Annals of the Institute of Statistical Mathematics, vol.60, no.4, pp.699–746, 2008.

Direct Importance Estimation for Covariate Shift Adaptation*

Masashi Sugiyama
Department of Computer Science, Tokyo Institute of Technology
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan
sugi@cs.titech.ac.jp

Taiji Suzuki
Department of Mathematical Informatics, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
s-taiji@stat.t.u-tokyo.ac.jp

Shinichi Nakajima
Nikon Corporation
201-9 Oaza-Miizugahara, Kumagaya-shi, Saitama 360-8559, Japan
nakajima.s@nikon.co.jp

Hisashi Kashima
IBM Research, Tokyo Research Laboratory
1623-14 Shimotsuruma, Yamato-shi, Kanagawa 242-8502, Japan
hkashima@jp.ibm.com

Paul von Bünau
Department of Computer Science, Technical University Berlin
Franklinstr. 28/29, 10587 Berlin, Germany
buenau@cs.tu-berlin.de

Motoaki Kawanabe
Fraunhofer FIRST.IDA
Kekuléstr. 7, 12489 Berlin, Germany
motoaki.kawanabe@first.fraunhofer.de

*This is an extended version of an earlier conference paper (Sugiyama et al., 2008). A MATLAB implementation of the proposed importance estimation method is available from 'http://sugiyama-www.cs.titech.ac.jp/~sugi/software/KLIEP'.

Abstract

A situation where training and test samples follow different input distributions is called *covariate shift*. Under covariate shift, standard learning methods such as maximum likelihood estimation are no longer consistent—weighted variants according to the ratio of test and training input densities are consistent. Therefore, accurately estimating the density ratio, called the *importance*, is one of the key issues in covariate shift adaptation. A naive approach to this task is to first estimate training and test input densities separately and then estimate the importance by taking the ratio of the estimated densities. However, this naive approach tends to perform poorly since density estimation is a hard task particularly in high dimensional cases. In this paper, we propose a direct importance estimation method that does not involve density estimation. Our method is equipped with a natural cross validation procedure and hence tuning parameters such as the kernel width can be objectively optimized. Furthermore, we give rigorous mathematical proofs for the convergence of the proposed algorithm. Simulations illustrate the usefulness of our approach.

Keywords

Covariate shift, Importance sampling, Model misspecification, Kullback-Leibler divergence, Likelihood cross validation.

1 Introduction

A common assumption in supervised learning is that training and test samples follow the *same* distribution. However, this basic assumption is often violated in practice and then standard machine learning methods do not work as desired. A situation where the input distribution $P(\boldsymbol{x})$ is different in the training and test phases but the conditional distribution of output values, $P(y|\boldsymbol{x})$, remains unchanged is called *covariate shift* (Shimodaira, 2000). In many real-world applications such as robot control (Sutton and Barto, 1998; Shelton, 2001; Hachiya et al., 2008), bioinformatics (Baldi and Brunak, 1998; Borgwardt et al., 2006), spam filtering (Bickel and Scheffer, 2007), brain-computer interfacing (Wolpaw et al., 2002; Sugiyama et al., 2007), or econometrics (Heckman, 1979), covariate shift is conceivable and thus learning under covariate shift is gathering a lot of attention these days.

The influence of covariate shift could be alleviated by weighting the log likelihood terms according to the *importance* (Shimodaira, 2000):

$$w(\boldsymbol{x}) := \frac{p_{\mathrm{te}}(\boldsymbol{x})}{p_{\mathrm{tr}}(\boldsymbol{x})},$$

where $p_{\mathrm{te}}(\boldsymbol{x})$ and $p_{\mathrm{tr}}(\boldsymbol{x})$ are test and training input densities. Since the importance is usually unknown, the key issue of covariate shift adaptation is how to accurately estimate the importance[1].

A naive approach to importance estimation would be to first estimate the training and test densities separately from training and test input samples, and then estimate the importance by taking the ratio of the estimated densities. However, density estimation is known to be a hard problem particularly in high-dimensional cases (Härdle et al., 2004). Therefore, this naive approach may not be effective—directly estimating the importance *without* estimating the densities would be more promising.

Following this spirit, the kernel mean matching (KMM) method has been proposed recently (Huang et al., 2007), which directly gives importance estimates without going through density estimation. KMM is shown to work well, given that tuning parameters such as the kernel width are chosen appropriately. Intuitively, model selection of importance estimation algorithms (such as KMM) is straightforward by cross validation (CV) over the performance of subsequent learning algorithms. However, this is highly unreliable since the ordinary CV score is heavily biased under covariate shift—for unbiased estimation of the prediction performance of subsequent learning algorithms, the CV procedure itself needs to be importance-weighted (Sugiyama et al., 2007). Since the importance weight has to have been fixed when model selection is carried out by importance weighted CV, it can not be used for model selection of importance estimation algorithms[2].

The above fact implies that model selection of importance estimation algorithms should be performed *within* the importance estimation step in an unsupervised manner. However, since KMM can only estimate the values of the importance at training input points, it can not be directly applied in the CV framework; an out-of-sample extension is needed, but this seems to be an open research issue currently.

In this paper, we propose a new importance estimation method which can overcome the above problems, i.e., the proposed method directly estimates the importance without density estimation *and* is equipped with a natural model selection procedure. Our basic idea is to find an importance estimate $\widehat{w}(\boldsymbol{x})$ such that the Kullback-Leibler divergence from the true test input density $p_{\mathrm{te}}(\boldsymbol{x})$ to its estimate $\widehat{p}_{\mathrm{te}}(\boldsymbol{x}) = \widehat{w}(\boldsymbol{x})p_{\mathrm{tr}}(\boldsymbol{x})$ is minimized. We propose an algorithm that can carry out this minimization without explicitly modeling $p_{\mathrm{tr}}(\boldsymbol{x})$ and $p_{\mathrm{te}}(\boldsymbol{x})$. We call the proposed method the *Kullback-Leibler Importance Estimation Procedure* (KLIEP). The optimization problem involved in KLIEP is convex,

[1] Covariate shift matters in parameter learning only when the model used for function learning is *misspecified* (i.e., the model is so simple that the true learning target function can not be expressed) (Shimodaira, 2000)—when the model is correctly (or overly) specified, ordinary maximum likelihood estimation is still consistent. Following this fact, there is a criticism that importance weighting is not needed; just the use of a complex enough model can settle the problem. However, too complex models result in huge variance and thus we practically need to choose a complex enough but not too complex model. For choosing such an appropriate model, we usually use a model selection technique such as cross validation (CV). However, the ordinary CV score is heavily biased due to covariate shift and we also need to importance-weight the CV score (or any other model selection criteria) for unbiasedness (Shimodaira, 2000; Sugiyama and Müller, 2005; Sugiyama et al., 2007). For this reason, estimating the importance is indispensable when covariate shift occurs.

[2] Once the importance weight has been fixed, importance weighted CV can be used for model selection of subsequent learning algorithms.

so the unique global solution can be obtained. Furthermore, the solution tends to be sparse, which contributes to reducing the computational cost in the test phase.

Since KLIEP is based on the minimization of the Kullback-Leibler divergence, its model selection can be naturally carried out through a variant of likelihood CV, which is a standard model selection technique in density estimation (Härdle et al., 2004). A key advantage of our CV procedure is that, not the training samples, but the *test input samples* are cross-validated. This highly contributes to improving the model selection accuracy when the number of training samples is limited but test input samples are abundantly available.

The simulation studies show that KLIEP tends to outperform existing approaches in importance estimation including the logistic regression based method (Bickel et al., 2007), and it contributes to improving the prediction performance in covariate shift scenarios.

2 New Importance Estimation Method

In this section, we propose a new importance estimation method.

2.1 Formulation and Notation

Let $\mathcal{D} \subset (\mathbb{R}^d)$ be the input domain and suppose we are given i.i.d. training input samples $\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$ from a training input distribution with density $p_{\mathrm{tr}}(\boldsymbol{x})$ and i.i.d. test input samples $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$ from a test input distribution with density $p_{\mathrm{te}}(\boldsymbol{x})$. We assume that $p_{\mathrm{tr}}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathcal{D}$. The goal of this paper is to develop a method of estimating the *importance* $w(\boldsymbol{x})$ from $\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$ and $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$: [3]

$$w(\boldsymbol{x}) := \frac{p_{\mathrm{te}}(\boldsymbol{x})}{p_{\mathrm{tr}}(\boldsymbol{x})}.$$

Our key restriction is that we avoid estimating densities $p_{\mathrm{te}}(\boldsymbol{x})$ and $p_{\mathrm{tr}}(\boldsymbol{x})$ when estimating the importance $w(\boldsymbol{x})$.

2.2 Kullback-Leibler Importance Estimation Procedure (KLIEP)

Let us model the importance $w(\boldsymbol{x})$ by the following linear model:

$$\widehat{w}(\boldsymbol{x}) = \sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}), \tag{1}$$

[3]Importance estimation is a pre-processing step of supervised learning tasks where training *output* samples $\{y_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$ at the training input points $\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$ are also available (Shimodaira, 2000; Sugiyama and Müller, 2005; Huang et al., 2007; Sugiyama et al., 2007). However, we do not use $\{y_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$ in the importance estimation step since they are irrelevant to the importance.

where $\{\alpha_\ell\}_{\ell=1}^{b}$ are parameters to be learned from data samples and $\{\varphi_\ell(\boldsymbol{x})\}_{\ell=1}^{b}$ are basis functions such that

$$\varphi_\ell(\boldsymbol{x}) \geq 0 \text{ for all } \boldsymbol{x} \in \mathcal{D} \text{ and for } \ell = 1, 2, \ldots, b.$$

Note that b and $\{\varphi_\ell(\boldsymbol{x})\}_{\ell=1}^{b}$ could be dependent on the samples $\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$ and $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$, i.e., *kernel* models are also allowed—we explain how the basis functions $\{\varphi_\ell(\boldsymbol{x})\}_{\ell=1}^{b}$ are chosen in Section 2.3.

Using the model $\widehat{w}(\boldsymbol{x})$, we can estimate the test input density $p_{\mathrm{te}}(\boldsymbol{x})$ by

$$\widehat{p}_{\mathrm{te}}(\boldsymbol{x}) = \widehat{w}(\boldsymbol{x}) p_{\mathrm{tr}}(\boldsymbol{x}).$$

We determine the parameters $\{\alpha_\ell\}_{\ell=1}^{b}$ in the model (1) so that the Kullback-Leibler divergence from $p_{\mathrm{te}}(\boldsymbol{x})$ to $\widehat{p}_{\mathrm{te}}(\boldsymbol{x})$ is minimized[4]:

$$\begin{aligned}
\mathrm{KL}[p_{\mathrm{te}}(\boldsymbol{x}) \| \widehat{p}_{\mathrm{te}}(\boldsymbol{x})] &= \int_{\mathcal{D}} p_{\mathrm{te}}(\boldsymbol{x}) \log \frac{p_{\mathrm{te}}(\boldsymbol{x})}{\widehat{w}(\boldsymbol{x}) p_{\mathrm{tr}}(\boldsymbol{x})} d\boldsymbol{x} \\
&= \int_{\mathcal{D}} p_{\mathrm{te}}(\boldsymbol{x}) \log \frac{p_{\mathrm{te}}(\boldsymbol{x})}{p_{\mathrm{tr}}(\boldsymbol{x})} d\boldsymbol{x} - \int_{\mathcal{D}} p_{\mathrm{te}}(\boldsymbol{x}) \log \widehat{w}(\boldsymbol{x}) d\boldsymbol{x}. \qquad (2)
\end{aligned}$$

Since the first term in Eq.(2) is independent of $\{\alpha_\ell\}_{\ell=1}^{b}$, we ignore it and focus on the second term. We denote it by J:

$$\begin{aligned}
J &:= \int_{\mathcal{D}} p_{\mathrm{te}}(\boldsymbol{x}) \log \widehat{w}(\boldsymbol{x}) d\boldsymbol{x} \qquad (3) \\
&\approx \frac{1}{n_{\mathrm{te}}} \sum_{j=1}^{n_{\mathrm{te}}} \log \widehat{w}(\boldsymbol{x}_j^{\mathrm{te}}) = \frac{1}{n_{\mathrm{te}}} \sum_{j=1}^{n_{\mathrm{te}}} \log \left(\sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}_j^{\mathrm{te}}) \right),
\end{aligned}$$

where the empirical approximation based on the test input samples $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$ is used from the first line to the second line above. This is our objective function to be maximized with respect to the parameters $\{\alpha_\ell\}_{\ell=1}^{b}$, which is concave (Boyd and Vandenberghe, 2004). Note that the above objective function only involves the test input samples $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$, i.e., we did not use the training input samples $\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$ yet. As shown below, $\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$ will be used in the constraint.

$\widehat{w}(\boldsymbol{x})$ is an estimate of the importance $w(\boldsymbol{x})$ which is non-negative by definition. Therefore, it is natural to impose $\widehat{w}(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathcal{D}$, which can be achieved by restricting

$$\alpha_\ell \geq 0 \text{ for } \ell = 1, 2, \ldots, b.$$

[4]One may also consider an alternative scenario where the inverse importance $w^{-1}(\boldsymbol{x})$ is parameterized and the parameters are learned so that the Kullback-Leibler divergence from $p_{\mathrm{tr}}(\boldsymbol{x})$ to $\widehat{p}_{\mathrm{tr}}(\boldsymbol{x})$ $(= \widehat{w}^{-1}(\boldsymbol{x}) p_{\mathrm{te}}(\boldsymbol{x}))$ is minimized. We may also consider using $\mathrm{KL}[\widehat{p}_{\mathrm{te}}(\boldsymbol{x}) \| p_{\mathrm{te}}(\boldsymbol{x})]$—however, this involves the model $\widehat{w}(\boldsymbol{x})$ in a more complex manner and does not seem to result in a simple optimization problem.

In addition to the non-negativity, $\widehat{w}(\boldsymbol{x})$ should be properly normalized since $\widehat{p}_{\mathrm{te}}(\boldsymbol{x})$ $(= \widehat{w}(\boldsymbol{x})p_{\mathrm{tr}}(\boldsymbol{x}))$ is a probability density function:

$$\begin{aligned} 1 &= \int_{\mathcal{D}} \widehat{p}_{\mathrm{te}}(\boldsymbol{x})d\boldsymbol{x} = \int_{\mathcal{D}} \widehat{w}(\boldsymbol{x})p_{\mathrm{tr}}(\boldsymbol{x})d\boldsymbol{x} \\ &\approx \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \widehat{w}(\boldsymbol{x}_i^{\mathrm{tr}}) = \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i^{\mathrm{tr}}), \end{aligned} \tag{4}$$

where the empirical approximation based on the training input samples $\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$ is used from the first line to the second line above.

Now our optimization criterion is summarized as follows.

$$\begin{aligned} &\underset{\{\alpha_\ell\}_{\ell=1}^{b}}{\text{maximize}} \left[\sum_{j=1}^{n_{\mathrm{te}}} \log \left(\sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}_j^{\mathrm{te}}) \right) \right] \\ &\qquad \text{subject to } \sum_{i=1}^{n_{\mathrm{tr}}} \sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i^{\mathrm{tr}}) = n_{\mathrm{tr}} \text{ and } \alpha_1, \alpha_2, \ldots, \alpha_b \geq 0. \end{aligned}$$

This is a convex optimization problem and the global solution can be obtained, e.g., by simply performing gradient ascent and feasibility satisfaction iteratively[5]. A pseudo code is described in Figure 1(a). Note that the solution $\{\widehat{\alpha}_\ell\}_{\ell=1}^{b}$ tends to be *sparse* (Boyd and Vandenberghe, 2004), which contributes to reducing the computational cost in the test phase. We refer to the above method as *Kullback-Leibler Importance Estimation Procedure* (KLIEP).

2.3 Model Selection by Likelihood Cross Validation

The performance of KLIEP depends on the choice of basis functions $\{\varphi_\ell(\boldsymbol{x})\}_{\ell=1}^{b}$. Here we explain how they can be appropriately chosen from data samples.

Since KLIEP is based on the maximization of the score J (see Eq.(3)), it would be natural to select the model such that J is maximized. The expectation over $p_{\mathrm{te}}(\boldsymbol{x})$ involved in J can be numerically approximated by *likelihood cross validation* (LCV) as follows: First, divide the test samples $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$ into R disjoint subsets $\{\mathcal{X}_r^{\mathrm{te}}\}_{r=1}^{R}$. Then obtain an importance estimate $\widehat{w}_r(\boldsymbol{x})$ from $\{\mathcal{X}_j^{\mathrm{te}}\}_{j \neq r}$ and approximate the score J using $\mathcal{X}_r^{\mathrm{te}}$ as

$$\widehat{J}_r := \frac{1}{|\mathcal{X}_r^{\mathrm{te}}|} \sum_{\boldsymbol{x} \in \mathcal{X}_r^{\mathrm{te}}} \log \widehat{w}_r(\boldsymbol{x}).$$

[5]If necessary, we may *regularize* the solution, e.g., by adding a penalty term (say, $\sum_{\ell=1}^{b} \alpha_\ell^2$) to the objective function or by imposing an upper bound on the solution. The normalization constraint (4) may also be weakened by allowing a small deviation. These modification is possible without sacrificing the convexity.

Input: $m = \{\varphi_\ell(\boldsymbol{x})\}_{\ell=1}^{b}$, $\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$, and $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$
Output: $\widehat{w}(\boldsymbol{x})$

$A_{j,\ell} \longleftarrow \varphi_\ell(\boldsymbol{x}_j^{\mathrm{te}}) \quad$ for $j = 1, 2, \ldots, n_{\mathrm{te}}$ and $\ell = 1, 2, \ldots, b$;
$b_\ell \longleftarrow \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \varphi_\ell(\boldsymbol{x}_i^{\mathrm{tr}}) \quad$ for $j = 1, 2, \ldots, n_{\mathrm{te}}$;
Initialize $\boldsymbol{\alpha}$ $(> \boldsymbol{0})$ and ε $(0 < \varepsilon \ll 1)$;
Repeat until convergence
    $\boldsymbol{\alpha} \longleftarrow \boldsymbol{\alpha} + \varepsilon \boldsymbol{A}^\top (\boldsymbol{1}./\boldsymbol{A}\boldsymbol{\alpha})$; % Gradient ascent
    $\boldsymbol{\alpha} \longleftarrow \boldsymbol{\alpha} + (1 - \boldsymbol{b}^\top \boldsymbol{\alpha})\boldsymbol{b}/(\boldsymbol{b}^\top \boldsymbol{b})$; % Constraint satisfaction
    $\boldsymbol{\alpha} \longleftarrow \max(\boldsymbol{0}, \boldsymbol{\alpha})$; % Constraint satisfaction
    $\boldsymbol{\alpha} \longleftarrow \boldsymbol{\alpha}/(\boldsymbol{b}^\top \boldsymbol{\alpha})$; % Constraint satisfaction
end
$\widehat{w}(\boldsymbol{x}) \longleftarrow \sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x})$;

(a) KLIEP main code

Input: $\mathcal{M} = \{m_k \mid m_k = \{\varphi_\ell^{(k)}(\boldsymbol{x})\}_{\ell=1}^{b^{(k)}}\}$, $\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$, and $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$
Output: $\widehat{w}(\boldsymbol{x})$

Split $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$ into R disjoint subsets $\{\mathcal{X}_r^{\mathrm{te}}\}_{r=1}^{R}$;
for each model $m \in \mathcal{M}$
    for each split $r = 1, 2, \ldots, R$
        $\widehat{w}_r(\boldsymbol{x}) \longleftarrow \mathrm{KLIEP}(m, \{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}, \{\mathcal{X}_j^{\mathrm{te}}\}_{j \neq r})$;
        $\widehat{J}_r(m) \longleftarrow \frac{1}{|\mathcal{X}_r^{\mathrm{te}}|} \sum_{\boldsymbol{x} \in \mathcal{X}_r^{\mathrm{te}}} \log \widehat{w}_r(\boldsymbol{x})$;
    end
    $\widehat{J}(m) \longleftarrow \frac{1}{R} \sum_{r=1}^{R} \widehat{J}_r(m)$;
end
$\widehat{m} \longleftarrow \mathrm{argmax}_{m \in \mathcal{M}} \widehat{J}(m)$;
$\widehat{w}(\boldsymbol{x}) \longleftarrow \mathrm{KLIEP}(\widehat{m}, \{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}, \{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}})$;

(b) Model selection by LCV

Figure 1: The KLIEP algorithm in pseudo code. './' indicates the element-wise division and $^\top$ denotes the transpose. Inequalities and the 'max' operation for vectors are applied element-wise. A MATLAB implementation of the KLIEP algorithm is available from '`http://sugiyama-www.cs.titech.ac.jp/~sugi/software/KLIEP`'.

We repeat this procedure for $r = 1, 2, \ldots, R$, compute the average of $\widehat{J}_r$ over all r, and use the average $\widehat{J}$ as an estimate of J:

$$\widehat{J} := \frac{1}{R} \sum_{r=1}^{R} \widehat{J}_r. \tag{5}$$

For model selection, we compute $\widehat{J}$ for all model candidates (the basis functions $\{\varphi_\ell(\boldsymbol{x})\}_{\ell=1}^{b}$ in the current setting) and choose the one that minimizes $\widehat{J}$. A pseudo code of the LCV procedure is summarized in Figure 1(b).

One of the potential limitations of CV in general is that it is not reliable in small sample cases since data splitting by CV further reduces the sample size. On the other hand, in our CV procedure, the data splitting is performed only over the *test input samples*, not over the training samples. Therefore, even when the number of training samples is small, our CV procedure does not suffer from the small sample problem as long as a large number of test input samples are available.

A good model may be chosen by the above CV procedure, given that a set of promising model candidates is prepared. As model candidates, we propose using a Gaussian kernel model centered at the *test* input points $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$, i.e.,

$$\widehat{w}(\boldsymbol{x}) = \sum_{\ell=1}^{n_{\mathrm{te}}} \alpha_\ell K_\sigma(\boldsymbol{x}, \boldsymbol{x}_\ell^{\mathrm{te}}),$$

where $K_\sigma(\boldsymbol{x}, \boldsymbol{x}')$ is the Gaussian kernel with kernel width σ:

$$K_\sigma(\boldsymbol{x}, \boldsymbol{x}') := \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma^2}\right). \tag{6}$$

The reason why we chose the test input points $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$ as the Gaussian centers, not the training input points $\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$, is as follows. By definition, the importance $w(\boldsymbol{x})$ tends to take large values if the training input density $p_{\mathrm{tr}}(\boldsymbol{x})$ is small and the test input density $p_{\mathrm{te}}(\boldsymbol{x})$ is large; conversely, $w(\boldsymbol{x})$ tends to be small (i.e., close to zero) if $p_{\mathrm{tr}}(\boldsymbol{x})$ is large and $p_{\mathrm{te}}(\boldsymbol{x})$ is small. When a function is approximated by a Gaussian kernel model, 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. Following this heuristic, we decided to allocate many kernels at high *test* input density regions, which can be achieved by setting the Gaussian centers at the test input points $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$.

Alternatively, we may locate $(n_{\mathrm{tr}}+n_{\mathrm{te}})$ Gaussian kernels at both $\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$ and $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$. However, in our preliminary experiments, this did not further improve the performance, but slightly increased the computational cost. When n_{te} is very large, just using all the test input points $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$ as Gaussian centers is already computationally rather demanding. To ease this problem, we practically propose using a subset of $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$ as Gaussian centers for computational efficiency, i.e.,

$$\widehat{w}(\boldsymbol{x}) = \sum_{\ell=1}^{b} \alpha_\ell K_\sigma(\boldsymbol{x}, \boldsymbol{c}_\ell), \tag{7}$$

where $\boldsymbol{c}_\ell$ is a template point randomly chosen from $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$ and b $(\leq n_{\mathrm{te}})$ is a prefixed number.

3 Theoretical Analyses

In this section, we investigate the convergence properties of the KLIEP algorithm. The theoretical statements we prove in this section are roughly summarized as follows.

- When a non-parametric model (e.g., kernel basis functions centered at test samples) is used for importance estimation, KLIEP converges to the optimal solution with convergence rate slightly slower than $\mathcal{O}_p(n^{-\frac{1}{2}})$ under $n = n_{\mathrm{tr}} = n_{\mathrm{te}}$ (Theorem 1 and Theorem 2).

- When a fixed set of basis functions is used for importance estimation, KLIEP converges to the optimal solution with convergence rate $\mathcal{O}_p(n^{-\frac{1}{2}})$. Furthermore, KLIEP has asymptotic normality around the optimal solution (Theorem 3 and Theorem 4).

3.1 Mathematical Preliminaries

Since we give rigorous mathematical convergence proofs, we first slightly change our notation for clearer mathematical exposition.

Below, we assume that the numbers of training and test samples are the same, i.e.,

$$n = n_{\mathrm{te}} = n_{\mathrm{tr}}.$$

We note that this assumption is just for simplicity; without this assumption, the convergence rate is solely determined by the sample size with the slower rate.

For arbitrary measure $\tilde{P}$ and $\tilde{P}$-integrable function f, we express its "expectation" as

$$\tilde{P}f := \int f \mathrm{d}\tilde{P}.$$

Let P and Q be the probability measures which generate test and training samples, respectively. In a similar fashion, we define the empirical distributions of test and training samples by P_n and Q_n, i.e.,

$$P_n f = \frac{1}{n}\sum_{j=1}^{n} f(\boldsymbol{x}_j^{\mathrm{te}}), \quad Q_n f = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i^{\mathrm{tr}}).$$

The set of basis functions is denoted by

$$\mathcal{F} := \{\varphi_\theta \mid \theta \in \Theta\},$$

where Θ is some parameter or index set. The set of basis functions at n samples are denoted using $\Theta_n \subseteq \Theta$ by

$$\mathcal{F}_n := \{\varphi_\theta \mid \theta \in \Theta_n\} \subset \mathcal{F},$$

which can behave stochastically. The set of finite linear combinations of $\mathcal{F}$ with positive coefficients and its bounded subset are denoted by

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

and their subsets at n samples are denoted by

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

Let $\hat{\mathcal{G}}_n$ be the feasible set of KLIEP:

$$\hat{\mathcal{G}}_n := \{ g \in \mathcal{G}_n \mid Q_n g = 1 \}.$$

Under the notations described above, the solution $\hat{g}_n$ of (generalized) KLIEP is given as follows:

$$\hat{g}_n := \operatorname*{arg\,max}_{g \in \hat{\mathcal{G}}_n} P_n \log (g).$$

For simplicity, we assume the optimal solution is uniquely determined. In order to derive the convergence rates of KLIEP, we make the following assumptions.

Assumption 1

1. P and Q are mutually absolutely continuous and have the following property:

$$0 < \eta_0 \leq \frac{\mathrm{d}P}{\mathrm{d}Q} \leq \eta_1$$

   on the support of P and Q. Let g_0 denote

$$g_0 := \frac{\mathrm{d}P}{\mathrm{d}Q}.$$

2. $\varphi_\theta \geq 0$ $(\forall \varphi_\theta \in \mathcal{F})$, and $\exists \epsilon_0,\ \xi_0 > 0$ such that

$$Q\varphi_\theta \geq \epsilon_0, \quad \|\varphi_\theta\|_\infty \leq \xi_0, \quad (\forall \varphi_\theta \in \mathcal{F}).$$

3. For some constants $0 < \gamma < 2$ and K,

$$\sup_{\tilde{Q}} \log N(\epsilon, \mathcal{G}^M, L_2(\tilde{Q})) \leq K \left(\frac{M}{\epsilon} \right)^\gamma, \tag{8}$$

   where the supremum is taken over all finitely discrete probability measures $\tilde{Q}$, or

$$\log N_{[]}(\epsilon, \mathcal{G}^M, L_2(Q)) \leq K \left(\frac{M}{\epsilon} \right)^\gamma. \tag{9}$$

   $N(\epsilon, \mathcal{F}, d)$ and $N_{[]}(\epsilon, \mathcal{F}, d)$ are the ϵ-covering number and the ϵ-bracketing number of $\mathcal{F}$ with norm d, respectively (van der Vaart and Wellner, 1996).

■

We define the (generalized) Hellinger distance with respect to Q as

$$h_Q(g, g') := \left(\int (\sqrt{g} - \sqrt{g'})^2 \mathrm{d}Q\right)^{1/2},$$

where g and g' are non-negative measurable functions (not necessarily probability densities). The lower bound of g_0 appeared in Assumption 1.1 will be used to ensure the existence of a Lipschitz continuous function that bounds the Hellinger distance from the true. The bound of g_0 is needed only on the support of P and Q. Assumption 1.3 controls the complexity of the model. By this complexity assumption, we can bound the tail probability of the difference between the empirical risk and the true risk uniformly over the function class $\mathcal{G}^M$.

3.2 Non-Parametric Case

First, we introduce a very important inequality that is a version of Talagrand's concentration inequality. The original form of Talagrand's concentration inequality is an inequality about the expectation of a general function $f(X_1, \ldots, X_n)$ of n variables, so the range of applications is quite large (Talagrand, 1996a,b).

Let

$$\sigma_P(\mathcal{F})^2 := \sup_{f \in \mathcal{F}} (Pf^2 - (Pf)^2).$$

For a functional $Y : \mathcal{G} \to \mathbb{R}$ defined on a set of measurable functions $\mathcal{G}$, we define its norm as

$$\|Y\|_{\mathcal{G}} := \sup_{g \in \mathcal{G}} |Y(g)|.$$

For a class $\mathcal{F}$ of measurable functions such that $\forall f \in \mathcal{F},\ \|f\|_\infty \leq 1$, the following bound holds, which we refer to as the *Bousquet bound* (Bousquet, 2002):

$$P\Big\{\|P_n - P\|_{\mathcal{F}} \geq \mathrm{E}\|P_n - P\|_{\mathcal{F}} + \sqrt{\frac{2t}{n}(\sigma_P(\mathcal{F})^2 + 2\mathrm{E}\|P_n - P\|_{\mathcal{F}})} + \frac{t}{3n}\Big\} \leq e^{-t}. \quad (10)$$

We can easily see that $\mathrm{E}\|P_n - P\|_{\mathcal{F}}$ and $\sigma_P(\mathcal{F})$ in the Bousquet bound can be replaced by other functions bounding from above. For example, $\mathrm{E}\|P_n - P\|_{\mathcal{F}}$ can be upper-bounded by the Rademacher complexity and $\sigma_P(\mathcal{F})$ can be bounded by using the $L_2(P)$-norm (Bartlett et al., 2005). By using the above inequality, we obtain the following theorem. The proof is summarized in Appendix A.

Theorem 1 *Let*

$$a_0^n := (Q_n g_0)^{-1},$$
$$\gamma_n := \max\{-P_n \log(\hat{g}_n) + P_n \log(a_0^n g_0), 0\}.$$

Then

$$h_Q(a_0^n g_0, \hat{g}_n) = \mathcal{O}_p(n^{-\frac{1}{2+\gamma}} + \sqrt{\gamma_n}).$$

The technical advantage of using the Hellinger distance instead of the KL-divergence is that the Hellinger distance is bounded from above by a Lipschitz continuous function while the KL-divergence is not Lipschitz continuous because $\log(x)$ diverges to $-\infty$ as $x \to 0$. This allows us to utilize uniform convergence results of empirical processes. See the proof for more details.

Remark 1 *If there exists N such that $\forall n \geq N$, $g_0 \in \mathcal{G}_n$, then $\gamma_n = 0$ $(\forall n \geq N)$. In this setting,*

$$h_Q(\hat{g}_n / a_0^n, g_0) = \mathcal{O}_p(n^{-\frac{1}{2+\gamma}}).$$

Remark 2 *a_0^n can be removed because*

$$\begin{aligned} h_Q(a_0^n g_0, g_0) &= \sqrt{\int g_0 (1 - \sqrt{a_0^n})^2 \mathrm{d}Q} \\ &= |1 - \sqrt{a_0^n}| = \mathcal{O}_p(1/\sqrt{n}) = \mathcal{O}_p(n^{-\frac{1}{2+\gamma}}). \end{aligned}$$

Thus,

$$h_Q(\hat{g}_n, g_0) \leq h_Q(\hat{g}_n, a_0^n g_0) + h_Q(a_0^n g_0, g_0) = \mathcal{O}_p(n^{-\frac{1}{2+\gamma}} + \sqrt{\gamma_n}).$$

We can derive another convergence theorem based on a different representation of the bias term from Theorem 1. The proof is also included in Appendix A.

Theorem 2 *In addition to Assumption 1, if there is $g_n^* \in \hat{\mathcal{G}}_n$ such that for some constant c_0, on the support of P and Q*

$$\frac{g_0}{g_n^*} \leq c_0^2,$$

then

$$h_Q(g_0, \hat{g}_n) = \mathcal{O}_p(n^{-\frac{1}{2+\gamma}} + h_Q(g_n^*, g_0)).$$

Example 1 *We briefly evaluate the convergence rate in a simple example in which $d = 1$, the support of P is $[0, 1] \subseteq \mathbb{R}$, $\mathcal{F} = \{K_1(x, x') \mid x' \in [0, 1]\}$, and $\mathcal{F}_n = \{K_1(x, x_j^{\mathrm{te}}) \mid j = 1, \ldots, n\}$ (for simplicity, we consider the case where the Gaussian width σ is 1, but we can apply the same argument to another choice of σ). Assume that P has a density $p(x)$ with a constant η_2 such that $p(x) \geq \eta_2 > 0$ $(\forall x \in [-1, 1])$. We also assume that the true importance g_0 is a mixture of Gaussian kernels, i.e.,*

$$g_0(x) = \int K_1(x, x') \mathrm{d}F(x') \quad (\forall x \in [0, 1]),$$

where F is a positive finite measure the support of which is contained in $[0, 1]$. For a measure F', we define $g_{F'}(x) := \int K_1(x, x') \mathrm{d}F'(x')$. By Lemma 3.1 of Ghosal and van der

Vaart (2001), for every $0 < \epsilon_n < 1/2$*, there exits a discrete positive finite measure* F' *on* $[0,1]$ *such that*

$$\|g_0 - g_{F'}\|_\infty \leq \epsilon_n, \qquad F'([0,1]) = F([0,1]).$$

Now divide $[0,1]$ *into bins with width* ϵ_n*, then the number of sample points* x_j^{te} *that fall in a bin is a binomial random variable. If* $\exp(-\eta_2 n\epsilon_n/4)/\epsilon_n \to 0$*, then by the Chernoff bound*[6]*, the probability of the event*

$$W_n := \{\max_j \min_{x\in\mathrm{supp}(F')} |x - x_j^{\mathrm{te}}| \leq \epsilon_n\}$$

converges to 1 ($\mathrm{supp}(F')$ *means the support of* F'*) because the density* $p(x)$ *is bounded from below across the support. One can show that* $|K_1(x,x_1) - K_1(x,x_2)| \leq |x_1 - x_2|/\sqrt{e} + |x_1 - x_2|^2/2$ $(\forall x)$ *because*

$$\begin{aligned}
&|K_1(x,x_1) - K_1(x,x_2)| \\
&\quad = \exp(-(x-x_1)^2/2)[1 - \exp(x(x_2 - x_1) + (x_1^2 - x_2^2)/2)] \\
&\quad \leq \exp(-(x-x_1)^2/2)|x(x_2 - x_1) + (x_1^2 - x_2^2)/2| \\
&\quad \leq \exp(-(x-x_1)^2/2)(|x - x_1||x_1 - x_2| + |x_1 - x_2|^2/2) \\
&\quad \leq |x_1 - x_2|/\sqrt{e} + |x_1 - x_2|^2/2.
\end{aligned}$$

Thus there exists $\tilde{\alpha}_j \geq 0$ $(j = 1,\ldots,n)$ *such that for* $\tilde{g}_n^* := \sum_j \tilde{\alpha}_j K_1(x, x_j^{\mathrm{te}})$*, the following is satisfied on the event* W_n*:* $\|\tilde{g}_n^* - g_{F'}\|_\infty \leq F'([0,1])(\epsilon_n/\sqrt{e} + \epsilon_n^2/2) = \mathcal{O}(\epsilon_n)$*. Now define*

$$g_n^* := \frac{\tilde{g}_n^*}{Q_n \tilde{g}_n^*}.$$

Then $g_n^* \in \hat{\mathcal{G}}_n$*.*

Set $\epsilon_n = 1/\sqrt{n}$*. Noticing* $|1 - Q_n\tilde{g}_n^*| = |1 - Q_n(\tilde{g}_n^* - g_{F'} + g_{F'} - g_0 + g_0)| \leq \mathcal{O}(\epsilon_n) + |1 - Q_n g_0| = \mathcal{O}_p(1/\sqrt{n})$*, we have*

$$\|g_n^* - \tilde{g}_n^*\|_\infty = \|g_n^*\|_\infty |1 - Q_n \tilde{g}_n^*| = \mathcal{O}_p(1/\sqrt{n}).$$

From the above discussion, we obtain

$$\|g_n^* - g_0\|_\infty = \mathcal{O}_p(1/\sqrt{n}).$$

This indicates

$$h_Q(g_n^*, g_0) = \mathcal{O}_p(1/\sqrt{n}),$$

and that $g_0/g_n^* \leq c_0^2$ *is satisfied with high probability.*

For the bias term of Theorem 1, set $\epsilon_n = C\log(n)/n$ *for sufficiently large* $C > 0$ *and replace* g_0 *with* $a_0^n g_0$*. Then we obtain* $\gamma_n = \mathcal{O}_p(\log(n)/n)$*.*

---

[6] Here we refer to the Chernoff bound as follows: let $\{X_i\}_{i=1}^n$ be independent random variables taking values on 0 or 1, then $P(\sum_{i=1}^n X_i < (1-\delta)\sum_{i=1}^n \mathrm{E}[X_i]) < \exp(-\delta^2 \sum_{i=1}^n \mathrm{E}[X_i]/2)$ for any $\delta > 0$.

As for the complexity of the model, a similar argument to Theorem 3.1 of Ghosal and van der Vaart (2001) gives

$$\log N(\epsilon, \mathcal{G}^M, \|\cdot\|_\infty) \le K \left(\log \frac{M}{\epsilon}\right)^2$$

for $0 < \epsilon < M/2$. *This gives both conditions (8) and (9) of Assumption 1.3 for arbitrary small* $\gamma > 0$ *(but the constant* K *depends on* γ*). Thus the convergence rate is evaluated as* $h_Q(g_0, \hat{g}_n) = \mathcal{O}_p(n^{-1/(2+\gamma)})$ *for arbitrary small* $\gamma > 0$.

3.3 Parametric Case

Next, we show asymptotic normality of KLIEP in a finite-dimensional case. We do not assume that g_0 is contained in the model, but it can be shown that KLIEP has asymptotic normality around the point that is "nearest" to the true. The finite-dimensional model we consider here is

$$\mathcal{F} = \mathcal{F}_n = \{\varphi_l \mid l = 1, \dots, b\} \quad (\forall n).$$

We define φ as

$$\varphi(x) := \begin{bmatrix} \varphi_1(x) \\ \vdots \\ \varphi_b(x) \end{bmatrix}.$$

$\mathcal{G}_n$ and $\mathcal{G}_n^M$ are independent of n and we can write them as

$$\begin{aligned} \mathcal{G}_n &= \mathcal{G} = \left\{\alpha^{\mathrm{T}}\varphi \mid \alpha \ge 0\right\}, \\ \mathcal{G}_n^M &= \mathcal{G}^M = \left\{\alpha^{\mathrm{T}}\varphi \mid \alpha \ge 0, \|\alpha^{\mathrm{T}}\varphi\|_\infty \le M\right\}. \end{aligned}$$

We define g_* as the optimal solution in the model, and α_* as the coefficient of g_*:

$$g_* := \mathop{\arg\max}_{g \in \mathcal{G}, Qg = 1} P \log g, \quad g_* = \alpha_*^{\mathrm{T}}\varphi. \tag{11}$$

In addition to Assumption 1, we assume the following conditions:

Assumption 2

1. $Q(\varphi\varphi^{\mathrm{T}}) \succ O$ (positive definite).
2. There exists $\eta_3 > 0$ such that $g_* \ge \eta_3$.

■

Let

$$\psi(\alpha)(x) = \psi(\alpha) := \log(\alpha^{\mathrm{T}}\varphi(x)).$$

Note that if $Q(\varphi\varphi^{\mathrm{T}}) \succ O$ is satisfied, then we obtain the following inequality:

$$\begin{aligned} \forall \beta \ne 0, \quad \beta^{\mathrm{T}}\nabla\nabla^{\mathrm{T}} P\psi(\alpha_*)\beta &= \beta^{\mathrm{T}}\nabla P \frac{\varphi^{\mathrm{T}}}{\alpha^{\mathrm{T}}\varphi}\Big|_{\alpha=\alpha_*}\beta = -\beta^{\mathrm{T}} P \frac{\varphi\varphi^{\mathrm{T}}}{(\alpha_*^{\mathrm{T}}\varphi)^2}\beta \\ &= -\beta^{\mathrm{T}} Q \left(\varphi\varphi^{\mathrm{T}}\frac{g_0}{g_*^2}\right)\beta \le -\beta^{\mathrm{T}} Q(\varphi\varphi^{\mathrm{T}})\beta\eta_0\epsilon_0^2/\xi_0^2 < 0. \end{aligned}$$

Thus, $-\nabla\nabla^{\mathrm{T}} P\psi(\alpha_*)$ is positive definite. We write it as

$$I_0 := -\nabla\nabla^{\mathrm{T}} P\psi(\alpha_*) \ \ (\succ O).$$

We set

$$\check{\alpha}_n := \frac{\hat{\alpha}_n}{a_*^n},$$

where $a_*^n := (Q_n g_*)^{-1}$ and $\hat{\alpha}_n^{\mathrm{T}}\varphi = \hat{g}_n$. We first show the $\sqrt{n}$-consistency of $\hat{\alpha}_n/a_*^n$ (i.e., $\|\check{\alpha}_n - \alpha_*\| = \mathcal{O}_p(1/\sqrt{n})$). From now on, let $\|\cdot\|_0$ denote a norm defined as

$$\|\alpha\|_0^2 := \alpha^{\mathrm{T}} I_0 \alpha.$$

By the positivity of I_0, there exist $0 < \xi_1 < \xi_2$ such that

$$\xi_1\|\alpha\| \leq \|\alpha\|_0 \leq \xi_2\|\alpha\|. \tag{12}$$

Lemma 1 *In a finite fixed dimensional model under Assumption 1 and Assumption 2, the KLIEP estimator satisfies*

$$\|\hat{\alpha}_n/a_*^n - \alpha_*\| = \|\check{\alpha}_n - \alpha_*\| = \mathcal{O}_p(1/\sqrt{n}).$$

From the relationship (12), this also implies $\|\check{\alpha}_n - \alpha_*\|_0 = \mathcal{O}_p(1/\sqrt{n})$*, which indicates*

$$h_Q(\hat{g}_n, a_*^n g_*) = \mathcal{O}_p(1/\sqrt{n}).$$

The proof is provided in Appendix B.

Next we discuss the asymptotic law of the KLIEP estimator. To do this we should introduce an approximating cone which is used to express the neighborhood of α_*. Let

$$\mathcal{S} := \{\alpha \mid Q\alpha^{\mathrm{T}}\varphi = 1, \alpha \geq \mathbf{0}\},$$
$$\mathcal{S}_n := \{\alpha \mid Q_n\alpha^{\mathrm{T}}\varphi = 1/a_*^n, \alpha \geq \mathbf{0}\}.$$

Note that $\alpha_* \in \mathcal{S}$ and $\check{\alpha}_n, \alpha_* \in \mathcal{S}_n$. Let the approximating cones of $\mathcal{S}$ and $\mathcal{S}_n$ at α_* be $\mathcal{C}$ and $\mathcal{C}_n$, where an approximating cone is defined in the following definition.

Definition 1 *Let D be a closed subset in $\mathbb{R}^k$ and $\theta \in D$ be a non-isolated point in D. If there is a closed cone A that satisfies the following conditions, we define A as an approximating cone at θ:*

- *For an arbitrary sequence $y_i \in D - \theta,\ y_i \to \mathbf{0}$*

$$\inf_{x\in A}\|x - y_i\| = o(\|y_i\|).$$

- *For an arbitrary sequence $x_i \in A,\ x_i \to \mathbf{0}$*

$$\inf_{y\in D-\theta}\|x_i - y\| = o(\|x_i\|).$$

Now $\mathcal{S}$ and $\mathcal{S}_n$ are convex polytopes, so that the approximating cones at α_* are also convex polytopes and

$$\mathcal{C} = \{\lambda(\alpha - \alpha_*) \mid \alpha \in \mathcal{S},\ \lambda \geq 0,\ \lambda \in \mathbb{R}\},$$
$$\mathcal{C}_n = \{\lambda(\alpha - \alpha_*) \mid \alpha \in \mathcal{S}_n,\ \lambda \geq 0,\ \lambda \in \mathbb{R}\},$$

for a sufficiently small ϵ. Without loss of generality, we assume for some j, $\alpha_{*,i} = 0$ ($i = 1, \ldots, j$) and $\alpha_{*,i} > 0$ ($i = j+1, \ldots, b$). Let $\nu_i := Q\varphi_i$. Then the approximating cone $\mathcal{C}$ is spanned by μ_i ($i = 1, \ldots, b-1$) defined as

$$\mu_1 := \left[1, 0, \ldots, 0, -\frac{\nu_1}{\nu_b}\right]^{\mathrm{T}}, \ldots, \mu_{b-1} := \left[0, \ldots, 0, 1, -\frac{\nu_{b-1}}{\nu_b}\right]^{\mathrm{T}}.$$

That is,

$$\mathcal{C} = \left\{\sum_{i=1}^{b-1} \beta_i \mu_i \mid \beta_i \geq 0\ (i \leq j),\ \beta_i \in \mathbb{R}\right\}.$$

Let $\mathcal{N}(\mu, \Sigma)$ be a multivariate normal distribution with mean μ and covariance Σ; we use the same notation for a degenerate normal distribution (i.e., the Gaussian distribution confined to the range of a rank deficient covariance matrix Σ). Then we obtain the asymptotic law of $\sqrt{n}(\check{\alpha}_n - \alpha_*)$.

Theorem 3 *Let*[7] $Z_1 \sim \mathcal{N}(0, I_0 - P(\varphi/g_*)P(\varphi/g_*)^{\mathrm{T}})$ *and* $Z_2 \sim \mathcal{N}(0, Q\varphi\varphi^{\mathrm{T}} - Q\varphi Q\varphi^{\mathrm{T}})$, *where* Z_1 *and* Z_2 *are independent. Further define* $Z := I_0^{-1}(Z_1 + Z_2)$ *and* $\lambda_* = \nabla P\psi(\alpha_*) - Q\varphi$. *Then*

$$\sqrt{n}(\check{\alpha}_n - \alpha_*) \rightsquigarrow \mathop{\arg\min}_{\delta \in \mathcal{C}, \lambda_*^{\mathrm{T}}\delta = 0} \|\delta - Z\|_0 \quad (\textit{convergence in law}).$$

The proof is provided in Appendix B. If $\alpha_* > \mathbf{0}$ (α_* is an inner point of the feasible set), asymptotic normality can be proven in a simpler way. Set R_n and R as follows:

$$R_n := I - \frac{Q_n\varphi Q_n\varphi^{\mathrm{T}}}{\|Q_n\varphi\|^2}, \qquad R := I - \frac{Q\varphi Q\varphi^{\mathrm{T}}}{\|Q\varphi\|^2}.$$

R_n and R are projection matrices to linear spaces $\mathcal{C}_n = \{\delta \mid \delta^{\mathrm{T}} Q_n \varphi = 0\}$ and $\mathcal{C} = \{\delta \mid \delta^{\mathrm{T}} Q\varphi = 0\}$ respectively. Note that $R_n(\check{\alpha}_n - \alpha_*) = \check{\alpha}_n - \alpha_*$. Now $\check{\alpha}_n \xrightarrow{p} \alpha_*$ indicates that the probability of the event $\{\check{\alpha}_n > \mathbf{0}\}$ goes to 1. Then on the event $\{\check{\alpha}_n > \mathbf{0}\}$, by the KKT condition

$$\begin{aligned}
\mathbf{0} &= \sqrt{n}R_n(\nabla P_n\psi(\check{\alpha}_n) - a_*^n Q_n\varphi) = \sqrt{n}R_n(\nabla P_n\psi(\check{\alpha}_n) - Q_n\varphi) \\
&= \sqrt{n}R(\nabla P_n\psi(\alpha_*) - Q_n\varphi) - \sqrt{n}RI_0R(\check{\alpha}_n - \alpha_*) + o_p(1) \\
\Rightarrow \sqrt{n}(\check{\alpha}_n - \alpha_*) &= \sqrt{n}(RI_0R)^{\dagger}R(\nabla P_n\psi(\alpha_*) - \nabla P\psi(\alpha_*) - Q_n\varphi + Q\varphi) + o_p(1) \\
&\rightsquigarrow (RI_0R)^{\dagger}RI_0Z, \qquad (13)
\end{aligned}$$

[7] Since $\alpha_*^{\mathrm{T}}(I_0 - P(\varphi/g_*)P(\varphi/g_*)^{\mathrm{T}})\alpha_* = 0$, Z_1 obeys a degenerate normal distribution.

where † means the Moore-Penrose pseudo-inverse and in the third equality we used the relation $\nabla P\psi(\alpha_*) - Q\varphi = \mathbf{0}$ according to the KKT condition. On the other hand, since $\delta = R\delta$ for $\delta \in \mathcal{C}$, we have

$$\begin{aligned}\|Z-\delta\|_0^2 =& (Z-\delta)^{\mathrm{T}} I_0 (Z-\delta) = (Z-R\delta)^{\mathrm{T}} I_0 (Z-R\delta)\\ =& (\delta - (RI_0R)^{\dagger} RI_0 Z)^{\mathrm{T}} RI_0R(\delta - (RI_0R)^{\dagger} RI_0 Z)\\ &+ (\text{the terms independent of } \delta).\end{aligned}$$

The minimizer of the right-hand side of the above equality in $\mathcal{C}$ is $\delta = (RI_0R)^{\dagger} RI_0 Z$. This and the result of Theorem 3 coincide with (13).

In addition to Theorem 3 we can show the asymptotic law of $\sqrt{n}(\hat{\alpha}_n - \alpha_*)$. The proof is also given in Appendix B.

Theorem 4 *Let Z, Z_2 and λ_* be as in Theorem 3. Then*

$$\sqrt{n}(\hat{\alpha}_n - \alpha_*) \rightsquigarrow \mathop{\arg\min}_{\delta \in \mathcal{C}, \lambda_*^{\mathrm{T}}\delta = 0} \|\delta - Z\|_0 + (Z^{\mathrm{T}} I_0 \alpha_*)\alpha_* \qquad (\textit{convergence in law}).$$

The second term of the right-hand side is expressed by $(Z^{\mathrm{T}} I_0 \alpha_*)\alpha_* = (Z_2^{\mathrm{T}} \alpha_*)\alpha_*$.

Remark 3 *By the KKT condition and the definition of I_0, it can be easily checked that*

$$\alpha_*^{\mathrm{T}} I_0 \delta = 0 \quad (\forall \delta \in \mathcal{C} \cap \{\delta' \mid \lambda_*^{\mathrm{T}}\delta' = 0\}), \quad \|\alpha_*\|_0 = \alpha_*^{\mathrm{T}} I_0 \alpha_* = 1.$$

Thus Theorem 4 gives an orthogonal decomposition of the asymptotic law of $\sqrt{n}(\hat{\alpha}_n - \alpha_*)$ to a parallel part and an orthogonal part to $\mathcal{C} \cap \{\delta' \mid \lambda_*^{\mathrm{T}}\delta' = 0\}$. Hence in particular, if $\alpha_* > \mathbf{0}$, then $\lambda_* = \mathbf{0}$ and $\mathcal{C}$ is a linear subspace so that

$$\sqrt{n}(\hat{\alpha}_n - \alpha_*) \rightsquigarrow Z.$$

4 Illustrative Examples

We have shown that the KLIEP algorithm has preferable convergence properties. In this section, we illustrate the behavior of the proposed KLIEP method and how it can be applied in covariate shift adaptation.

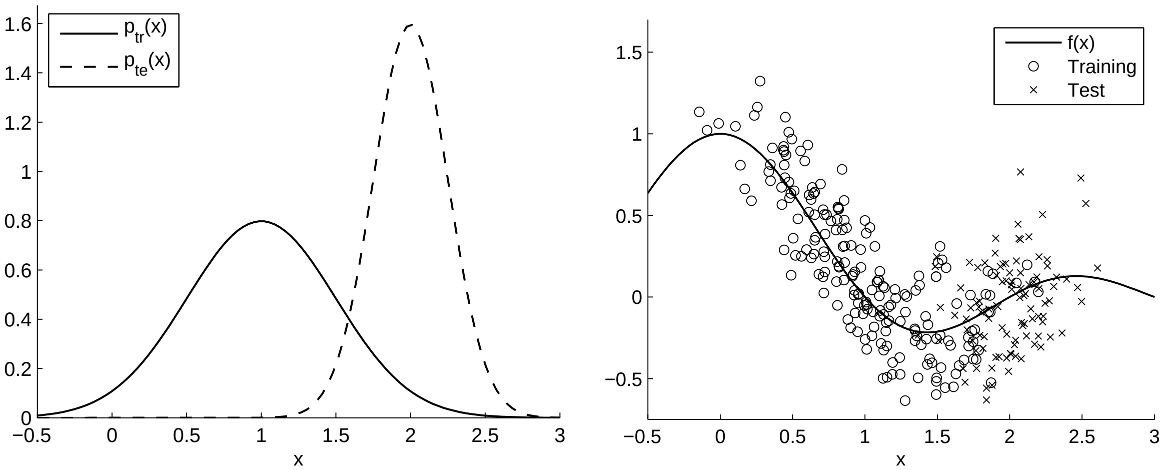
4.1 Setting

Let us consider a one-dimensional toy regression problem of learning

$$f(x) = \mathrm{sinc}(x).$$

Let the training and test input densities be

$$\begin{aligned} p_{\mathrm{tr}}(x) &= \mathcal{N}(x; 1, (1/2)^2),\\ p_{\mathrm{te}}(x) &= \mathcal{N}(x; 2, (1/4)^2),\end{aligned}$$



(a) Training input density $p_{\mathrm{tr}}(x)$ and test input density $p_{\mathrm{te}}(x)$

(b) Target function $f(x)$, training samples $\{(x_i^{\mathrm{tr}}, y_i^{\mathrm{tr}})\}_{i=1}^{n_{\mathrm{tr}}}$, and test samples $\{(x_j^{\mathrm{te}}, y_j^{\mathrm{te}})\}_{j=1}^{n_{\mathrm{te}}}$.

Figure 2: Illustrative example.

where $\mathcal{N}(x;\mu,\sigma^2)$ denotes the Gaussian density with mean μ and variance σ^2. We create training output value $\{y_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$ by

$$y_i^{\mathrm{tr}} = f(x_i^{\mathrm{tr}}) + \epsilon_i^{\mathrm{tr}},$$

where the noise $\{\epsilon_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$ has density $\mathcal{N}(\epsilon;0,(1/4)^2)$. Test output value $\{y_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$ are also generated in the same way. Let the number of training samples be $n_{\mathrm{tr}} = 200$ and the number of test samples be $n_{\mathrm{te}} = 1000$. The goal is to obtain a function $\widehat{f}(x)$ such that the following *generalization error* G (or the mean test error) is minimized:

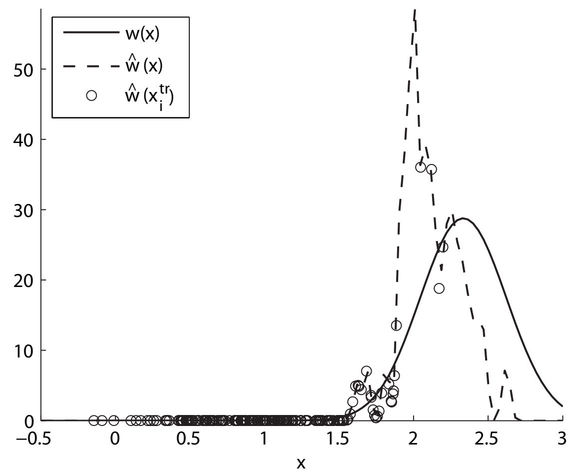
$$G := \frac{1}{n_{\mathrm{te}}} \sum_{j=1}^{n_{\mathrm{te}}} \left(\widehat{f}(x_j^{\mathrm{te}}) - y_j^{\mathrm{te}} \right)^2. \tag{14}$$

This setting implies that we are considering a (weak) extrapolation problem (see Figure 2, where only 100 test samples are plotted for clear visibility).

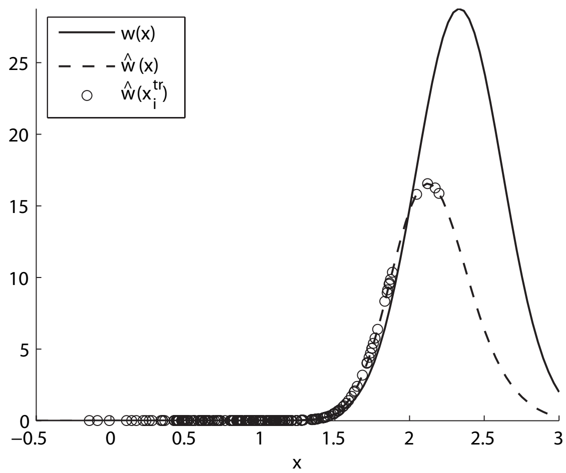
4.2 Importance Estimation by KLIEP

First, we illustrate the behavior of KLIEP in importance estimation, where we only use $\{x_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$ and $\{x_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$.

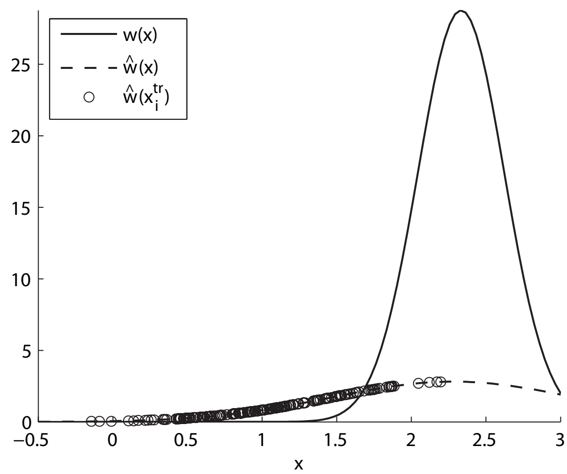
Figure 3 depicts the true importance and its estimates by KLIEP; the Gaussian kernel model (7) with $b = 100$ is used and three different Gaussian widths are tested. The graphs show that the performance of KLIEP is highly dependent on the Gaussian width; the estimated importance function $\widehat{w}(x)$ is highly fluctuated when σ is small, while it is overly smoothed when σ is large. When σ is chosen appropriately, KLIEP seems to work reasonably well for this example.



(a) Gaussian width $\sigma = 0.02$



(b) Gaussian width $\sigma = 0.2$



(c) Gaussian width $\sigma = 0.8$

Figure 3: Results of importance estimation by KLIEP. $w(x)$ is the true importance function and $\widehat{w}(x)$ is its estimation obtained by KLIEP.

Figure 4 depicts the values of the true J (see Eq.(3)) and its estimate by 5-fold LCV (see Eq.(5)); the means, the 25 percentiles, and the 75 percentiles over 100 trials are plotted as functions of the Gaussian width σ. This shows that LCV gives a very good estimate of J, which results in an appropriate choice of σ.

4.3 Covariate Shift Adaptation by IWLS and IWCV

Next, we illustrate how the estimated importance could be used for covariate shift adaptation. Here we use $\{(x_i^{\mathrm{tr}}, y_i^{\mathrm{tr}})\}_{i=1}^{n_{\mathrm{tr}}}$ and $\{x_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$ for learning; the test output values $\{y_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$ are used only for evaluating the generalization performance.

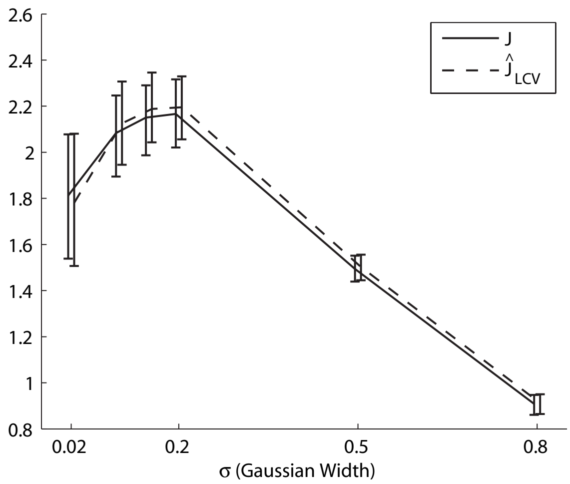


Figure 4: Model selection curve for KLIEP. J is the true score of an estimated importance (see Eq.(3)) and $\widehat{J}_{\mathrm{LCV}}$ is its estimate by 5-fold LCV (see Eq.(5)).

We use the following polynomial regression model:

$$\widehat{f}(x;\boldsymbol{\theta}) := \sum_{\ell=0}^{t} \theta_i x^{\ell}, \tag{15}$$

where t is the order of polynomials. The parameter vector $\boldsymbol{\theta}$ is learned by *importance-weighted least-squares* (IWLS):

$$\widehat{\boldsymbol{\theta}}_{\mathrm{IWLS}} := \mathop{\mathrm{argmin}}_{\boldsymbol{\theta}} \left[\sum_{i=1}^{n_{\mathrm{tr}}} \widehat{w}(x_i^{\mathrm{tr}}) \left(\widehat{f}(x_i^{\mathrm{tr}};\boldsymbol{\theta}) - y_i^{\mathrm{tr}} \right)^2 \right].$$

It is known that IWLS is consistent when the true importance $w(x_i^{\mathrm{tr}})$ is used as weights—ordinary LS is not consistent due to covariate shift, given that the model $\widehat{f}(x;\boldsymbol{\theta})$ is not correctly specified[8] (Shimodaira, 2000). For the linear regression model (15), the above minimizer $\widehat{\boldsymbol{\theta}}_{\mathrm{IWLS}}$ is given analytically by

$$\widehat{\boldsymbol{\theta}}_{\mathrm{IWLS}} = (\boldsymbol{X}^{\top}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\widehat{\boldsymbol{W}}\boldsymbol{y},$$

where

$$\begin{aligned} [\boldsymbol{X}]_{i,\ell} &= (x_i^{\mathrm{tr}})^{\ell-1}, \\ \widehat{\boldsymbol{W}} &= \mathrm{diag}\left(\widehat{w}(x_1^{\mathrm{tr}}), \widehat{w}(x_2^{\mathrm{tr}}), \ldots, \widehat{w}(x_{n_{\mathrm{tr}}}^{\mathrm{tr}})\right), \\ \boldsymbol{y} &= (y_1^{\mathrm{tr}}, y_2^{\mathrm{tr}}, \ldots, y_{n_{\mathrm{tr}}}^{\mathrm{tr}})^{\top}. \end{aligned} \tag{16}$$

[8] A model $\widehat{f}(x;\boldsymbol{\theta})$ is said to be *correctly specified* if there exists a parameter $\boldsymbol{\theta}^*$ such that $\widehat{f}(\boldsymbol{x};\boldsymbol{\theta}^*) = f(\boldsymbol{x})$.

diag $(a, b, \ldots, c)$ denotes the diagonal matrix with diagonal elements $a, b, \ldots, c$.

We choose the order t of polynomials based on *importance-weighted CV* (IWCV) (Sugiyama et al., 2007). More specifically, we first divide the training samples $\{z_i^{\mathrm{tr}} \mid z_i^{\mathrm{tr}} = (x_i^{\mathrm{tr}}, y_i^{\mathrm{tr}})\}_{i=1}^{n_{\mathrm{tr}}}$ into R disjoint subsets $\{\mathcal{Z}_r^{\mathrm{tr}}\}_{r=1}^{R}$. Then we learn a function $\widehat{f}_r(\boldsymbol{x})$ from $\{\mathcal{Z}_j^{\mathrm{tr}}\}_{j \neq r}$ by IWLS and compute its mean test error for the remaining samples $\mathcal{Z}_r^{\mathrm{tr}}$:

$$\widehat{G}_r := \frac{1}{|\mathcal{Z}_r^{\mathrm{tr}}|} \sum_{(x,y) \in \mathcal{Z}_r^{\mathrm{tr}}} \widehat{w}(x) \left(\widehat{f}_r(x) - y\right)^2.$$

We repeat this procedure for $r = 1, 2, \ldots, R$, compute the average of $\widehat{G}_r$ over all r, and use the average $\widehat{G}$ as an estimate of G:

$$\widehat{G} := \frac{1}{R} \sum_{r=1}^{R} \widehat{G}_r. \tag{17}$$

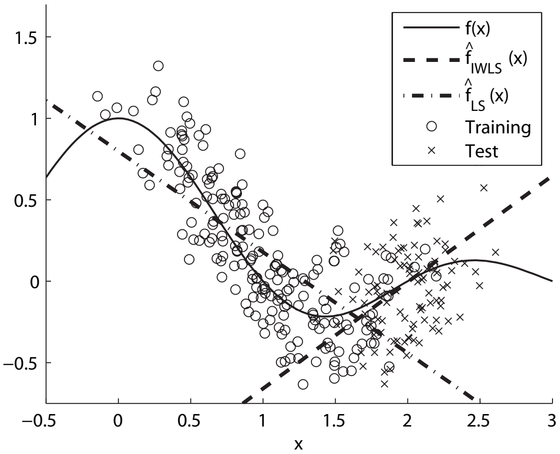
For model selection, we compute $\widehat{G}$ for all model candidates (the order t of polynomials in the current setting) and choose the one that minimizes $\widehat{G}$. We set the number of folds in IWCV at $R = 5$. IWCV is shown to be unbiased, while ordinary CV with misspecified models is biased due to covariate shift (Sugiyama et al., 2007).

Figure 5 depicts the functions learned by IWLS with different orders of polynomials. The results show that for all cases, the learned functions reasonably go through the test samples (note that the test *output* points are not used for obtaining the learned functions). Figure 6(a) depicts the true generalization error of IWLS and its estimate by IWCV; the means, the 25 percentiles, and the 75 percentiles over 100 runs are plotted as functions of the order of polynomials. This shows that IWCV roughly grasps the trend of the true generalization error. For comparison purposes, we also include the results by ordinary LS and ordinary CV in Figure 5 and Figure 6. Figure 5 shows that the functions obtained by ordinary LS go through the training samples, but not through the test samples. Figure 6 shows that the scores of ordinary CV tend to be biased, implying that model selection by ordinary CV is not reliable.

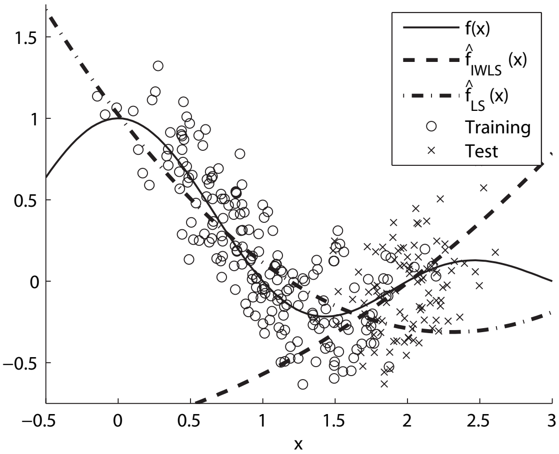
Finally, we compare the generalization error obtained by IWLS/LS and IWCV/CV, which is summarized in Figure 7 as box plots. This shows that IWLS+IWCV tends to outperform other methods, illustrating the usefulness of the proposed approach in covariate shift adaptation.

5 Discussion

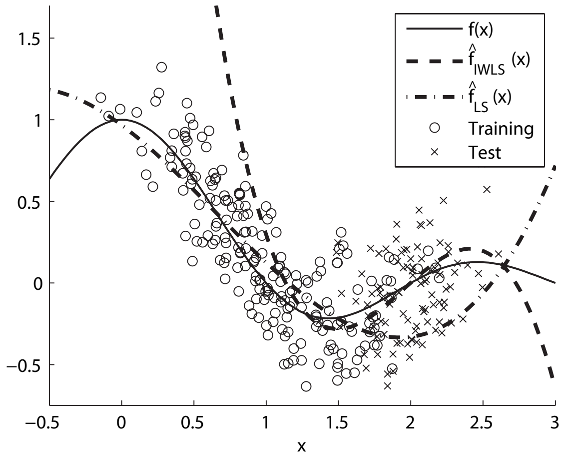
In this section, we discuss the relation between KLIEP and existing approaches.



(a) Polynomial of order 1



(b) Polynomial of order 2



(c) Polynomial of order 3

Figure 5: Learned functions obtained by IWLS and LS, which are denoted by $\widehat{f}_{\mathrm{IWLS}}(x)$ and $\widehat{f}_{\mathrm{LS}}(x)$, respectively.

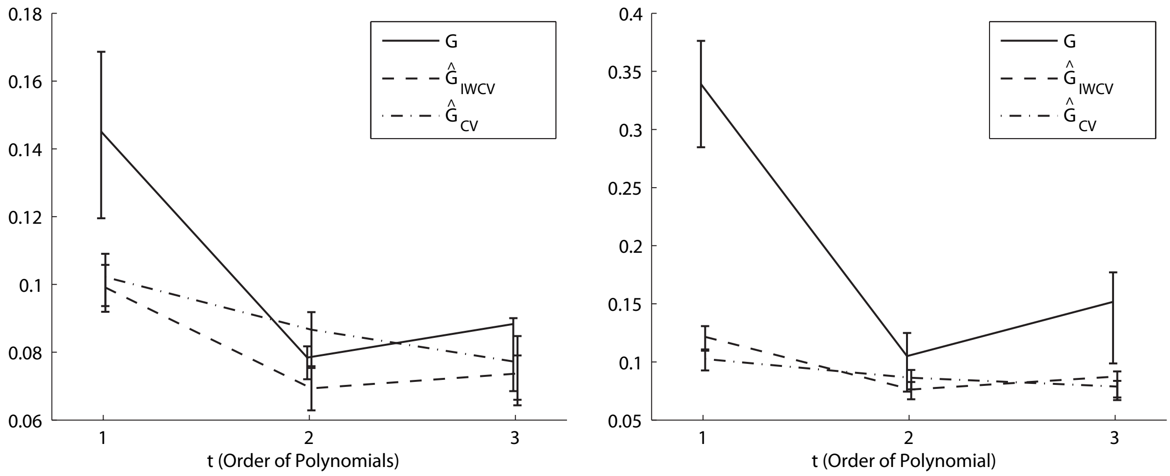
5.1 Kernel Density Estimator

The *kernel density estimator* (KDE) is a non-parametric technique to estimate a density $p(\boldsymbol{x})$ from its i.i.d. samples $\{\boldsymbol{x}_k\}_{k=1}^{n}$. For the Gaussian kernel, KDE is expressed as

$$\widehat{p}(\boldsymbol{x}) = \frac{1}{n(2\pi\sigma^2)^{d/2}} \sum_{k=1}^{n} K_\sigma(\boldsymbol{x}, \boldsymbol{x}_k), \tag{18}$$

where $K_\sigma(\boldsymbol{x}, \boldsymbol{x}')$ is the Gaussian kernel (6) with width σ.

The estimation performance of KDE depends on the choice of the kernel width σ, which can be optimized by LCV (Härdle et al., 2004)—a subset of $\{\boldsymbol{x}_k\}_{k=1}^{n}$ is used for density estimation and the rest is used for estimating the likelihood of the held-out samples. Note



(a) IWLS (b) LS

Figure 6: Model selection curves for IWLS/LS and IWCV/CV. G denotes the true generalization error of a learned function (see Eq.(14)), while $\widehat{G}_{\mathrm{IWCV}}$ and $\widehat{G}_{\mathrm{CV}}$ denote its estimate by 5-fold IWCV and 5-fold CV, respectively (see Eq.(17)).

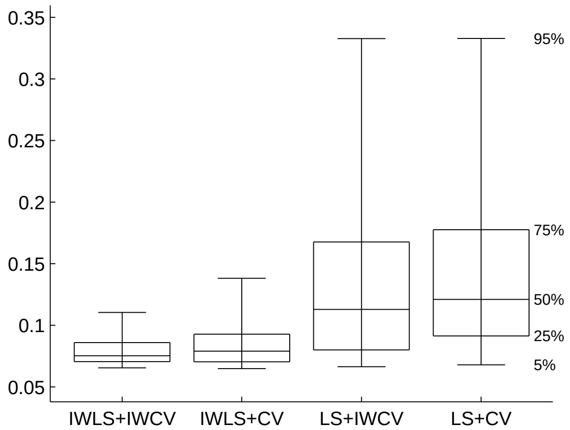


Figure 7: Box plots of generalization errors.

that model selection based on LCV corresponds to choosing σ such that the Kullback-Leibler divergence from $p(\boldsymbol{x})$ to $\widehat{p}(\boldsymbol{x})$ is minimized.

KDE can be used for importance estimation by first estimating $\widehat{p}_{\mathrm{tr}}(\boldsymbol{x})$ and $\widehat{p}_{\mathrm{te}}(\boldsymbol{x})$ separately from $\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$ and $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$, and then estimating the importance by $\widehat{w}(\boldsymbol{x}) = \widehat{p}_{\mathrm{te}}(\boldsymbol{x})/\widehat{p}_{\mathrm{tr}}(\boldsymbol{x})$. A potential limitation of this approach is that KDE suffers from the *curse of dimensionality* (Härdle et al., 2004), i.e., the number of samples needed to maintain the same approximation quality grows exponentially as the dimension of the input space increases. Furthermore, model selection by LCV is unreliable in small sample cases since data splitting in the CV procedure further reduces the sample size. Therefore, the KDE-based approach may not be reliable in high-dimensional cases.

5.2 Kernel Mean Matching

The *kernel mean matching* (KMM) method avoids density estimation and directly gives an estimate of the importance at training input points (Huang et al., 2007).

The basic idea of KMM is to find $\widehat{w}(\boldsymbol{x})$ such that the mean discrepancy between nonlinearly transformed samples drawn from $p_{\mathrm{te}}(\boldsymbol{x})$ and $p_{\mathrm{tr}}(\boldsymbol{x})$ is minimized in a *universal reproducing kernel Hilbert space* (Steinwart, 2001). The Gaussian kernel (6) is an example of kernels that induce universal reproducing kernel Hilbert spaces and it has been shown that the solution of the following optimization problem agrees with the true importance:

$$\begin{aligned}&\min_{w(\boldsymbol{x})} \left\| \int K_\sigma(\boldsymbol{x},\cdot)p_{\mathrm{te}}(\boldsymbol{x})d\boldsymbol{x} - \int K_\sigma(\boldsymbol{x},\cdot)w(\boldsymbol{x})p_{\mathrm{tr}}(\boldsymbol{x})d\boldsymbol{x}\right\|_{\mathcal{H}}^2\\ &\text{subject to } \int w(\boldsymbol{x})p_{\mathrm{tr}}(\boldsymbol{x})d\boldsymbol{x} = 1 \ \text{ and } \ w(\boldsymbol{x}) \geq 0,\end{aligned}$$

where $\|\cdot\|_{\mathcal{H}}$ denotes the norm in the Gaussian reproducing kernel Hilbert space and $K_\sigma(\boldsymbol{x},\boldsymbol{x}')$ is the Gaussian kernel (6) with width σ.

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

$$\begin{aligned}&\min_{\{w_i\}_{i=1}^{n_{\mathrm{tr}}}} \left[\frac{1}{2}\sum_{i,i'=1}^{n_{\mathrm{tr}}} w_i w_{i'} K_\sigma(\boldsymbol{x}_i^{\mathrm{tr}},\boldsymbol{x}_{i'}^{\mathrm{tr}}) - \sum_{i=1}^{n_{\mathrm{tr}}} w_i\kappa_i\right]\\ &\text{subject to } \left|\sum_{i=1}^{n_{\mathrm{tr}}} w_i - n_{\mathrm{tr}}\right| \leq n_{\mathrm{tr}}\epsilon \ \text{ and } \ 0 \leq w_1, w_2, \ldots, w_{n_{\mathrm{tr}}} \leq B,\end{aligned}$$

where

$$\kappa_i := \frac{n_{\mathrm{tr}}}{n_{\mathrm{te}}}\sum_{j=1}^{n_{\mathrm{te}}} K_\sigma(\boldsymbol{x}_i^{\mathrm{tr}},\boldsymbol{x}_j^{\mathrm{te}}).$$

B (≥ 0) and ϵ (≥ 0) are tuning parameters which control the regularization effects. The solution $\{\widehat{w}_i\}_{i=1}^{n_{\mathrm{tr}}}$ is an estimate of the importance at the training input points $\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$.

Since KMM does not involve density estimation, it is expected to work well even in high-dimensional cases. However, the performance is dependent on the tuning parameters B, ϵ, and σ, and they can not be simply optimized, e.g., by CV since estimates of the importance are available only at the training input points. Thus, an out-of-sample extension is needed to apply KMM in the CV framework, but this seems to be an open research issue currently.

A relation between KMM and a variant of KLIEP has been studied in Tsuboi et al. (2008).

5.3 Logistic Regression

Another approach to directly estimating the importance is to use a probabilistic classifier. Let us assign a selector variable $\delta = -1$ to training input samples and $\delta = 1$ to test input samples, i.e., the training and test input densities are written as

$$p_{\mathrm{tr}}(\boldsymbol{x}) = p(\boldsymbol{x}|\delta = -1),$$
$$p_{\mathrm{te}}(\boldsymbol{x}) = p(\boldsymbol{x}|\delta = 1).$$

An application of the Bayes theorem immediately yields that the importance can be expressed in terms of δ as follows (Bickel et al., 2007):

$$w(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\delta = 1)}{p(\boldsymbol{x}|\delta = -1)} = \frac{p(\delta = -1)}{p(\delta = 1)} \frac{p(\delta = 1|\boldsymbol{x})}{p(\delta = -1|\boldsymbol{x})}.$$

The probability ratio of test and training samples may be simply estimated by the ratio of the numbers of samples:

$$\frac{p(\delta = -1)}{p(\delta = 1)} \approx \frac{n_{\mathrm{tr}}}{n_{\mathrm{te}}}.$$

The conditional probability $p(\delta|\boldsymbol{x})$ could be approximated by discriminating test samples from training samples using a *logistic regression* (LogReg) classifier, where δ plays the role of a class variable. Below, we briefly explain the LogReg method.

The LogReg classifier employs a parametric model of the following form for expressing the conditional probability $p(\delta|\boldsymbol{x})$:

$$\widehat{p}(\delta|\boldsymbol{x}) := \frac{1}{1 + \exp\left(-\delta \sum_{\ell=1}^{u} \beta_\ell \phi_\ell(\boldsymbol{x})\right)},$$

where u is the number of basis functions and $\{\phi_\ell(\boldsymbol{x})\}_{\ell=1}^{u}$ are fixed basis functions. The parameter $\boldsymbol{\beta}$ is learned so that the negative log-likelihood is minimized:

$$\widehat{\boldsymbol{\beta}} := \mathop{\mathrm{argmin}}_{\boldsymbol{\beta}} \Bigg[\sum_{i=1}^{n_{\mathrm{tr}}} \log\left(1 + \exp\left(\sum_{\ell=1}^{u} \beta_\ell \phi_\ell(\boldsymbol{x}_i^{\mathrm{tr}})\right)\right) + \sum_{j=1}^{n_{\mathrm{te}}} \log\left(1 + \exp\left(-\sum_{\ell=1}^{u} \beta_\ell \phi_\ell(\boldsymbol{x}_j^{\mathrm{tr}})\right)\right) \Bigg].$$

Since the above objective function is convex, the global optimal solution can be obtained by standard nonlinear optimization methods such as Newton's method, conjugate gradient, or the BFGS method (Minka, 2007). Then the importance estimate is given by

$$\widehat{w}(\boldsymbol{x}) = \frac{n_{\mathrm{tr}}}{n_{\mathrm{te}}} \exp\left(\sum_{\ell=1}^{u} \widehat{\beta}_\ell \phi_\ell(\boldsymbol{x})\right).$$

An advantage of the LogReg method is that model selection (i.e., the choice of basis functions $\{\phi_\ell(\boldsymbol{x})\}_{\ell=1}^{u}$) is possible by standard CV, since the learning problem involved above is a standard supervised classification problem.

6 Experiments

In this section, we compare the experimental performance of KLIEP and existing approaches.

6.1 Importance Estimation for Artificial Datasets

Let $p_{\mathrm{tr}}(\boldsymbol{x})$ be the d-dimensional Gaussian density with mean $(0,0,\ldots,0)^\top$ and covariance identity and $p_{\mathrm{te}}(\boldsymbol{x})$ be the d-dimensional Gaussian density with mean $(1,0,\ldots,0)^\top$ and covariance identity. The task is to estimate the importance at training input points:

$$w_i := w(\boldsymbol{x}_i^{\mathrm{tr}}) = \frac{p_{\mathrm{te}}(\boldsymbol{x}_i^{\mathrm{tr}})}{p_{\mathrm{tr}}(\boldsymbol{x}_i^{\mathrm{tr}})} \quad \text{for } i = 1, 2, \ldots, n_{\mathrm{tr}}.$$

We compare the following methods:

KLIEP(σ): $\{w_i\}_{i=1}^{n_{\mathrm{tr}}}$ are estimated by KLIEP with the Gaussian kernel model (7). The number of template points is fixed at $b = 100$. Since the performance of KLIEP is dependent on the kernel width σ, we test several different values of σ.

KLIEP(CV): The kernel width σ in KLIEP is chosen based on 5-fold LCV (see Section 2.3).

KDE(CV): $\{w_i\}_{i=1}^{n_{\mathrm{tr}}}$ are estimated by KDE with the Gaussian kernel (18). The kernel widths for the training and test densities are chosen separately based on 5-fold LCV (see Section 5.1).

KMM(σ): $\{w_i\}_{i=1}^{n_{\mathrm{tr}}}$ are estimated by KMM (see Section 5.2). The performance of KMM is dependent on B, ϵ, and σ. We set $B = 1000$ and $\epsilon = (\sqrt{n_{\mathrm{tr}}} - 1)/\sqrt{n_{\mathrm{tr}}}$ following Huang et al. (2007), and test several different values of σ. We used the *CPLEX* software for solving quadratic programs in the experiments.

LogReg(σ): Gaussian kernels (7) are used as basis functions, where kernels are put at all training and test input points[9]. Since the performance of LogReg is dependent on the kernel width σ, we test several different values of σ. We used the *LIBLINEAR* implementation of logistic regression for the experiments (Lin et al., 2007).

LogReg(CV): The kernel width σ in LogReg is chosen based on 5-fold CV.

We fixed the number of test input points at $n_{\mathrm{te}} = 1000$ and consider the following two settings for the number n_{tr} of training samples and the input dimension d:

(a) $n_{\mathrm{tr}} = 100$ and $d = 1, 2, \ldots, 20$,

[9] We also tested another LogReg model where only 100 Gaussian kernels are used and the Gaussian centers are chosen randomly from the test input points. Our preliminary experiments showed that this does not degrade the performance.

(b) $d = 10$ and $n_{\mathrm{tr}} = 50, 60, \ldots, 150$.

We run the experiments 100 times for each d, each n_{tr}, and each method, and evaluate the quality of the importance estimates $\{\widehat{w}_i\}_{i=1}^{n_{\mathrm{tr}}}$ by the *normalized mean squared error* (NMSE):

$$\mathrm{NMSE} := \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \left(\frac{\widehat{w}_i}{\sum_{i'=1}^{n_{\mathrm{tr}}} \widehat{w}_{i'}} - \frac{w_i}{\sum_{i'=1}^{n_{\mathrm{tr}}} w_{i'}} \right)^2 .$$

NMSEs averaged over 100 trials are plotted in log scale in Figure 8. Figure 8(a) shows that the error of KDE(CV) sharply increases as the input dimension grows, while KLIEP, KMM, and LogReg with appropriate kernel widths tend to give smaller errors than KDE(CV). This would be the fruit of directly estimating the importance without going through density estimation. The graph also shows that the performance of KLIEP, KMM, and LogReg is dependent on the kernel width σ—the results of KLIEP(CV) and LogReg(CV) show that model selection is carried out reasonably well. Figure 9(a) summarizes the results of KLIEP(CV), KDE(CV), and LogReg(CV), where, for each input dimension, the best method in terms of the mean error and comparable ones based on the *t-test* at the significance level 5% are indicated by '$\circ$'; the methods with significant difference from the best method are indicated by '$\times$'. This shows that KLIEP(CV) works significantly better than KDE(CV) and LogReg(CV).

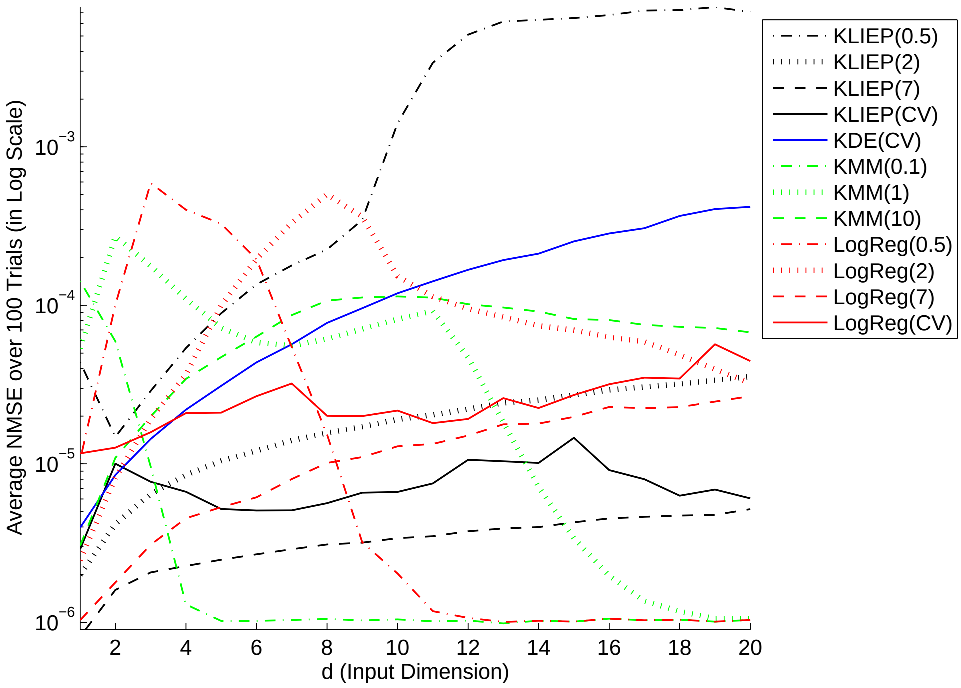
Figure 8(b) shows that the errors of all methods tend to decrease as the number of training samples grows. Again, KLIEP, KMM, and LogReg with appropriate kernel widths tend to give smaller errors than KDE(CV), and model selection in KLIEP(CV) and LogReg(CV) is shown work reasonably well. Figure 9(b) shows that KLIEP(CV) tends to give significantly smaller errors than KDE(CV) and LogReg(CV).

Overall, KLIEP(CV) is shown to be a useful method in importance estimation.

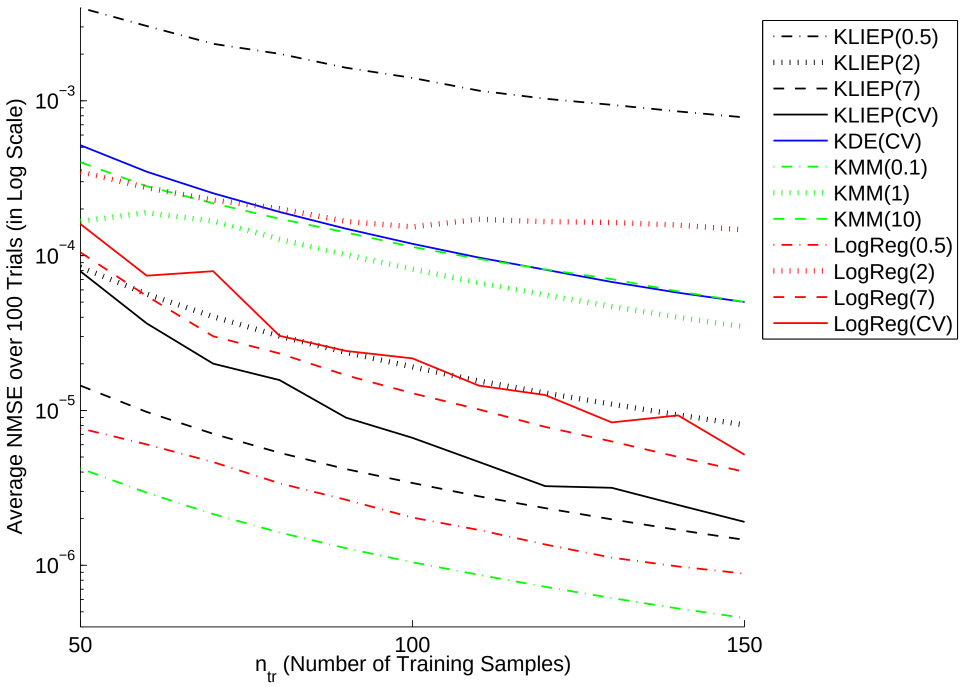
6.2 Covariate Shift Adaptation with Regression and Classification Benchmark Datasets

Here we employ importance estimation methods for covariate shift adaptation in regression and classification benchmark problems (see Table 1).

Each dataset consists of input/output samples $\{(\boldsymbol{x}_k, y_k)\}_{k=1}^{n}$. We normalize all the input samples $\{\boldsymbol{x}_k\}_{k=1}^{n}$ into $[0,1]^d$ and choose the test samples $\{(\boldsymbol{x}_j^{\mathrm{te}}, y_j^{\mathrm{te}})\}_{j=1}^{n_{\mathrm{te}}}$ from the pool $\{(\boldsymbol{x}_k, y_k)\}_{k=1}^{n}$ as follows. We randomly choose one sample $(\boldsymbol{x}_k, y_k)$ from the pool and accept this with probability $\min(1, 4(x_k^{(c)})^2)$, where $x_k^{(c)}$ is the c-th element of $\boldsymbol{x}_k$ and c is randomly determined and fixed in each trial of experiments; then we remove $\boldsymbol{x}_k$ from the pool regardless of its rejection or acceptance, and repeat this procedure until we accept n_{te} samples. We choose the training samples $\{(\boldsymbol{x}_i^{\mathrm{tr}}, y_i^{\mathrm{tr}})\}_{i=1}^{n_{\mathrm{tr}}}$ uniformly from the rest. Intuitively, in this experiment, the test input density tends to be lower than the training input density when $x_k^{(c)}$ is small. We set the number of samples at $n_{\mathrm{tr}} = 100$ and $n_{\mathrm{te}} = 500$ for all datasets. Note that we only use $\{(\boldsymbol{x}_i^{\mathrm{tr}}, y_i^{\mathrm{tr}})\}_{i=1}^{n_{\mathrm{tr}}}$ and $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$ for training regressors or classifiers; the test output values $\{y_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$ are used only for evaluating the generalization performance.

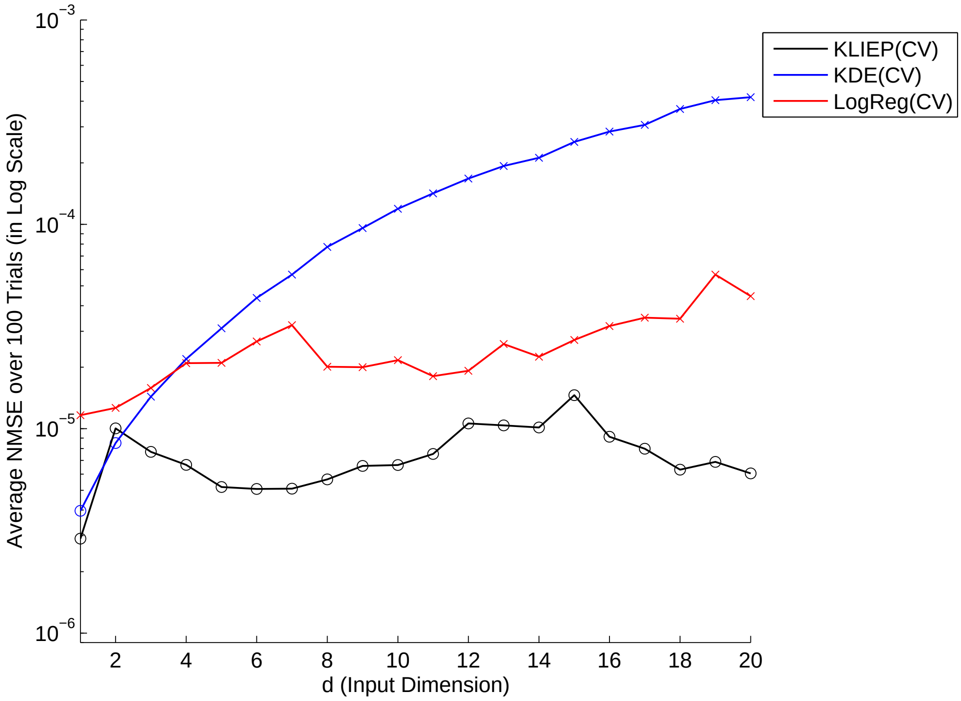


(a) When input dimension is changed

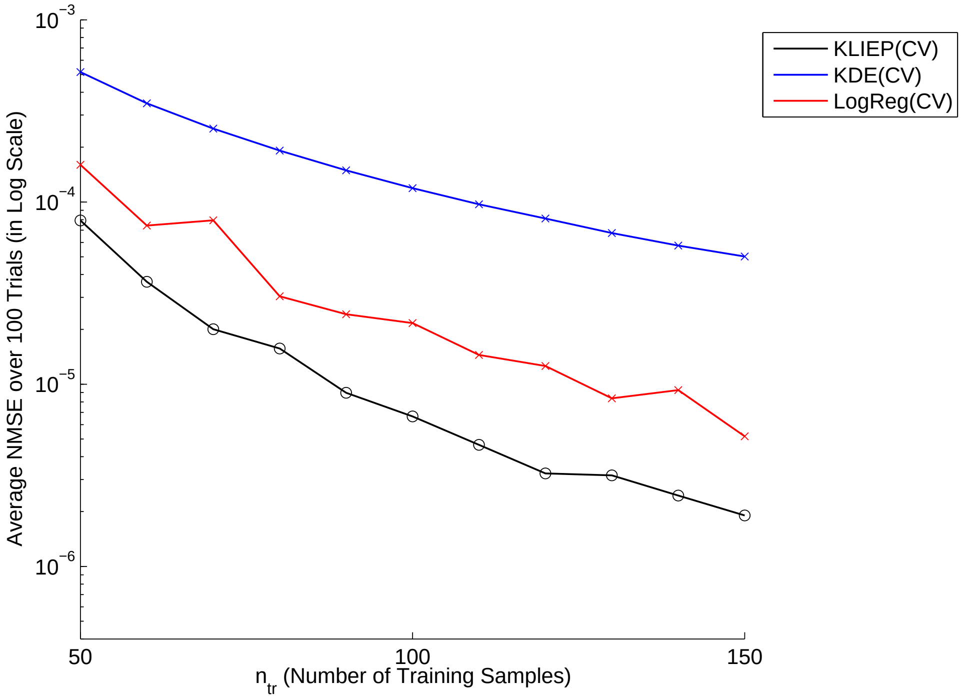


(b) When training sample size is changed

Figure 8: NMSEs averaged over 100 trials in log scale.



(a) When input dimension is changed



(b) When training sample size is changed

Figure 9: NMSEs averaged over 100 trials in log scale. For each dimension/number of training samples, the best method in terms of the mean error and comparable ones based on the *t-test* at the significance level 5% are indicated by '◦'; the methods with significant difference from the best method are indicated by '×'.

We use the following kernel model for regression or classification:

$$\widehat{f}(\boldsymbol{x};\boldsymbol{\theta}) := \sum_{\ell=1}^{t} \theta_\ell K_h(\boldsymbol{x}, \boldsymbol{m}_\ell),$$

where $K_h(\boldsymbol{x}, \boldsymbol{x}')$ is the Gaussian kernel (6) with width h and $\boldsymbol{m}_\ell$ is a template point randomly chosen from $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$. We set the number of kernels[10] at $t = 50$. We learn the parameter $\boldsymbol{\theta}$ by *importance-weighted regularized least-squares* (IWRLS) (Sugiyama et al., 2007):

$$\widehat{\boldsymbol{\theta}}_{\mathrm{IWRLS}} := \mathop{\mathrm{argmin}}_{\boldsymbol{\theta}} \left[\sum_{i=1}^{n_{\mathrm{tr}}} \widehat{w}(\boldsymbol{x}_i^{\mathrm{tr}}) \left(\widehat{f}(\boldsymbol{x}_i^{\mathrm{tr}};\boldsymbol{\theta}) - y_i^{\mathrm{tr}} \right)^2 + \lambda \|\boldsymbol{\theta}\|^2 \right]. \tag{19}$$

The solution $\widehat{\boldsymbol{\theta}}_{\mathrm{IWRLS}}$ is analytically given by

$$\widehat{\boldsymbol{\theta}}_{\mathrm{IWRLS}} = (\boldsymbol{K}^\top \widehat{\boldsymbol{W}} \boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \boldsymbol{K}^\top \widehat{\boldsymbol{W}} \boldsymbol{y},$$

where $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{y}$ is defined by Eq.(16), and

$$\begin{aligned} [\boldsymbol{K}]_{i,\ell} &:= K_h(\boldsymbol{x}_i^{\mathrm{tr}}, \boldsymbol{m}_\ell), \\ \widehat{\boldsymbol{W}} &:= \mathrm{diag}\left(\widehat{w}_1, \widehat{w}_2, \ldots, \widehat{w}_{n_{\mathrm{tr}}} \right). \end{aligned}$$

The kernel width h and the regularization parameter λ in IWRLS (19) are chosen by 5-fold IWCV. We compute the IWCV score by

$$\frac{1}{5} \sum_{r=1}^{5} \frac{1}{|\mathcal{Z}_r^{\mathrm{tr}}|} \sum_{(\boldsymbol{x},y) \in \mathcal{Z}_r^{\mathrm{tr}}} \widehat{w}(\boldsymbol{x}) L\left(\widehat{f}_r(\boldsymbol{x}), y \right),$$

where $\mathcal{Z}_r^{\mathrm{tr}}$ is the r-th held-out sample set (see Section 4.3) and

$$L\left(\widehat{y}, y\right) := \begin{cases} (\widehat{y} - y)^2 & \text{(Regression)}, \\ \frac{1}{2}(1 - \mathrm{sign}\{\widehat{y} y\}) & \text{(Classification)}. \end{cases}$$

We run the experiments 100 times for each dataset and evaluate the *mean test error*:

$$\frac{1}{n_{\mathrm{te}}} \sum_{j=1}^{n_{\mathrm{te}}} L\left(\widehat{f}(\boldsymbol{x}_j^{\mathrm{te}}), y_j^{\mathrm{te}} \right).$$

The results are summarized in Table 1, where 'Uniform' denotes uniform weights, i.e., no importance weight is used. The table shows that KLIEP(CV) compares favorably with Uniform, implying that the importance weighting techniques combined with KLIEP(CV)

[10] We fixed the number of kernels at a rather small number since we are interested in investigating the prediction performance under model misspecification; for over-specified models, importance-weighting methods have no advantage over the no importance method.

Table 1: Mean test error averaged over 100 trials. The numbers in the brackets are the standard deviation. All the error values are normalized so that the mean error by 'Uniform' (uniform weighting, or equivalently no importance weighting) is one. For each dataset, the best method and comparable ones based on the *Wilcoxon signed rank test* at the significance level 5% are described in bold face. The upper half are regression datasets taken from DELVE (Rasmussen et al., 1996) and the lower half are classification datasets taken from IDA (Rätsch et al., 2001). 'KMM(σ)' denotes KMM with kernel width σ.

Data	Dim	Uniform	KLIEP (CV)	KDE (CV)	KMM (0.01)	KMM (0.3)	KMM (1)	LogReg (CV)
kin-8fh	8	1.00(0.34)	**0.95(0.31)**	1.22(0.52)	1.00(0.34)	1.12(0.37)	1.59(0.53)	1.38(0.40)
kin-8fm	8	1.00(0.39)	**0.86(0.35)**	1.12(0.57)	1.00(0.39)	**0.98(0.46)**	1.95(1.24)	1.38(0.61)
kin-8nh	8	**1.00(0.26)**	**0.99(0.22)**	1.09(0.20)	**1.00(0.27)**	1.04(0.17)	1.16(0.25)	1.05(0.17)
kin-8nm	8	**1.00(0.30)**	**0.97(0.25)**	1.14(0.26)	1.00(0.30)	1.09(0.23)	1.20(0.22)	1.14(0.24)
abalone	7	1.00(0.50)	**0.97(0.69)**	1.02(0.41)	1.01(0.51)	**0.96(0.70)**	**0.93(0.39)**	**0.90(0.40)**
image	18	**1.00(0.51)**	**0.94(0.44)**	0.98(0.45)	**0.97(0.50)**	**0.97(0.45)**	1.09(0.54)	**0.99(0.47)**
ringnorm	20	1.00(0.04)	0.99(0.06)	0.87(0.04)	1.00(0.04)	**0.87(0.05)**	**0.87(0.05)**	0.93(0.08)
twonorm	20	1.00(0.58)	**0.91(0.52)**	1.16(0.71)	0.99(0.50)	**0.86(0.55)**	**0.99(0.70)**	**0.92(0.56)**
waveform	21	**1.00(0.45)**	**0.93(0.34)**	1.05(0.47)	**1.00(0.44)**	**0.93(0.32)**	0.98(0.31)	**0.94(0.33)**
Average		1.00(0.38)	0.95(0.35)	1.07(0.40)	1.00(0.36)	0.98(0.37)	1.20(0.47)	1.07(0.36)

are useful for improving the prediction performance under covariate shift. KLIEP(CV) works much better than KDE(CV); actually KDE(CV) tends to be worse than Uniform, which may be due to high dimensionality. We tested 10 different values of the kernel width σ for KMM and described three representative results in the table. KLIEP(CV) is slightly better than KMM with the best kernel width. Finally, LogReg(CV) is overall shown to work reasonably well, but it performs very poorly for some datasets. As a result, the average performance is not good.

Overall, we conclude that the proposed KLIEP(CV) is a promising method for covariate shift adaptation.

7 Conclusions

In this paper, we addressed the problem of estimating the importance for covariate shift adaptation. The proposed method, called KLIEP, does not involve density estimation so it is more advantageous than a naive KDE-based approach particularly in high-dimensional problems. Compared with KMM which also directly gives importance estimates, KLIEP is practically more useful since it is equipped with a model selection procedure. Our experiments highlighted these advantages and therefore KLIEP is shown to be a promising method for covariate shift adaptation.

In KLIEP, we modeled the importance function by a linear (or kernel) model, which resulted in a convex optimization problem with a sparse solution. However, our framework allows the use of any models. An interesting future direction to pursue would be to search for a class of models which has additional advantages, e.g., faster optimization (Tsuboi et al., 2008).

LCV is a popular model selection technique in density estimation and we used a variant of LCV for optimizing the Gaussian kernel width in KLIEP. In density estimation, however, it is known that LCV is not consistent under some condition (Schuster and Gregory, 1982; Hall, 1987). Thus it is important to investigate whether a similar inconsistency phenomenon is observed also in the context of importance estimation.

We used IWCV for model selection of regressors or classifiers under covariate shift. IWCV has smaller bias than ordinary CV and the model selection performance was shown to be improved by IWCV. However, the variance of IWCV tends to be larger than ordinary CV (Sugiyama et al., 2007) and therefore model selection by IWCV could be rather unstable. In practice, slightly regularizing the importance weight involved in IWCV can ease the problem, but this introduces an additional tuning parameter. Our important future work in this context is to develop a method to optimally regularize IWCV, e.g., following the line of Sugiyama et al. (2004).

Finally, the range of application of importance weights is not limited to covariate shift adaptation. For example, the density ratio could be used for anomaly detection, feature selection, independent component analysis, and conditional density estimation. Exploring possible application areas will be important future directions.

A Proofs of Theorem 1 and Theorem 2

A.1 Proof of Theorem 1

The proof follows the line of Nguyen et al. (2007). From the definition of γ_n, it follows that

$$-P_n \log \hat{g}_n \leq -P_n \log(a_0^n g_0) + \gamma_n.$$

Then, by the convexity of $-\log(\cdot)$, we obtain

$$\begin{aligned}
& - P_n \log\left(\frac{\hat{g}_n + a_0^n g_0}{2}\right) \leq \frac{-P_n \log \hat{g}_n - P_n \log a_0^n g_0}{2} \leq -P_n \log a_0^n g_0 + \frac{\gamma_n}{2} \\
\Leftrightarrow & - P_n \log\left(\frac{\hat{g}_n + a_0^n g_0}{2 a_0^n g_0}\right) - \frac{\gamma_n}{2} \leq 0.
\end{aligned}$$

$\log(g/g')$ is unstable when g is close to 0, while $\log\left(\frac{g+g'}{2g'}\right)$ is a slightly increasing function with respect to $g \geq 0$, its minimum is attained at $g = 0$, and $-\log(2) > -\infty$. Therefore, the above expression is easier to deal with than $\log(\hat{g}_n/g_0)$. Note that this technique can be found in van der Vaart and Wellner (1996) and van de Geer (2000).

We set $g' := \frac{a_0^n g_0 + \hat{g}_n}{2a_0^n}$. Since $Q_n g' = Q_n g_0 = 1/a_0^n$,

$$\begin{aligned}
& - P_n \log\left(\frac{\hat{g}_n + a_0^n g_0}{2a_0^n g_0}\right) - \frac{\gamma_n}{2} \leq 0 \\
\Rightarrow \quad & (Q_n - Q)(g' - g_0) - (P_n - P)\log\left(\frac{g'}{g_0}\right) - \frac{\gamma_n}{2} \\
& \leq -Q(g' - g_0) + P\log\left(\frac{g'}{g_0}\right) \\
& \leq 2P\left(\sqrt{\frac{g'}{g_0}} - 1\right) - Q(g' - g_0) = Q\left(2\sqrt{g' g_0} - 2g_0\right) - Q(g' - g_0) \\
& = Q\left(2\sqrt{g' g_0} - g' - g_0\right) = -h_Q(g', g_0)^2. \qquad (20)
\end{aligned}$$

The Hellinger distance between $\hat{g}_n/a_0^n$ and g_0 has the following bound (see Lemma 4.2 in van de Geer, 2000):

$$\frac{1}{16} h_Q(\hat{g}_n/a_0^n, g_0) \leq h_Q(g', g_0).$$

Thus it is sufficient to bound $|(Q_n - Q)(g' - g_0)|$ and $|(P_n - P)\log\left(\frac{g'}{g_0}\right)|$ from above.

From now on, we consider the case where the inequality (8) in Assumption 1.3 is satisfied. The proof for the setting of the inequality (9) can be carried out along the line of Nguyen et al. (2007). We will utilize the Bousquet bound (10) to bound $|(Q_n - Q)(g' - g_0)|$ and $|(P_n - P)\log\left(\frac{g'}{g_0}\right)|$. In the following, we prove the assertion in 4 steps. In the first and second steps, we derive upper bounds of $|(Q_n - Q)(g' - g_0)|$ and $|(P_n - P)\log\left(\frac{g'}{g_0}\right)|$, respectively. In the third step, we bound the ∞-norm of $\hat{g}_n$ which is needed to prove the convergence. Finally, we combine the results of Steps 1 to 3 and obtain the assertion. The following statements heavily rely on Koltchinskii (2006).

Step 1. Bounding $|(Q_n - Q)(g' - g_0)|$.

Let

$$\iota(g) := \frac{g + g_0}{2},$$

and

$$\mathcal{G}_n^M(\delta) := \{\iota(g) \mid g \in \mathcal{G}_n^M, Q(\iota(g) - g_0) - P\log(\iota(g)/g_0) \leq \delta\} \cup \{g_0\}.$$

Let $\phi_n^M(\delta)$ be

$$\phi_n^M(\delta) := ((M + \eta_1)^{\gamma/2}\delta^{1-\gamma/2}/\sqrt{n}) \vee ((M + \eta_1)n^{-2/(2+\gamma)}) \vee (\delta/\sqrt{n}).$$

Then applying Lemma 2 to $\mathcal{F} = \{2(g - g_0)/(M + \eta_1) \mid g \in \mathcal{G}_n^M(\delta)\}$, we obtain that there is a constant C that only depends on K and γ such that

$$\mathrm{E}_Q\left[\sup_{g \in \mathcal{G}_n^M, \|g - g_0\|_{Q,2} \leq \delta} |(Q_n - Q)(g - g_0)|\right] \leq C\phi_n^M(\delta), \qquad (21)$$

where $\|f\|_{Q,2} := \sqrt{Qf^2}$.

Next, we define the "diameter" of a set $\{g - g_0 \mid g \in \mathcal{G}_n^M(\delta)\}$ as

$$\tilde{D}^M(\delta) := \sup_{g\in\mathcal{G}_n^M(\delta)} \sqrt{Q(g-g_0)^2} = \sup_{g\in\mathcal{G}_n^M(\delta)} \|g-g_0\|_{Q,2}.$$

It is obvious that

$$\tilde{D}^M(\delta) \geq \sup_{g\in\mathcal{G}_n^M(\delta)} \sqrt{Q(g-g_0)^2 - (Q(g-g_0))^2}.$$

Note that for all $g \in \mathcal{G}_n^M(\delta)$,

$$\begin{aligned} Q(g-g_0)^2 &= Q(\sqrt{g}-\sqrt{g_0})^2(\sqrt{g}+\sqrt{g_0})^2 \\ &\leq (M+3\eta_1)Q(\sqrt{g}-\sqrt{g_0})^2 = (M+3\eta_1)h_Q(g,g_0)^2. \end{aligned}$$

Thus from the inequality (20), it follows that

$$\begin{aligned} \forall g \in \mathcal{G}_n^M(\delta), \quad \delta &\geq Q(g-g_0) - P\log(g/g_0) \\ &\geq h_Q(g,g_0)^2 \geq \|g-g_0\|_{Q,2}^2/(M+3\eta_1), \end{aligned}$$

which implies

$$\tilde{D}^M(\delta) \leq \sqrt{(M+3\eta_1)\delta} =: D^M(\delta).$$

So, by the inequality (21), we obtain

$$\begin{aligned} \mathrm{E}_Q\left[\sup_{g\in\mathcal{G}_n^M(\delta)} |(Q_n-Q)(g-g_0)|\right] &\leq C\phi_n^M(D^M(\delta)) \\ &\leq C_M\left(\frac{\delta^{(1-\gamma/2)/2}}{\sqrt{n}} \vee n^{-2/(2+\gamma)} \vee \frac{\delta^{1/2}}{\sqrt{n}}\right), \end{aligned}$$

where C_M is a constant depending on M, γ, η_1, and K.

Let $q > 1$ be an arbitrary constant. For some $\delta > 0$, let $\delta_j := q^j\delta$, where j is an integer, and let

$$\mathcal{H}_\delta^M := \bigcup_{\delta_j\geq\delta} \{\frac{\delta}{\delta_j}(g-g_0) \mid g \in \mathcal{G}_n^M(\delta_j)\}.$$

Then, by Lemma 3, there exists K_M for all $M > 1$ such that for

$$U_{n,t}^M(\delta) := K_M\left[\phi_n^M(D^M(\delta)) + \sqrt{\frac{t}{n}}D^M(\delta) + \frac{t}{n}\right],$$

and an event E_δ^M

$$E_{n,\delta}^M := \left\{\sup_{g\in\mathcal{H}_\delta^M} |(Q_n-Q)g| \leq U_{n,t}^M(\delta)\right\},$$

the following is satisfied:

$$Q(E_\delta^M) \geq 1 - e^{-t}.$$

Step 2. Bounding $|(P_n - P)(\log(g'/g_0)|$.

Along the same arguments with Step 1 using the Lipschitz continuity of the function $g \mapsto \log(\frac{g+g_0}{2g_0})$ on the support of P, we also obtain a similar inequality for

$$\tilde{\mathcal{H}}_{n,\delta}^M := \bigcup_{\delta_j \geq \delta} \left\{ \frac{\delta}{\delta_j} \log\left(\frac{g}{g_0}\right) \mid g \in \mathcal{G}_n^M(\delta_j) \right\},$$

i.e., there exists a constant $\tilde{K}_M$ that depends on K, M, γ, η_1, and η_0 such that

$$P(\tilde{E}_\delta^M) \geq 1 - e^{-t},$$

where $\tilde{E}_\delta^M$ is an event defined by

$$\tilde{E}_{n,\delta}^M := \left\{ \sup_{f \in \tilde{\mathcal{H}}_\delta^M} |(P_n - P)f| \leq \tilde{U}_{n,t}^M(\delta) \right\},$$

and

$$\tilde{U}_{n,t}^M(\delta) := \tilde{K}_M \left[\phi_n^M(D^M(\delta)) + \sqrt{\frac{t}{n}} D^M(\delta) + \frac{t}{n} \right].$$

Step 3. Bounding the ∞-norm of $\hat{g}_n/a_0^n$.

We can show that all elements of $\hat{\mathcal{G}}_n$ are uniformly bounded from above with high probability. Let

$$S_n := \left\{ \inf_{\varphi \in \mathcal{F}_n} Q_n \varphi \geq \epsilon_0/2 \right\} \cap \{3/4 < a_0^n < 5/4\}.$$

Then by Lemma 4, we can take a sufficiently large $\bar{M}$ such that $g/a_0^n \in \mathcal{G}_n^{\bar{M}}$ $(\forall g \in \hat{\mathcal{G}}_n)$ on the event S_n and $Q(S_n) \to 1$.

Step 4. Combining Steps 1,2, and 3.

We consider an event

$$E_n := E_{n,\delta}^{\bar{M}} \cap \tilde{E}_{n,\delta}^{\bar{M}} \cap S_n.$$

On the event E_n, $\hat{g}_n \in \mathcal{G}_n^{\bar{M}}$. For $\psi : \mathbb{R}_+ \to \mathbb{R}_+$, we define the $\#$-transform and the $\flat$-transform as follows (Koltchinskii, 2006):

$$\psi^\flat(\delta) := \sup_{\sigma \geq \delta} \frac{\psi(\sigma)}{\sigma}, \quad \psi^\#(\epsilon) := \inf\{\delta > 0 \mid \psi^\flat(\delta) \leq \epsilon\}.$$

Here we set

$$\delta_n^M(t) := (U_{n,t}^M)^\#(1/4q), \quad V_{n,t}^M(\delta) := (U_{n,t}^M)^\flat(\delta),$$
$$\tilde{\delta}_n^M(t) := (\tilde{U}_{n,t}^M)^\#(1/4q), \quad \tilde{V}_{n,t}^M(\delta) := (\tilde{U}_{n,t}^M)^\flat(\delta).$$

Then on the event E_n,

$$\sup_{g \in \mathcal{G}_n^{\bar{M}}(\delta_j)} |(Q_n - Q)(g - g_0)| \leq \frac{\delta_j}{\delta} U_{n,t}^{\bar{M}}(\delta) \leq \delta_j V_{n,t}^{\bar{M}}(\delta), \tag{22}$$

$$\sup_{g \in \mathcal{G}_n^{\bar{M}}(\delta_j)} \left| (P_n - P) \log\left(\frac{g}{g_0}\right) \right| \leq \frac{\delta_j}{\delta} \tilde{U}_{n,t}^{\bar{M}}(\delta) \leq \delta_j \tilde{V}_{n,t}^{\bar{M}}(\delta). \tag{23}$$

Take arbitrary j and δ such that

$$\delta_j \geq \delta \geq \delta_n^{\bar{M}}(t) \vee \tilde{\delta}_n^{\bar{M}}(t) \vee 2q\gamma_n.$$

Let

$$\mathcal{G}_n^{\bar{M}}(a, b) := \mathcal{G}_n^{\bar{M}}(b) \backslash \mathcal{G}_n^{\bar{M}}(a) \quad (a < b).$$

Here, we assume $\iota(\hat{g}_n / a_0^n) \in \mathcal{G}_n^{\bar{M}}(\delta_{j-1}, \delta_j)$. Then we will derive a contradiction. In these settings, for $g' := \iota(\hat{g}_n / a_0^n)$,

$$\begin{aligned} \delta_{j-1} \leq |Q(g' - g_0) + P \log \frac{g'}{g_0}| &\leq |(Q_n - Q)(g' - g_0)| + |(P_n - P) \log \frac{g'}{g_0}| + \frac{\gamma_n}{2} \\ &\leq \delta_j V_{n,t}^{\bar{M}}(\delta) + \delta_j \tilde{V}_{n,t}^{\bar{M}}(\delta) + \frac{\gamma_n}{2}, \end{aligned}$$

which implies

$$\frac{3}{4q} \leq \frac{1}{q} - \frac{\gamma_n}{2\delta_j} \leq V_{n,t}^{\bar{M}}(\delta) + \tilde{V}_{n,t}^{\bar{M}}(\delta). \tag{24}$$

So, either $V_{n,t}^{\bar{M}}(\delta)$ or $\tilde{V}_{n,t}^{\bar{M}}(\delta)$ is greater than $\frac{3}{8q}$. This contradicts the definition of the #-transform.

We can show that $\delta_n^{\bar{M}}(t) \vee \tilde{\delta}_n^{\bar{M}}(t) = O(n^{-\frac{2}{2+\gamma}} t)$. To see this, for some $s > 0$, set

$$\begin{aligned} &\hat{\delta}_1 = \left(\frac{\delta^{(1-\gamma/2)/2}}{\sqrt{n}}\right)^{\#}(s), \quad \hat{\delta}_2 = \left(n^{-2/(2+\gamma)}\right)^{\#}(s), \quad \hat{\delta}_3 = \left(\frac{\delta^{1/2}}{\sqrt{n}}\right)^{\#}(s), \\ &\hat{\delta}_4 = \left(\sqrt{\frac{t}{n}\delta}\right)^{\#}(s), \quad \hat{\delta}_5 = \left(\frac{t}{n}\right)^{\#}(s), \end{aligned}$$

where all the #-transforms are taken with respect to δ. Then they satisfy

$$s = \frac{\hat{\delta}_1^{(1-\gamma/2)/2}/\sqrt{n}}{\hat{\delta}_1}, \quad s = \frac{n^{-2/(2+\gamma)}}{\hat{\delta}_2}, \quad s = \frac{\hat{\delta}_3^{1/2}/\sqrt{n}}{\hat{\delta}_3}, \quad s = \frac{\sqrt{\hat{\delta}_4 t/n}}{\hat{\delta}_4}, \quad s = \frac{t/n}{\hat{\delta}_5}.$$

Thus, by using some constants $c_1, \ldots, c_4$, we obtain

$$\hat{\delta}_1 = c_1 n^{-2/(2+\gamma)}, \quad \hat{\delta}_2 = c_2 n^{-2/(2+\gamma)}, \quad \hat{\delta}_3 = c_3 n^{-1}, \quad \hat{\delta}_4 = c_4 t/n, \quad \hat{\delta}_5 = c_5 t/n.$$

Following the line of Koltchinskii (2006), for $\epsilon = \epsilon_1 + \cdots + \epsilon_m$, we have

$$(\psi_1 + \cdots + \psi_m)^{\#}(\epsilon) \leq \psi_1^{\#}(\epsilon_1) \vee \cdots \vee \psi_m^{\#}(\epsilon_m).$$

Thus we obtain $\delta_n^{\bar{M}}(t) \vee \tilde{\delta}_n^{\bar{M}}(t) = O(n^{-\frac{2}{2+\gamma}}t)$.

The above argument results in

$$\frac{1}{16}h_Q(\hat{g}_n/a_0^n, g_0) \leq h_Q(g', g_0) = \mathcal{O}_p(n^{-\frac{1}{2+\gamma}} + \sqrt{\gamma_n}).$$

■

In the following, we show lemmas used in the proof of Theorem 1. We use the same notations as those in the proof of Theorem 1.

Lemma 2 *Consider a class $\mathcal{F}$ of functions such that $-1 \leq f \leq 1$ for all $f \in \mathcal{F}$ and $\sup_{\tilde{Q}} \log N(\epsilon, \mathcal{F}, L_2(\tilde{Q})) \leq \frac{T}{\epsilon^\gamma}$, where the supremum is taken over all finitely discrete probability measures. Then there is a constant $C_{T,\gamma}$ depending on γ and T such that for $\delta^2 = \sup_{f\in\mathcal{F}} Qf^2$,*

$$\mathrm{E}[\|Q_n - Q\|_{\mathcal{F}}] \leq C_{T,\gamma}\left[(n^{-\frac{2}{2+\gamma}}) \vee (\delta^{1-\gamma/2}/\sqrt{n}) \vee (\delta/\sqrt{n})\right]. \tag{25}$$

Proof

This lemma can be shown along a similar line to Mendelson (2002), but we shall pay attention to the point that $\mathcal{F}$ may not contain the constant function 0. Let $(\epsilon_i)_{1\leq i\leq n}$ be i.i.d. Rademacher random variables, i.e., $P(\epsilon_i = 1) = P(\epsilon_i = -1) = 1/2$, $R_n(\mathcal{F})$ be the Rademacher complexity of $\mathcal{F}$ defined as

$$R_n(\mathcal{F}) = \frac{1}{n}\mathrm{E}_Q\mathrm{E}_\epsilon \sup_{f\in\mathcal{F}} |\sum_{i=1}^n \epsilon_i f(\boldsymbol{x}_i^{\mathrm{tr}})|.$$

Then by Talagrand (1994),

$$\mathrm{E}_Q \sup_{f\in\mathcal{F}} \|Q_n f^2\| \leq \sup_{f\in\mathcal{F}} Qf^2 + 8R_n(\mathcal{F}). \tag{26}$$

Set $\hat{\delta}^2 = \sup_{f\in\mathcal{F}} Q_n f^2$. Then noticing that $\log N(\epsilon, F \cup \{0\}, L_2(Q_n)) \leq \frac{T}{\epsilon^\gamma} + 1$, it can be shown that there is a universal constant C such that

$$\begin{aligned}\frac{1}{n}\mathrm{E}_\epsilon \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \epsilon_i f(\boldsymbol{x}_i^{\mathrm{tr}})\right| &\leq \frac{C}{\sqrt{n}}\int_0^{\hat{\delta}} \sqrt{1 + \log N(\epsilon, F, L_2(Q_n))}d\epsilon \\ &\leq \frac{C}{\sqrt{n}}\left(\frac{\sqrt{T}}{1-\gamma/2}\hat{\delta}^{1-\gamma/2} + \hat{\delta}\right). \end{aligned}\tag{27}$$

See van der Vaart and Wellner (1996) for detail. Taking the expectation with respect Q and employing Jensen's inequality and (26), we obtain

$$R_n(\mathcal{F}) \leq \frac{C_{T,\gamma}}{\sqrt{n}}\left[\left(\delta^2 + R_n(\mathcal{F})\right)^{(1-\gamma/2)/2} + \left(\delta^2 + R_n(\mathcal{F})\right)^{1/2}\right],$$

where $C_{T,\gamma}$ is a constant depending on T and γ. Thus we have

$$R_n(\mathcal{F}) \leq C_{T,\gamma}\left[(n^{-\frac{2}{2+\gamma}}) \vee (\delta^{1-\gamma/2}/\sqrt{n}) \vee (\delta/\sqrt{n})\right]. \tag{28}$$

By the symmetrization argument (van der Vaart and Wellner, 1996), we have

$$\mathrm{E}[\sup_{f\in\mathcal{F}}|(Q_n - Q)f|] \leq 2R_n(\mathcal{F}). \tag{29}$$

Combining (28) and (29), we obtain the assertion. ∎

Lemma 3 *For all $M > 1$, there exists K_M depending on γ, η_1, q, and K such that*

$$Q\left(\sup_{g\in\mathcal{H}_\delta^M}|(Q_n - Q)g| \geq K_M\left[\phi_n^M(D^M(\delta)) + \sqrt{\frac{t}{n}}D^M(\delta) + \frac{t}{n}\right]\right) \leq e^{-t}.$$

Proof

Since $\phi_n^M(D^M(\delta))/\delta$ and $D^M(\delta)/\delta$ are monotone decreasing, we have

$$\begin{aligned}
&\mathrm{E}\left[\sup_{f\in\mathcal{H}_\delta^M}|(Q_n - Q)f|\right] \leq \sum_{\delta_j\geq\delta}\frac{\delta}{\delta_j}\mathrm{E}\left[\sup_{g\in\mathcal{G}_n^M(\delta_j)}|(Q_n - Q)(g - g_0)|\right] \\
&\leq \sum_{\delta_j\geq\delta}\frac{\delta}{\delta_j}C\phi_n^M(D^M(\delta_j)) \leq \sum_{\delta_j\geq\delta}\frac{\delta}{\delta_j^{1-\gamma'}}C\frac{\phi_n^M(D^M(\delta_j))}{\delta_j^{\gamma'}} \\
&\leq \sum_{\delta_j\geq\delta}\frac{\delta}{\delta_j^{1-\gamma'}}C\frac{\phi_n^M(D^M(\delta))}{\delta^{\gamma'}} = C\phi_n^M(D^M(\delta))\sum_{\delta_j\geq\delta}\frac{\delta^{1-\gamma'}}{\delta_j^{1-\gamma'}} \\
&\leq C\phi_n^M(D^M(\delta))\sum_{j\geq 0}q^{-j(1-\gamma')} = c_{\gamma,q}\phi_n^M(D^M(\delta)),
\end{aligned} \tag{30}$$

where $c_{\gamma,q}$ is a constant that depends on γ, K, and q, and

$$\begin{aligned}
\sup_{f\in\mathcal{H}_\delta^M}\sqrt{Qf^2} &\leq \sup_{\delta_j\geq\delta}\frac{\delta}{\delta_j}\sup_{g\in\mathcal{G}_n^M(\delta_j)}\sqrt{Q(g - g_0)^2} \\
&\leq \delta\sup_{\delta_j\geq\delta}\frac{D^M(\delta_j)}{\delta_j} \leq \delta\frac{D^M(\delta)}{\delta} = D^M(\delta).
\end{aligned} \tag{31}$$

Using the Bousquet bound, we obtain

$$Q\left(\sup_{g\in\mathcal{H}_\delta^M}|(Q_n - Q)g|/M \geq C\left[c_{\gamma,q}\frac{\phi_n^M(D^M(\delta))}{M} + \sqrt{\frac{t}{n}}\frac{D^M(\delta)}{M} + \frac{t}{n}\right]\right) \leq e^{-t},$$

where C is some universal constant. Thus, there exists K_M for all $M > 1$ such that

$$Q\left(\sup_{g\in\mathcal{H}_\delta^M}|(Q_n - Q)g| \geq K_M\left[\phi_n^M(D^M(\delta)) + \sqrt{\frac{t}{n}}D^M(\delta) + \frac{t}{n}\right]\right) \leq e^{-t}.$$

∎

Lemma 4 *For an event* $S_n := \{\inf_{\varphi \in \mathcal{F}_n} Q_n\varphi \geq \epsilon_0/2\} \cap \{3/4 < a_0^n < 5/4\}$*, we have*

$$Q(S_n) \to 1.$$

Moreover, there exists a sufficiently large $\bar{M} > 0$ *such that* $g/a_0^n \in \mathcal{G}_n^{\bar{M}}$ $(\forall g \in \hat{\mathcal{G}}_n)$ *on the event* S_n*.*

Proof

It is obvious that

$$(Q_n - Q)g_0 = \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right).$$

Thus, because of $Qg_0 = 1$,

$$a_0^n = 1 + \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right).$$

Moreover, Assumption 1.3 implies

$$\|Q_n - Q\|_{\mathcal{F}_n} = \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right).$$

Thus,

$$\inf_{\varphi \in \mathcal{F}_n} Q_n\varphi \geq \epsilon_0 - \mathcal{O}_p(1/\sqrt{n}),$$

implying

$$Q(\bar{S}_n) \to 1 \text{ for } \bar{S}_n := \left\{\inf_{\varphi \in \mathcal{F}_n} Q_n\varphi \geq \epsilon_0/2\right\}.$$

On the event S_n, all the elements of $\hat{\mathcal{G}}_n$ is uniformly bounded from above:

$$\begin{aligned} 1 = Q_n(\sum_l \alpha_l\varphi_l) &= \sum_l \alpha_l Q_n(\varphi_l) \geq \sum_l \alpha_l\epsilon_0/2 \\ \Rightarrow \sum_l \alpha_l &\leq 2/\epsilon_0. \end{aligned}$$

Set $\tilde{M} = 2\xi_0/\epsilon_0$, then on the event S_n, $\hat{\mathcal{G}}_n \subset \mathcal{G}_n^{\tilde{M}}$ is always satisfied. Since a_0^n is bounded from above and below on the event S_n, we can take a sufficiently large $\bar{M} > \tilde{M}$ such that $g/a_0^n \in \mathcal{G}_n^{\bar{M}}$ $(\forall g \in \hat{\mathcal{G}}_n)$. ■

A.2 Proof of Theorem 2

The proof is a version of Theorem 10.13 in van de Geer (2000). We set $g' := \frac{g_n^* + \hat{g}_n}{2}$. Since $Q_n g' = Q_n \hat{g}_n = 1$,

$$
\begin{aligned}
& - P_n \log\left(\frac{\hat{g}_n + g_n^*}{2 g_n^*}\right) \leq 0 \\
\Rightarrow \quad & \delta_n := (Q_n - Q)(g' - g_n^*) - (P_n - P)\log\left(\frac{g'}{g_n^*}\right) \leq 2P\left(\sqrt{\frac{g'}{g_n^*}} - 1\right) - Q(g' - g_n^*) \\
& = 2P\left[\left(1 - \frac{g_n^*}{g_0}\right)\left(\sqrt{\frac{g'}{g_n^*}} - 1\right)\right] + 2P\left[\frac{g_n^*}{g_0}\left(\sqrt{\frac{g'}{g_n^*}} - 1\right)\right] - Q(g' - g_n^*) \\
& = 2Q\left(\sqrt{g_0} - \sqrt{g_n^*}\right)\left(\sqrt{\frac{g_0}{g_n^*}} + 1\right)\left(\sqrt{g'} - \sqrt{g_n^*}\right) - h_Q(g', g_n^*)^2 \\
& \leq (1 + c_0) h_Q(g_0, g_n^*) h_Q(g', g_n^*) - h_Q(g', g_n^*)^2. \qquad (32)
\end{aligned}
$$

If $(1 + c_0) h_Q(g_0, g_n^*) h_Q(g', g_n^*) \geq |\delta_n|$, the assertion immediately follows. Otherwise we can apply the same arguments as Theorem 1 replacing g_0 with g_n^*. ■

B Proofs of Lemma 1, Theorem 3 and Theorem 4

B.1 Proof of Lemma 1

First we prove the consistency of $\check{\alpha}_n$. Note that for $g' = \frac{g_* + \hat{g}_n / a_*^n}{2}$

$$
P \log\left(\frac{g'}{Q(g') g_*}\right) \leq 0, \quad - P_n \log\left(\frac{g'}{g_*}\right) \leq 0.
$$

Thus , we have

$$
- \log Q g' - (P_n - P)\log\left(\frac{g'}{g_*}\right) \leq P \log\left(\frac{g'}{Q(g') g_*}\right) \leq 0. \qquad (33)
$$

In a finite dimensional situation, the inequality (8) is satisfied with arbitrary $\gamma > 0$; see Lemma 2.6.15 in van der Vaart and Wellner (1996). Thus, we can show that the left-hand side of (33) converges to 0 in probability in a similar way to the proof of Theorem 1. This and $\nabla\nabla P \log\left(\frac{\alpha^{\mathrm{T}}\varphi + g_*}{2 g_*}\right)\Big|_{\alpha = \alpha_*} = -I_0/4 \prec O$ give $\hat{\alpha}_n \xrightarrow{p} \alpha_*$.

Next we prove $\sqrt{n}$-consistency. By the KKT condition, we have

$$
\nabla P_n \psi(\hat{\alpha}_n) - \hat{\lambda} + \hat{s}(Q_n \varphi) = \mathbf{0}, \quad \hat{\lambda}^{\mathrm{T}} \hat{\alpha}_n = 0, \quad \hat{\lambda} \leq \mathbf{0}, \qquad (34)
$$

$$
\nabla P \psi(\alpha_*) - \lambda_* + s_*(Q\varphi) = \mathbf{0}, \quad \lambda_*^{\mathrm{T}} \alpha_* = 0, \quad \lambda_* \leq \mathbf{0}, \qquad (35)
$$

with the Lagrange multiplier $\hat{\lambda}, \lambda_* \in \mathbb{R}^b$ and $\hat{s}, s_* \in \mathbb{R}$ (note that KLIEP "maximizes" $P_n\psi(\alpha)$, thus $\hat{\lambda} \le \mathbf{0}$). Noticing that $\nabla\psi(\alpha) = \frac{\varphi}{\alpha^{\mathrm{T}}\varphi}$, we obtain

$$\hat{\alpha}_n^{\mathrm{T}} \nabla P_n \psi(\hat{\alpha}_n) + \hat{s}(Q_n \hat{\alpha}_n^{\mathrm{T}} \varphi) = 1 + \hat{s} = 0. \tag{36}$$

Thus we have $\hat{s} = -1$. Similarly we obtain $s_* = -1$. This gives

$$\hat{\lambda} = \nabla P_n \psi(\hat{\alpha}_n) - Q_n\varphi, \qquad \lambda_* = \nabla P\psi(\alpha_*) - Q\varphi. \tag{37}$$

Therefore, $\hat{\alpha}_n \xrightarrow{p} \alpha_*$ and $g_* \ge \eta_3 > 0$ gives

$$\hat{\lambda} \quad \xrightarrow{p} \quad \lambda_*.$$

Thus the probability of $\{i \mid \hat{\lambda}_i < 0\} \supseteq \{i \mid \lambda_{*,i} < 0\}$ goes to 1 ($\hat{\lambda}_i$ and $\lambda_{*,i}$ mean the i-th element of $\hat{\lambda}$ and λ_* respectively). Recalling the complementary condition $\hat{\lambda}^{\mathrm{T}}\hat{\alpha}_n = 0$, the probability of $\{i \mid \hat{\alpha}_{n,i} = 0\} \supseteq \{i \mid \lambda_{*,i} < 0\}$ goes to 1. Again by the complementary condition $\lambda_*^{\mathrm{T}}\alpha_* = 0$, the probability of

$$(\check{\alpha}_n - \alpha_*)^{\mathrm{T}}\lambda_* = 0$$

goes to 1. In particular $(\check{\alpha}_n - \alpha_*)^{\mathrm{T}}\lambda_* = o_p(1/n)$.

Set $Z_n' := \sqrt{n}(\nabla P_n\psi(\alpha_*) - Q_n\varphi - (\nabla P\psi(\alpha_*) - Q\varphi))$. By the optimality and consistency of $\check{\alpha}_n$, we obtain

$$\begin{aligned}
0 \le& P_n\psi(\check{\alpha}_n) - P_n\psi(\alpha_*) \\
=& (\check{\alpha}_n - \alpha_*)^{\mathrm{T}} \nabla P_n\psi(\alpha_*) - \frac{1}{2}(\check{\alpha}_n - \alpha_*)^{\mathrm{T}} I_0 (\check{\alpha}_n - \alpha_*) + o_p\left(\|\check{\alpha}_n - \alpha_*\|^2\right) \\
=& (\check{\alpha}_n - \alpha_*)^{\mathrm{T}} (\lambda_* + \frac{Z_n'}{\sqrt{n}}) - \frac{1}{2}(\check{\alpha}_n - \alpha_*)^{\mathrm{T}} I_0 (\check{\alpha}_n - \alpha_*) + o_p\left(\|\check{\alpha}_n - \alpha_*\|^2\right) \\
=& (\check{\alpha}_n - \alpha_*)^{\mathrm{T}} \frac{Z_n'}{\sqrt{n}} - \frac{1}{2}(\check{\alpha}_n - \alpha_*)^{\mathrm{T}} I_0 (\check{\alpha}_n - \alpha_*) + o_p\left(\|\check{\alpha}_n - \alpha_*\|^2 + 1/n\right)
\end{aligned} \tag{38}$$

because $\nabla\nabla^{\mathrm{T}} P_n\psi(\alpha_*) = -I_0 + o_p(1)$ and $(\check{\alpha}_n - \alpha_*)^{\mathrm{T}}\lambda_* = o_p(1/n)$. Thus noticing $Z_n/\sqrt{n} = \mathcal{O}_p(1/\sqrt{n})$, we obtain the assertion. ■

B.2 Proof of Theorem 3

The proof relies on Self and Liang (1987) and Fukumizu et al. (2004), but we shall pay attention to the fact that the feasible parameter set stochastically behaves and the true importance g_0 may not be contained in the model. Set

$$Z_n := \sqrt{n} I_0^{-1} \left(\nabla P_n\psi(\alpha_*) - Q_n\varphi - (\nabla P\psi(\alpha_*) - Q\varphi)\right).$$

By Lemma 1 and the inequality (38), we obtain

$$\begin{aligned}
0 &\le (\check{\alpha}_n - \alpha_*)^{\mathrm{T}} \nabla P_n\psi(\alpha_*) - \frac{1}{2}(\check{\alpha}_n - \alpha_*)^{\mathrm{T}} I_0 (\check{\alpha}_n - \alpha_*) + o_p\left(1/n\right) \\
&= -\frac{1}{2}\|\check{\alpha}_n - \alpha_* - Z_n/\sqrt{n}\|_0^2 + \frac{1}{2}\|Z_n/\sqrt{n}\|_0^2 + o_p\left(1/n\right).
\end{aligned}$$

We define

$$\rho(\alpha) := \|\alpha - \alpha_* - Z_n/\sqrt{n}\|_0^2,$$
$$\tilde{\alpha}_n := \mathop{\arg\min}_{\alpha\in\mathcal{S}_n, \lambda_*^{\mathrm{T}}\alpha=0} \rho(\alpha), \qquad \ddot{\alpha}_n := \mathop{\arg\min}_{\alpha\in\mathcal{S}, \lambda_*^{\mathrm{T}}\alpha=0} \rho(\alpha).$$

In the following, we show (Step 1) $\sqrt{n}(\breve{\alpha}_n - \tilde{\alpha}_n) = o_p(1)$, (Step 2) $\sqrt{n}(\tilde{\alpha}_n - \ddot{\alpha}_n) = o_p(1)$, and finally (Step 3) derive the asymptotic law of $\sqrt{n}(\ddot{\alpha}_n - \alpha_*)$ and simultaneously it gives the asymptotic law of $\sqrt{n}(\breve{\alpha}_n - \alpha_*)$.

Step 1. Derivation of $\sqrt{n}(\breve{\alpha}_n - \tilde{\alpha}_n) = o_p(1)$.

$\rho(\alpha_*) \geq \rho(\tilde{\alpha}_n)$ implies

$$\|\tilde{\alpha}_n - \alpha_*\|_0 \leq \|\tilde{\alpha}_n - \alpha_* - Z_n/\sqrt{n}\|_0 + \|Z_n/\sqrt{n}\|_0 \leq 2\|Z_n/\sqrt{n}\|_0 = \mathcal{O}_p(1/\sqrt{n}).$$

As shown in the proof of Lemma 1, the probability of $\lambda_*^{\mathrm{T}}\breve{\alpha}_n = 0$ goes to 1. This and the optimality of $\tilde{\alpha}_n$ gives

$$-\frac{1}{2}\rho(\tilde{\alpha}_n) \geq -\frac{1}{2}\rho(\breve{\alpha}_n) - o_p(1/n). \tag{39}$$

Due to the optimality of $\breve{\alpha}_n$, and applying the Taylor expansion of log-likelihood as in (38) to $\tilde{\alpha}_n$ instead of $\breve{\alpha}_n$ we have

$$-\frac{1}{2}\rho(\tilde{\alpha}_n) \leq -\frac{1}{2}\rho(\breve{\alpha}_n) + o_p\left(1/n\right). \tag{40}$$

The condition $\lambda_*^{\mathrm{T}}\tilde{\alpha}_n = 0$ is needed to ensure this inequality. If this condition is not satisfied, we cannot assure more than $\lambda_*^{\mathrm{T}}(\tilde{\alpha}_n - \alpha_*) = \mathcal{O}_p(1/\sqrt{n})$. Combining (39) and (40), we obtain

$$-o_p(1/n) \leq \frac{1}{2}\left(\rho(\breve{\alpha}_n) - \rho(\tilde{\alpha}_n)\right) \leq o_p(1/n).$$

By the optimality of $\tilde{\alpha}_n$ and the convexity of $\mathcal{S}_n$, we obtain

$$\begin{aligned}\|\sqrt{n}(\breve{\alpha}_n - \tilde{\alpha}_n)\|_0^2 &\leq \|\sqrt{n}(\breve{\alpha}_n - \alpha_*) - Z_n\|_0^2 - \|\sqrt{n}(\tilde{\alpha}_n - \alpha_*) - Z_n\|_0^2 \\ &= o_p(1).\end{aligned} \tag{41}$$

Step 2. Derivation of $\sqrt{n}(\tilde{\alpha}_n - \ddot{\alpha}_n) = o_p(1)$.

In a similar way to the case of $\tilde{\alpha}_n$, we can show

$$\ddot{\alpha}_n - \alpha_* = \mathcal{O}_p(1/\sqrt{n}).$$

Let $\tilde{\alpha}'_n$ and $\ddot{\alpha}'_n$ denote the projection of $\tilde{\alpha}_n$ to $\mathcal{S}$ and $\ddot{\alpha}_n$ to $\mathcal{S}_n$:

$$\tilde{\alpha}'_n := \mathop{\arg\min}_{\alpha\in\mathcal{S}, \lambda_*^{\mathrm{T}}\alpha=0} \|\tilde{\alpha}_n - \alpha\|_0, \qquad \ddot{\alpha}'_n := \mathop{\arg\min}_{\alpha\in\mathcal{S}_n, \lambda_*^{\mathrm{T}}\alpha=0} \|\ddot{\alpha}_n - \alpha\|_0.$$

Then

$$\begin{aligned}\|\sqrt{n}(\ddot{\alpha}_n - \alpha_*) - Z_n\|_0 &\geq \|\sqrt{n}(\ddot{\alpha}'_n - \alpha_*) - Z_n\|_0 - \|\sqrt{n}(\ddot{\alpha}'_n - \ddot{\alpha}_n)\|_0 \\ &\geq \|\sqrt{n}(\tilde{\alpha}_n - \alpha_*) - Z_n\|_0 - \|\sqrt{n}(\ddot{\alpha}'_n - \ddot{\alpha}_n)\|_0,\end{aligned}$$

and similarly

$$\|\sqrt{n}(\tilde{\alpha}_n - \alpha_*) - Z_n\|_0 \geq \|\sqrt{n}(\ddot{\alpha}_n - \alpha_*) - Z_n\|_0 - \|\sqrt{n}(\tilde{\alpha}'_n - \tilde{\alpha}_n)\|_0.$$

Thus

$$\begin{aligned}-\|\sqrt{n}(\tilde{\alpha}'_n - \tilde{\alpha}_n)\|_0 &\leq \|\sqrt{n}(\tilde{\alpha}_n - \alpha_*) - Z_n\|_0 - \|\sqrt{n}(\ddot{\alpha}_n - \alpha_*) - Z_n\|_0 \\ &\leq \|\sqrt{n}(\ddot{\alpha}'_n - \ddot{\alpha}_n)\|_0.\end{aligned}$$

So, if we can show

$$\|\sqrt{n}(\tilde{\alpha}'_n - \tilde{\alpha}_n)\|_0 = o_p(1), \quad \|\sqrt{n}(\ddot{\alpha}'_n - \ddot{\alpha}_n)\|_0 = o_p(1), \tag{42}$$

then

$$\begin{aligned}\|\sqrt{n}(\ddot{\alpha}_n - \tilde{\alpha}_n)\|_0 &= \|\sqrt{n}(\ddot{\alpha}_n - \alpha_*) - \sqrt{n}(\tilde{\alpha}'_n - \alpha_*) + \sqrt{n}(\tilde{\alpha}'_n - \tilde{\alpha}_n)\|_0 \\ &\leq \|\sqrt{n}(\ddot{\alpha}_n - \alpha_*) - \sqrt{n}(\tilde{\alpha}'_n - \alpha_*)\|_0 + \|\sqrt{n}(\tilde{\alpha}'_n - \tilde{\alpha}_n)\|_0 \\ &\leq \sqrt{\|\sqrt{n}(\tilde{\alpha}'_n - \alpha_*) - Z_n\|_0^2 - \|\sqrt{n}(\ddot{\alpha}_n - \alpha_*) - Z_n\|_0^2} + o_p(1) \\ &\leq \sqrt{o_p(1) + \|\sqrt{n}(\tilde{\alpha}_n - \alpha_*) - Z_n\|_0^2 - \|\sqrt{n}(\ddot{\alpha}_n - \alpha_*) - Z_n\|_0^2} + o_p(1) \\ &\leq o_p(1).\end{aligned} \tag{43}$$

Thus it is sufficient to prove (42).

Note that as $n \to \infty$, the probabilities of $\ddot{\alpha}_n \in \alpha_* + \mathcal{C}$ and $\tilde{\alpha}_n \in \alpha_* + \mathcal{C}_n$ tend to 1 because $\|\tilde{\alpha}_n - \alpha_*\|, \|\ddot{\alpha}_n - \alpha_*\| = o_p(1)$. Similar to μ_i, we define $\hat{\mu}_i$ using $\hat{\nu}_i := Q_n\varphi_i$ instead of ν_i. It can be easily seen that

$$\hat{\mu}_i \xrightarrow{p} \mu_i,$$

and with high probability

$$\mathcal{C}_n = \left\{ \sum_{i=1}^{b-1} \beta_i \hat{\mu}_i \mid \beta_i \geq 0 \ (i \leq j), \ \beta_i \in \mathbb{R} \right\},$$

where j is the number satisfying $\alpha_{*,i} = 0$ $(i = 1, \ldots, j)$ and $\alpha_{*,i} > 0$ $(i = j+1, \ldots, b)$.

As mentioned above, $\tilde{\alpha}_n - \alpha_* \in \mathcal{C}_n$ and $\ddot{\alpha}_n - \alpha_* \in \mathcal{C}$ with high probability. Thus, $\tilde{\alpha}_n$ and $\ddot{\alpha}_n$ can be expressed as $\tilde{\alpha}_n - \alpha_* = \sum \tilde{\beta}_i \hat{\mu}_i$ and $\ddot{\alpha}_n - \alpha_* = \sum \ddot{\beta}_i \mu_i$. Moreover $\tilde{\alpha}_n - \alpha_* = \mathcal{O}_p(1/\sqrt{n})$ and $\ddot{\alpha}_n - \alpha_* = \mathcal{O}_p(1/\sqrt{n})$ imply $\tilde{\beta}_i, \ddot{\beta}_i = \mathcal{O}_p(1/\sqrt{n})$. Since $\tilde{\alpha}_n, \ddot{\alpha}_n, \alpha_* \in \{\alpha \mid \lambda_*^{\mathrm{T}}\alpha = 0\}$, $\tilde{\beta}_i = 0$ and $\ddot{\beta}_i = 0$ for all i such that $\lambda_{*,i} \neq 0$. This gives

$$\sum \tilde{\beta}_i \mu_i \in \mathcal{C} \cap \{\delta \mid \lambda_*^{\mathrm{T}}\delta = 0\}, \quad \sum \ddot{\beta}_i \hat{\mu}_i \in \mathcal{C}_n \cap \{\delta \mid \lambda_*^{\mathrm{T}}\delta = 0\}.$$

Thus, with high probability, the following is satisfied:

$$\sqrt{n}\|\tilde{\alpha}_n - \tilde{\alpha}'_n\|_0 \le \sqrt{n}\left\|\sum \tilde{\beta}_i\hat{\mu}_i - \sum \tilde{\beta}_i\mu_i\right\|_0 \le \sqrt{n}\sum |\tilde{\beta}_i|\|\hat{\mu}_i - \mu_i\|_0 = o_p(1),$$
$$\sqrt{n}\|\ddot{\alpha}_n - \ddot{\alpha}'_n\|_0 \le \sqrt{n}\left\|\sum \ddot{\beta}_i\mu_i - \sum \ddot{\beta}_i\hat{\mu}_i\right\|_0 \le \sqrt{n}\sum |\ddot{\beta}_i|\|\hat{\mu}_i - \mu_i\|_0 = o_p(1),$$

which imply (42). Consequently (43) is obtained.

Step 3. Derivation of the asymptotic law of $\sqrt{n}(\check{\alpha}_n - \alpha_*)$.

By (41) and (43), we have obtained

$$\sqrt{n}\|\check{\alpha}_n - \ddot{\alpha}_n\|_0 = o_p(1). \tag{44}$$

By the central limit theorem,

$$\sqrt{n}(\nabla P_n\psi(\alpha_*) - \nabla P\psi(\alpha_*)) \rightsquigarrow Z_1, \quad \sqrt{n}(Q_n\varphi - Q\varphi) \rightsquigarrow Z_2.$$

The independence of Z_1 and Z_2 follows from the independence of P_n and Q_n. Thus by the continuous mapping theorem, we have

$$Z_n \rightsquigarrow I_0^{-1}(Z_1 + Z_2).$$

A projection to a closed convex set is a continuous map. Thus, by the continuous mapping theorem, it follows that

$$\sqrt{n}(\ddot{\alpha}_n - \alpha_*) \rightsquigarrow \mathop{\arg\min}_{\delta\in\mathcal{C},\lambda_*^{\mathrm{T}}\delta=0} \|\delta - Z\|_0.$$

By (44) and Slusky's lemma,

$$\sqrt{n}(\check{\alpha}_n - \alpha_*) \rightsquigarrow \mathop{\arg\min}_{\delta\in\mathcal{C},\lambda_*^{\mathrm{T}}\delta=0} \|\delta - Z\|_0.$$

This concludes the proof. ■

B.3 Proof of Theorem 4

Note that

$$\sqrt{n}(\hat{\alpha}_n - \alpha_*) - \sqrt{n}(\check{\alpha}_n - \alpha_*) = \sqrt{n}(1 - 1/a_*^n)\hat{\alpha}_n.$$

From the definition, $\sqrt{n}(1/a_*^n - 1) = \sqrt{n}(Q_n(g_*) - 1) \rightsquigarrow \alpha_*^{\mathrm{T}} Z_2$. Now $\alpha_*^{\mathrm{T}}(I_0 - P(\varphi/g_*)P(\varphi^{\mathrm{T}}/g_*))\alpha_* = 0$ which implies $\alpha_*^{\mathrm{T}} Z_1 = 0$ $(a.s.)$, thus $\alpha_*^{\mathrm{T}} Z_2 = \alpha_*^{\mathrm{T}} I_0 Z$ $(a.s.)$. Recalling $\hat{\alpha}_n \xrightarrow{p} \alpha_*$, we obtain the assertion by Slusky's lemma and the continuous mapping theorem. ■

Acknowledgments

This work was supported by MEXT (17700142 and 18300057), the Okawa Foundation, the Microsoft CORE3 Project, and the IBM Faculty Award.

References

P. Baldi and S. Brunak. *Bioinformatics: The Machine Learning Approach*. MIT Press, Cambridge, 1998.

P. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33:1487–1537, 2005.

S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th International Conference on Machine Learning*, 2007.

S. Bickel and T. Scheffer. Dirichlet-enhanced spam filtering based on biased samples. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.

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

O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical process. *C. R. Acad. Sci. Paris Ser. I Math.*, 334:495–500, 2002.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

K. Fukumizu, T. Kuriki, K. Takeuchi, and M. Akahira. *Statistics of Singular Models*. The Frontier of Statistical Science 7. Iwanami Syoten, 2004. in Japanese.

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.

H. Hachiya, T. Akiyama, M. Sugiyama, and J. Peters. Adaptive importance sampling with automatic model selection in value function approximation. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI-08)*, Chicago, USA, Jul. 13–17 2008.

P. Hall. On Kullback-Leibler loss and density estimation. *The Annals of Statistics*, 15(4): 1491–1519, 1987.

W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models*. Springer Series in Statistics. Springer, Berlin, 2004.

J. J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–162, 1979.

J. Huang, A. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 601–608. MIT Press, Cambridge, MA, 2007.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 34:2593–2656, 2006.

C.-J. Lin, R. C. Weng, and S. S. Keerthi. Trust region Newton method for large-scale logistic regression. Technical report, Department of Computer Science, National Taiwan University, 2007. URL `http://www.csie.ntu.edu.tw/~cjlin/liblinear/`.

S. Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48:1977–1991, 2002.

T. P. Minka. A comparison of numerical optimizers for logistic regression. Technical report, Microsoft Research, 2007. URL `http://research.microsoft.com/~minka/papers/logreg/minka-logreg.pdf`.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functions and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems 20*, 2007.

C. E. Rasmussen, R. M. Neal, G. E. Hinton, D. van Camp, M. Revow, Z. Ghahramani, R. Kustra, and R. Tibshirani. The DELVE manual, 1996. URL `http://www.cs.toronto.edu/~delve/`.

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

E. Schuster and C. Gregory. On the non-consistency of maximum likelihood nonparametric density estimators. In W. F. Eddy, editor, *In Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, pages 295–298, New York, NY, USA, 1982. Springer.

S. G. Self and K. Y. Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82:605–610, 1987.

C. R. Shelton. *Importance Sampling for Reinforcement Learning with Multiple Objectives*. PhD thesis, Massachusetts Institute of Technology, 2001.

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

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

M. Sugiyama, M. Kawanabe, and K.-R. Müller. Trading variance reduction with unbiasedness: The regularized subspace information criterion for robust model selection in kernel regression. *Neural Computation*, 16(5):1077–1104, 2004.

M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, May 2007.

M. Sugiyama and K.-R. Müller. Input-dependent estimation of generalization error under covariate shift. *Statistics & Decisions*, 23(4):249–279, 2005.

M. Sugiyama, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems 20*, Cambridge, MA, 2008. MIT Press.

R. S. Sutton and G. A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Annals of Probability*, 22:28–76, 1994.

M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126:505–563, 1996a.

M. Talagrand. A new look at independence. *Annals of Statistics*, 24:1–34, 1996b.

Y. Tsuboi, H. Kashima, S. Hido, S. Bickel, and M. Sugiyama. Direct density ratio estimation for large-scale covariate shift adaptation. In M. J. Zaki, K. Wang, C. Apte, and H. Park, editors, *Proceedings of the Eighth SIAM International Conference on Data Mining (SDM2008)*, pages 443–454, Atlanta, Georgia, USA, Apr. 24–26 2008.

S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes. with Applications to Statistics*. Springer, New York, 1996.

J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan. Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6):767–791, 2002.