Mutual Information Approximation via Maximum Likelihood Estimation of Density Ratio

Taiji Suzuki
Dept. of Mathematical Informatics,
The University of Tokyo
Email: s-taiji@stat.t.u-tokyo.ac.jp

Masashi Sugiyama
Dept. of Computer Science,
Tokyo Institute of Technology.
Email: sugi@cs.titech.ac.jp

Toshiyuki Tanaka
Dept. of Systems Science,
Kyoto University
Email: tt@i.kyoto-u.ac.jp

***Abstract*—We propose a new method of approximating mutual information based on maximum likelihood estimation of a *density ratio* function. The proposed method, *Maximum Likelihood Mutual Information* (MLMI), possesses useful properties, e.g., it does not involve density estimation, the global optimal solution can be efficiently computed, it has suitable convergence properties, and model selection criteria are available. Numerical experiments show that MLMI compares favorably with existing methods.**

I. Introduction

Mutual information (MI) between two random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$

$$I(\boldsymbol{X},\boldsymbol{Y}) := \iint p_{\boldsymbol{XY}}(\boldsymbol{x},\boldsymbol{y}) \log \frac{p_{\boldsymbol{XY}}(\boldsymbol{x},\boldsymbol{y})}{p_{\boldsymbol{X}}(\boldsymbol{x}) p_{\boldsymbol{Y}}(\boldsymbol{y})} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \quad (1)$$

plays a central role in information theory and statistics since it vanishes if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent. A great deal of effort has been made to estimate MI from samples. Such MI estimators are useful in testing independence, and have been applied to various machine learning and signal processing tasks such as feature selection and independent component analysis.

In this paper, we propose a new MI estimator, *Maximum Likelihood Mutual Information* (MLMI), that can overcome the limitations of existing approaches. MLMI does not involve density estimation, but directly learns the density ratio

$$w(\boldsymbol{x},\boldsymbol{y}) := \frac{p_{\boldsymbol{XY}}(\boldsymbol{x},\boldsymbol{y})}{p_{\boldsymbol{X}}(\boldsymbol{x}) p_{\boldsymbol{Y}}(\boldsymbol{y})} \quad (2)$$

via maximum likelihood (ML) estimation. Given a density-ratio model $g(\boldsymbol{x},\boldsymbol{y})$, MI can simply be approximated by

$$\widehat{I}(X,Y) := \frac{1}{n}\sum_{i=1}^{n} \log g(\boldsymbol{x}_i,\boldsymbol{y}_i). \quad (3)$$

We prove consistency of the MLMI procedure and elucidate its rate of convergence. Estimation of the density ratio is formulated in the ML framework as a convex optimization problem. Therefore, the unique global optimal solution can be obtained efficiently. Furthermore, cross validation and information criteria are available for model selection, so that values of tuning parameters, such as the kernel width, can be adaptively determined in an objective manner. Numerical experiments show that MLMI compares favorably with existing methods.

II. New MI Estimator

In this section, we formulate the MI estimation problem as a density ratio estimation problem and propose a new MI estimation method.

Formulation: Let $\mathcal{D}_{\boldsymbol{X}} \subset \mathbb{R}^{d_{\boldsymbol{X}}}$ and $\mathcal{D}_{\boldsymbol{Y}} \subset \mathbb{R}^{d_{\boldsymbol{Y}}}$ be data domains. Suppose we are given n independent and identically distributed (i.i.d.) paired samples $\{(\boldsymbol{x}_i,\boldsymbol{y}_i) \in \mathcal{D}_{\boldsymbol{X}} \times \mathcal{D}_{\boldsymbol{Y}}\}_{i=1}^{n}$ drawn from a joint distribution with density $p_{\boldsymbol{XY}}(\boldsymbol{x},\boldsymbol{y})$. Let $p_{\boldsymbol{X}}(\boldsymbol{x})$ and $p_{\boldsymbol{Y}}(\boldsymbol{y})$ denote the marginal densities of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. The goal is to estimate MI defined by (1).

A key factor of our approach is to avoid density estimation, since density estimation is harder than estimation of MI; instead, we estimate the *density ratio* $w(\boldsymbol{x},\boldsymbol{y})$ defined by (2). An obtained density-ratio model $g(\boldsymbol{x},\boldsymbol{y})$ is then plugged into (3) to estimate MI. We model the density ratio by a linear model $g(\boldsymbol{x},\boldsymbol{y}) := \boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{x},\boldsymbol{y})$, where $\boldsymbol{\alpha} := (\alpha_1,\ldots,\alpha_b)^\top$ are parameters to be learned from samples, where $^\top$ denotes the transpose of a vector, and where $\boldsymbol{\varphi}(\boldsymbol{x},\boldsymbol{y}) := (\varphi_1(\boldsymbol{x},\boldsymbol{y}),\ldots,\varphi_b(\boldsymbol{x},\boldsymbol{y}))^\top$ are basis functions such that $\boldsymbol{\varphi}(\boldsymbol{x},\boldsymbol{y}) \geq \boldsymbol{0}_b$ for all $(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{D}_{\boldsymbol{X}} \times \mathcal{D}_{\boldsymbol{Y}}$. $\boldsymbol{0}_b$ denotes the b-dimensional all-zero vector. The basis functions $\boldsymbol{\varphi}(\boldsymbol{x},\boldsymbol{y})$ and their number b could be dependent on the samples $\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^{n}$, i.e., *kernel* models are also allowed.

Maximum Likelihood Estimation of Density Ratio: Given a density-ratio model $g(\boldsymbol{x},\boldsymbol{y})$, one may estimate the joint density $p_{\boldsymbol{XY}}(\boldsymbol{x},\boldsymbol{y})$ by $\widehat{p}_{\boldsymbol{XY}}(\boldsymbol{x},\boldsymbol{y}) := g(\boldsymbol{x},\boldsymbol{y}) p_{\boldsymbol{X}}(\boldsymbol{x}) p_{\boldsymbol{Y}}(\boldsymbol{y})$, based on which we propose to estimate the parameter $\boldsymbol{\alpha}$ of $g(\boldsymbol{x},\boldsymbol{y})$ from the samples in the ML framework: Ignoring irrelevant constants, it is equivalent to maximizing

$$\widehat{J}(\boldsymbol{\alpha}) := \frac{1}{n}\sum_{i=1}^{n} \log\left(\boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{x}_i,\boldsymbol{y}_i)\right).$$

This is our objective function, which is concave.

The density-ratio model $g(\boldsymbol{x},\boldsymbol{y})$ should be non-negative by definition. It is therefore natural to impose the constraint $g(\boldsymbol{x},\boldsymbol{y}) \geq 0$ for all $(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{D}_X \times \mathcal{D}_Y$. This can be achieved by imposing $\boldsymbol{\alpha} \geq \boldsymbol{0}_b$. In addition to non-negativity, $g(\boldsymbol{x},\boldsymbol{y})$ should be properly normalized since $\widehat{p}_{\boldsymbol{XY}}(\boldsymbol{x},\boldsymbol{y})$ is a density:

$$1 = \iint \widehat{p}_{\boldsymbol{XY}}(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \approx \frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n} \boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{x}_i,\boldsymbol{y}_j),$$

where we used the *U-statistic* for obtaining the empirical estimator. Let $\widehat{\mathcal{S}} := \{\boldsymbol{\alpha} \mid \frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n} \boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{x}_i,\boldsymbol{y}_j) =$

$1, \boldsymbol{\alpha} \geq \mathbf{0}_b\}$. Then our optimization criterion is summarized as

$$\widehat{\boldsymbol{\alpha}} := \arg\max_{\boldsymbol{\alpha} \in \widehat{\mathcal{S}}} \widehat{J}(\boldsymbol{\alpha}).$$

Let $\widehat{g}(\boldsymbol{x}, \boldsymbol{y}) := \widehat{\boldsymbol{\alpha}}^\top \boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y})$. Then our MI estimator is given by $\widehat{I}^{(\mathrm{MLMI})}(\boldsymbol{X}, \boldsymbol{Y}) := \frac{1}{n} \sum_{i=1}^n \log \widehat{g}(\boldsymbol{x}_i, \boldsymbol{y}_i)$. We call this method *Maximum Likelihood Mutual Information (MLMI)*.

III. Rates of Convergence

Parametric Cases: We start with parametric cases where the number of basis functions b is fixed. We derive the asymptotic distribution of learned parameter $\widehat{\boldsymbol{\alpha}}$, the convergence rate of MI estimation, and the learning curve of MLMI. We assume the same regularity conditions as those used in [11], which is a common setup in developing asymptotic parametric theory.

First, we define some notations. Let J and $\mathcal{S}$ be the 'ideal' objective function and feasible set of MLMI, respectively:

$$J(\boldsymbol{\alpha}) := \iint p_{\boldsymbol{XY}}(\boldsymbol{x}, \boldsymbol{y}) \log(\boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y})) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y},$$
$$\mathcal{S} := \{\boldsymbol{\alpha} \mid \iint p_{\boldsymbol{X}}(\boldsymbol{x}) p_{\boldsymbol{Y}}(\boldsymbol{y}) \boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} = 1, \boldsymbol{\alpha} \geq 0\}.$$

Let $\boldsymbol{\alpha}^* := \arg\max_{\boldsymbol{\alpha} \in \mathcal{S}} J(\boldsymbol{\alpha})$ and $g^*(\boldsymbol{x}, \boldsymbol{y}) := {\boldsymbol{\alpha}^*}^\top \boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y})$. $\mathcal{S}$ is a convex polytope and the *approximating cone* $\mathcal{C}$ at $\boldsymbol{\alpha}^*$ is defined as $\mathcal{C} := \{\lambda(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \mid \boldsymbol{\alpha} \in \mathcal{S},\ \lambda \geq 0\}$. Let $\nabla_{\boldsymbol{\alpha}}$ be a partial derivative operator with respect to $\boldsymbol{\alpha}$. We denote $\nabla_{\boldsymbol{\alpha}} \log g|_{g=\boldsymbol{\alpha}^{*\top}\boldsymbol{\varphi}}$ by $\nabla_{\boldsymbol{\alpha}} \log g^*$; similarly we use $\nabla_{\boldsymbol{\alpha}} \log \widehat{g}$. Let

$$\mathcal{C}_\perp := \mathcal{C} \cap \{\boldsymbol{\delta} \mid \boldsymbol{\delta}^\top \iint p_{\boldsymbol{XY}}(\boldsymbol{x}, \boldsymbol{y}) \nabla_{\boldsymbol{\alpha}} \log g^*(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} = 0\},$$
$$\boldsymbol{G} := \iint p_{\boldsymbol{XY}}(\boldsymbol{x}, \boldsymbol{y}) \nabla_{\boldsymbol{\alpha}} \log g^*(\boldsymbol{x}, \boldsymbol{y}) \nabla_{\boldsymbol{\alpha}}^\top \log g^*(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y},$$
$$\boldsymbol{Q} := -\iint p_{\boldsymbol{XY}}(\boldsymbol{x}, \boldsymbol{y}) \nabla_{\boldsymbol{\alpha}} \nabla_{\boldsymbol{\alpha}}^\top \log g^*(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y},$$
$$c := \frac{1}{n} \iiint p_{\boldsymbol{X}}(\boldsymbol{x}) p_{\boldsymbol{X}}(\boldsymbol{x}') p_{\boldsymbol{Y}}(\boldsymbol{y}) g^*(\boldsymbol{x}, \boldsymbol{y}) g^*(\boldsymbol{x}', \boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{x}' \mathrm{d}\boldsymbol{y}$$
$$+ \frac{2}{n} \iiiint p_{\boldsymbol{XY}}(\boldsymbol{x}, \boldsymbol{y}) p_{\boldsymbol{X}}(\boldsymbol{x}') p_{\boldsymbol{Y}}(\boldsymbol{y}') g^*(\boldsymbol{x}, \boldsymbol{y}') g^*(\boldsymbol{x}', \boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} \mathrm{d}\boldsymbol{x}' \mathrm{d}\boldsymbol{y}'$$
$$+ \frac{1}{n} \iiint p_{\boldsymbol{X}}(\boldsymbol{x}) p_{\boldsymbol{Y}}(\boldsymbol{y}') p_{\boldsymbol{Y}}(\boldsymbol{y}) g^*(\boldsymbol{x}, \boldsymbol{y}) g^*(\boldsymbol{x}, \boldsymbol{y}') \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} \mathrm{d}\boldsymbol{y}' - \frac{4}{n}$$
$$\boldsymbol{\phi} := \iint p_{\boldsymbol{XY}}(\boldsymbol{x}, \boldsymbol{y}) p_{\boldsymbol{Y}}(\boldsymbol{y}') \frac{\boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y})}{g^*(\boldsymbol{x}, \boldsymbol{y})} g^*(\boldsymbol{x}, \boldsymbol{y}') \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} \mathrm{d}\boldsymbol{y}'$$
$$+ \iint p_{\boldsymbol{XY}}(\boldsymbol{x}, \boldsymbol{y}) p_{\boldsymbol{X}}(\boldsymbol{x}') \frac{\boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y})}{g^*(\boldsymbol{x}, \boldsymbol{y})} g^*(\boldsymbol{x}', \boldsymbol{y}) \mathrm{d}\boldsymbol{x}' \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y}.$$

Let $\boldsymbol{R}$ be the orthogonal projection matrix onto the linear hull of $\mathcal{C}_\perp$. Let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then we have the following theorem.

Theorem 1: $\sqrt{n}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)$ converges in law to $\arg\min_{\boldsymbol{\delta} \in \mathcal{C}_\perp} \|\boldsymbol{\delta} - Z\|_{\boldsymbol{Q}} + \boldsymbol{\alpha}^* Z'$, where $\|\boldsymbol{\alpha}\|_{\boldsymbol{Q}}^2 := \boldsymbol{\alpha}^\top \boldsymbol{Q} \boldsymbol{\alpha}$, and where $Z \in \mathbb{R}^b$ and $Z' \in \mathbb{R}$ are random variables such that $\begin{bmatrix} Z \\ Z' \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}_{b+1}, \begin{bmatrix} \boldsymbol{R}\boldsymbol{Q}^{-1}\boldsymbol{R}\boldsymbol{G}\boldsymbol{R}\boldsymbol{Q}^{-1}\boldsymbol{R} & \boldsymbol{R}\boldsymbol{\phi} \\ \boldsymbol{\phi}^\top \boldsymbol{R} & c \end{bmatrix}\right)$.

We analyze the convergence rate of our MI estimator. Let

$$I(g) := \iint p_{\boldsymbol{XY}}(\boldsymbol{x}, \boldsymbol{y}) \log g(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y},$$
$$\widehat{I}(g) := \frac{1}{n} \sum_{i=1}^n \log g(\boldsymbol{x}_i, \boldsymbol{y}_i).$$

Let $\mathcal{O}_p$ be the asymptotic order in probability. Then, we have

Theorem 2: $|\widehat{I}(\widehat{g}) - I(w)| = |I(g^*) - I(w)| + \mathcal{O}_p(n^{-\frac{1}{2}})$.

Corollary 1: If the parametric model contains the true density ratio, i.e., $g^* = w$, then $|\widehat{I}(\widehat{g}) - I(w)| = \mathcal{O}_p(n^{-\frac{1}{2}})$.

The corollary means that MLMI retains optimality in terms of the order of convergence in n, since $\mathcal{O}_p(n^{-\frac{1}{2}})$ is the optimal convergence rate in the parametric setup.

Let $\mathbb{E}$ denote expectation over random samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$. The behavior of $\mathbb{E}[I(\widehat{g})]$ as a function of the number of samples n is called the *learning curve*. The learning curve has been a quantity of interest in statistics since it can be used for deriving *information criteria*. Here, in order to avoid technical difficulties, we assume in the analysis of the learning curve of MLMI that $\mathcal{C}_\perp$ is a linear space. Then the learning curve can be expressed as follows.

Theorem 3: $\mathbb{E}[I(\widehat{g})] = I(g^*) - \frac{\mathrm{tr}(\boldsymbol{R}\boldsymbol{G}\boldsymbol{R}(\boldsymbol{R}\boldsymbol{Q}\boldsymbol{R})^\dagger)}{2n} + \frac{c}{2n} + O(n^{-\frac{3}{2}})$, where $\dagger$ denotes the Moore-Penrose pseudo-inverse.

Non-Parametric Cases: Now we go on to elucidation of the non-parametric convergence rates.

First, we define some notations. The set of basis functions is denoted by $\mathcal{F} := \{\varphi_\theta \mid \theta \in \Theta\}$, where Θ is a parameter/index set. The set of basis functions used for estimation with n samples is characterized by a subset of the parameter set $\Theta_n \subseteq \Theta$ and denoted by $\mathcal{F}_n := \{\varphi_\theta \mid \theta \in \Theta_n\} \subset \mathcal{F}$, which can be stochastic. The set of finite linear combinations of $\mathcal{F}$ with non-negative coefficients and its bounded subset are denoted respectively by

$$\mathcal{G} := \{\textstyle\sum_l \alpha_l \varphi_{\theta_l} \mid \alpha_l \geq 0,\ \varphi_{\theta_l} \in \mathcal{F}\},$$
$$\mathcal{G}^M := \{g \in \mathcal{G} \mid \|g\|_\infty \leq M\}.$$

Their subsets used for estimation with n samples are denoted by $\mathcal{G}_n := \{\sum_l \alpha_l \varphi_{\theta_l} \mid \alpha_l \geq 0,\ \varphi_{\theta_l} \in \mathcal{F}_n\} \subset \mathcal{G}$. Let $\widehat{\mathcal{G}}_n$ and $\widehat{g}_n$ be the feasible set and the solution of MLMI:

$$\widehat{\mathcal{G}}_n := \{g \in \mathcal{G}_n \mid \tfrac{1}{n(n-1)} \textstyle\sum_{1 \leq i \neq j \leq n} g(\boldsymbol{x}_i, \boldsymbol{y}_j) = 1\},$$
$$\widehat{g}_n := \arg\max_{g \in \widehat{\mathcal{G}}_n} \left[\tfrac{1}{n} \textstyle\sum_{i=1}^n \log g(\boldsymbol{x}_i, \boldsymbol{y}_i)\right].$$

For simplicity, we assume that the solution $\widehat{g}_n$ is unique.

Here we use the (generalized) *Hellinger distance* with respect to $p_{\boldsymbol{X}} p_{\boldsymbol{Y}}$ as the error metric, since this allows us to avoid some technical difficulties:

$$h_Q(g, g') := \left(\iint p_{\boldsymbol{X}}(\boldsymbol{x}) p_{\boldsymbol{Y}}(\boldsymbol{y}) (\sqrt{g} - \sqrt{g'})^2 \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y}\right)^{1/2},$$

where g and g' are non-negative measurable functions (not necessarily probability densities). We further make the following assumptions.

1. On the support of $p_{\boldsymbol{XY}}$, there exists a constant $\eta < \infty$ such that the true density ratio w is upper-bounded as $w(\boldsymbol{x}, \boldsymbol{y}) \leq \eta$.
2. All basis functions are non-negative, and there exist constants $\epsilon,\ \xi > 0$ such that $\forall \varphi \in \mathcal{F},\ \|\varphi\|_\infty \leq \xi$ and $\iint p_{\boldsymbol{X}}(\boldsymbol{x}) p_{\boldsymbol{Y}}(\boldsymbol{y}) \varphi(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} \geq \epsilon$.
3. There exist constants $0 \leq \gamma < 2$ and $K > 0$ such that $\log N_{[]}(\epsilon, \mathcal{G}^M, h_Q) \leq K(\sqrt{M}/\epsilon)^\gamma$, where $N_{[]}(\epsilon, \mathcal{F}, h)$ is the *ϵ-bracketing number* of $\mathcal{F}$ with distance h [9].

Assumption 3 ensures that the model is not so complicated. Gaussian mixture models satisfy this condition [2]. Let $g_n^* := \arg\max_{g \in \widehat{\mathcal{G}}_n} \iint p_{\boldsymbol{XY}}(\boldsymbol{x}, \boldsymbol{y}) \log g(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y}$. Then we have the following theorems.

Theorem 4: If there exist c_0, c_1 such that $P_{\{(x_i, y_i)\}}(c_0 \leq \frac{w}{g_n^*} \leq c_1) \to 1$, then $h_Q(\widehat{g}_n, w) = \mathcal{O}_p(n^{-\frac{1}{2+\gamma}} + h_Q(g_n^*, w))$.

Theorem 5: If there exists $\delta > 0$ such that $g(\boldsymbol{x}, \boldsymbol{y}) \geq \delta$ for $\forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}_{\boldsymbol{X}} \times \mathcal{D}_{\boldsymbol{Y}}$ and $\forall g \in \widehat{\mathcal{G}}_n$, then $|\widehat{I}(\widehat{g}_n) - I(w)| = |I(g_n^*) - I(w)| + \mathcal{O}_p(n^{-\frac{1}{2+\gamma}})$.

Corollary 2: If there exists N such that $w \in \mathcal{G}_n$ for $\forall n \geq N$, and if there exists $\delta > 0$ such that $g(\boldsymbol{x}, \boldsymbol{y}) \geq \delta$ for $\forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}_{\boldsymbol{X}} \times \mathcal{D}_{\boldsymbol{Y}}$ and $\forall g \in \widehat{\mathcal{G}}_n$, then $|\widehat{I}(\widehat{g}_n) - I(w)| = \mathcal{O}_p(n^{-\frac{1}{2+\gamma}})$.

This corollary shows that the convergence rate of non-parametric MLMI is slightly slower than the parametric counterpart, but the non-parametric method requires a milder model assumption for eliminating the modeling error. According to [5], the above convergence rate achieves the optimal minimax rate under some setup. Thus the convergence property of non-parametric MLMI would be optimal in the same sense.

IV. Model Selection

We have shown that MLMI possesses preferable convergence properties. However, the practical performance of MLMI may strongly depend on the choice of the basis functions $\boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y})$. Here we show how model selection of MLMI can be carried out.

Parametric Cases: For parametric models, we provide an information criterion which can be used for model selection. Assume that $\mathcal{C}_\perp$ is a linear space for simplicity. Let $\widehat{\mathcal{H}}$ be the inclusionwise minimal face of $\widehat{\mathcal{S}}$ which contains $\widehat{\boldsymbol{\alpha}}$, and let $\widehat{\mathcal{C}}_\perp$ be the linear space defined by $\widehat{\mathcal{C}}_\perp := \{\lambda(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}) \mid \lambda \in \mathbb{R},\ \boldsymbol{\alpha} \in \widehat{\mathcal{H}}\}$. Let $\widehat{\boldsymbol{R}}$ be the orthogonal projection onto $\widehat{\mathcal{C}}_\perp$ and let

$$\widehat{I}^{(\mathrm{IC})}(\widehat{g}) := \widehat{I}(\widehat{g}) - \tfrac{1}{n}\mathrm{tr}(\widehat{\boldsymbol{R}}\widehat{\boldsymbol{G}}\widehat{\boldsymbol{R}}(\widehat{\boldsymbol{R}}\widehat{\boldsymbol{Q}}\widehat{\boldsymbol{R}})^\dagger),$$
$$\widehat{\boldsymbol{G}} := \tfrac{1}{n}\textstyle\sum_{i=1}^n \nabla_{\boldsymbol{\alpha}} \log \widehat{g}(\boldsymbol{x}_i, \boldsymbol{y}_i) \nabla_{\boldsymbol{\alpha}}^\top \log \widehat{g}(\boldsymbol{x}_i, \boldsymbol{y}_i),$$
$$\widehat{\boldsymbol{Q}} := -\tfrac{1}{n}\textstyle\sum_{i=1}^n \nabla_{\boldsymbol{\alpha}} \nabla_{\boldsymbol{\alpha}}^\top \log \widehat{g}(\boldsymbol{x}_i, \boldsymbol{y}_i).$$

Then we have the following theorem:

Theorem 6: $\mathbb{E}[\widehat{I}^{(\mathrm{IC})}(\widehat{g})] = \mathbb{E}[I(\widehat{g})] + O(n^{-\frac{3}{2}})$.

The above theorem shows that $\widehat{I}^{(\mathrm{IC})}$ is an asymptotic unbiased estimator of $\mathbb{E}[I(\widehat{g})]$ up to $O(n^{-1})$. On the other hand, the naive empirical estimator $\widehat{I}(\widehat{g})$ is asymptotically unbiased only up to $O(n^{-\frac{1}{2}})$: $\mathbb{E}[\widehat{I}(\widehat{g})] = \mathbb{E}[I(\widehat{g})] + O(n^{-1})$. Thus $\widehat{I}^{(\mathrm{IC})}$ would be a more accurate estimator of $I(\widehat{g})$ than the naive one.

For model selection, we prepare a set of model candidates (the basis functions $\boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y})$ in the current setting) and choose the one that has the largest value of $\widehat{I}^{(\mathrm{IC})}$ as the 'best' model.

Non-Parametric Cases: For non-parametric models, we cannot unfortunately utilize the theoretical results given in the previous section since the coefficient of the convergence rate is not explicitly given. Here, we propose to use numerical estimation via cross validation for model selection.

First, the samples $\{\boldsymbol{z}_i \,|\, \boldsymbol{z}_i = (\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ are divided into K disjoint subsets $\{\mathcal{Z}_k\}_{k=1}^K$ of (approximately) the same size. Then a density ratio estimator $\widehat{g}_{\mathcal{Z}_k}(\boldsymbol{x}, \boldsymbol{y})$ is obtained using $\{\mathcal{Z}_j\}_{j \neq k}$ (i.e., without $\mathcal{Z}_k$) and the score for the hold-out samples $\mathcal{Z}_k$ is computed as

$$\widehat{I}_{\mathcal{Z}_k}^{(K\text{-CV})} = \tfrac{1}{|\mathcal{Z}_k|}\textstyle\sum_{(\boldsymbol{x}', \boldsymbol{y}') \in \mathcal{Z}_k} \log \widehat{g}_{\mathcal{Z}_k}(\boldsymbol{x}', \boldsymbol{y}'),$$

where $|\mathcal{Z}_k|$ denotes the number of sample pairs in the set $\mathcal{Z}_k$. Repeat this procedure for $k = 1, \ldots, K$ and output its average

$$\widehat{I}^{(K\text{-CV})} = \tfrac{1}{K}\textstyle\sum_{k=1}^K \widehat{I}_{\mathcal{Z}_k}^{(K\text{-CV})}.$$

Then we have the following theorem.

Theorem 7: The leave-one-out cross-validation score $\widehat{I}^{(n\text{-CV})}$ is an unbiased estimate of MI learned from $(n-1)$ samples: $\mathbb{E}[\widehat{I}^{(n\text{-CV})}] = \mathbb{E}[I(\widehat{g}_{n-1})]$.

Given that $\mathbb{E}[I(\widehat{g}_{n-1})] \approx \mathbb{E}[I(\widehat{g}_n)]$, cross validation gives an *almost* unbiased estimate of MI: Similar almost unbiasedness can be established for general K-fold cross validation, but the bias may be larger. For model selection, we compute $\widehat{I}^{(K\text{-CV})}$ for all model candidates and choose the one that maximizes the cross-validation score.

Basis Function Design: A good model may be chosen by an information criterion or cross validation, given that a set of promising model candidates is prepared. As model candidates, we propose to use a Gaussian kernel model:

$$\varphi_\ell(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\tfrac{\|\boldsymbol{x} - \boldsymbol{u}_\ell\|^2}{2\sigma^2}) \exp(-\tfrac{\|\boldsymbol{y} - \boldsymbol{v}_\ell\|^2}{2\sigma^2}),$$

where $\{(\boldsymbol{u}_\ell, \boldsymbol{v}_\ell)\}_{\ell=1}^b$ are Gaussian centers; we choose the centers randomly from $\{\boldsymbol{z}_i \,|\, \boldsymbol{z}_i = (\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$.

By definition, the density ratio $\frac{p_{\boldsymbol{XY}}(\boldsymbol{x}, \boldsymbol{y})}{p_{\boldsymbol{X}}(\boldsymbol{x}) p_{\boldsymbol{Y}}(\boldsymbol{y})}$ tends to take large values if $p_{\boldsymbol{XY}}(\boldsymbol{x}, \boldsymbol{y})$ is large and $p_{\boldsymbol{X}}(\boldsymbol{x}) p_{\boldsymbol{Y}}(\boldsymbol{y})$ is small; conversely, the ratio tends to be small (i.e., close to zero) if $p_{\boldsymbol{X}}(\boldsymbol{x}) p_{\boldsymbol{Y}}(\boldsymbol{y})$ is large and $p_{\boldsymbol{XY}}(\boldsymbol{x}, \boldsymbol{y})$ is small. When a non-negative function is approximated by a Gaussian mixture model in general, many kernels may be needed in the region where the output of the target function is large; on the other hand, only a small number of kernels would be enough in the region where the output of the target function is close to zero. On the basis of this idea, we propose a heuristic to allocate many kernels in regions where $p_{\boldsymbol{XY}}(\boldsymbol{x}, \boldsymbol{y})$ is large, which can be achieved by setting the Gaussian centers at $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$.

Alternatively, we might locate Gaussian kernels on $\{(\boldsymbol{x}_i, \boldsymbol{y}_j)\}_{i,j=1}^n$, which, however, requires n^2 basis functions and therefore is computationally prohibitive when n is large. Our preliminary experiments showed that using n^2 kernels did not improve the performance, but significantly increased the computation time. For this reason, in the experiments to follow, we decided to use the above heuristic with the number of basis functions fixed to $b = \min(200, n)$ and to choose the Gaussian width σ by cross validation.

V. Relation to Existing Methods

In this section, we discuss the characteristics of existing and the proposed approaches.

Legendre-Fenchel Duality of KL-Divergence: MI (or, more generally, KL-divergence for arbitrary two probability densities) can be characterized by *Legendre-Fenchel duality* of the convex function '$-\log$' [7]. More precisely, $I(\boldsymbol{X}, \boldsymbol{Y})$ can be characterized as the solution of the following concave maximization problem [5]:

$$I(\boldsymbol{X}, \boldsymbol{Y}) = \sup_{g \geq 0}[\textstyle\int (-p_{\boldsymbol{X}}(\boldsymbol{x}) p_{\boldsymbol{Y}}(\boldsymbol{y}) g(\boldsymbol{x}, \boldsymbol{y}) + p_{\boldsymbol{XY}}(\boldsymbol{x}, \boldsymbol{y}) \log g(\boldsymbol{x}, \boldsymbol{y})) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} + 1],$$

where the supremum is taken over all non-negative measurable functions. In [5], an empirical version of this optimization

problem is solved by restricting the search space within a (Gaussian) reproducing kernel Hilbert space with a regularizer.

Our MLMI method is closely related to the method given by [5]: Indeed, if the linear model assumption and the normalization constraint are imposed, and if the expectation is approximated by the sample average, the above formulation is reduced to MLMI. However, our approach would be more advantageous in the following respects.

• Our ML formulation is equipped with built-in regularization effects due to non-negativity and normalization constraints. This allows us to avoid introducing an additional regularization parameter and contributes to reducing the computational cost.
• A non-parametric convergence rate was investigated for general KL divergence estimation in [5]. On the other hand, we elucidated the non-parametric convergence rate specifically in the context of MI estimation; this includes proper treatment of the 'decoupling' effect in $p_{\boldsymbol{X}}(\boldsymbol{x})p_{\boldsymbol{Y}}(\boldsymbol{y})$. Furthermore, the strong positivity of the true density ratio w is not required in our proof and our result also covers a situation where the true density ratio is not contained in the model.
• We derived an asymptotic distribution of the estimator in the parametric setup and elucidated the parametric convergence rate of MI estimation and the asymptotic learning curve.
• We gave a practical heuristic for designing basis functions for approximating the density ratio. This allows us to reduce the computational cost significantly.

Kernel Density Estimation (KDE) and Adaptive Histogram Methods: KDE is a non-parametric technique to estimate a density function $p(\boldsymbol{x})$ from its i.i.d. samples $\{\boldsymbol{x}_i\}_{i=1}^n$. For the Gaussian kernel, KDE is expressed as

$$\widehat{p}(\boldsymbol{x}) = \frac{1}{n(2\pi\sigma^2)^{d/2}} \sum_{i=1}^n \exp(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2\sigma^2}).$$

The performance of KDE depends on the choice of the kernel width σ, which can be optimized by cross validation. KDE-based estimation of MI can be performed using density estimates $\widehat{p}_{\boldsymbol{XY}}(\boldsymbol{x},\boldsymbol{y})$, $\widehat{p}_{\boldsymbol{X}}(\boldsymbol{x})$, and $\widehat{p}_{\boldsymbol{Y}}(\boldsymbol{y})$ obtained from $\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^n$, $\{\boldsymbol{x}_i\}_{i=1}^n$, and $\{\boldsymbol{y}_i\}_{i=1}^n$, respectively as

$$\widehat{I} = \frac{1}{n}\sum_{i=1}^n \log \frac{\widehat{p}_{\boldsymbol{XY}}(\boldsymbol{x}_i,\boldsymbol{y}_i)}{\widehat{p}_{\boldsymbol{X}}(\boldsymbol{x}_i)\widehat{p}_{\boldsymbol{Y}}(\boldsymbol{y}_i)}.$$

However, density estimation itself is known to be a hard problem, and division by estimated densities may increase estimation errors. For this reason, the KDE-based approach may not be reliable in practice.

Histogram-based estimators with data-dependent partition would be more adaptive density estimation schemes. In the context of KL divergence estimation, consistency properties of histogram-based methods, which could be regarded as implicitly estimating the ratio $\frac{p_{\boldsymbol{XY}}(\boldsymbol{x},\boldsymbol{y})}{p_{\boldsymbol{X}}(\boldsymbol{x})p_{\boldsymbol{Y}}(\boldsymbol{y})}$, have been investigated in [10], [8]. MI estimation following this line has been explored in [1]. However, such histogram-based methods may seriously suffer from the *curse of dimensionality*, and are therefore not reliable in high-dimensional problems. Furthermore, the convergence rate seems to be unexplored yet.

K-**Nearest Neighbor (KNN) Based Method:** If estimates of the entropies H are obtained, MI can be estimated via the MI-entropy identity as

$$I(\boldsymbol{X},\boldsymbol{Y}) = H(\boldsymbol{X}) + H(\boldsymbol{Y}) - H(\boldsymbol{X},\boldsymbol{Y}),$$
$$H(\boldsymbol{X}) := -\int p_{\boldsymbol{X}}(\boldsymbol{x}) \log p_{\boldsymbol{X}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

In [4], an entropy estimator that utilizes the KNN distance has been developed. Let us define the norm of $\boldsymbol{z} = (\boldsymbol{x},\boldsymbol{y})$ by $\|\boldsymbol{z}\|_{\mathrm{z}} := \max\{\|\boldsymbol{x}\|, \|\boldsymbol{y}\|\}$, where $\|\cdot\|$ denotes the Euclidean norm. Let $N_k(i)$ be the set of KNN samples of $(\boldsymbol{x}_i,\boldsymbol{y}_i)$ with respect to the norm $\|\cdot\|_{\mathrm{z}}$, and let

$$\epsilon_{\boldsymbol{X}}(i) := \max\{\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\| \mid (\boldsymbol{x}_{i'},\boldsymbol{y}_{i'}) \in N_k(i)\},$$
$$n_{\boldsymbol{X}}(i) := |\{\boldsymbol{z}_{i'} \mid \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\| \le \epsilon_{\boldsymbol{X}}(i)\}|,$$
$$\epsilon_{\boldsymbol{Y}}(i) := \max\{\|\boldsymbol{y}_i - \boldsymbol{y}_{i'}\| \mid (\boldsymbol{x}_{i'},\boldsymbol{y}_{i'}) \in N_k(i)\},$$
$$n_{\boldsymbol{Y}}(i) := |\{\boldsymbol{z}_{i'} \mid \|\boldsymbol{y}_i - \boldsymbol{y}_{i'}\| \le \epsilon_{\boldsymbol{Y}}(i)\}|.$$

Then the KNN-based MI estimator is given by

$$\widehat{I} = \psi(k) + \psi(n) - \tfrac{1}{k} - \tfrac{1}{n}\textstyle\sum_{i=1}^n [\psi(n_{\boldsymbol{X}}(i)) + \psi(n_{\boldsymbol{Y}}(i))],$$

where ψ is the *digamma* function.

An advantage of the above KNN-based method is that it does not simply replace entropies with their estimates, but it is designed to cancel errors of individual entropy estimates. A practical drawback of the KNN-based approach is that the estimation accuracy depends on the value of k and there seems no systematic strategy to choose the value of k appropriately.

Recently a KL divergence estimator utilizing KNN density estimation is proposed in [6] and its consistency has been investigated. A notable property of this estimator is that consistency of density estimation is not necessary to establish consistency of the KL divergence estimator. However, the rate of convergence seems to be still an open research issue.

Edgeworth Expansion (EDGE) Based Method: In [3], an entropy estimator based on the *Edgeworth expansion* was proposed. The basic idea is to approximate the entropy by that of the normal distribution and additional higher-order correction terms. More specifically, for a d-dimensional distribution, an estimator $\widehat{H}$ of the entropy H is given by

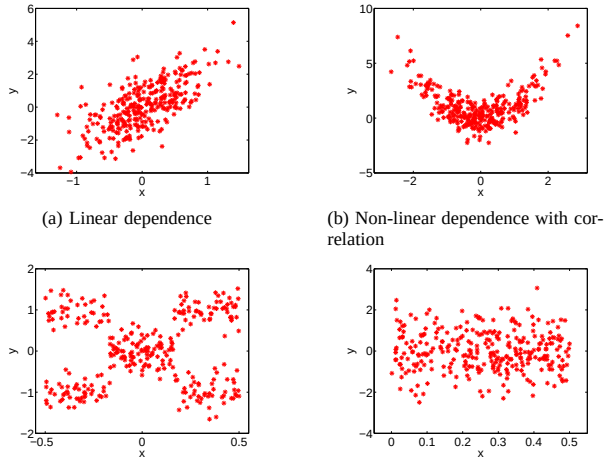
$$\widehat{H} = H^* - \textstyle\sum_{i=1}^d \frac{\kappa_{i,i,i}^2}{12} - \sum_{i,j=1,i\neq j}^d \frac{\kappa_{i,i,j}^2}{4} - \sum_{i,j,k=1,i<j<k}^d \frac{\kappa_{i,j,k}^2}{72},$$

where H^* is the entropy of the normal distribution with covariance matrix equal to the target distribution and $\kappa_{i,j,k}$ is the standardized third cumulant of the target distribution. An estimate of MI can be obtained via the MI-entropy identity. In practice, all the cumulants should be estimated from samples.

If the underlying distribution is close to normal, the above approximation is accurate and the EDGE-based method works well. However, if the distributions are far from the normal distribution, the approximation error becomes large and therefore the EDGE-based method is no longer reliable.

VI. Numerical Experiments

In this section, we experimentally investigate the performance of the proposed and existing MI estimators. The task is to estimate MI between $X \in \mathbb{R}$ and $Y \in \mathbb{R}$. We used the following four datasets (see Figure 1):



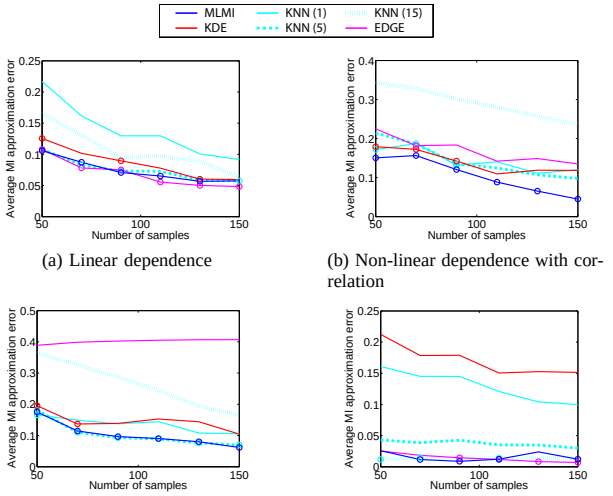
(a) Linear dependence

(b) Non-linear dependence with correlation

(c) Non-linear dependence without correlation

(d) Independence

Fig. 1. Datasets used in experiments.



(a) Linear dependence

(b) Non-linear dependence with correlation

(c) Non-linear dependence without correlation

(d) Independence

Fig. 2. MI approximation error measured by $|\widehat{I} - I|$ averaged over 100 trials as a function of the sample size n. The symbol '$\circ$' on a line means that the corresponding method is the best in terms of the average error or is judged to be comparable to the best method by the *t-test* at the significance level 1%.

(a) Linear dependence: Y has a linear dependence on X as $X \sim \mathcal{N}(0, 0.5)$ and $Y|X \sim \mathcal{N}(3X, 1)$.
(b) Non-linear dependence with correlation: Y has a quadratic dependence on X as $X \sim \mathcal{N}(0, 1)$ and $Y|X \sim \mathcal{N}(X^2, 1)$.
(c) Non-linear dependence without correlation: Y has a lattice-structured dependence on X as $X \sim \mathcal{U}(-\frac{1}{2}, \frac{1}{2})$, $Y|X \sim \mathcal{N}(0, \frac{1}{3})$ if $X \leq |\frac{1}{6}|$, and $Y|X \sim \frac{1}{2}(\mathcal{N}(1, \frac{1}{3}) + \mathcal{N}(-1, \frac{1}{3}))$ otherwise, where $\mathcal{U}(a, b)$ is the uniform density on (a, b).
(d) Independence: X and Y are independent of each other as $X \sim \mathcal{U}(0, \frac{1}{2})$ and $Y|X \sim \mathcal{N}(0, 1)$.

We compared MLMI, KDE, KNN ($k = 1, 5, 15$), and EDGE. Figure 2 depicts the average approximation errors—MLMI, KDE, KNN with $k = 5$, and EDGE performed well on the dataset (a), MLMI tends to outperform the other methods on the dataset (b), MLMI and KNN with $k = 5$ showed the best performance against the other methods on the dataset (c), and MLMI, EDGE, and KNN with $k = 15$ performed well on the dataset (d). KDE worked moderately well on the datasets (a)–(c), while it performed poorly on the dataset (d). This instability would be ascribed to division by estimated densities, which tends to magnify estimation errors. KNN seems to work well on all four datasets if the value of k is chosen optimally. As already mentioned, however, there is no systematic model selection strategy for KNN, so that KNN would be unreliable in practice. EDGE worked well on the datasets (a), (b), and (d), which possess high normality. However, for the dataset (c), where normality of the target distributions is low, the EDGE method performed poorly. In contrast, MLMI with cross validation performed reasonably well for all four datasets in a stable manner.

These experimental results have shown that MLMI nicely compensates for the weaknesses of the existing methods, and we therefore conclude that MLMI should be regarded as a useful alternative to the existing methods of MI estimation.

VII. Conclusions

We have proposed a new method of estimating mutual information. The proposed method, called MLMI, has several useful properties, e.g., it is a single-shot procedure, density estimation is not involved, it is equipped with a cross-validation procedure for model selection, and the unique global solution can be computed efficiently. We have provided a rigorous convergence analysis of the proposed algorithm as well as numerical experiments illustrating the usefulness of the proposed method.

A MATLAB® implementation of MLMI is available from 'http://sugiyama-www.cs.titech.ac.jp/˜sugi/software/MLMI/index.html'. TS acknowledges support by MEXT GCOE Program (G05). MS acknowledges supports by MEXT (2068000), JFE 21st Century Foundation, AOARD, and SCAT.

References

[1] G. A. Darbellay and I. Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321, 1999.
[2] S. Ghosal and A. W. van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Annals of Statistics*, 29:1233–1263, 2001.
[3] M. M. Van Hulle. Edgeworth approximation of multivariate differential entropy. *Neural Computation*, 17(9):1903–1910, 2005.
[4] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69:066138, 2004.
[5] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Nonparametric estimation of the likelihood ratio and divergence functionals. *IEEE International Symposium on Information Theory*, pages 2016–2020, Nice, France, 2007.
[6] F. Pérez-Cruz. Kullback-Leibler divergence estimation of continuous distributions. *IEEE International Symposium on Information Theory*, pages 1666–1670, Toronto, Canada, 2008.
[7] G. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.
[8] J. Silva and S. Narayanan. Universal consistency of data-driven partitions for divergence estimation. *IEEE International Symposium on Information Theory*, pages 2021–2025, Nice, France, 2007.
[9] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes. With Applications to Statistics*. Springer, New York, 1996.
[10] Q. Wang, S. R. Kulkarmi, and S. Verdú. Divergence estimation of contiunous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, 51(9):3064–3074, 2005.
[11] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.