A and the condition number $\kappa(A)$ of the problem is studied. A formula such as $c(A) = O(\kappa(A)^\alpha)$ is obtained, where α is a constant depending on the algorithm. At the second stage, the probability distribution of $\kappa(A)$ is estimated, for example, in the form of $\Pr(\kappa(A) \geq x) \leq x^{-\beta}$, where the probability is designed to represent a "practical" input distribution. As a result, the average computational cost of the algorithm can be evaluated. For details of this approach, see Blum and Shub (1986); Renegar (1987); Demmel (1988); Kostlan (1988); Edelman (1988); Edelman (1992); Shub (1993); Shub and Smale (1994); Shub and Smale (1996); Cheung and Cucker (2002); Cucker and Wschebor (2002); Beltran and Pardo (2006); Bürgisser et al. (2010).

1.3 Smoothed Analysis

The "average" performance is often controversial, because it is hard to identify the input probability distribution in real-world problems. Spielman and Teng (2004) proposed the *smoothed analysis* to refine the second stage of the above scheme for obtaining more meaningful probabilistic upper complexity bounds. Smoothed analysis is a hybrid of the worst and average-case analyses. Consider the averaged computational cost $\mathrm{E}_P[c(A)]$, where $c(A)$ is the cost of an algorithm for input A and $\mathrm{E}_P[\,\cdot\,]$ denotes the expectation with respect to the probability P over the input space. Let $\mathcal{P}$ be a set of probability distributions on the input space. Then, in the smoothed analysis, the performance of the algorithm is measured by $\max_{P\in\mathcal{P}} \mathrm{E}_P[c(A)]$, i.e., the worst-case evaluation of the expected computational cost over a set of probability distributions.

The smoothed analysis was successfully employed in understanding the practical efficiency of the simplex algorithm for linear programming problems (Spielman & Teng, 2004; Bürgisser et al., 2006a). In the context of machine learning, the smoothed analysis was applied to elucidate the complexity of learning algorithms such as the perceptron algorithm and the k-means method; see Vershynin (2006); Blum and Dunagan (2002); Becchetti et al. (2006); Röglin and Vöcking (2007); Manthey and Röglin (2009); Bürgisser et al.

(2006b); Bürgisser and Cucker (2010); Sankar et al. (2006) for more applications of the smoothed analysis technique.

The concept of the smoothed analysis, i.e., the worst-case evaluation of the expected computational cost over a set of probability distributions, is compatible with many problem setups in machine learning and statistics. A typical assumption in statistical inference is that training samples are distributed according to a probability distribution in a probability distribution set. The probability distribution may be specified by a finite-dimensional parameter, or an infinite-dimensional space may be introduced to deal with a probability distribution set.

1.4 Our Contributions

In this study, we apply the concept of smoothed analysis for studying the computational cost of density-ratio estimation algorithms. In our analysis, we define the probability distribution on the basis of training samples, and study the optimal choice of the loss functions for M-estimators.

More specifically, we consider the optimization problems associated with the M-estimators. There are some definitions of condition numbers to measure the complexity of optimization problems (Bürgisser et al., 2006c; Renegar, 1995; Todd et al., 2001). In unconstrained non-linear optimization problems, the condition number defined from the Hessian matrix of the loss function plays a crucial role, because it determines the convergence rate of optimization and the numerical stability (Luenberger & Ye, 2008; Nocedal & Wright, 1999). When a loss function to be optimized depends on random samples, the computational cost will be affected by the distribution of the condition number. Therefore, we study the distribution of condition numbers for randomly perturbed matrices. Next, we derive the loss function that has the smallest condition number among all M-estimators in the min-max sense. We also give a probabilistic evaluation of the condition number. Finally, we verify these theoretical findings through numerical experiments.

There are many important aspects to the computational cost of numerical algorithms such as memory requirements, the role of stopping conditions, and the scalability to large data sets. In this study, we evaluate the computational cost and stability of learning problems on the basis of the condition number of the loss function, because the condition number is a major parameter to quantify the difficulty of the numerical computation as explained above.

1.5 Structure of the Paper

The remainder of this paper is structured as follows. In Section 2, we formulate the problem of density-ratio estimation and briefly review existing methods. In Section 3, a kernel-based density-ratio estimator is introduced. Section 4 is the main contribution of this paper, i.e., the presentation of condition number analyses of density-ratio estimation methods. In Section 5, we further investigate the possibility of reducing the condition number of loss functions. In Section 6, we experimentally investigate the behavior of

condition numbers, confirming the validity of our theoretical analysis. In Section 7, we conclude by summarizing our contributions and indicating possible future research directions. Technical details are presented in Appendix.

2 Estimation of Density Ratios

In this section, we formulate the problem of density-ratio estimation and briefly review existing methods.

2.1 Formulation and Notations

Consider two probability distributions P and Q on a probability space $\mathcal{Z}$. Let 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 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. 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.

2.2 M-Estimator Based on φ-Divergence

An estimator of the density ratio based on the φ-divergence (Ali & Silvey, 1966; Csiszár, 1967) has been proposed by Nguyen et al. (2010). Let $\varphi : \Re \to \Re$ be a convex function, and suppose that $\varphi(1) = 0$. Then, the φ-divergence between P and Q is defined by the integral

$$I(P, Q) = \int \varphi(q/p) dP.$$

Setting $\varphi(z) = -\log z$, we obtain the Kullback-Leibler divergence as an example of the φ-divergence. Let ψ be the conjugate dual function of φ, i.e.,

$$\psi(z) = \sup_{u \in \Re} \{zu - \varphi(u)\} = -\inf_{u \in \Re} \{\varphi(u) - zu\}.$$

When φ is a convex function, we also have

$$\varphi(z) = -\inf_{u\in\Re}\{\psi(u) - zu\}. \tag{2}$$

We assume ψ is differentiable. See Section 12 and 26 of Rockafellar (1970) for details on the conjugate dual function. Substituting (2) into the φ-divergence, we obtain the expression,

$$I(P,Q) = -\inf_{w}\left[\int\psi(w)dP - \int wdQ\right], \tag{3}$$

where the infimum is taken over all measurable functions $w : \mathcal{Z} \to \Re$. The infimum is attained at the function w satisfying

$$\frac{q(x)}{p(x)} = \psi'(w(x)), \tag{4}$$

where ψ' is the derivative of ψ.

Approximating (3) with the empirical distribution, we obtain the empirical loss function,

$$\inf_{w}\ \frac{1}{n}\sum_{i=1}^{n}\psi(w(X_i)) - \frac{1}{m}\sum_{j=1}^{m}w(Y_j).$$

A parametric or non-parametric model is assumed for the function w. This estimator is referred to as the *M-estimator* of the density ratio (Nguyen et al., 2010). The M-estimator based on the Kullback-Leibler divergence is derived from $\psi(z) = -1 - \log(-z)$. Sugiyama et al. (2008a) have studied the estimator in detail using the Kullback-Leibler divergence, and proposed a practical method that includes basis function selection by cross-validation. Kanamori et al. (2009) proposed *unconstrained Least-Squares Importance Fitting* (uLSIF) which is derived from the quadratic function $\psi(z) = z^2/2$.

3 Kernel-Based M-Estimator

In this study, we consider kernel-based estimators of density ratios because the kernel methods provide a powerful and unified framework for statistical inference (Schölkopf & Smola, 2002). Let $\mathcal{H}$ be an RKHS endowed with the kernel function k defined on $\mathcal{Z}\times\mathcal{Z}$. Then, based on (3), we minimize the following loss function over $\mathcal{H}$.

$$\inf_{w}\ \frac{1}{n}\sum_{i=1}^{n}\psi(w(X_i)) - \frac{1}{m}\sum_{j=1}^{m}w(Y_j) + \frac{\lambda}{2}\|w\|_{\mathcal{H}}^2, \quad w\in\mathcal{H}, \tag{5}$$

where the regularization term $\frac{\lambda}{2}\|w\|_{\mathcal{H}}^2$ with the regularization parameter λ is introduced to avoid overfitting. Then, an estimator of the density ratio w_0 is given by $\psi'(\widehat{w}(x))$,

where $\widehat{w}$ is the minimizer of (5). Statistical convergence properties of the kernel estimator using the Kullback-Leibler divergence have been investigated in Nguyen et al. (2010) and Sugiyama et al. (2008b), and similar analysis for the squared-loss was given in Kanamori et al. (2012).

In the RKHS $\mathcal{H}$, the representer theorem (Kimeldorf & Wahba, 1971) is applicable, and the optimization problem on $\mathcal{H}$ is reduced to a finite-dimensional optimization problem. A detailed analysis leads us to a specific form of the solution as follows.

Lemma 1. *Suppose the samples* (1) *are observed and assume that the function* ψ *in* (5) *is a differentiable convex function, and that* $\lambda > 0$. *Let* $v(\alpha, \beta) \in \Re^n$ *be the vector-valued function defined by*

$$v(\alpha,\beta)_i = \psi'\bigg(\sum_{j=1}^{n}\alpha_j k(X_i, X_j) + \sum_{\ell=1}^{m}\beta_\ell k(X_i, Y_\ell)\bigg), \quad i = 1,\dots,n,$$

for $\alpha \in \Re^n$ *and* $\beta \in \Re^m$, *where* ψ' *denotes the derivative of* ψ. *Let* $\mathbf{1}_m = (1,\dots,1)^\top \in \Re^m$ *for a positive integer* m *and suppose that there exists a vector* $\bar{\alpha} = (\bar{\alpha}_1,\dots,\bar{\alpha}_n) \in \Re^n$ *such that*

$$\frac{1}{n}v(\bar{\alpha}, \mathbf{1}_m/m\lambda) + \lambda\bar{\alpha} = 0. \tag{6}$$

Then, the estimator $\widehat{w}$, *an optimal solution of* (5), *has the form*

$$\widehat{w}(z) = \sum_{i=1}^{n}\bar{\alpha}_i k(z, X_i) + \frac{1}{m\lambda}\sum_{j=1}^{m}k(z, Y_j). \tag{7}$$

The proof is deferred to Appendix A, which can be regarded as an extension of the proof for the least-squares estimator (Kanamori et al., 2012) to general M-estimators. This theorem implies that it is sufficient to find n variables $\bar{\alpha}_1,\dots,\bar{\alpha}_n$ to obtain the estimator $\widehat{w}$.

Using Lemma 1, we can obtain the estimator based on the φ-divergence by solving the following optimization problem

$$\begin{aligned}&\inf_{w}\ \frac{1}{n}\sum_{i=1}^{n}\psi(w(X_i)) - \frac{1}{m}\sum_{j=1}^{m}w(Y_j) + \frac{\lambda}{2}\|w\|_{\mathcal{H}}^2,\\ &\quad \text{s.t.}\quad w(\cdot) = \sum_{i=1}^{n}\alpha_i k(\cdot, X_i) + \frac{1}{m\lambda}\sum_{j=1}^{m}k(\cdot, Y_j), \quad \alpha_1,\dots,\alpha_n \in \Re.\end{aligned} \tag{8}$$

Though the problem (8) is a constrained optimization problem with respect to the parameter $\alpha = (\alpha_1,\dots,\alpha_n)^\top$, it can be easily rewritten as an unconstrained one. In this paper, our main concern is to study which ψ we should use as the loss function of the M-estimator. In Section 4 and 5, we will show that the quadratic function is a preferable choice from a computational efficiency viewpoint.

Consider the condition (6) for the quadratic function, $\psi(z) = z^2/2$. 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, for the quadratic loss $\psi(z) = z^2/2$, we have

$$v(\alpha, \beta) = K_{11}\alpha + K_{12}\beta,$$

and thus, there exists a vector $\bar{\alpha}$ that satisfies the equation (6). For $\psi(z) = z^2/2$, the problem (8) is reduced to

$$\min_{\alpha} \ \frac{1}{2}\alpha^\top \left(\frac{1}{n}K_{11}^2 + \lambda K_{11}\right)\alpha + \frac{1}{nm\lambda}\mathbf{1}_m^\top K_{21}K_{11}\alpha, \quad \alpha \in \Re^n, \tag{9}$$

by ignoring the term that is independent of the parameter α. The density-ratio estimator obtained by solving (9) is referred to as the *kernelized uLSIF* (KuLSIF) (Kanamori et al., 2012).

When the matrix K_{11} is non-degenerate, the optimal solution of (9) is equal to

$$\begin{aligned} -\frac{1}{nm\lambda}\left(\frac{1}{n}K_{11}^2 + \lambda K_{11}\right)^{-1} K_{11}K_{12}\mathbf{1}_m &= -\frac{1}{nm\lambda}\left(\frac{1}{n}K_{11} + \lambda I_n\right)^{-1} K_{11}^{-1}K_{11}K_{12}\mathbf{1}_m \\ &= -\frac{1}{nm\lambda}\left(\frac{1}{n}K_{11} + \lambda I_n\right)^{-1} K_{12}\mathbf{1}_m. \end{aligned} \tag{10}$$

It is straightforward to confirm that the optimal solution of the problem

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

is the same as (10). The estimator given by solving the optimization problem (11) is denoted by *Reduced-KuLSIF* (R-KuLSIF). Though the objective functions in KuLSIF and R-KuLSIF are different, the optimal solution is the same. In Section 5, we show that R-KuLSIF is more preferable than the other M-estimators (including KuLSIF) from a numerical computation viewpoint.

4 Condition Number Analysis for Density-Ratio Estimation

In this section, we study the condition number of loss functions for density-ratio estimation. Through the analysis of condition numbers, we elucidate the computational efficiency of the M-estimator, which is the main contribution of this study.

4.1 Condition Number in Numerical Analysis and Optimization

The condition number plays a crucial role in numerical analysis and optimization (Demmel, 1997; Luenberger & Ye, 2008; Sankar et al., 2006). The main concepts are briefly reviewed here.

Let A be a symmetric positive definite matrix. Then, the condition number of A is defined by $\lambda_{\max}/\lambda_{\min}\,(\geq 1)$, where $\lambda_{\max}$ and $\lambda_{\min}$ are the maximum and minimum eigenvalues of A, respectively[1]. The condition number of A is denoted by $\kappa(A)$.

In numerical analysis, the condition number governs the round-off error of the solution of a linear equation $Ax = b$. A matrix A with a large condition number results in a large upper bound on the relative error of the solution x. More precisely, in the perturbed linear equation

$$(A + \delta A)(x + \delta x) = b + \delta b,$$

the relative error of the solution is given as (Demmel, 1997, Section 2.2)

$$\frac{\|\delta x\|}{\|x\|} \;\leq\; \frac{\kappa(A)}{1 - \kappa(A)\|\delta A\|/\|A\|}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right),$$

where $\|A\|$ is the *operator norm* for the matrix A defined by

$$\|A\| = \max_{x \in \Re^n \setminus \{0\}} \frac{\|Ax\|}{\|x\|}.$$

Hence, a small condition number is preferred in numerical computation.

In optimization problems, the condition number provides an upper bound of the convergence rate for optimization algorithms. Let us consider a minimization problem $\min_x f(x),\ x \in \Re^n$, where $f : \Re^n \to \Re$ is a differentiable function and let x_0 be a local optimal solution. We consider an iterative algorithm that generates a sequence $\{x_i\}_{i=1}^{\infty}$. Let ∇f be the gradient vector of f. In various iterative algorithms, the sequence is generated in the following form

$$x_{i+1} = x_i - \eta_i\, H_i^{-1} \nabla f(x_i), \quad i = 1, 2, \ldots, \tag{12}$$

where η_i is a non-negative number appropriately determined and H_i is a symmetric positive definite matrix which approximates the Hessian matrix of f at x_0, i.e., $\nabla^2 f(x_0)$. Then, under a mild assumption, the sequence $\{x_i\}_{i=1}^{\infty}$ converges to a local minimizer x_0.

We introduce convergence rates of some optimization methods. According to the 'modified Newton method' theorem (Luenberger & Ye, 2008, Section 10.1), the convergence rate of (12) is given by

$$\|x_k - x_0\| = O\Bigg(\prod_{i=1}^{k} \frac{\kappa_i - 1}{\kappa_i + 1}\Bigg), \tag{13}$$

[1] In general, the condition number for a (possibly non-symmetric) matrix is defined through singular values. However, the above simple definition is sufficient for our purpose.

where κ_i is the condition number of $H_i^{-1/2}(\nabla^2 f(x_0))H_i^{-1/2}$. Though the modified Newton method theorem is shown only for convex quadratic functions (Luenberger & Ye, 2008), the rate-of-convergence behavior is essentially the same for general nonlinear objective functions. In terms of non-quadratic functions, details are presented in Section 8.6 of Luenberger and Ye (2008). Equation (13) implies that the convergence rate of the sequence $\{x_k\}$ is fast if $\kappa_i, i = 1, 2, \ldots$ are small. In the conjugate gradient method, the convergence rate is expressed by (13) with $\sqrt{\kappa(\nabla^2 f(x_0))}$ instead of κ_i (Nocedal & Wright, 1999, Section 5.1). Even in proximal-type methods, the convergence rate is described by a quantity similar to the condition number, when the objective function is strongly convex. See Proposition 3 and 4 in Schmidt et al. (2011) for details.

A pre-conditioning technique is often applied to speed up the convergence rate of the optimization algorithm. The idea behind pre-conditioning is to perform a change of variables $x = S\bar{x}$, where S is an invertible matrix. An iterative algorithm is applied to the function $\bar{f}(\bar{x}) = f(S\bar{x})$ in the coordinate system $\bar{x}$. Then a local optimal solution $\bar{x}_0$ of $\bar{f}(\bar{x})$ is pulled back to $x_0 = S\bar{x}_0$.

The pre-conditioning technique is useful, if the conditioning of $\bar{f}(\bar{x})$ is preferable to $f(x)$. However, in general, there are some difficulties in obtaining a suitable pre-conditioning. Consider the iterative algorithm (12) with $H_i = I$ in the coordinate $\bar{x}$, i.e., $\bar{x}_{i+1} = \bar{x}_i - \eta_i \nabla \bar{f}(\bar{x}_i)$. The Hessian matrix is given as $\nabla^2 \bar{f}(\bar{x}_0) = S^\top \nabla^2 f(x_0) S$. Then, the best change of variables is given by $S = (\nabla^2 f(x_0))^{-1/2}$. This is also confirmed by the fact that the gradient descent method with respect to $\bar{x}$ is represented as $x_{i+1} = x_i - \eta_i S S^\top \nabla f(x_i)$ in the coordinate system x. In this case, there are at least two drawbacks:

1. There is no unified strategy to find a good change of variables $x = S\bar{x}$.

2. Under the best change of variables $S = (\nabla^2 f(x_0))^{-1/2}$, the computation of the variable change can be expensive and unstable, when the condition number of $\nabla^2 f(x_0)$ is large.

Similar drawbacks appear in the conjugate gradient methods (Hager & Zhang, 2006; Nocedal & Wright, 1999).

The first drawback is obvious. To find a good change of variables, it is necessary to estimate the shape of the function f around the local optimal solution x_0 *before* solving the problem. Except for a specific type of problems such as discretized partial differential equations, finding a good change of variables is difficult (Benzi et al., 2011; Axelsson & Neytcheva, 2002; Badia et al., 2009). Though there are some general-purpose pre-conditioners such as the incomplete Cholesky decomposition and banded pre-conditioners, their degree of success varies from problem to problem (Nocedal & Wright, 1999, Chap. 5).

To remedy the second drawback, one can use a matrix S with a moderate condition number. When $\kappa(S)$ is moderate, the computation of the variable change is stable. In the optimization toolbox in MATLAB®, gradient descent methods are implemented by the function `fminunc`. The default method in `fminunc` is the BFGS quasi-Newton method, and the Cholesky factorization of the approximate Hessian is used as the transformation

matrix S at each step of the algorithm. When the modified Cholesky factorization is used, the condition number of S is guaranteed to be bounded from above by some constant C. See Moré and Sorensen (1984) for more details.

When the variable change $x = S\bar{x}$ with a bounded condition number is used, there is a trade-off between the numerical accuracy and convergence rate. The trade-off is summarized as

$$\min_{S:\kappa(S)\leq C} \kappa(S^\top(\nabla^2 f(x_0))S) = \max\left\{ \frac{\kappa(\nabla^2 f(x_0))}{C^2},\, 1 \right\}. \tag{14}$$

The proof of this equality is given in Appendix B. When C in (14) is small, the computation of the variable change is stable. Conversely, the convergence speed will be slow because the right-hand side of (14) is large. Thus, the formula (14) presents the trade-off between the numerical stability and the convergence speed. This implies that the convergence rate and stable computation are not consistent when the condition number of the original problem is large. If $\kappa(\nabla^2 f(x_0))$ is small, however, the right-hand side of (14) will not be too large. In this case, the trade-off is not significant and thus the numerical stability and convergence speed can be consistent.

Therefore, it is preferable that the condition number of the original problem is kept as small as possible, despite the fact that some scaling or pre-conditioning techniques are available. In the following section, we pursue a loss function of the density-ratio estimator whose Hessian matrix has a small condition number.

4.2 Condition Number Analysis of M-Estimators

In this section, we study the condition number of the Hessian matrix associated with the minimization problem in the φ-divergence approach, and show that KuLSIF is optimal among all M-estimators. More specifically, we will provide two kinds of condition numbers analyses: a min-max evaluation (Section 4.2.1) and a probabilistic evaluation (Section 4.2.2).

4.2.1 Min-Max Evaluation

We assume that a universal RKHS $\mathcal{H}$ (Steinwart, 2001) endowed with a kernel function k on a compact set $\mathcal{Z}$ is used for density-ratio estimation. The M-estimator is obtained by solving the problem (8). The Hessian matrix of the loss function (8) is equal to

$$\frac{1}{n} K_{11} D_{\psi,w} K_{11} + \lambda K_{11}, \tag{15}$$

where $D_{\psi,w}$ is the n-by-n diagonal matrix defined as

$$D_{\psi,w} = \begin{pmatrix} \psi''(w(X_1)) & & \\ & \ddots & \\ & & \psi''(w(X_n)) \end{pmatrix}, \tag{16}$$

and ψ'' denotes the second-order derivative of ψ. The condition number of the above Hessian matrix is denoted by $\kappa_0(D_{\psi,w})$:

$$\kappa_0(D_{\psi,w}) = \kappa\left(\frac{1}{n}K_{11}D_{\psi,w}K_{11} + \lambda K_{11}\right).$$

In KuLSIF, the equality $\psi'' = 1$ holds, and thus the condition number is equal to $\kappa_0(I_n)$. Now we analyze the relation between $\kappa_0(I_n)$ and $\kappa_0(D_{\psi,w})$.

Theorem 1 (Min-max Evaluation)**.** *Suppose that $\mathcal{H}$ is a universal RKHS, and that K_{11} is non-singular. Let c be a positive constant. Then, the equality*

$$\inf_{\substack{\psi:\Re\to\Re,\\ \psi''((\psi')^{-1}(1))=c}} \sup_{w\in\mathcal{H}} \kappa_0(D_{\psi,w}) = \kappa_0(cI_n) \tag{17}$$

holds, where the infimum is taken over all convex second-order continuously differentiable functions satisfying $\psi''((\psi')^{-1}(1)) = c$.

The proof is deferred to Appendix C.

Both $\psi(z) = z^2/2$ and $\psi(z) = -1 - \log(-z)$ satisfy the constraint $\psi''((\psi')^{-1}(1)) = 1$, and KuLSIF using $\psi(z) = z^2/2$ minimizes the worst-case condition number, because of the fact that the condition number of KuLSIF does not depend on the optimal solution. Note that, because both sides of (17) depend on the samples $X_1, \ldots, X_n$, KuLSIF achieves the min-max solution for *each* observation.

By introducing the constraint $\psi''((\psi')^{-1}(1)) = c$, the balance between the loss term and the regularization term in the objective function of (8) is adjusted. Suppose that $q(x) = p(x)$, i.e., the density ratio is a constant. Then, according to the equality (4), the optimal $w \in \mathcal{H}$ satisfies $1 = \psi'(w(x))$, if the constant $(\psi')^{-1}(1)$ is included in $\mathcal{H}$. In this case, the diagonal of $D_{\psi,w}$ is equal to $\psi''(w(X_i)) = \psi''((\psi')^{-1}(1)) = c$. Thus, the Hessian matrix (15) is equal to $\frac{c}{n}K_{11}^2 + \lambda K_{11}$, which is independent of ψ as long as ψ satisfies $\psi''((\psi')^{-1}(1)) = c$. Then, the constraint $\psi''((\psi')^{-1}(1)) = c$ adjusts the scaling of the loss term at the constant density ratio. Under the adjustment, the quadratic function $\psi(z) = cz^2/2$ is optimal up to a linear term in the min-max sense.

4.2.2 Probabilistic Evaluation

Next, we present a probabilistic evaluation of condition numbers. As shown in (15), the Hessian matrix at the estimated function $\widehat{w}$ (which is the minimizer of (8)) is given as

$$H = \frac{1}{n}K_{11}D_{\psi,\widehat{w}}K_{11} + \lambda K_{11}.$$

Let us define the random variable T_n as

$$T_n = \max_{1\le i\le n} \psi''(\widehat{w}(X_i)). \tag{18}$$

Since ψ is convex, T_n is a non-negative random variable. Let F_n be the distribution function of T_n. The notations T_n and F_n imply that they depend on n. To be precise, T_n and F_n actually depend on both n and m. Here we suppose that m is fixed to a natural number including infinity, or m is a function of n as $m = m_n$. Then, T_n and F_n depend only on n.

Below, we first compute the distribution of the condition number $\kappa(H)$. Then we investigate the relation between the function ψ and the distribution of the condition number $\kappa(H)$. To this end, we need to study eigenvalues and condition numbers of random matrices. For the Wishart distribution, the probability distribution of condition numbers has been investigated by Edelman (1988) and Edelman and Sutton (2005). Recently, the condition number of matrices perturbed by additive Gaussian noise has been investigated under the name of *smoothed analysis* (Sankar et al., 2006; Spielman & Teng, 2004; Tao & Vu, 2007). However, the statistical property of the above-defined matrix H is more complicated than those studied in the existing literature. In our problem, the probability distribution of each element will be far from well-known, and elements are correlated to each other through the kernel function.

Now, we briefly introduce the core idea of the smoothed analysis (Spielman & Teng, 2004), and discuss its relation with our study. Consider the averaged computational cost $\mathrm{E}_P[c(X)]$, where $c(X)$ is the cost of an algorithm for input X, and $\mathrm{E}_P[\,\cdot\,]$ denotes the expectation with respect to the probability P over the input space. Let $\mathcal{P}$ be a set of probabilities on the input space. In the smoothed analysis, the performance of the algorithm is measured by $\max_{P\in\mathcal{P}} \mathrm{E}_P[c(X)]$. The set of Gaussian distributions is a popular choice for $\mathcal{P}$.

Conversely, in our theoretical analysis, we consider the probabilistic order of condition numbers $O_p(\kappa(H))$, as a measure of computational costs. The worst-case evaluation of the computational complexity is measured by $\max_{P,Q} O_p(\kappa(H))$, where the sample distributions P and Q vary in an appropriate set of distributions. The quantity, $\max_{P,Q} O_p(\kappa(H))$, is the counterpart of the worst-case evaluation of the averaged computational cost $\mathrm{E}_P[c(X)]$ in the smoothed analysis. The probabilistic order of $\kappa(H)$ depends on the loss function ψ. Then, we suggest that the loss function that achieves the optimal solution of the min-max problem, $\min_\psi \max_{P,Q} O_p(\kappa(H))$, is the optimal choice. The details are shown below, where our concern is not only to provide the worst-case computational cost, but also to find the optimal loss function for the M-estimator.

Theorem 2 (Probabilistic Evaluation)**.** *Let $\mathcal{H}$ be an RKHS endowed with a kernel function $k : \mathcal{Z}\times\mathcal{Z} \to \Re$ satisfying the boundedness condition, $\sup_{x,x'\in\mathcal{Z}} k(x,x') < \infty$. Assume that the Gram matrix K_{11} is almost surely positive definite in terms of the probability measure P. Suppose that, for the regularization parameter $\lambda_{n,m}$, the boundedness condition $\limsup_{n\to\infty} \lambda_{n,m} < \infty$ is satisfied. Let $U = \sup_{x,x'\in\mathcal{Z}} k(x,x')$ and t_n be a sequence such that*

$$\lim_{n\to\infty} F_n(t_n/U) = 1, \tag{19}$$

where F_n is the probability distribution of T_n defined in (18)*. Then, we have*

$$\lim_{n\to\infty} \Pr\left(\kappa(H) \leq \kappa(K_{11})\big(1 + \frac{t_n}{\lambda_{n,m}}\big)\right) = 1, \tag{20}$$

where H is defined as $H = \frac{1}{n}K_{11}D_{\psi,\widehat{w}}K_{11} + \lambda K_{11}$. The probability $\Pr(\cdot)$ is defined from the distribution of samples $X_1, \ldots, X_n$, $Y_1, \ldots, Y_m$.

The proof of Theorem 2 is deferred to Appendix D.

Remark 1. *The Gaussian kernel on a compact set $\mathcal{Z}$ meets the condition of Theorem 2 under a mild assumption on the probability P. Suppose that $\mathcal{Z}$ is included in the ball $\{x \in \Re^d \mid \|x\| \leq R\}$. Then, for $k(x, x') = \exp\{-\gamma\|x - x'\|^2\}$ with $x, x' \in \mathcal{Z}$ and $\gamma > 0$, we have $e^{-4\gamma R^2} \leq k(x, x') \leq 1$. If the distribution P of samples $X_1, \ldots, X_n$ is absolutely continuous with respect to the Lebesgue measure, the Gram matrix of the Gaussian kernel is almost surely positive definite because K_{11} is positive definite if $X_i \neq X_j$ for $i \neq j$.*

When ψ is the quadratic function, $\psi(z) = z^2/2$, the distribution function F_n is given by $F_n(t) = \mathbf{1}[t \geq 1]$, where $\mathbf{1}[\,\cdot\,]$ is the indicator function. By choosing $t_n = 1$ in Theorem 2, an upper bound of $\kappa(H)$ for $\psi(z) = z^2/2$ is asymptotically given as $\kappa(K_{11})(1 + \lambda_{n,m}^{-1})$. Conversely, for the M-estimator with the Kullback-Leibler divergence (Nguyen et al., 2010), the function ψ is defined as $\psi(z) = -1 - \log(-z)$, $z < 0$, and thus, $\psi''(z) = 1/z^2$ holds. Then we have $T_n = \max_{1\leq i\leq n}(\widehat{w}(X_i))^{-2}$. Note that there is a possibility that $(\widehat{w}(X_i))^2$ takes a very small value, and thus it is expected that T_n is of a larger than constant order. As a result, t_n would diverge to infinity for $\psi(z) = -1 - \log(-z)$. Results of the above theoretical analysis are confirmed by numerical studies in Section 6.

Using the above argument, we show that the quadratic loss is approximately an optimal loss function in the sense of probabilistic upper bounds in Theorem 2. Suppose that the true density ratio $q(z)/p(z)$ is well approximated by the estimator $\psi'(\widehat{w}(z))$. Instead of T_n, we study an approximation $\sup_{z\in\mathcal{Z}} \psi''((\psi')^{-1}(q(z)/p(z)))$. Then, for any loss function ψ such that $\psi''((\psi')^{-1}(1)) = 1$, the inequality

$$\sup_{p,q}\sup_{z\in\mathcal{Z}} \psi''((\psi')^{-1}(q(z)/p(z))) \geq 1 = \inf_{\psi}\sup_{p,q}\sup_{z\in\mathcal{Z}} \psi''((\psi')^{-1}(q(z)/p(z)))$$

holds, where p and q are probability densities such that $(\psi')^{-1}(q/p) \in \mathcal{H}$. The equality holds for the quadratic loss. The meaning of the constraint $\psi''((\psi')^{-1}(1)) = 1$ is presented in Section 4.2.1. Thus, $t_n = 1$ provided by the quadratic loss function is expected to approximately attain the minimum upper bound in (20). The quantity $\sup_{p,q}\sup_{z\in\mathcal{Z}} \psi''((\psi')^{-1}(q(z)/p(z)))$ is the counterpart of $\max_{P\in\mathcal{P}} E_P[c(X)]$ in the smoothed analysis. We expect that the loss function attaining the infimum of this quantity provides a computationally efficient learning algorithm.

5 Reduction of Condition Numbers

In the previous section, we showed that KuLSIF is preferable in terms of computational efficiency and numerical stability. In this section, we study the reduction of condition numbers.

Let $L_{\mathrm{KuLSIF}}(\alpha)$ and $L_{\text{R-KuLSIF}}(\alpha)$ be loss functions of KuLSIF (9) and R-KuLSIF (11), respectively. The Hessian matrices of $L_{\mathrm{KuLSIF}}(\alpha)$ and $L_{\text{R-KuLSIF}}(\alpha)$ are given by

$$H_{\mathrm{KuLSIF}} = \nabla^2 L_{\mathrm{KuLSIF}}(\alpha) = \frac{1}{n} K_{11}^2 + \lambda K_{11}, \tag{21}$$

$$H_{\text{R-KuLSIF}} = \nabla^2 L_{\text{R-KuLSIF}}(\alpha) = \frac{1}{n} K_{11} + \lambda I_n. \tag{22}$$

Because of the equality $\kappa(H_{\mathrm{KuLSIF}}) = \kappa(K_{11})\kappa(H_{\text{R-KuLSIF}})$, we have the inequality

$$\kappa(H_{\text{R-KuLSIF}}) \ \leq \ \kappa(H_{\mathrm{KuLSIF}}).$$

This inequality implies that the loss function $L_{\mathrm{KuLSIF}}(\alpha)$ can be transformed to $L_{\text{R-KuLSIF}}(\alpha)$ without changing the optimal solution, whereas the condition number is reduced. Hence, R-KuLSIF will be more preferable than KuLSIF in the sense of both convergence speed and numerical stability as explained in Section 4.1. Though the loss function of R-KuLSIF is not a member of the regularized M-estimator (8), KuLSIF can be transformed to R-KuLSIF without any computational effort.

Below, we study whether the same reduction of condition numbers is possible in the general φ-divergence approach. If there are M-estimators other than KuLSIF whose condition numbers are reducible, we should compare them with R-KuLSIF and pursue more computationally efficient density-ratio estimators. Our conclusion is that among all of the φ-divergence approaches, the condition number is reducible only for KuLSIF. Thus, the reduction of condition numbers by R-KuLSIF is a special property that makes R-KuLSIF particularly attractive for practical use.

We now show why the condition number of KuLSIF is reducible from $\kappa(H_{\mathrm{KuLSIF}})$ to $\kappa(H_{\text{R-KuLSIF}})$ without changing the optimal solution. Solving an unconstrained optimization problem is equivalent to finding a zero of the gradient vector of the loss function. For the loss functions $L_{\text{R-KuLSIF}}(\alpha)$ and $L_{\mathrm{KuLSIF}}(\alpha)$, the equality

$$\nabla L_{\text{R-KuLSIF}}(\alpha) = K_{11}^{-1} \nabla L_{\mathrm{KuLSIF}}(\alpha)$$

holds for any α. Hence, for non-degenerate K_{11}, zeros of $\nabla L_{\text{R-KuLSIF}}(\alpha)$ and $\nabla L_{\mathrm{KuLSIF}}(\alpha)$ are the same. In general, for the quadratic convex loss functions $L_1(\alpha)$ and $L_2(\alpha)$ that share the same optimal solution, there exists a matrix C such that $\nabla L_1 = C \nabla L_2$. Indeed, for $L_1(\alpha) = (\alpha - \alpha^*)^\top A_1 (\alpha - \alpha^*)$ and $L_2(\alpha) = (\alpha - \alpha^*)^\top A_2 (\alpha - \alpha^*)$, the matrix $C = A_1 A_2^{-1}$ yields the equality $\nabla L_1 = C \nabla L_2$. Based on this fact, one can obtain the quadratic loss function that shares the same optimal solution with a smaller condition number without further computational cost.

Now, we study loss functions of general M-estimators. Let $L_\psi(\alpha)$ be the loss function of the M-estimator (8), and let $L(\alpha)$ be any other function. Suppose that $\nabla L(\alpha^*) = 0$ holds if and only if $\nabla L_\psi(\alpha^*) = 0$. This implies that extremal points of $L_\psi(\alpha)$ and $L(\alpha)$ are the same. Then, there exists a matrix-valued function $C(\alpha) \in \Re^{n \times n}$ such that

$$\nabla L(\alpha) = C(\alpha) \nabla L_\psi(\alpha), \tag{23}$$

where $C(\alpha)$ is non-degenerate for any α. Suppose $C(\alpha)$ is differentiable. Then, the derivative of the above equation at the extremal point α^* leads to the equality

$$\nabla^2 L(\alpha^*) = C(\alpha^*)\nabla^2 L_\psi(\alpha^*).$$

When $\kappa(\nabla^2 L(\alpha^*)) \leq \kappa(\nabla^2 L_\psi(\alpha^*))$, $L(\alpha)$ will be preferable to $L_\psi(\alpha)$ for numerical computation.

We require a careful treatment for the choice of the matrix $C(\alpha)$ or the loss function $L(\alpha)$. If there is no restriction on the matrix-valued function $C(\alpha)$, the most preferable choice of $C(\alpha^*)$ is given by $C(\alpha^*) = (\nabla^2 L_\psi(\alpha^*))^{-1}$. However this is clearly meaningless for the purpose of numerical computation because the transformation requires the knowledge of the optimal solution. Even if the function $L_\psi(\alpha)$ is quadratic, finding $(\nabla^2 L_\psi(\alpha^*))^{-1}$ is computationally equivalent to solving the optimization problem. To obtain a suitable loss function $L(\alpha)$ without additional computational effort, we need to impose a meaningful constraint on $C(\alpha)$. Below, we assume that the matrix-valued function $C(\alpha)$ is a constant function[2].

As shown in the proof of Lemma 1, the gradient of the loss function $L_\psi(\alpha)$ is equal to

$$\nabla L_\psi(\alpha) = \frac{1}{n} K_{11} v(\alpha, \mathbf{1}_m/m\lambda) + \lambda K_{11}\alpha,$$

where the function v is defined in Lemma 1. Let $C \in \Re^{n\times n}$ be a constant matrix, and suppose that the $\Re^n$-valued function $C\nabla L_\psi(\alpha)$ is represented as the gradient of a function L, i.e., there exists an L such that $\nabla L = C\nabla L_\psi$. Then, the function $C\nabla L_\psi$ is called *integrable* (Nakahara, 2003). We now require a ψ for which there exists a non-identity matrix C such that $C\nabla L_\psi(\alpha)$ is integrable. According to the Poincaré lemma (Nakahara, 2003; Spivak, 1979), the necessary and sufficient condition of integrability is that the Jacobian matrix of $C\nabla L_\psi(\alpha)$ is symmetric. The Jacobian matrix of $C\nabla L_\psi(\alpha)$ is given by

$$J_{\psi,C}(\alpha) = \frac{1}{n} C K_{11} D_{\psi,\alpha} K_{11} + \lambda C K_{11},$$

where $D_{\psi,\alpha}$ is the n-by-n diagonal matrix with diagonal elements

$$(D_{\psi,\alpha})_{ii} = \psi''\left(\sum_{j=1}^{n} \alpha_j k(X_i, X_j) + \frac{1}{m\lambda}\sum_{\ell=1}^{m} k(X_i, Y_\ell)\right), \quad i = 1, \ldots, n.$$

In terms of the Jacobian matrix $J_{\psi,C}(\alpha)$, we have the following theorem.

Theorem 3. *Let c be a constant value in $\Re$ and the function ψ be second-order continuously differentiable. Suppose that the Gram matrix K_{11} is non-singular, and that K_{11} does not have any zero elements. If there exists a non-singular matrix $C \neq cI_n$ such that $J_{\psi,C}(\alpha)$ is symmetric for any $\alpha \in \Re^n$, then, ψ'' is a constant function.*

[2]We must admit that this is a rather strict condition. It is an important future work to investigate the relaxation of the condition in a feasible way.

The proof is provided in Appendix E.

Theorem 3 implies that for the non-quadratic function ψ, the gradient $C\nabla L_\psi(\alpha)$ cannot be integrable unless $C = cI_n,\ c \in \Re$. As a result, the condition number of loss functions is reducible only when ψ is a quadratic function[3]. The same procedure works for kernel ridge regression (Chapelle, 2007; Ratliff & Bagnell, 2007) and kernel PCA (Mika et al., 1999). However, there exists no similar procedure for M-estimators with non-quadratic functions.

In general, the change of variables is a standard and useful approach to reducing the condition number of loss functions. However, we need a good prediction of the Hessian matrix at the optimal solution to obtain good conditioning. Moreover, additional computation including matrix manipulation will be required for the coordinate transformation. Conversely, an advantage of the transformation considered in this section is that it does not require any effort to predict the Hessian matrix or to manipulate the matrix.

Remark 2. *We summarize our theoretical results on condition numbers. Let $H_{\psi\text{-}div}$ be the Hessian matrix of the loss function* (8)*. Then, the following inequalities hold:*

$$\kappa(H_{\text{R-KuLSIF}}) \leq \kappa(H_{\text{KuLSIF}}) = \sup_{w \in \mathcal{H}} \kappa(H_{\text{KuLSIF}}) \leq \sup_{w \in \mathcal{H}} \kappa(H_{\psi\text{-}div}).$$

Based on a probabilistic evaluation, the inequality

$$\kappa(H_{\text{KuLSIF}}) \ \leq \ \kappa(H_{\psi\text{-}div})$$

will also hold with high probability.

6 Simulation Study

In this section, we experimentally investigate the relation between the condition number and the convergence rate. All computations are conducted using a Xeon X5482 (3.20GHz) and 32GB physical memory with CentOS Linux release 5.2. For optimization problems, we applied the gradient descent method and quasi Newton methods instead of the Newton method, since the Newton method does not efficiently work for high-dimensional problems (Luenberger & Ye, 2008, introduction of Chap. 10).

6.1 Synthetic Data

In the M-estimator based on the φ-divergence, the Hessian matrix involved in the optimization problem (8) is given as

$$H = \frac{1}{n} K_{11} D_{\psi,w} K_{11} + \lambda K_{11} \in \Re^{n \times n}. \tag{24}$$

[3]The linear function does not provide a consistent estimator of density ratios, because ψ' is constant.

For the estimator using the Kullback-Leibler divergence (Nguyen et al., 2010; Sugiyama et al., 2008a), the function $\varphi(z)$ is given as $\varphi(z) = -\log z$, and thus, $\psi(z) = -1 - \log(-z)$, $z < 0$. Then, $\psi'(z) = -1/z$ and $\psi''(z) = 1/z^2$ for $z < 0$. Thus, for the optimal solution $w_\psi(x)$ under the population distribution, we have $\psi''(w_\psi(x)) = \psi''((\psi')^{-1}(w_0(x))) = w_0(x)^2$, where w_0 is the true density ratio q/p. Then the Hessian matrix at the target function w_ψ is given as

$$H_{\mathrm{KL}} = \frac{1}{n} K_{11} \mathrm{diag}(w_0(X_1)^2, \ldots, w_0(X_n)^2) K_{11} + \lambda K_{11} \in \Re^{n \times n}.$$

Conversely, in KuLSIF, the Hessian matrix is given by H_{KuLSIF} defined in (21), and the Hessian matrix of R-KuLSIF, $H_{\text{R-KuLSIF}}$, is shown in (22).

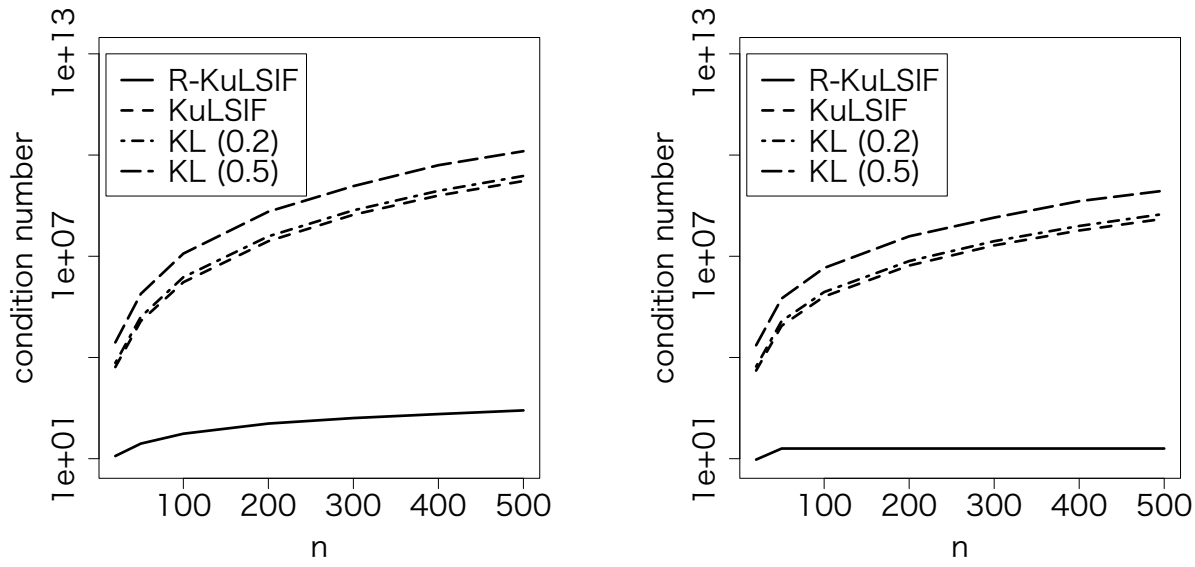
The condition numbers of Hessian matrices, H_{KL}, H_{KuLSIF}, and $H_{\text{R-KuLSIF}}$, are numerically compared. In addition, the condition number of K_{11} is computed. The probability distributions P and Q are set to the normal distribution on the 10-dimensional Euclidean space with the identity variance-covariance matrix I_{10}. The mean vectors of P and Q are set to $0 \times \mathbf{1}_{10}$ and $\mu \times \mathbf{1}_{10}$ with $\mu = 0.2$ or $\mu = 0.5$, respectively. Note that the mean value μ only affects the condition number of the KL method, not R-KuLSIF and KuLSIF. The true density-ratio w_0 is determined by P and Q. In the kernel-based estimators, we use the Gaussian kernel with width $\sigma = 4$, where $\sigma = 4$ is close to the median of the distance $\|X_i - X_j\|$ between samples. Using the median distance as the kernel width is a popular heuristic (Caputo et al., 2002; Schölkopf & Smola, 2002). We study two setups: In the first setup, the sample size from P is equal to that from Q, that is, $n = m$, and in the second setup, the sample size from Q is fixed to $m = 50$ and n is varied from 20 to 500. The regularization parameter λ is set to $\lambda_{n,m} = 1/(n \wedge m)^{0.9}$, where $n \wedge m = \min\{n, m\}$.

In each setup, the samples $X_1, \ldots, X_n$ are randomly generated and the condition number is computed. Figure 1 shows the condition number average over 1000 runs. We see that for all cases, the condition number of R-KuLSIF is significantly smaller than that of the other methods. Thus, it is expected that R-KuLSIF converges faster than the other methods and that R-KuLSIF is robust against numerical degeneracy.

Figure 2 and Table 1 show the average number of iterations and average computation time for solving the optimization problems over 50 runs. The probability distributions P and Q are the same as those in the above experiments, and the mean vector of Q is set to $0.5 \times \mathbf{1}_{10}$. The number of samples from each probability distribution is set to $n = m = 100, \ldots, 6000$, and the regularization parameter is set to $\lambda = 1/(n \wedge m)^{0.9}$. Note that n is equal to the number of parameters to be optimized. R-KuLSIF, KuLSIF, and the method based on the Kullback-Leibler divergence (KL) are compared. In addition, the computation time for solving the linear equation

$$\left(\frac{1}{n} K_{11} + \lambda I_n\right) \alpha = -\frac{1}{nm\lambda} K_{12} \mathbf{1}_m \tag{25}$$

instead of optimizing (11) is also shown as "direct" in the plot. The kernel parameter σ is determined based on the median of $\|X_i - X_j\|$. To solve the optimization problems for M-estimators, we use two optimization methods: one is the BFGS quasi-Newton method



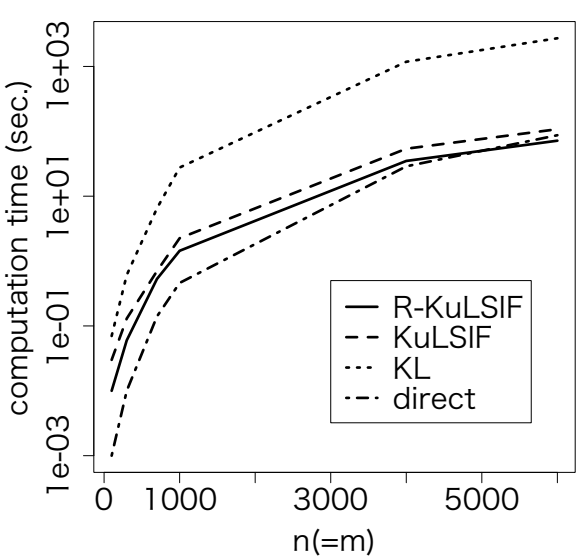
(a) $n = m$, $\sigma = 4$.

(b) $m = 50$, $\sigma = 4$.

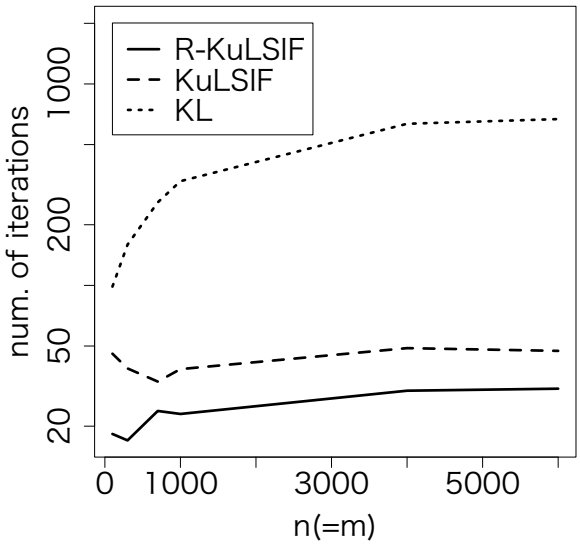
Figure 1: Average condition number of Hessian matrix over 1000 runs. Left panel shows condition number in case $n = m$ and $\sigma = 4$, and right panel shows result when sample size from Q is fixed to $m = 50$ and σ is set to 4. KL(μ) denotes condition number of H_{KL}, when mean vector of probability distribution Q is specified by μ. Note that condition number of R-KuLSIF and KuLSIF does not depend on μ.

implemented in the `optim` function in R (R Development Core Team, 2009), and the other is the steepest descent method. Furthermore, for the "direct" method, we use the `solve` function in R. Figure 2 shows the result for the BFGS method and Table 1 shows the result for the steepest descent method. In the numerical experiments for the steepest descent method, the maximum number of iterations is limited to 4000, and the KL method reaches the limit. The numerical results indicate that the number of iterations in the optimization procedure is highly correlated with the condition number of the Hessian matrices.

Although the practical computational time would depend on various issues such as stopping rules, our theoretical results in Section 4 are shown to be in good agreement with the empirical results for the synthetic data. We observed that numerical optimization methods such as the quasi-Newton method are competitive with numerical algorithms for solving linear equations using LU decomposition or Cholesky decomposition, especially when the sample size n (which is equal to the number of optimization parameters in the current setup) is large. This implies that the theoretical result obtained in this study will be useful in large sample cases, which is common in practical applications.



(a) Computation time (sec.)



(b) Number of iterations in the BFGS method.

Figure 2: Average computation time and average number of iterations in BFGS method over 50 runs.

Table 1: Average computation time and average number of iterations in steepest descent method over 50 runs. "$>$" means that actual computation time is longer than number described in table.

	$n = 100,\ m = 100$		$n = 300,\ m = 300$		$n = 700,\ m = 700$	
Estimator	Comput. time (sec.)	Number of iterations	Comput. time (sec.)	Number of iterations	Comput. time (sec.)	Number of iterations
R-KuLSIF	0.07	21.0	0.50	33.7	4.46	49.0
KuLSIF	1.23	288.0	10.16	487.5	78.21	640.4
KL	11.58	1941.6	> 111.83	> 4000	> 539.72	> 4000

6.2 Benchmark Data

Next, we apply the density-ratio estimation to benchmark data sets, and compare the computational cost. The statistical performance of each estimator for a linear model has been extensively compared on benchmark data sets in Kanamori et al. (2009), Kanamori et al. (2012), and Hido et al. (2011). Therefore, here, we focus on the numerical efficiency of each method.

Let us consider an outlier detection problem of finding irregular samples in a data set ("evaluation data set") based on another data set ("model data set") that only contains regular samples (Hido et al., 2011). Defining the density ratio over the two sets of samples,

we can see that the density-ratio values for regular samples are close to one, while those for outliers tend to deviate significantly from one. Since the evaluation data set usually has a wider support than the model data set, we regard the evaluation data set as samples corresponding to the denominator of the density ratio and the model data set as samples corresponding to the numerator. Then the target density-ratio $w_0(x)$ is approximately equal to one in a wide range of the data domain, and will take small values around outliers.

The data sets provided by IDA (Rätsch et al., 2001) are used. These are binary classification data sets consisting of positive/negative and training/test samples. We allocate all positive training samples to the "model" set, while all positive test samples and 5% of negative test samples are assigned to the "evaluation set." Thus, we regard the positive samples as inliers and negative samples as outliers.

Table 2 shows the average computation time and average number of iterations over 20 runs for `image` and `splice` and over 50 runs for the other data sets. In the same way as the simulations in Section 6.1, we compare R-KuLSIF, KuLSIF, and the M-estimator with the Kullback-Leibler divergence (KL). In addition, the computation time of solving the linear equation (25) is shown as "direct" in the table. For the optimization, we use the BFGS method implemented in the `optim` function in R (R Development Core Team, 2009), and we use the `solve` function in R for the "direct" method. The kernel parameter σ is determined based on the median of $\|X_i - X_j\|$ which is computed by the function `sigest` in the kernlab library (Karatzoglou et al., 2004). The average number of samples is shown in the second column, and the regularization parameter is set to $\lambda = 1/(n \wedge m)^{0.9}$.

The numerical results show that, when the sample size is balanced (i.e., n and m are comparable to each other), the number of iterations for R-KuLSIF is the smallest, which agrees well with our theoretical analysis. On the other hand, for `titanic`, `waveform`, `banana`, `ringnorm`, and `twonorm`, the number of iterations for each method is almost the same. In these data sets, m is much smaller than n, and thus the second term λK_{11} in the Hessian matrix (24) for the M-estimator will govern the convergence property, since the order of $\lambda_{n,m}$ is larger than $O(1/n)$. This tendency is explained by the result in Theorem 2. Based on (20), we see that a large $\lambda_{n,m}$ will provide a smaller upper bound of $\kappa(H)$.

Next, we investigate the number of iterations when n and m are comparable to each other. The data sets, `titanic`, `waveform`, `banana`, `ringnorm`, and `twonorm` are used. We consider two setups: In the first series of experiments, the evaluation data set consists of all positive test samples, and the model data set is defined by all negative test samples. Therefore, the target density-ratio may be far from the constant function $w_0(x) = 1$. Table 3 shows the average computation time and average number of iterations over 20 runs. In this case, the number of iterations for optimization agrees with our theoretical result, that is, R-KuLSIF yields low computational costs for all experiments. In the second series of experiments, both model samples and evaluation samples are randomly chosen from all (i.e., both positive and negative) test samples. Thus, the target density-ratio is almost equal to the constant function $w_0(x) = 1$. Table 4 shows the average computation time and the average number of iterations over 20 runs. The number of iterations for "KL" is much smaller than that for the first setup shown in Table 3. This

Table 2: Average computation time (s) and average number of iterations for benchmark data sets are shown. BFGS quasi-Newton method in `optim` function of R environment is used to obtain numerical solution. Data sets are arranged in ascending order of sample size n. Results of method having lowest mean are described in bold face.

(a) Computation time (s)

data set	n	m	R-KuLSIF	KuLSIF	KL	direct
thyroid	26	43	0.008	0.015	0.015	**0.001**
b-cancer	27	58	0.008	0.012	0.013	**0.001**
heart	49	76	0.01	0.016	0.021	**0.001**
german	104	211	0.02	0.03	0.05	**0.002**
diabetes	118	165	0.02	0.04	0.07	**0.002**
f-solar	241	368	0.05	0.11	0.24	**0.01**
image	625	746	0.85	2.19	6.02	**0.15**
titanic	767	47	0.98	0.96	1.11	**0.28**
splice	1153	483	1.66	3.59	6.50	**0.84**
waveform	1746	131	4.06	3.96	5.95	**2.50**
banana	2437	184	11.51	10.77	14.18	**6.69**
ringnorm	3816	198	18.27	**12.77**	29.97	24.92
twonorm	3850	203	22.10	**15.70**	30.14	26.69

(b) Number of iterations

data set	n	m	R-KuLSIF	KuLSIF	KL
thyroid	26	43	**14.1**	39.8	36.1
b-cancer	27	58	**13.1**	30.8	29.8
heart	49	76	**14.0**	35.0	42.4
german	104	211	**15.5**	39.1	48.8
diabetes	118	165	**14.8**	44.3	65.3
f-solar	241	368	**14.7**	30.8	61.1
image	625	746	**22.1**	61.3	135.3
titanic	767	47	20.7	**16.4**	19.7
splice	1153	483	**15.0**	28.8	49.9
waveform	1746	131	20.3	**17.7**	28.2
banana	2437	184	28.4	**23.2**	30.6
ringnorm	3816	198	20.3	**13.5**	31.7
twonorm	3850	203	22.2	**13.2**	26.9

Table 3: For balanced sample size, average computation time (s) and average number of iterations for benchmark data sets are presented. Titanic, waveform, banana, ringnorm, and twonorm are used as data sets. Evaluation data set consists of all positive test samples, and model data set is defined by all negative test samples, i.e., density ratio will be far from constant function. BFGS quasi-Newton method in `optim` function of R environment is used to obtain numerical solution. Data sets are arranged in ascending order of sample size n. Results of method having lowest mean are described in bold face.

(a) Computation time (s)

data set	n	m	R-KuLSIF	KuLSIF	KL	direct
titanic	1327	2775	6.11	6.24	15.93	**1.45**
waveform	3032	6168	52.74	155.30	676.53	**16.96**
banana	4383	5417	97.64	248.08	1466.94	**52.97**
ringnorm	6933	7067	**145.37**	169.08	3374.45	177.96
twonorm	7002	6998	**145.61**	206.12	3243.83	226.20

(b) Number of iterations

data set	n	m	R-KuLSIF	KuLSIF	KL
titanic	1327	2775	**20.9**	21.6	54.1
waveform	3032	6168	**36.3**	132.7	425.6
banana	4383	5417	**40.0**	110.2	487.2
ringnorm	6933	7067	**34.5**	48.6	595.1
twonorm	7002	6998	**28.6**	48.7	545.0

is because the condition number of the Hessian matrix (24) is likely to be small when the true density-ratio w_0 is close to the constant function. R-KuLSIF is, however, still the preferable approach. Furthermore, the computation time of R-KuLSIF is comparable to that of a direct method such as the Cholesky decomposition when the sample size (i.e., the number of variables) is large.

In summary, the numerical experiments showed that the convergence rate for optimization is well explained by the condition number of the Hessian matrix. The relation between the loss function ψ and condition number was discussed in Section 4, and our theoretical result implies that R-KuLSIF is computationally an effective way to estimate density ratios. The numerical results in this section also indicated that our theoretical result is useful to obtain practical and computationally efficient estimators.

7 Conclusions

We considered the problem of estimating the ratio of two probability densities and investigated theoretical properties of the kernel least-squares estimator called KuLSIF. More

Table 4: For balanced sample size, average computation time (s) and average number of iterations for benchmark data sets are presented. Titanic, waveform, banana, ringnorm, and twonorm are used as data sets. Evaluation data set and model data set are randomly generated from all (i.e., both positive and negative) test samples, i.e., density ratio is close to constant function. BFGS quasi-Newton method in `optim` function of R environment is used to obtain numerical solution. Data sets are arranged in ascending order of sample size n. Results of method having lowest mean are described in bold face.

(a) Computation time (s)

data set	n	m	R-KuLSIF	KuLSIF	KL	direct
titanic	2052	2050	10.20	11.42	19.93	**5.13**
waveform	4600	4600	63.55	124.41	536.46	**58.55**
banana	4900	4900	112.21	130.62	328.56	**78.08**
ringnorm	7000	7000	135.70	**124.79**	1694.38	258.03
twonorm	7000	7000	**133.44**	153.27	1199.00	243.46

(b) Number of iterations

data set	n	m	R-KuLSIF	KuLSIF	KL
titanic	2052	2050	18.9	**17.9**	33.7
waveform	4600	4600	**25.4**	57.0	170.9
banana	4900	4900	**34.0**	39.9	86.7
ringnorm	7000	7000	**23.1**	23.4	227.7
twonorm	7000	7000	**24.1**	29.8	184.4

specifically, we theoretically studied the condition number of Hessian matrices, because the condition number is closely related to the convergence rate of optimization and the numerical stability. We found that KuLSIF has a smaller condition number than the other methods. Therefore, KuLSIF will have preferable computational properties. We further showed that R-KuLSIF, which is an alternative formulation of KuLSIF, possesses an even smaller condition number. Numerical experiments showed that practical numerical properties of optimization algorithms could be well explained by our theoretical analysis of condition numbers, even though the condition number only provides an upper bound of the rate of convergence. A theoretical issue to be further investigated is the derivation of a tighter probabilistic order of the condition number.

Density-ratio estimation was shown to provide new approaches to solving various machine learning problems (Sugiyama et al., 2009; Sugiyama et al., 2012), including covariate shift adaptation (Shimodaira, 2000; Zadrozny, 2004; Sugiyama & Müller, 2005; Gretton et al., 2009; Sugiyama et al., 2007; Bickel et al., 2009; Quiñonero-Candela et al., 2009; Sugiyama & Kawanabe, 2012), multi-task learning (Bickel et al., 2008; Simm et al., 2011), inlier-based outlier detection (Hido et al., 2008; Smola et al., 2009; Hido et al., 2011), change detection in time-series (Kawahara & Sugiyama, 2011), divergence estimation

(Nguyen et al., 2010), two-sample testing (Sugiyama et al., 2011a), mutual information estimation (Suzuki et al., 2008; Suzuki et al., 2009b), feature selection (Suzuki et al., 2009a), sufficient dimension reduction (Sugiyama et al., 2010a), independence testing (Sugiyama & Suzuki, 2011), independent component analysis (Suzuki & Sugiyama, 2011), causal inference (Yamada & Sugiyama, 2010), object matching (Yamada & Sugiyama, 2011), clustering (Kimura & Sugiyama, 2011), conditional density estimation (Sugiyama et al., 2010b), and probabilistic classification (Sugiyama, 2010). In future work, we will develop practical algorithms for a wide range of applications on the basis of theoretical guidance provided in this study.

Acknowledgment

The authors are grateful to 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 partially supported by MEXT Kakenhi 22700289, Global COE Program "The Research and Training Center for New Development in Mathematics," and the Aihara Project, the FIRST program from JSPS, initiated by CSTP. M. Sugiyama was supported by SCAT, AOARD, and the JST PRESTO program.

A Proof of Lemma 1

Proof. We consider the minimization of the loss function,

$$\frac{1}{n}\sum_{i=1}^{n}\psi(w(X_i)) - \frac{1}{m}\sum_{j=1}^{m} w(Y_j) + \frac{\lambda}{2}\|w\|_{\mathcal{H}}^2, \quad w \in \mathcal{H}. \tag{26}$$

Applying the representer theorem (Kimeldorf & Wahba, 1971), we see that an optimal solution of (26) has the form of

$$w = \sum_{j=1}^{n}\alpha_j k(\cdot, X_j) + \sum_{\ell=1}^{m}\beta_\ell k(\cdot, Y_\ell). \tag{27}$$

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 (26) under the constraint (27) is given as

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

If α and β satisfy the above conditions, they are the optimal solution because the loss function is convex in α and β. Substituting $\beta = \frac{1}{m\lambda}\mathbf{1}_m$, we obtain

$$\frac{1}{n}K_{11}v(\alpha, \mathbf{1}_m/m\lambda) + \lambda K_{11}\alpha = 0, \quad \text{and}$$
$$\frac{1}{n}K_{21}v(\alpha, \mathbf{1}_m/m\lambda) + \lambda K_{21}\alpha = 0.$$

For $\alpha = \bar{\alpha}$, the above equalities are satisfied, since

$$\frac{1}{n}v(\bar{\alpha}, \mathbf{1}_m/m\lambda) + \lambda\bar{\alpha} = 0$$

is assumed. Therefore, $\alpha = \bar{\alpha}$ and $\beta = \mathbf{1}_m/m\lambda$ with (27) provide the minimizer of (26). □

B Proof of Equation (14)

Let $\kappa(A)$ be the condition number of the symmetric positive definite matrix A, then we shall prove the equality

$$\min_{S:\kappa(S)\leq C} \kappa(S^\top AS) = \max\left\{\frac{\kappa(A)}{C^2},\ 1\right\},$$

where S is symmetric and positive definite. The same equality holds, when S is non-symmetric and the condition number of S is defined through singular values. We prove the case that S is a symmetric positive definite matrix for simplicity.

Proof. First, we prove $\min_{S:\kappa(S)\leq C} \kappa(SAS) \geq \max\{\frac{\kappa(A)}{C^2},\ 1\}$.

The matrix A is symmetric positive definite, thus, there exists an orthogonal matrix Q and a diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ such that $A = Q\Lambda Q^\top$. The eigenvalues are arranged in the decreasing order, i.e., $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$. In the similar way, let S be $PDP^\top$, where P is an orthogonal matrix and $D = \mathrm{diag}(d_1, \ldots, d_n)$ is a diagonal matrix such that $d_1 \geq d_2 \geq \cdots \geq d_n > 0$ and $d_1/d_n \leq C$. Hence,

$$\kappa(SAS) = \kappa(PDP^\top Q\Lambda Q^\top PDP^\top) = \kappa(DP^\top Q\Lambda Q^\top PD).$$

Let $Q^\top P$ be $R^\top$ which is also an orthogonal matrix. Then the maximum eigenvalue of $DR\Lambda R^\top D$ is given as

$$\max_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^\top DR\Lambda R^\top D\boldsymbol{x}.$$

Let $R = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)$, where $\boldsymbol{r}_i \in \Re^n$, and we choose $\boldsymbol{x}_1$ such that $\boldsymbol{r}_i^\top D\boldsymbol{x}_1 = 0$ for $i = 2, \ldots, n$ and $\|\boldsymbol{x}_1\| = 1$. Then,

$$\max_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^\top DR\Lambda R^\top D\boldsymbol{x} \ \geq\ \boldsymbol{x}_1^\top DR\Lambda R^\top D\boldsymbol{x}_1 = \lambda_1(\boldsymbol{x}_1^\top D\boldsymbol{r}_1)^2.$$

From the assumption on $\boldsymbol{x}_1$, $D\boldsymbol{x}_1$ is represented as $c\boldsymbol{r}_1$ for some $c \in \Re$, and we have $(\boldsymbol{x}_1^\top D\boldsymbol{r}_1)^2 = c^2 = \boldsymbol{x}_1^\top D^2\boldsymbol{x}_1 \geq d_n^2$. Hence, we have

$$\max_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^\top SAS\boldsymbol{x} \geq \lambda_1 d_n^2.$$

On the other hand, the minimum eigenvalue of $DR\Lambda R^\top D$ is given as

$$\min_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^\top DR\Lambda R^\top D\boldsymbol{x}.$$

We choose $\boldsymbol{x}_n$ such that $\boldsymbol{r}_i^\top D\boldsymbol{x}_n = 0$ for $i = 1, \ldots, n-1$ and $\|\boldsymbol{x}_n\| = 1$. Then,

$$\begin{aligned}
\min_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^\top DR\Lambda R^\top D\boldsymbol{x} \;&\leq\; \boldsymbol{x}_n^\top DR\Lambda R^\top D\boldsymbol{x}_n \\
&= \lambda_n (\boldsymbol{x}_n^\top D\boldsymbol{r}_n)^2 \\
&\leq\; \lambda_n \boldsymbol{x}_n^\top D^2\boldsymbol{x}_n \quad \text{(Schwarz inequality)} \\
&\leq\; \lambda_n d_1^2.
\end{aligned}$$

As a result, the condition number of SAS is bounded below as

$$\kappa(SAS) \geq \frac{\lambda_1 d_n^2}{\lambda_n d_1^2} = \frac{\kappa(A)}{(d_1/d_n)^2} \geq \frac{\kappa(A)}{C^2}.$$

Next, we prove $\min_{S:\kappa(S)\leq C} \kappa(SAS) \leq \max\{\frac{\kappa(A)}{C^2},\ 1\}$. If $\kappa(A) \leq C^2$, the inequality $\min_{S:\kappa(S)\leq C} \kappa(SAS) = 1$ holds, because we can choose $S = A^{-1/2}$. Then, we prove $\min_{S:\kappa(S)\leq C} \kappa(SAS) \leq \frac{\kappa(A)}{C^2}$ when $1 \leq C^2 \leq \kappa(A)$ holds.

Let $S = Q\Gamma Q^\top$ with Γ be a diagonal matrix $\mathrm{diag}(\gamma_1, \ldots, \gamma_n)$, then $\kappa(SAS) = \kappa(\mathrm{diag}(\gamma_1^2\lambda_1, \ldots, \gamma_n^2\lambda_n))$ holds. Let $\gamma_1 = 1$ and $\gamma_n = C$. Since $1 \leq C^2 \leq \kappa(A) = \lambda_1/\lambda_n$ holds, for $k = 2, \ldots, n-1$ we have

$$1 \leq \min\left\{C,\ \sqrt{\frac{\lambda_1}{\lambda_k}}\right\}, \qquad C\sqrt{\frac{\lambda_n}{\lambda_k}} \leq \min\left\{C,\ \sqrt{\frac{\lambda_1}{\lambda_k}}\right\}$$

and thus, we obtain

$$\max\left\{1, C\sqrt{\frac{\lambda_n}{\lambda_k}}\right\} \leq \min\left\{C,\ \sqrt{\frac{\lambda_1}{\lambda_k}}\right\}, \qquad k = 2, \ldots, n-1.$$

Hence, there exists $\gamma_k,\ k = 2, \ldots, n-1$ such that

$$\max\left\{1,\ C\sqrt{\frac{\lambda_n}{\lambda_k}}\right\} \;\leq\; \gamma_k \;\leq\; \min\left\{C, \sqrt{\frac{\lambda_1}{\lambda_k}}\right\}.$$

Thus, $1 \leq \gamma_k \leq C$ holds for all $k = 2, \ldots, n-1$. Moreover, $C^2\lambda_n \leq \gamma_k^2\lambda_k \leq \lambda_1$ also holds. These inequalities imply $\kappa(S) = C$ and $\kappa(SAS) = \lambda_1/(C^2\lambda_n) = \kappa(A)/C^2$. Therefore $\min_{S:\kappa(S)\leq C} \kappa(SAS) \leq \frac{\kappa(A)}{C^2}$ holds if $1 \leq C^2 \leq \kappa(A)$. □

C Proof of Theorem 1

We show the proof of Theorem 1.

Proof. For a fixed function ψ satisfying the assumption in the theorem, let b be a real number in the domain of ψ such that $\psi'(b) = 1$. Then, we have $\psi''(b) = c$. Let w_b be the constant function taking b over $\mathcal{Z}$. In a universal RKHS, for any $\delta > 0$, there exists $w \in \mathcal{H}$ such that $\|w_b - w\|_\infty \leq \delta$. According to Appendix D in Horn and Johnson (1985), eigenvalues of a matrix are continuous on its entries, and thus the same thing holds for the minimal and maximal eigenvalues and the condition number as long as the condition number is well-defined. Then, for any $\varepsilon > 0$ there exists $w \in \mathcal{H}$ such that

$$|\kappa_0(D_{\psi,w}) - \kappa_0(cI_n)| = |\kappa_0(D_{\psi,w}) - \kappa_0(D_{\psi,w_b})| \leq \varepsilon, \tag{28}$$

since $\psi''(w_b) = \psi''(b) = c$. For any ψ satisfying the assumption, we show that (28) leads to the inequality

$$\sup\{\kappa_0(D_{\psi,w}) \mid w \in \mathcal{H}\} \;\geq\; \kappa_0(cI_n) \tag{29}$$

for fixed samples $X_1, \ldots, X_n$. We prove (29) by contradiction. Suppose that $\sup\{\kappa_0(D_{\psi,w}) \mid w \in \mathcal{H}\} < \kappa_0(cI_n)$ holds, and let $\delta = \kappa_0(cI_n) - \sup\{\kappa_0(D_{\psi,w}) \mid w \in \mathcal{H}\}$. Then, δ is positive. The inequality $\kappa_0(D_{\psi,w}) \leq \sup\{\kappa_0(D_{\psi,w}) \mid w \in \mathcal{H}\}$ leads to the inequality $\kappa_0(cI_n) - \kappa_0(D_{\psi,w}) \geq \kappa_0(cI_n) - \sup\{\kappa_0(D_{\psi,w}) \mid w \in \mathcal{H}\} = \delta > \delta/2 > 0$ for all $w \in \mathcal{H}$. This inequality contradicts (28), because the inequality (28) guarantees that there exists $w \in \mathcal{H}$ such that $|\kappa_0(D_{\psi,w}) - \kappa_0(cI_n)| \leq \delta/2$ holds. Hence, the inequality (29) should hold.

In addition, for the quadratic function $\psi(z) = cz^2/2$, the equality

$$\sup\{\kappa_0(D_{\psi,w}) \mid w \in \mathcal{H}\} = \kappa_0(cI_n).$$

holds. Thus, we obtain (17). □

D Proof of Theorem 2

The following lemma is the key to prove Theorem 2.

Lemma 2. *Let k be a kernel function on $\mathcal{Z} \times \mathcal{Z}$ satisfying the boundedness condition, $\sup_{x,x' \in \mathcal{Z}} k(x, x') < \infty$, and U be $U = \sup_{x,x' \in \mathcal{Z}} k(x, x')$. Suppose that the Gram matrix $(K_{11})_{ij} = k(X_i, X_j)$ is almost surely positive definite in terms of the probability measure P. Then, the inequality*

$$^\forall\delta > 0, \qquad \Pr\left(\kappa(H) > \kappa(K_{11})\big(1 + \frac{\delta}{\lambda}\big)\right) \;\leq\; 1 - F_n(\delta/U) \tag{30}$$

holds, where H is defined by $H = \frac{1}{n} K_{11} D_{\psi,\widehat{w}} K_{11} + \lambda K_{11}$. In the above expressions, the probability $\Pr(\cdots)$ is defined from the distribution of all samples $X_1, \ldots, X_n$, $Y_1, \ldots, Y_m$.

Proof. Let k_i be the i-th column vector of the Gram matrix K_{11}, and d_i be $\psi''(\widehat{w}(X_i))$. Then the matrix H is represented as

$$H = \frac{1}{n}\sum_{i=1}^{n} d_i k_i k_i^\top + \lambda K_{11} \in \Re^{n\times n}.$$

Let us define

$$W_n = \min_{\|a\|=1} a^\top H a, \quad Z_n = \max_{\|a\|=1} a^\top H a,$$

i.e., W_n and Z_n are the minimal and maximal eigenvalues of H. Then, the condition number of H is given as $\kappa(H) = Z_n/W_n$. Let τ_1 and τ_n be the maximal and minimal eigenvalues of K_{11}. Since all diagonal elements of K_{11} are less than or equal to U, we have

$$0 < \tau_1 \leq \operatorname{Tr} K_{11} \leq Un.$$

Then, we have a lower bound of W_n and an upper bound of Z_n as follows:

$$\begin{aligned}
W_n &= \min_{\|a\|=1} \frac{1}{n}\sum_{i=1}^{n} d_i (k_i^\top a)^2 + \lambda a^\top K_{11} a \;\geq\; \lambda\tau_n, \\
Z_n &= \max_{\|a\|=1} \frac{1}{n}\sum_{i=1}^{n} d_i (k_i^\top a)^2 + \lambda a^\top K_{11} a \\
&\leq \frac{\max_j d_j}{n} \max_{\|a\|=1} \sum_{i=1}^{n} (k_i^\top a)^2 + \lambda\tau_1 \\
&= \frac{\max_j d_j}{n}\tau_1^2 + \lambda\tau_1 \\
&\leq U\tau_1 \max_j d_j + \lambda\tau_1,
\end{aligned}$$

where the last inequality for Z_n follows from $\tau_1 \leq Un$. Therefore, for any $\delta > 0$, we have

$$\begin{aligned}
\Pr\Big(\kappa(H) > \kappa(K_{11})\big(1+\frac{\delta}{\lambda}\big)\Big) \;\leq\; &\Pr\left(\frac{U\tau_1 \max_j d_j + \lambda\tau_1}{\lambda\tau_n} > \kappa(K_{11})\big(1+\frac{\delta}{\lambda}\big)\right) \\
=&\Pr\left(\max_j d_j > \delta/U\right) \\
=&1 - \Pr\left(\max_j d_j \leq \delta/U\right) \\
=&1 - F_n(\delta/U).
\end{aligned}$$

□

In Lemma 2, the distributions of W_n and Z_n are separately computed. This idea is borrowed from smoothed analysis of the condition numbers (Sankar et al., 2006).

Below, we show the proof of Theorem 2.

proof of Theorem 2. Using (30) in Lemma 2, we have

$$\lim_{n\to\infty} \Pr\left(\kappa(H) > \kappa(K_{11})\big(1+\frac{t_n}{\lambda}\big)\right) \leq 1 - \lim_{n\to\infty} F_n(t_n/U) = 0.$$

□

E Proof of Theorem 3

We show the proof of Theorem 3

Proof. Assume that $\psi''(z)$ is not a constant function. Since K_{11} is non-singular, the vector $K_{11}\alpha + \frac{1}{m\lambda}K_{12}\mathbf{1}_m$ takes an arbitrary value in $\Re^n$ by varying $\alpha \in \Re^n$. Hence, each diagonal element of $D_{\psi,\alpha}$ can take arbitrary values in an open subset $S \subset \Re$. We consider $(CK_{11})^{-1}J_{\psi,C}(\alpha)((CK_{11})^\top)^{-1}$ instead of $J_{\psi,C}$. Suppose that there exists a matrix C such that the matrix

$$(CK_{11})^{-1}J_{\psi,C}(\alpha)((CK_{11})^\top)^{-1} = \frac{1}{n}\operatorname{diag}(s_1,\ldots,s_n)K_{11}(K_{11}C^\top)^{-1} + \lambda(K_{11}C^\top)^{-1} \quad (31)$$

is symmetric for any $(s_1,\ldots,s_n) \in S^n$. Let a_{ij} be the (i,j) element of $K_{11}(K_{11}C^\top)^{-1}$, and t_{ij} be the (i,j) element of $(K_{11}C^\top)^{-1}$. Then, the (i,j) and (j,i) elements of (31) are equal to $\frac{1}{n}s_ia_{ij} + \lambda t_{ij}$ and $\frac{1}{n}s_ja_{ji} + \lambda t_{ji}$, respectively. Due to the assumption, the equality

$$\frac{1}{n}s_ia_{ij} + \lambda t_{ij} = \frac{1}{n}s_ja_{ji} + \lambda t_{ji}$$

holds for any $s_i, s_j \in S$. When $i \neq j$, we obtain $a_{ij} = a_{ji} = 0$ and $t_{ij} = t_{ji}$. Thus, $K_{11}(K_{11}C^\top)^{-1}$ should be equal to some diagonal matrix, and $(K_{11}C^\top)^{-1}$ is a symmetric matrix. There exists a diagonal matrix $Q = \operatorname{diag}(q_1,\ldots,q_n)$ such that $K_{11} = Q(K_{11}C^\top)$ holds. As a result, we have $(K_{11})_{ij} = q_i(K_{11}C^\top)_{ij}$, $(K_{11})_{ji} = q_j(K_{11}C^\top)_{ji}$, $(K_{11}C^\top)_{ij} = (K_{11}C^\top)_{ji}$,and $(K_{11})_{ij} = (K_{11})_{ji}$. Hence we obtain

$$(K_{11})_{ij} = q_i(K_{11}C^\top)_{ij} = q_j(K_{11}C^\top)_{ij},$$

and then, $q_i = q_j$ or $(K_{11}C^\top)_{ij} = 0$ holds for any i and j. Since $(K_{11})_{ij}$ is non-zero element, the only possibility is $q_1 = q_2 = \cdots = q_n \neq 0$. Therefore, the diagonal matrix Q should be proportional to the identity matrix and there exists a constant $c \in \Re$ such that the equality $C = cI_n$ holds. This equality contradicts the assumption. □

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.

Axelsson, O., & Neytcheva, M. (2002). Robust preconditioners for saddle point problems. *Numerical Methods and Application* (pp. 158–166).

Badia, S., Nobile, F., & Vergara, C. (2009). Robin-robin preconditioned Krylov methods for fluid-structure interaction problems. *Computer Methods in Applied Mechanics and Engineering*, *198*, 2768–2784.

Becchetti, L., Leonardi, S., Marchetti-Spaccamela, A., Schafer, G., & Vredeveld, T. (2006). *Average-case and smoothed competitive analysis of the multilevel feedback algorithm*Open Access publications from Maastricht University urn:nbn:nl:ui:27-17093). Maastricht University.

Beltran, C., & Pardo, L. M. (2006). Estimates on the distribution of the condition number of singular matrices. *Foundations of Computational Mathematics*, *7*, 87–134.

Benzi, M., Haber, E., & Taralli, L. (2011). A preconditioning technique for a class of PDE-constrained optimization problems. *Adv. Comput. Math.*, *35*, 149–173.

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

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

Blum, A., & Dunagan, J. (2002). Smoothed analysis of the perceptron algorithm for linear programming. *Proc. of the 13th Annual ACM-SIAM Symp. on Discrete Algorithms* (pp. 905–914).

Blum, L., & Shub, M. (1986). Evaluating rational functions: Infinite precision is finite cost and tractable on average. *SIAM J. Comput.*, *15*, 384–398.

Bürgisser, P., & Cucker (2010). Smoothed analysis of moore-penrose inversion. *SIAM J. Matrix Anal. & Appl.*, *31*, 2769–2783.

Bürgisser, P., Cucker, F., & de Naurois, P. (2006a). The complexity of semilinear problems in succinct representation. *Computational Complexity*, *15*, 197–235.

Bürgisser, P., Cucker, F., & Lotz, M. (2006b). General formulas for the smoothed analysis of condition numbers. *C. R. Acad. Sci. Paris, Ser. I*, *343*, 145–150.

Bürgisser, P., Cucker, F., & Lotz, M. (2006c). Smoothed analysis of complex conic condition numbers. *Journal de Mathématiques Pures et Appliquées*, *86*, 293–309.

Bürgisser, P., Cucker, F., & Lotz, M. (2010). Coverage processes on spheres and condition numbers for linear programming. *The Annals of Probability*, *38*, 570–604.

Caputo, B., Sim, K., Furesjo, F., & Smola, A. (2002). Appearance-based object recognition using SVMs: Which kernel should I use? *Proceedings of NIPS Workshop on Statistical Methods for Computational Experiments in Visual Processing and Computer Vision*.

Chapelle, O. (2007). Training a support vector machine in the primal. *Neural Computation*, *19*, 1155–1178.

Cheung, D., & Cucker, F. (2002). Probabilistic analysis of condition numbers for linear programming. *Journal of Optimization Theory and Applications*, *114*, 55–67.

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

Cucker, F., & Wschebor, M. (2002). On the expected condition number of linear programming problems. *Numerische Mathematik*, *94*, 94–419.

Demmel, J. (1988). The probability that a numerical analysis problem is difficult. *Mathematics of Computation*, *50*, 449–480.

Demmel, J. W. (1997). *Applied numerical linear algebra*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics.

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, *1*, 211–218.

Edelman, A. (1988). Eigenvalues and condition numbers of random matrices. *SIAM Journal on Matrix Analysis and Applications*, *9*, 543–560.

Edelman, A. (1992). On the distribution of a scaled condition nubmer. *Math. Comp.*, *58*, 185–190.

Edelman, A., & Sutton, B. D. (2005). Tails of condition number distributions. *SIAM Journal on Matrix Analysis and Applications*, *27*, 547–560.

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.

Hager, W. W., & Zhang, H. (2006). A survey of the nonlinear conjugate gradient methods. *Pacific Journal of Optimization*, *2*, 35–58.

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.

Horn, R., & Johnson, C. (1985). *Matrix analysis*. Cambridge University 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. (2012). Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, *86*, 335–367.

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, *11*, 1–20.

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

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

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

Kostlan, E. (1988). Complexity theory of numerical linear algebra. *Journal of Computational and Applied Mathematics*, *22*, 219–230.

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

Manthey, B., & Röglin, H. (2009). Worst-case and smoothed analysis of k-means clustering with Bregman divergences. *ISAAC* (pp. 1024–1033).

Mika, S., Schölkopf, B., Smola, A., Müller, K.-R., Scholz, M., & Rätsch, G. (1999). Kernel PCA and de-noising in feature spaces. *Proceedings of the 1998 conference on Advances in neural information processing systems II* (pp. 536–542). Cambridge, MA, USA: MIT Press.

Moré, J. J., & Sorensen, D. C. (1984). Newton's method. In G. H. Golub (Ed.), *Studies in numerical analysis*. pub-MATH-ASSOC-AMER.

Nakahara, M. (2003). *Geometry, topology and physics, second edition*. Taylor & Francis.

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.

Nocedal, J., & Wright, S. J. (1999). *Numerical optimization*. Springer.

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.

Ratliff, N., & Bagnell, J. D. (2007). Kernel conjugate gradient for fast kernel machines. *International Joint Conference on Artificial Intelligence*.

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

Renegar, J. (1987). On the efficiency of newton's method in approximating all zeros of a system of complex polynomials. *Math. Oper. Res.*, *12*, 121–148.

Renegar, J. (1995). Incorporating condition measures into the complexity theory of linear programming. *SIAM Journal on Optimization*, *5*.

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

Röglin, H., & Vöcking, B. (2007). Smoothed analysis of integer programming. *Math. Program.*, *110*, 21–56.

Sankar, A., Spielman, D. A., & Teng, S.-H. (2006). Smoothed analysis of the condition numbers and growth factors of matrices. *SIAM Journal on Matrix Analysis and Applications*, *28*, 446–476.

Schmidt, M., Le Roux, N., & Bach, F. (2011). Convergence rates of inexact proximal-gradient methods for convex optimization. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira and K. Weinberger (Eds.), *Advances in neural information processing systems 24*, 1458–1466.

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.

Shub, M. (1993). Some remarks on Bézout's theorem and complexity theory. *From Topology to Computation: Proceedings of the Smalefest* (pp. 443–455). Springer.

Shub, M., & Smale, S. (1994). Complexity of Bézout's theorem. V: Polynomial time. *Theoretical Computer Science*, *133*.

Shub, M., & Smale, S. (1996). Complexity of Bézout's theorem. IV: Probability of success; extensions. *SIAM J. of Numer. Anal.*, *33*, 128–148.

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

Smale, S. (1981). The fundamental theorem of algebra and complexity theory. *Bull. Amer. Math. Soc.*, *4*, 1–36.

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

Spielman, D. A., & Teng, S.-H. (2004). Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM*, *51*, 385–463.

Spivak, M. (1979). *A comprehensive introduction to differential geometry. Vol. I.* Publish or Perish Inc. Second edition.

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

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. (2012). *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. (2011). Least-squares independence test. *IEICE Transactions on Information and Systems*, *E94-D*, 1333–1336.

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

Sugiyama, M., Suzuki, T., & Kanamori, T. (2011b). Density ratio matching under the Bregman divergence: A unified framework of density ratio estimation. *Annals of the Institute of Statistical Mathematics*. to appear.

Sugiyama, M., Suzuki, T., & 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., Kanamori, T., Suzuki, T., Hachiya, H., & Okanohara, D. (2010a). Conditional density estimation via least-squares density ratio estimation. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS2010)* (pp. 781–788). Sardinia, Italy.

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

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

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

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. (2009b). 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.

Tao, T., & Vu, V. H. (2007). The condition number of a randomly perturbed matrix. *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing* (pp. 248–255). New York, NY, USA: ACM.

Todd, M. J., Tunçel, L., & Ye, Y. (2001). Characterizations, bounds, and probabilistic analysis of two complexity measures for linear programming problems. *Mathematical Programming*, *90*, 59–69.

Turing, A. M. (1948). Rounding-off errors in matrix processes. *Quarterly Journal of Mechanics and Applied Mathematics*, *1*, 287–308.

Vershynin, R. (2006). Beyond Hirsch conjecture: walks on random polytopes and smoothed complexity of the simplex method. *FOCS 2006 (47th Annual Symposium on Foundations of Computer Science* (pp. 133–142).

von Neumann, J., & Goldstine, H. (1947). Numerical inverting of matrices of high order. *Bull. Amer. Math. Soc.*, *53*, 1021–1099.

Yamada, M., & Sugiyama, M. (2010). Dependence minimizing regression with model selection for non-linear causal inference under non-Gaussian noise. *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI2010)* (pp. 643–648). Atlanta, Georgia, USA: The AAAI Press.

Yamada, M., & Sugiyama, M. (2011). Cross-domain object matching with model selection. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS2011)* (pp. 807–815). Fort Lauderdale, Florida, USA.

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.