Johann Radon Institute for
Computational and Applied Mathematics
Austrian Academy of Sciences (ÖAW)

On a regularization of unsupervised domain adaptation in RKHS

E.R. Gizewski, L. Mayer, B.A. Moser,
D.H. Nguyen, S. Pereverzyev Jr,
S.V. Pereverzyev, N. Shepeleva,
W. Zellinger

RICAM-Report 2021-17

On a regularization of unsupervised domain adaptation in RKHS

Elke R. Gizewski[*1,2], Lukas Mayer[†3], Bernhard A. Moser[‡4], Duc Hoan Nguyen[§5], Sergiy Pereverzyev Jr[¶1,2], Sergei V. Pereverzyev[‖5], Natalia Shepeleva[**4] and Werner Zellinger[††4]

[1]*Department of Neuroradiology, Medical University of Innsbruck, Anichstrasse 35, 6020 Innsbruck, Austria*
[2]*Neuroimaging Research Core Facility,Medical University of Innsbruck, Anichstrasse 35, 6020 Innsbruck, Austria*
[3]*Department of Neurology, Medical University of Innsbruck, Anichstrasse 35, 6020 Innsbruck, Austria*
[4]*Software Competence Center Hagenberg, Softwarepark 21, 4232 Hagenberg, Austria*
[5]*Johann Radon Institute, Altenberger Strasse 69, 4040 Linz, Austria*

Abstract

We analyze the use of the so-called general regularization scheme in the scenario of unsupervised domain adaptation under the covariate shift assumption. Learning algorithms arising from the above scheme are generalizations of importance weighted regularized least squares method, which up to now is among the most used approaches in the covariate shift setting. We explore a link between the considered domain adaptation scenario and estimation of Radon-Nikodym derivatives in reproducing kernel Hilbert spaces, where the general regularization

[*]Elke.Gizewski@i-med.ac.at
[†]lukas.mayer@i-med.ac.at
[‡]Bernhard.moser@scch.at
[§]duc.nguyen@ricam.oeaw.ac.at
[¶]sergiy.pereverzyev@i-med.ac.at
[‖]sergei.pereverzyev@oeaw.ac.at
[**]Natalia.shepeleva@scch.at
[††]Werner.Zellinger@scch.at

scheme can also be employed and is a generalization of the kernelized unconstrained least-squares importance fitting. We estimate the convergence rates of the corresponding regularized learning algorithms and discuss how to resolve the issue with the tuning of their regularization parameters. The theoretical results are illustrated by numerical examples, one of which is based on real data collected for automatic stenosis detection in cervical arteries.

Keywords— Unsupervised domain adaptation; Covariate shift; Reproducing kernel Hilbert spaces; General regularization scheme; Radon-Nikodym numerical differentiation; Tuning of regularization parameters.

1 Introduction

This paper is focused on the use of regularized kernel methods in the context of unsupervised domain adaptation under covariate shift.

In statistical learning theory, the domain adaptation scenario arises when one studies two relationships between the explanatory (input) variable $x \in \mathbf{X} \subset \mathbb{R}^d$ and the response (output) variable $y \in \mathbf{Y} \subset \mathbb{R}$ under the assumption that they are governed by different probabilistic laws with respect to measures $p(x, y)$ and $q(x, y)$ on $\mathbf{X} \times \mathbf{Y}$.

This means that an input $x \in \mathbf{X}$ does not determine uniquely an output $y \in \mathbf{Y}$, but rather a conditional probability $\rho(y|x)$ of y given x, which is assumed to be unknown. Then one uses a training data sample $\mathbf{z} = \{(x_i, y_i), x_i \in \mathbf{X}, y_i \in \mathbf{Y}, i = 1, 2, \ldots, n\}, |\mathbf{z}| = n$, drawn independently and identically (i.i.d) from one of the measures, say $p(x, y)$, to reduce the expected risk of the prediction y from x over the other measure $q(x, y)$. In the context of domain adaptation, $p(x, y)$ and $q(x, y)$ are called, respectively, as the source probability and the target probability.

In general, the domain adaptation problem with different source and target probabilities is unsolvable, as $p(x, y)$, $q(x, y)$ could be arbitrarily far apart. Therefore, in the present study we follow [1], [2] and rely on the so-called covariate shift assumption, where only probabilities of inputs in the source (S) and the target (T) domains (marginal probabilities) $\rho_S(x)$ and $\rho_T(x)$ differs, while the conditional probability $\rho(y|x)$ is the same under both the source and the target probabilities. This means that the joint probabilities $p(x, y)$, $q(x, y)$ can be factorized as the following products

$$p(x, y) = \rho(y|x)\rho_S(x), \ q(x, y) = \rho(y|x)\rho_T(x). \tag{1}$$

In this article we restrict ourselves to learning with least squares loss, where the expected risk of the prediction of y from x by means of a function $f : \mathbf{X} \to \mathbf{Y}$ is defined in the target domain as

$$\mathcal{R}_q(f) := \int_{\mathbf{X}\times\mathbf{Y}} (f(x) - y)^2 dq(x, y).$$

It is easy to check that $\mathcal{R}_q(f)$ attains its minimum at the so-called regression function

$$f(x) = f_q(x) = \int_{\mathbf{Y}} y d\rho(y|x). \tag{2}$$

But, in unsupervised domain adaptation setting, neither $\mathcal{R}_q(f)$ nor $f_q(x)$ can be computed, because the information about underlying probability $q(x, y)$ is only provided in the form of a set $X' = (x'_1, x'_2, \ldots, x'_m), |X'| = m$, of unlabeled examples x'_i of inputs drawn i.i.d. from the target marginal probability measure $\rho_T(x)$. Then, the goal is to use this information, together with training data $\mathbf{z}$, to approximate the ideal minimizer f_q by an empirical estimator $f_{\mathbf{z}}$ in the sense of excess risk

$$\mathcal{R}_q(f_{\mathbf{z}}) - \mathcal{R}_q(f_q) = \|f_{\mathbf{z}} - f_q\|^2_{L_{2,\rho_T}};$$

here L_{2,ρ_T} is the space of square integrable functions $f : \mathbf{X} \to \mathbb{R}$ with respect to the marginal probability measure ρ_T.

Observe that in the standard supervised learning setting [3] one would use training data $\mathbf{z}$ to approximate the ideal minimizer f_p of the expected risk $\mathcal{R}_p(f)$ over the source probability p. Then, from (1), (2) it follows that in the context of unsupervised domain adaptation under covariate shift, we are aiming at approximating the same regression function $f^*(x) = f_p(x) = f_q(x)$ given by (2) as in the standard supervised learning. The main difference between supervised learning and domain adaptation is now that we are interested in an empirical estimator $f_{\mathbf{z}}$ committing as little error as possible not in the space L_{2,ρ_S} associated with marginal source measure ρ_S, but in the space L_{2,ρ_T} generated by the target measure ρ_T.

On the other hand, since in unsupervised domain adaptation under covariate shift the aim of approximation is the same function $f^*(x)$ as in the standard supervised learning, it is natural to adjust the methods developed there to the domain adaptation scenario.

Note that supervised learning in reproducing kernel Hilbert spaces (RKHS) is one of the most well-developed parts of statistical learning theory, and regularized kernel ridge regression is one of the most well-understood supervised learning algorithm. This algorithm has been already employed in unsupervised domain adaptation in combination with sample reweighting [4], [5], but to the best of our knowledge, no risk bounds were known for this combination, even under covariate shift assumption.

Note also that regularized kernel ridge regression is just a particular example of a linear regularization scheme in RKHS. At the same time, a large class of regularization scheme in RKHS, collectively known as spectral regularization, has been extensively studied in supervised learning setting (see,e.g, [6], [7], [8], and the references therein). First contribution of the present study is that in the next section we show how the analysis of [6] can be extended to the setting of domain adaptation with covariate shift. Moreover, in Section 5 using a toy example we demonstrate a potential advantage of the use of general regularization scheme compared to a combination of regularized kernel ridge regression and sample reweighting.

Since we have no direct access to the target probability measure ρ_T and to the space L_{2,ρ_T} in which we are going to approximate the regression function $f^* = f_q$, some additional assumptions should be imposed on the relationship between the source probability ρ_S and the target probability ρ_T. In the present study we follow [2] and assume that there is a function $\beta : \mathbf{X} \to \mathbb{R}_+$ such that

$$d\rho_T(x) = \beta(x) d\rho_S(x).$$

Then $\beta(x)$ can be viewed as the Radon-Nikodym derivative $\frac{d\rho_T}{d\rho_S}$ of the target measure with respect to the source measure. In the next section, we assume that $\beta(x)$ is exactly given and discuss how supervised learning algorithms based on the general regularization scheme can be adjusted to the context of domain adaptation.

In practice, however, neither ρ_S nor $\beta = \frac{d\rho_T}{d\rho_S}$ is known. Therefore, in Section 3 we at first discuss how $\beta(x)$ can be approximately reconstructed from unlabeled examples of inputs drawn according to the source and target probabilities. In the above reconstruction, the general regularization scheme in RKHS can be employed again. Our results in this direction extend and specify recent results of [9]. Then we discuss how to employ the general regularization scheme in unsupervised domain adaptation without knowing the exact values of Radon-Nikodym derivative $\beta(x)$ and estimate the accuracy of the corresponding approximations.

The problem of domain adaptation has been tackled by many approaches, and a number of surveys has been created on this topic. Here we refer to a recent survey [10]. Most domain adaptation algorithms depend on the so-called hyperparameters that change the performance of the algorithm and need to be tuned. Usually, algorithm performance variation can be attributed to just a few hyperparameters, such as a regularization parameter in kernel ridge regression. In spite of its importance, the question of selecting these parameters has not been much studied in the context of domain adaptation.

Note that usually an adaptive (data-driven) choice of a regularization parameter is made by trying several combinations of its values, and then by selecting one of them according to some performance criteria, such as cross-validation, for example.

Even leaving aside that the use of cross-validation is problematic in unsupervised domain adaptation, in the approach above one selects only one element from a family of approximants computed by an employed domain adaptation algorithm. Of course, the other approximants corresponding to the tried parameter values are used in the selection process, but then they are rejected, in spite of the numerical expenses made for their construction. At the same time, the rejected approximants may also contribute to the improvement of approximation accuracy, as it will be demonstrated in the last section.

In Section 4 we explore the idea to use the computed approximations in the construction of a new one. More precisely, the idea is to construct a new approximant in the form of a linear combination of approximants $f_1, f_2, \ldots, f_l$, computed for all tried parameter values, in a way that it should mimic the best approximation by such linear combinations. There are several implementations of this idea in the context of supervised learning [11], [12], [13], and [14], where it is called as the

aggregation by the linear functional strategy, because it is based on the technique [15] originally developed in the regularization theory.

In Section 4 we extend the above idea to the case of domain adaptation. Note that the aggregation by the linear functional strategy presupposes that the aggregated elements $f_1, f_2, \ldots, f_l$ belong to some RKHS $\mathcal{H} = \mathcal{H}_K$ with a positive-definite function $K : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ as reproducing kernel. This assumption is natural for approximants $f_1, f_2, \ldots, f_l$ constructed by means of kernel ridge regression. Moreover, in view of recently proposed kernel framework for analysing deep networks [16] the above assumption does not pose a significant limitation when dealing with approximants resulting from neural networks algorithms. The space $\mathcal{H}_K$ just should be wide enough, such as a Sobolev sapace W_2^γ with a moderate index of smoothness γ, or RKHS generated by a sum of several, possibly universal [17], reproducing kernels.

Therefore, the approach presented in Section 4 may potentially be used not only for the algorithms that are based on the general regularization scheme in RKHS, but in the present study we restrict ourselves only to that class of algorithms.

Finally, in Section 5 we present some numerical tests illustrating the theoretical results.

2 Risk bounds under the assumption of knowing the Radon-Nikodym derivative.

2.1 Assumptions and auxiliaries

From now on we assume that the regression function $f^* = f_p = f_q$, minimizing the expected risks $\mathcal{R}_p(f)$, $\mathcal{R}_q(f)$, belongs to a specified reproducing kernel Hilbert space $\mathcal{H}_K$. Such assumption is rather common in supervised learning, where it is referred to as "well-specified" or "inner regularity" case, see [18], [19] and [20].

Let $J_T : \mathcal{H}_K \hookrightarrow L_{2,\rho_T}$ and $J_S : \mathcal{H}_K \hookrightarrow L_{2,\rho_S}$ be the inclusion operators. Recall that the information about the source and target marginal measures are only provided in the form of samples $X_S = \{x_1, x_2, \ldots, x_n\}$ and $X_T = \{x'_1, x'_2, \ldots, x'_m\}$ drawn independently and identically (i.i.d) from ρ_S and ρ_T respectively. In the sequel, we distinguish two sample operators

$$S_{X_T} f = (f(x'_1), f(x'_2), \ldots, f(x'_m)) \in \mathbb{R}^m,$$
$$S_{X_S} f = (f(x_1), f(x_2), \ldots, f(x_n)) \in \mathbb{R}^n,$$

acting from $\mathcal{H}_K$ to $\mathbb{R}^m$ and $\mathbb{R}^n$, where the norms in later spaces are m^{-1}-times and n^{-1}-times the standard Euclidean norms, such that the adjoint operators $S^*_{X_T}$:

$\mathbb{R}^m \to \mathcal{H}_K$ and $S^*_{X_S} : \mathbb{R}^n \to \mathcal{H}_K$ are given as

$$S^*_{X_T} u(\cdot) = \frac{1}{m} \sum_{j=1}^{m} K(\cdot, x'_j) u_j, \qquad u = (u_1, u_2, \ldots, u_m) \in \mathbb{R}^m,$$

$$S^*_{X_S} v(\cdot) = \frac{1}{n} \sum_{i=1}^{n} K(\cdot, x_i) v_i, \qquad u = (v_1, v_2, \ldots, v_n) \in \mathbb{R}^n.$$

In this section, we assume that we have access to the values $\beta(x_i)$ of the Radon-Nikodym derivative $\beta(x) = \frac{d\rho_T(x)}{d\rho_S(x)}$ at the points $x_i, i = 1, 2, \ldots, n$, drawn i.i.d from $\rho_S(x)$, and consider a diagonal $n \times n$ matrix $B = diag(\beta(x_1), \beta(x_2), \ldots, \beta(x_n))$.

Moreover, as in [2], we assume that $\beta(x)$ is uniformly bounded on $\mathbf{X}$, such that $|\beta(x)| \leq b_0$ for some $b_0 > 0$ and any $x \in \mathbf{X}$.

The subsequent analysis is based on two others assumptions, both of which are quite common and not restrictive. Namely, in what follows we always assume that $K : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ is a continuous and bounded kernel that is for any $x \in \mathbf{X}$

$$\|K(\cdot, x)\|_{\mathcal{H}_K} = \langle K(\cdot, x), K(\cdot, x) \rangle^{\frac{1}{2}}_{\mathcal{H}_K} = [K(x, x)]^{\frac{1}{2}} \leq \kappa_0 < \infty.$$

In addition, we assume that for any input $x \in \mathbf{X}$ the corresponding output $y \in \mathbf{Y} \subset \mathbb{R}$ satisfies the bound $|y| \leq y_0$ for some $y_0 > 0$.

Moreover, in the sequel we adopt the convention that c denotes a generic positive coefficient, which can vary from appearance to appearance and may only depend on basic parameter such as ρ_S, ρ_T, κ_0, b_0, y_0 and others introduced below.

We will need the following statement.

Lemma 1. *With probability at least $1 - \delta$ we have*

$$\left\| S^*_{X_T} S_{X_T} - S^*_{X_S} B S_{X_S} \right\|_{\mathcal{H}_K \to \mathcal{H}_K} \leq c \left(\log^{\frac{1}{2}} \frac{1}{\delta} \right) \left(m^{-\frac{1}{2}} + n^{-\frac{1}{2}} \right),$$

$$\left\| S^*_{X_T} S_{X_T} f^* - S^*_{X_S} B \bar{y} \right\|_{\mathcal{H}_K} \leq c \left(\log^{\frac{1}{2}} \frac{1}{\delta} \right) \left(m^{-\frac{1}{2}} + n^{-\frac{1}{2}} \right),$$

where $\bar{y} = (y_1, y_2, \ldots, y_n)$ is the vector of outputs corresponding to the inputs $X_S = \{x_1, x_2, \ldots, x_n\}$.

The proof of Lemma 1 is based on Lemma 4 of [2], which we formulate in our notations as follows

Lemma 2. *([2]) Let ϕ be a map from $\mathbf{X}$ into $\mathcal{H}_K$ such that $\|\phi(x)\|_{\mathcal{H}_K} \leq R$ for all $x \in \mathbf{X}$. Then with probability at least $1 - \delta$ it holds*

$$\left\| \frac{1}{m} \sum_{j=1}^{m} \phi(x'_j) - \frac{1}{n} \sum_{i=1}^{n} \beta(x_i) \phi(x_i) \right\|_{\mathcal{H}_K} \leq \left(1 + \sqrt{2 \log \frac{2}{\delta}} \right) R \sqrt{\frac{b_0^2}{n} + \frac{1}{m}}$$

Moreover, we will need a concentration inequality that follows from [21], see also [22].

Lemma 3 (Concentration lemma)**.** *If $\xi_1, \xi_2, \ldots, \xi_n$ are zero mean independent random variables with values in a separable Hilbert space, such as, say, RKHS $\mathcal{H}_K$, and for some $D > 0$ one has $\|\xi_i\|_{\mathcal{H}_K} \leq D$, $i = 1, 2, \ldots, n$, then the following bound*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i \right\|_{\mathcal{H}_K} \leq \frac{D\sqrt{2 \log \frac{2}{\delta}}}{\sqrt{n}}$$

holds true with probability at least $1 - \delta$.

Proof of Lemma 1

Proof. For any $f \in \mathcal{H}_K$ we define a map $\phi = \phi_f : \mathbf{X} \to \mathcal{H}_K$ as $\phi_f(x) = K(\cdot, x)f(x)$, $x \in \mathbf{X}$. It clear that

$$\|\phi_f(x)\|_{\mathcal{H}_K} = |f(x)| \, \|K(\cdot, x)\|_{\mathcal{H}_K} \leq \kappa_0 |f(x)|.$$

Moreover,

$$|f(x)| = |\langle K(\cdot, x), f(x) \rangle_{\mathcal{H}_K}| \leq \|K(\cdot, x)\|_{\mathcal{H}_K} \|f\|_{\mathcal{H}_K} \leq \kappa_0 \|f\|_{\mathcal{H}_K},$$

such that for the map $\phi = \phi_f$ the condition of the above Lemma ([2]) is satisfied with $R = \kappa_0^2 \|f\|_{\mathcal{H}_K}$. Then directly from that lemma for any $f \in \mathcal{H}_K$ we have

$$\begin{aligned} \left\| S^*_{X_T} S_{X_T} f - S^*_{X_S} B S_{X_S} f \right\|_{\mathcal{H}_K} &= \left\| \frac{1}{m} \sum_{j=1}^{m} \phi_f(x'_j) - \frac{1}{n} \sum_{i=1}^{n} \beta(x_i) \phi_f(x_i) \right\|_{\mathcal{H}_K} \\ &\leq \left(1 + \sqrt{2 \log \frac{2}{\delta}} \right) \left(\sqrt{\frac{b_0^2}{n} + \frac{1}{m}} \right) \kappa_0^2 \|f\|_{\mathcal{H}_K} \\ &\leq c \left(\log^{\frac{1}{2}} \frac{1}{\delta} \right) \left(m^{-\frac{1}{2}} + n^{-\frac{1}{2}} \right) \|f\|_{\mathcal{H}_K} \end{aligned}$$

that proves the first statement of Lemma 1.
To prove the second statement we observe that from just proved bound one has

$$\begin{aligned} \left\| S^*_{X_T} S_{X_T} f^* - S^*_{X_S} B \bar{y} \right\|_{\mathcal{H}_K} \leq & \left\| S^*_{X_T} S_{X_T} f^* - S^*_{X_S} B S_{X_S} f^* \right\|_{\mathcal{H}_K} \\ & + \left\| S^*_{X_S} B S_{X_S} f^* - S^*_{X_S} B \bar{y} \right\|_{\mathcal{H}_K} \\ \leq & c \log^{\frac{1}{2}} \frac{1}{\delta} \left(m^{-\frac{1}{2}} + n^{-\frac{1}{2}} \right) \|f\|_{\mathcal{H}_K} \\ & + \left\| S^*_{X_S} B S_{X_S} f^* - S^*_{X_S} B \bar{y} \right\|_{\mathcal{H}_K} \end{aligned} \tag{3}$$

Consider now the map $\xi : \mathbf{X} \times \mathbf{Y} \to \mathcal{H}_K$ defined by

$$\xi(x, y) = K(\cdot, x)(f^*(x) - y)\beta(x).$$

It is clear that

$$\begin{aligned}\|\xi(x,y)\|_{\mathcal{H}_K} &= \|K(\cdot,x)\|_{\mathcal{H}_K}\left|\int_{\mathbf{Y}} y' d\rho(y'|x) - y\right| |\beta(x)| \\ &\leq 2y_0 b_0 \kappa_0.\end{aligned}$$

Moreover, for $p(x,y) = \rho(y|x)\rho_S(x)$ we have

$$\begin{aligned}\int_{\mathbf{X}\times\mathbf{Y}} \xi(x,y) dp(x,y) &= \int_{\mathbf{X}} K(\cdot,x)\beta(x) \int_{\mathbf{Y}} \left(\int_{\mathbf{Y}} y' d\rho(y'|x) - y\right) d\rho(y|x) d\rho_S(x) \\ &= 0,\end{aligned}$$

such that for (x_i, y_i), $i = 1, 2, \ldots, n$, drawn i.i.i from the measure $p(x,y)$ the corresponding values $\xi_i = \xi(x_i, y_i)$ are zero mean independent random variables in $\mathcal{H}_K$. Then for the just defined $\xi_i = K(\cdot, x_i)(f^*(x_i) - y_i)\beta(x_i)$ the conditions of Concentration lemma are satisfied with $D = 2y_0 b_0 \kappa_0$, such that

$$\left\|S^*_{X_S} B S_{X_S} f^* - S^*_{X_S} B\bar{y}\right\|_{\mathcal{H}_K} = \left\|\frac{1}{n}\sum_{i=1}^n \xi_i\right\|_{\mathcal{H}_K} \leq \frac{2y_0 b_0 \kappa_0 \sqrt{2\log\frac{2}{\delta}}}{\sqrt{n}}.$$

This bound together with (3) gives us the second statement of Lemma 1. □

2.2 General regularization scheme in covariate shift domain adaptation problem

One of the most popular approaches to the approximation of the minimizer $f^* = f_q$ of the target expected risk $\mathcal{R}_q(f)$ by using the data $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^n$ sampled from the source measure $p(x,y)$ is penalized least squares regression combined with sample reweighting, that is also called as importance weighted regularized least squares (IWRLS), see, e.g [23], [4] and [5]. If the approximation is performed in $\mathcal{H}_K$, and as above, we assume that we have access to the values $\beta_i = \beta(x_i)$, $i = 1, 2, \ldots, n$, of the Radon-Nikodym derivative, then within IWRLS-approach the approximant $f_{\mathbf{z}} = f^{\lambda}_{\mathbf{z}}$ of $f^* = f_q$ is constructed as the minimizer of weighted and penalized empirical risk

$$\mathcal{R}_{\mathbf{z},\lambda,\beta}(f) = \frac{1}{n}\sum_{i=1}^n \beta_i \left(f(x_i) - y_i\right)^2 + \lambda\|f\|^2_{\mathcal{H}_K}.$$

Since $\beta_i = \beta(x_i)$ are assumed to be non-negative, we can use our notations and rewrite $\mathcal{R}_{\mathbf{z},\lambda,\beta}$ in the form of the so-called Tikhonov regularization functional

$$\mathcal{R}_{\mathbf{z},\lambda,\beta}(f) = \left\|B^{\frac{1}{2}} S_{X_S} f - B^{\frac{1}{2}}\bar{y}\right\|^2_{\mathbb{R}^n} + \lambda\|f\|^2_{\mathcal{H}_K},$$

$B^{\frac{1}{2}} = diag(\sqrt{\beta_1}, \sqrt{\beta_2}, \ldots, \sqrt{\beta_n})$, such that its minimizer admits the following representation

$$f^{\lambda}_{\mathbf{z}} = (\lambda\mathbf{I} + S^*_{X_S} B S_{X_S})^{-1} S^*_{X_S} B\bar{y} \tag{4}$$

It is well-known that Tikhonov regularization scheme can be profitably used in the standard supervised learning context. Here we refer to [24], where Tikhonov regularization was analysed as a supervised learning algorithm in RKHS, and the best-known kernel independent convergence rates were obtained for this scheme. Then in [6] it has been shown that the same type of results are true for a large class of supervised learning algorithms which are essentially all the linear regularization schemes. Below we show how the analysis of [6] can be extended to the setting of domain adaptation with covariate shift, such that IWRLS-approach (4) will be covered.

Recall (see, e.g., Definition 2.2 in [25]) that regularization schemes can be indexed by parametrized functions $g_\lambda : [0, c] \to \mathbb{R}$, $\lambda > 0$. The only requirements are that there are positive constants $\gamma_0, \gamma_{-\frac{1}{2}}, \gamma_{-1}$ for which

$$\begin{aligned} \sup_{0<t\leq c} |1 - tg_\lambda(t)| &\leq \gamma_0, \\ \sup_{0<t\leq c} \sqrt{t}|g_\lambda(t)| &\leq \frac{\gamma_{-\frac{1}{2}}}{\sqrt{\lambda}}, \\ \sup_{0<t\leq c} \sqrt{t}|g_\lambda(t)| &< \frac{\gamma_{-1}}{\lambda}. \end{aligned} \tag{5}$$

Qualification of the regularization scheme indexed by g_λ is the maximal $\nu > 0$ for which

$$\sup_{0<t\leq c} t^\nu |1 - tg_\lambda(t)| \leq \gamma_\nu \lambda^\nu, \tag{6}$$

where γ_ν does not depend on λ. Following Definition 2.3 of [25] we also say that qualification ν covers a non-decreasing function $\varphi : [0, c] \to \mathbb{R}$, $\varphi(0) = 0$, if the function $t \to \frac{t^\nu}{\varphi(t)}$ is non-decreasing for $t \in (0, c]$.

Keeping in mind that $S^*_{S_X} B S_{S_X}$ is a self-adjoint, non-negative and compact operator on RKHS $\mathcal{H}_K$ one can use the operator functional calculus to represent IWRLS-approximant (4) in terms of the function $g_\lambda(t) = (\lambda + t)^{-1}$ indexing Tikhonov regularization, such that

$$f^\lambda_{\mathbf{z}} = g_\lambda(S^*_{X_S} B S_{X_S}) S^*_{X_S} B \bar{y} \tag{7}$$

It is easy to check that for $g_\lambda(t) = (\lambda + t)^{-1}$ the requirements (5) are satisfied with $\gamma_0 = \gamma_{-1} = 1$, $\gamma_{-\frac{1}{2}} = \frac{1}{2}$. Moreover, the qualification ν of the Tikhonov regularization scheme is equal to 1, and such a small qualification is the main drawback of this scheme.

At the same time, the whole arsenal of regularization schemes $g_\lambda(t)$ can potentially be used to construct approximations $f_{\mathbf{z}} = f^\lambda_{\mathbf{z}}$ of the minimizer $f^* = f_q$ of the target expected risk $\mathcal{R}_q(f)$ in the form (7) from the data $X_S, \bar{y}$ that are sampled from the source measure p. For example, the qualification of the regularization can be increased if one employs the so-called iterated Tikhonov regularization, according to which IWRLS-approach needs to be repeated such that the approximation

$f_{\mathbf{z}}^{\lambda} = f_{\mathbf{z},l}^{\lambda}$ obtained in the previous $l-th$ step plays the role of an initial guess for the next approximation $f_{\mathbf{z}}^{\lambda} = f_{\mathbf{z},l+1}^{\lambda}$ constructed as the minimizer of weighted and penalized empirical risk

$$\mathcal{R}_{\mathbf{z},\lambda,\beta}^{l+1}(f) = \frac{1}{n}\sum_{i=1}^{n}\beta_i(f(x_i)-y_i)^2 + \lambda\left\|f - f_{\mathbf{z},l}^{\lambda}\right\|_{\mathcal{H}_K}^2, \quad f_{\mathbf{z},0}^{\lambda} = 0.$$

After ν such iterations we obtained the approximation $f_{\mathbf{z}}^{\lambda} = f_{\mathbf{z},\nu}^{\lambda}$ that can be represented in the form (7) with

$$g_\lambda(t) = g_{\lambda,\nu}(t) = \frac{1-\frac{\lambda^\nu}{(\lambda+t)^\nu}}{t}.$$

The regularization indexed by $g_{\lambda,\nu}(t)$ has the qualification ν that can be taken as large as desired. Moreover, for $g_\lambda(t) = g_{\lambda,\nu}(t)$ the requirements (5), (6) are satisfied with $\gamma_0 = 1, \gamma_{-\frac{1}{2}} = \nu^{\frac{1}{2}}, \gamma_{-1} = \nu, \gamma_\nu = 1$.

The Landweber iteration is another example of an iteration procedure that is used as a regularization, but in this scheme the number of iteration steps l defines the value of the regularization parameter λ, such that $\lambda = \frac{1}{l}$.

The Landweber iteration is indexed by the function $g_\lambda : [0, c] \to \mathbb{R}$ of the form

$$g_\lambda(t) = \frac{1-(1-\mu t)^l}{t}, \quad \lambda = \frac{1}{l}, \quad 0 < \mu < c^{-1},$$

which satisfies the requirements (5) with $\gamma_0 = 1$, $\gamma_{-\frac{1}{2}} = \sqrt{\mu}, \gamma_{-1} = 1$. Moreover, the Landweber iteration can be considered as a scheme with arbitrary high qualification, but it should be noted that in (6) $\gamma_\nu = \left(\frac{\nu}{\mu e}\right)^\nu \to \infty$ with $\nu \to \infty$. Note also that the algorithm (7) with $B = diag(1, 1, \ldots, 1)$ and $g_\lambda(t)$ indexing the Landweber iteration is known in the standard supervised learning setting as gradient descent learning.

Note that the approximants (7) result from the application of the regularization schemes g_λ to the finite-dimensional equation

$$S_{X_S}^* B S_{X_S} = S_{X_S}^* B\bar{y} \tag{8}$$

However, we are not interested in solving this equation. Instead, we intent to approximate a solution of the equation arising from the minimization of the excess risk

$$\mathcal{R}_q(f) - \mathcal{R}_q(f_q) = \|f - f_q\|_{L_{2,\rho_T}}^2 \tag{9}$$

In RKHS $\mathcal{H}_K$ the above minization can be written in terms of the inclusion operator $J_T : \mathcal{H}_K \to L_{2,\rho_T}$ as $\|J_T f - f_q\|_{L_{2,\rho_T}} \to min$, and it leads to the infinite-dimensional normal equation

$$J_T^* J_T f = J_T^* f_q. \tag{10}$$

Because of compactness of the operator $J_T^* J_T$, its inverse $(J_T^* J_T)^{-1}$ cannot be a bounded operator in $\mathcal{H}_K$, and this makes the equation (10) ill-posed, but since f_q is assumed to be in $\mathcal{H}_K = Range(J_T)$, the Moore-Penrose generalized solution $f^\dagger$ of (10) coincides in $\mathcal{H}_K$ with f_q, or $J_T f^\dagger = f_q$ in L_{2,ρ_T}.

Of course, the equation (10) is not accessible because neither q nor f_q are known, but the result [26] of the regularization theory tells us that there is always a continuous, strictly increasing function $\varphi : [0, \|J_T^* J_T\|_{\mathcal{H}_K}] \to \mathbb{R}$ that obeys $\varphi(0) = 0$ and allows the representation of $f^\dagger = f_q$ in terms of the so-called source condition:

$$f_q = \varphi(J_T^* J_T)\nu_q, \quad \nu_q \in \mathcal{H}_K. \tag{11}$$

The function φ above is usually called index function. Moreover, for every $\epsilon > 0$ one can find such φ that (11) holds true for ν_q with

$$\|\nu_q\|_{\mathcal{H}_K} \le (1+\epsilon)\|f_q\|_{\mathcal{H}_K}.$$

Note that since the operator $J_T^* J_T$ is not accessible, there is a reason to restrict ourselves to consideration of such index functions φ, which allow us to control perturbations in the operators involved in the definition of source conditions. In the context of supervised learning, a class of such index functions has been discussed in [6], and here we follow that study. Namely, we consider the class $\mathcal{F} = \mathcal{F}(0, c)$ of index functions $\varphi : [0, c] \to \mathbb{R}_+$ allowing splitting $\varphi(t) = v(t)\psi(t)$ into monotone Lipschitz part $v, v(t) = 0$, with the Lipschitz constant equal to 1, and an operator monotone part $\psi, \psi(0) = 0$.

Recall that a function ψ is operator monotone on $[0, c]$ if for any pair of self-adjoint operators U, V with spectra in $[0, c]$ such that $U \le V$ (i.e. $V - U$ is an non-negative operator) we have $\psi(U) \le \psi(V)$.

Examples of operator monotone index functions are $\psi(t) = t^\nu$, $\psi(t) = \log^{-\nu}\left(\frac{1}{t}\right)$, $\psi(t) = \log^{-\nu}\left(\log\frac{1}{t}\right), 0 < \nu \le 1$, while an example of a function φ from the above defined class $\mathcal{F}$ is $\varphi(t) = t^r \log^{-\nu}\left(\frac{1}{t}\right), r > 1, 0 < \nu \le 1$, since it can be splitted in a Lipschitz part $v(t) = t^r$ and an operator monotone part $\psi(t) = \log^{-\nu}\left(\frac{1}{t}\right)$.

The following lemma can be proved by repeating line by line the argument of the proof of Theorem 10 in [6] (see also Proposition 4.1 in [25]), where the items denoted there as T_x and $S_x^* y$ should be substituted by $\tilde{T}$ and $\tilde{f}$.

Lemma 4. *Let J be the canonical inclusion of RKHS $\mathcal{H}_K = \mathcal{H}_K(\mathbf{X})$ on $\mathbf{X}$ into $L_{2,\rho} = L_{2,\rho}(\mathbf{X})$, and $T = J^* J$. Consider $f = \varphi(T)v$, where $v \in \mathcal{H}_K$, $\varphi \in \mathcal{F}(0, c)$ and c is large enough. Assume that a self-adjoint, non-negative operator $\tilde{T} : \mathcal{H}_K \to \mathcal{H}_K$ and $\tilde{f} \in \mathcal{H}_K$ are such that for some sufficiently small $\Delta \in (0, 1)$ with probability at least $1 - \delta$ we have*

$$\|T - \tilde{T}\|_{\mathcal{H}_K \to \mathcal{H}_K} \le c\Delta \log^{\frac{1}{2}} \frac{1}{\delta}, \quad \|\tilde{T} f - \tilde{f}\|_{\mathcal{H}_K} \le c\Delta \log^{\frac{1}{2}} \frac{1}{\delta}.$$

If a regularization scheme indexed by $g_\lambda(t)$ has a qualification ν that covers the function φ and $\Delta \le \lambda < 1$, then with probability at least $1 - \delta$ it holds

$$\|f - g_\lambda(\tilde{T})\tilde{f}\|_{\mathcal{H}_K} \le c\left(\varphi(\lambda) + \frac{\Delta}{\lambda}\right) \log \frac{1}{\delta}.$$

If, in addition, the qualification ν covers the function $\varphi(t)\sqrt{t}$, then

$$\|f - g_\lambda(\tilde{T})\tilde{f}\|_{L_{2,\rho}} \le c\left(\varphi(\lambda)\sqrt{\lambda} + \frac{\Delta}{\sqrt{\lambda}}\right)\log\frac{1}{\delta}.$$

The values of coefficients c in the above inequalities do not depend on λ, Δ, δ.

Theorem 1. *Assume that the source condition (11) is satisfied with $\varphi \in \mathcal{F}(0, c)$ and c is large enough. Consider the approximant $f_{\mathbf{z}}^{\lambda}$ given by (7), where the regularization scheme indexed by $g_\lambda(t)$ has the qualification ν that covers the function $\varphi(t)\sqrt{t}$. Consider also the function $\theta(t) = \varphi(t)t$ and choose $\lambda = \lambda_{m,n} = \theta^{-1}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}})$. Then for sufficiently large m and n with probability at least $1 - \delta$ it holds*

$$\|f_q - f_{\mathbf{z}}^{\lambda_{m,n}}\|_{L_{2,\rho_T}} \le c\log\frac{1}{\delta}\,\varphi(\theta^{-1}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}}))\sqrt{\theta^{-1}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}})},$$
$$\|f_q - f_{\mathbf{z}}^{\lambda_{m,n}}\|_{\mathcal{H}_K} \le c\log\frac{1}{\delta}\,\varphi(\theta^{-1}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}}))$$

The values of the coefficients c in the above inequalities do not depend on δ, m, n.

Proof. It is well-known [27] that under the considered assumptions with probability at least $1 - \delta$ it holds

$$\|J_T^* J_T - S_{X_T}^* S_{X_T}\|_{\mathcal{H}_K \to \mathcal{H}_K} \le \frac{2\sqrt{2}\kappa_0 \log^{\frac{1}{2}}\frac{1}{\delta}}{m^{\frac{1}{2}}}.$$

The combination of this bound and the first inequality of Lemma 1 yields that with probability $1 - \delta$ one has

$$\|J_T^* J_T - S_{X_S}^* B S_{X_S}\|_{\mathcal{H}_K \to \mathcal{H}_K} \le c\log^{\frac{1}{2}}\frac{1}{\delta}\,(m^{-\frac{1}{2}} + n^{-\frac{1}{2}}).$$

Moreover, from Lemma 1 with probability $1 - \delta$ we have

$$\begin{aligned}\|S_{X_S}^* B S_{X_S} f_q - S_{X_S}^* B\bar{y}\|_{\mathcal{H}_K} \le& \|S_{X_T}^* S_{X_T} - S_{X_S}^* B S_{X_S}\|_{\mathcal{H}_K \to \mathcal{H}_K}\|f_q\|_{\mathcal{H}_K} \\ &+ \|S_{X_T}^* S_{X_T} f_q - S_{X_S}^* B\bar{y}\|_{\mathcal{H}_K} \\ \le& c\log^{\frac{1}{2}}\frac{1}{\delta}\,(m^{-\frac{1}{2}} + n^{-\frac{1}{2}})\end{aligned}$$

Now we observe that the assumptions of Lemma 4 are satisfied for $T = J_T^* J_T$, $f = f_q$, $\bar{T} = S_{X_S}^* B S_{X_S}$, $\bar{f} = S_{X_S}^* B\bar{y}$, $\Delta = m^{-\frac{1}{2}} + n^{-\frac{1}{2}}$ and

$$\lambda = \lambda_{m,n} = \theta^{-1}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}}) = \theta^{-1}(\Delta) \ge \Delta.$$

Moreover, for $\lambda = \lambda_{m,n}$ and $\Delta = (m^{-\frac{1}{2}} + n^{-\frac{1}{2}})$ we have $\varphi(\lambda)\lambda = \Delta \Rightarrow \varphi(\lambda) = \frac{\Delta}{\lambda}$, $\varphi(\lambda)\sqrt{\lambda} = \frac{\Delta}{\sqrt{\lambda}}$. Then the statement of the theorem follows from Lemma 4. □

Remark 1. *To the best of our knowledge, up to now in the setting of domain adaptation with covariate shift no error bounds were known even for IWRLS-approach (4) that corresponds to Tikhonov regularization scheme* $g_\lambda(t) = (\lambda + t)^{-1}$. *On the other hand, in the standard supervised learning setting this scheme has been analysed in [24] uniformly for whole class of RKHS* $\mathcal{H}_K$ *under the assumption, which in our terms can be written as* $\|(J_T J_T^*)^{-r} f_q\|_{L_{2,\rho_T}} \leq c$ *with* $r > \frac{1}{2}$. *From Proposition 3.2 of [28] we know that the above assumption can be equivalently written as the source condition (11) with* $\varphi(t) = t^{r-\frac{1}{2}}$. *For this index function our Theorem 1 gives respectively the error bounds of orders* $O\left((m^{-\frac{1}{2}} + n^{-\frac{1}{2}})^{\frac{2r}{2r+1}}\right)$ *and* $O\left((m^{-\frac{1}{2}} + n^{-\frac{1}{2}})^{\frac{2r-1}{2r+1}}\right)$ *in* L_{2,ρ_T} *and* $\mathcal{H}_K$. *For a sufficiently large number* $m \geq n$ *of unlabeled inputs* $x'_1, x'_2, \ldots, x'_m$ *sampled from the target measure* ρ_T *the above results match the orders of the bounds [24] in the standard supervised learning setting. The comparison of Theorem 1 with the results [24] (for Tikhonov regularization) and [6] (for general regularization scheme) allows the conclusion that in the scenario of domain adaptation with covariate shift one can guarantee the same order of the error as in the standard supervised learning setting, provided that the number of unlabeled target inputs is big enough, and the values of the Radon-Nikodym derivative at that inputs are known. The later assumption is seldom satisfied in practice. Therefore, in the next section we discuss approximate Radon-Nikodym differentiation and its use in the context of domain adaptation.*

3 Approximate domain adaptation

3.1 Regularized Radon-Nikodym numerical differentiation in RKHS.

In this section our goal is to approximate the Radon-Nikodym derivative $\beta(x) = \frac{d\rho_T}{d\rho_S}$ by some function $\tilde{\beta}(x)$ and then use this approximation within the regularization (7), where the matrix $B = diag(\beta(x_1), \beta(x_2), \ldots, \beta(x_n))$ will be substituted by a matrix $\tilde{B} = diag(\tilde{\beta}(x_1), \tilde{\beta}(x_2), \ldots, \tilde{\beta}(x_n))$. This in fact means that we need a strategy that ensures a good pointwise approximation to the derivatives $\beta(x)$. Then it seems to be natural to approximate $\beta(x)$ in the norm of some reproducing kernel Hilbert space, where pointwise evaluations are well-defined. Such space does not need to be the same as the one we used in the previous section for domain adaptation, but with some abuse of notations we still denote it as $\mathcal{H}_K$.

In the literature various RKHS-based approaches are available for a Radon-Nikodym derivative estimation. Here we may refer to [29] and to references therein. Conceptually, several of the above approaches can be derived from a regularization of an integral equation, which can be written in our terms as

$$J_S^* J_S \beta = J_T^* J_T \mathbf{1} \tag{12}$$

and which is ill-pose similar to (10). Here $\mathbf{1}$ is the constant function that takes the value 1 everywhere, and almost without loss of generality we assume that $\mathbf{1} \in \mathcal{H}_K$,

because otherwise the kernel $K_1(x, x') = 1 + K(x, x')$ will, for example, be used to generate a suitable RKHS containing all constant functions.

Equation (12) originates from the observation that for any bounded and continuous function f its expected value with respect to measure ρ_T coincides with the expected value of $f\beta$ with respect to measure ρ_S, i.e,

$$\int_{\mathbf{X}} f(x')\beta(x')d\rho_S(x') = \int_{\mathbf{X}} f(x')d\rho_T(x').$$

By replacing the function $f(x')$ by $K(x, x')$ we obtain that for any $x \in \mathbf{X}$ it holds

$$\int_{\mathbf{X}} K(x, x')\beta(x')d\rho_S(x') = J_S^*\beta = \int_{\mathbf{X}} K(x, x')d\rho_T(x') = J_T^*\mathbf{1} = J_T^*J_T\mathbf{1}. \quad (13)$$

However, if following [29] and [9] we assume that $\beta = \frac{d\rho_T}{d\rho_S} \in \mathcal{H}_K$, then $J_S^*\beta = J_S^*J_S\beta$ and from (13) we arrive at the equation (12).

Note that $\beta \in \mathcal{H}_K$ is essentially a model assumption that is only needed for theoretical analysis presented below.

Just as the equation (10) is inaccessible, so is the equation (12). But in contrast to (10), the reduction of (12) to a finite-dimensional problem does not require any labels, such as $\bar{y}$, that were necessary for dealing with (8). Since in practice, the amount of unlabeled inputs is usually much greater than that of labeled ones, we assume that the sizes M and N of i.i.d. samples $(x'_1, x'_2, \ldots, x'_M)$ and $(x_1, x_2, \ldots, x_N)$ drawn respectively from ρ_T and ρ_S are much larger than m and n appearing in Theorem 1.

Then we consider two sample operators

$$S_{M,T}f = (f(x'_1), f(x'_2), \ldots, f(x'_M)) \in \mathbb{R}^M,$$
$$S_{N,S}f = (f(x_1), f(x_2), \ldots, f(x_N)) \in \mathbb{R}^N,$$

and the finite-dimensional problem

$$S_{N,S}^*S_{N,S}\beta = S_{M,T}^*S_{M,T}\mathbf{1}, \quad (14)$$

which is an empirical version of the equation (12), where, similar to the above notations the operators $S_{N,S}^* : \mathbb{R}^N \to \mathcal{H}_K$, $S_{M,T}^* : \mathbb{R}^M \to \mathcal{H}_K$ are given as

$$S_{N,S}^*v(\cdot) = \frac{1}{N}\sum_{i=1}^{N} K(\cdot, x_i)v_i, \qquad u = (v_1, v_2, \ldots, v_N) \in \mathbb{R}^N,$$
$$S_{M,T}^*u(\cdot) = \frac{1}{M}\sum_{j=1}^{M} K(\cdot, x'_j)u_j, \qquad u = (u_1, u_2, \ldots, u_M) \in \mathbb{R}^M.$$

A regularization of equations (12), (14) may serve as a starting point for several approaches of estimating the Radon-Nikodym derivative β. For example, as it has been observed in [29], the known kernel mean matching (KMM) method [2] can be

viewed as the regularization of (13), (12) by the method of quasi (least-squares) solutions, originally proposed by Valentin Ivanov (1963) and also known as Ivanov regularization (see, e.g., [30] and [31] for its use in the context of learning). In KMM an Ivanov-type regularization is applied to the empirical version (14) that allows one to explicitly control a constrained approximation of the values $\beta_i = \beta(x_i)$, $i = 1, 2, \ldots, n, n < N$, but it leads to a quadratic problem in variables β_i.

At the same time, the kernelized unconstrained least-squares importance fitting (KuLSIF) proposed in [29] allows an analytic-form solution and can be reduced to solving a linear problem with respect to corresponding variables.

From Theorem 1 of [29] it follows that in KuLSIF the approximation $\tilde{\beta}$ of the Radon-Nikodym derivative $\beta = \frac{d\rho_T}{d\rho_S}$ is in fact constructed by application of the Tikhonov regularization scheme to the empirical version (14) of the equation (12), that is in KuLSIF we have

$$\tilde{\beta} = \beta^{\lambda}_{M,N} = g_\lambda(S^*_{N,S}S_{N,S})S^*_{M,S}S_{M,S}\mathbf{1}, \tag{15}$$

where $g_\lambda(t) = (\lambda + t)^{-1}$.

Though there are several studies devoted to KMM and KuLSIF, to the best of our knowledge there has been no study of pointwise approximation error $\beta(x)-\tilde{\beta}(x)$, which is of interest in the analysis of regularized domain adaptation methods, such as IWRLS. For example, in [29] and [32] (see Type I setting there) the statistical consistency and accuracy of KuLSIF have been analysed in the space L_{2,ρ_S}, where pointwise evaluations are undefined. We can also mention the study [9], where KuLSIF represented as (15) with $g_\lambda(t) = (\lambda + t)^{-1}$ was discussed in a RKHS, but only convergence of $\tilde{\beta}$ to β was proved, without quantifying its rate.

At the same time, using again Lemma 4 and the concept of source conditions naturally appearing because of equation (12), we can obtain the following statement

Theorem 2. *Assume that $\beta = \frac{d\rho_T}{d\rho_S}$ meets source condition $\beta = \phi(J^*_S J_S)\nu_\beta$, where $\phi \in \mathcal{F}(0, c)$, and c is large enough. Consider the approximant $\beta^{\lambda}_{M,N}$ given by (15), where the regularization scheme indexed by $g_\lambda(t)$ has the qualification ν that covers the index function $\phi(t)$. Let $\lambda = \lambda_{M,N} = \theta^{-1}_\phi(M^{-\frac{1}{2}} + N^{-\frac{1}{2}})$, where $\theta_\phi(t) = \phi(t)t$. Then for sufficiently large M an N with probability at least $1 - \delta$ it holds*

$$\left\|\beta - \beta^{\lambda_{M,N}}_{M,N}\right\|_{\mathcal{H}_K} \le c\left(\log\frac{1}{\delta}\right)\phi\left(\theta^{-1}_\phi(M^{-\frac{1}{2}} + N^{-\frac{1}{2}})\right).$$

Proof. Referring again to [27] we have with probability $1 - \delta$ that

$$\left\|J^*_T J_T - S^*_{M,T}S_{M,T}\right\|_{\mathcal{H}_K\to\mathcal{H}_K} \le \frac{c\log^{\frac{1}{2}}\frac{1}{\delta}}{\sqrt{M}},$$

$$\left\|J^*_S J_S - S^*_{N,S}S_{N,S}\right\|_{\mathcal{H}_K\to\mathcal{H}_K} \le \frac{c\log^{\frac{1}{2}}\frac{1}{\delta}}{\sqrt{N}}.$$

Then with the same probability it holds that

$$\begin{aligned}\left\|S^*_{N,S}S_{N,S}\beta - S^*_{M,T}S_{M,T}\mathbf{1}\right\|_{\mathcal{H}_K} \leq& \left\|J^*_S J_S - S^*_{N,S}S_{N,S}\right\|_{\mathcal{H}_K\to\mathcal{H}_K}\|\beta\|_{\mathcal{H}_K}\\ &+ \|J^*_T J_T - S^*_{M,T}S_{M,T}\|_{\mathcal{H}_K\to\mathcal{H}_K}\|\mathbf{1}\|_{\mathcal{H}_K}\\ \leq& c\left(N^{-\frac{1}{2}} + M^{-\frac{1}{2}}\right)\log^{\frac{1}{2}}\frac{1}{\delta},\end{aligned}$$

where we also have used that $\beta \in \mathcal{H}_K$ and β solves (12).

Now the statement of the theorem follows from Lemma 4 if we observe that its assumptions are satisfied for $\varphi = \phi$, $T = J^*_S J_S$, $f = \beta$, $\tilde{T} = S^*_{N,S}S_{N,S}$, $\tilde{f} = S^*_{M,T}S_{M,T}\mathbf{1}$, $\Delta = M^{-\frac{1}{2}} + N^{-\frac{1}{2}}$, and, moreover, that for $\lambda = \lambda_{M,N} = \theta^{-1}_\phi(M^{-\frac{1}{2}} + N^{-\frac{1}{2}}) = \theta^{-1}_\phi(\Delta) \geq \Delta$ it holds

$$\phi(\lambda_{M,N}) = \frac{\Delta}{\lambda_{M,N}} = \phi\left(\theta^{-1}_\phi(M^{-\frac{1}{2}} + N^{-\frac{1}{2}})\right).$$

□

Substituting in (7) the matrix B by the matrix

$$B_{M,N} = diag(\beta^{\lambda_{M,N}}_{M,N}(x_1), \beta^{\lambda_{M,N}}_{M,N}(x_2), \ldots, \beta^{\lambda_{M,N}}_{M,N}(x_n))$$

we can employ general regularization scheme in unsupervised domain adaptation without knowing the exact values of Radon-Nikodym derivative $\beta(x)$.

To estimate the accuracy of the approximation

$$f^\lambda_{\mathbf{z},M,N} = g_\lambda(S^*_{X_S}B_{M,N}S_{X_S})S^*_{X_S}B_{M,N}\,\bar{y} \tag{16}$$

we at first observed that for any $f \in \mathcal{H}_K$

$$\begin{aligned}\left\|S^*_{X_S}BS_{X_S}f - S^*_{X_S}B_{M,N}S_{X_S}f\right\|_{\mathcal{H}_K} &= \frac{1}{n}\left\|\sum_{i=1}^n K(\cdot,x_i)\left(\beta(x_i) - \beta^{\lambda_{M,N}}_{M,N}(x_i)\right)f(x_i)\right\|_{\mathcal{H}_K}\\ &\leq \kappa_0^3\left\|\beta - \beta^{\lambda_{M,N}}_{M,N}\right\|_{\mathcal{H}_K}\|f\|_{\mathcal{H}_K}\end{aligned}$$

Then in view of Theorem 2 with probability $1-\delta$ we have

$$\begin{aligned}&\|J^*_T J_T - S^*_{X_S}B_{M,N}S_{X_S}\|_{\mathcal{H}_K\to\mathcal{H}_K}\\ &\quad\leq \left\|J^*_T J_T - S^*_{X_S}BS_{X_S}\right\|_{\mathcal{H}_K\to\mathcal{H}_K} + \left\|S^*_{X_S}BS_{X_S} - S^*_{X_S}B_{M,N}S_{X_S}\right\|_{\mathcal{H}_K\to\mathcal{H}_K}\\ &\quad\leq c\left(\log^{\frac{1}{2}}\frac{1}{\delta}\right)(m^{-\frac{1}{2}} + n^{-\frac{1}{2}}) + \kappa_0^3\left\|\beta - \beta^{\lambda_{M,N}}_{M,N}\right\|_{\mathcal{H}_K}\\ &\quad\leq c\left(\log\frac{1}{\delta}\right)\left[m^{-\frac{1}{2}} + n^{-\frac{1}{2}} + \phi\left(\theta^{-1}_\phi(M^{-\frac{1}{2}} + N^{-\frac{1}{2}})\right)\right]\end{aligned} \tag{17}$$

In a similar way one can easily check that

$$\left\|S^*_{X_S}B\bar{y} - S^*_{X_S}B_{M,N}\bar{y}\right\|_{\mathcal{H}_K} \leq \kappa_0^2 y_0 \left\|\beta - \beta^{\lambda_{M,N}}_{M,N}\right\|_{\mathcal{H}_K},$$

and then Theorem 2 and the argument from the proof of Theorem 1 with probability $1-\delta$ give us

$$\begin{aligned}
&\|S^*_{X_S}B_{M,N}S_{X_S}f_q - S^*_{X_S}B_{M,N}\bar{y}\|_{\mathcal{H}_K} \\
&\quad\leq \left\|S^*_{X_S}BS_{X_S}f_q - S^*_{X_S}B\bar{y}\right\|_{\mathcal{H}_K} + \left\|S^*_{X_S}BS_{X_S} - S^*_{X_S}B_{M,N}S_{X_S}\right\|_{\mathcal{H}_K\to\mathcal{H}_K}\|f_q\|_{\mathcal{H}_K} \\
&\qquad + \left\|S^*_{X_S}B\bar{y} - S^*_{X_S}B_{M,N}\bar{y}\right\|_{\mathcal{H}_K} \\
&\quad\leq c(m^{-\frac{1}{2}} + n^{-\frac{1}{2}})\log^{\frac{1}{2}}\frac{1}{\delta} + \left\|\beta - \beta^{\lambda_{M,N}}_{M,N}\right\|_{\mathcal{H}_K}\left(\kappa_0^3\|f_q\|_{\mathcal{H}_K} + \kappa_0^2 y_0\right) \\
&\quad\leq c\left(\log\frac{1}{\delta}\right)\left[m^{-\frac{1}{2}} + n^{-\frac{1}{2}} + \phi\left(\theta_\phi^{-1}(M^{-\frac{1}{2}} + N^{-\frac{1}{2}})\right)\right] \qquad (18)
\end{aligned}$$

Now we are in position to formulate a statement which links regularized domain adaptation and Radon-Nikodym numerical differentiation.

Theorem 3. *Let assumptions and conditions of Theorems 1 and 2 be satisfied. Then with probability at least* $1-\delta$ *for*

$$\lambda_\delta = \theta^{-1}\left(\left(\log^{\frac{1}{2}}\frac{1}{\delta}\right)\left(m^{-\frac{1}{2}} + n^{-\frac{1}{2}} + \phi(\theta_\phi^{-1}(M^{-\frac{1}{2}} + N^{-\frac{1}{2}}))\right)\right),$$

it holds

$$\left\|f_q - f^{\lambda_\delta}_{\mathbf{z},M,N}\right\|_{\mathcal{H}_K} \leq c\left(\log^{\frac{3}{2}}\frac{1}{\delta}\right)\varphi\left(\theta^{-1}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}} + \phi(\theta_\phi^{-1}(M^{-\frac{1}{2}} + N^{-\frac{1}{2}})))\right),$$

$$\left\|f_q - f^{\lambda_\delta}_{\mathbf{z},M,N}\right\|_{L_{2,\rho_T}} \leq c\left(\log^{\frac{3}{2}}\frac{1}{\delta}\right)\varphi\left(\theta^{-1}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}} + \phi(\theta_\phi^{-1}(M^{-\frac{1}{2}} + N^{-\frac{1}{2}})))\right)\zeta_0,$$

where $\zeta_0 = \sqrt{\theta^{-1}(m^{-\frac{1}{2}} + n^{-\frac{1}{2}} + \phi(\theta_\phi^{-1}(M^{-\frac{1}{2}} + N^{-\frac{1}{2}})))}$.

Proof. From (17), (18) it follows that the assumptions of Lemma 4 are satisfied for $T = J_T^*J_T$, $f = f_q$, $\tilde{T} = S^*_{X_S}B_{M,N}S_{X_S}$, $\tilde{f} = S^*_{X_S}B_{M,N}\tilde{y}$, and $\Delta = \left(m^{-\frac{1}{2}} + n^{-\frac{1}{2}} + \phi(\theta_\phi(M^{-\frac{1}{2}} + N^{-\frac{1}{2}}))\right)\log^{\frac{1}{2}}\frac{1}{\delta}$.

Therefore, the statement of the theorem follows from Lemma 4 in the same way as in the proofs of Theorems 1 and 2. □

Remark 2. *As has been emphasized in Remark 1, the main massage of Theorem 1 is that for sufficiently large number of unlabeled data sampled from the target domain (marginal probability* ρ_T*) one may potentially guarantee the same order of error bounds as in the standard supervised learning. But for this one needs to know the exact values of the corresponding Radon-Nikodym derivative. From such*

perspective, Theorem 3 continues the above message by saying that in unsupervised domain adaptation error bounds of the same order as in the standard supervised learning may potentially be guaranteed provided that there are big enough amounts of unlabeled data sampled from both target and source domains. To estimate how big these amounts have to be, let us consider the case where KuLSIF-approach is employed in Radon-Nikodym numerical differentiation. As we already noted, it corresponds to (15) with $g_\lambda(t) = (\lambda + t)^{-1}$ *indexing the Tikhonov regularization scheme. Since this is a scheme with the qualification* $\nu = 1$*, its full regularization and approximation capacity will be realized when* $\beta = \frac{d\rho_T}{d\rho_S}$ *meets source condition* $\beta = \phi(J_S^* J_S)\nu_\beta$ *with* $\phi(t) = t$*. Then* $\phi(\theta_\phi^{-1}(M^{-\frac{1}{2}} + N^{-\frac{1}{2}})) = (M^{-\frac{1}{2}} + N^{-\frac{1}{2}})^{\frac{1}{2}}$*, and the error bounds guaranteed by Theorem 3 will be of the same order as the ones in Theorem 1, if* M *and* N *are of order* n^2*. This, for example, means that in unsupervised domain adaptation performed by a combination of IWRLS- and KuLSIF-approaches an amount of unlabeled data should be at least as big as the squared amount of labeled ones to potentially allow an accuracy of the same order as in the standard supervised learning.*

4 Resolving the regularization parameter issue by an aggregation

The choice of the regularization parameters $\lambda_{m,n}, \lambda_{M,N}, \lambda_\delta$ suggested by Theorems 1, 2, 3 crucially relies on the knowledge of the index functions φ, ϕ describing the smoothness of $f_q, \beta = \frac{d\rho_T}{d\rho_S}$ in terms of the corresponding source conditions. Since such smoothness is usually unknown, one faces the issue of how to choose the values of the regularization parameter λ for constructing the approximations (15), (16). Note that the issue with the choice of λ in (15) is easier, because in the regularization theory one can find several parameter choice rules that can guarantee an accuracy of order $\phi\left(\theta_\phi^{-1}(M^{-\frac{1}{2}} + N^{-\frac{1}{2}})\right)$ under the assumption of Theorem 2 and do not require any knowledge of the index functions ϕ. One of such rules is the so-called balancing principle (see, e.g., Section 1.1.5 in [25]), which has been already used in the context of kernel learning in [33], [8]. In particular, Proposition 4.5 from [25] allows us to assume that we already have in our disposal an approximation $\tilde{\beta}_{M,N} = \beta^\lambda_{M,N}$ of $\beta = \frac{d\rho_T}{d\rho_S} \in Range(\phi(J_S^* J_S))$ such that with probability $1 - \delta$ it holds

$$\left\|\beta - \tilde{\beta}_{M,N}\right\|_{\mathcal{H}_K} \leq c\left(\log\frac{1}{\delta}\right)\phi\left(\theta_\phi^{-1}(M^{-\frac{1}{2}} + N^{-\frac{1}{2}})\right)$$

Then the argument from the previous section allows the calculation of such matrix $B_{M,N} = diag(\tilde{\beta}_{M,N}(x_1), \tilde{\beta}_{M,N}(x_2), \ldots, \tilde{\beta}_{M,N}(x_n))$ that (17), (18) hold. As a result, with probability $1 - \delta$ we have that

$$\left\|J_T^* J_T f_q - S^*_{X_S} B_{M,N}\bar{y}\right\|_{\mathcal{H}_K} \leq c\left(\log\frac{1}{\delta}\right)\left(m^{-\frac{1}{2}} + n^{-\frac{1}{2}} + \phi(\theta_\phi^{-1}(M^{-\frac{1}{2}} + N^{-\frac{1}{2}}))\right) \tag{19}$$

The above bound will be used in the present section in resolving the issue with the choice of λ in (16). Note that this issue is more sophisticated all around, because, as Theorem 3 hints, a choice of λ in (16) is coupled with the value of the regularization parameter in (15), which is usually not given a priori. In this situation, the use of known regularization parameter choice rules seems to be problematic.

Note that according to that rules, one would select only one element, say $f^{\lambda_\mu}_{\mathbf{z},M,N}$, from a family of approximants $f^{\lambda}_{\mathbf{z},M,N}$ given by (16) for several values of the regularization parameter $\lambda = \lambda_1, \lambda_2, \ldots, \lambda_l$. Other elements $f^{\lambda_k}_{\mathbf{z},M,N}, k \neq \mu$, would be left aside, in spite of efforts spent for their construction.

In contrast, in the present section we discuss the use of a linear combination

$$f_{\mathbf{z}} = \sum_{k=1}^{l} c_k f^{\lambda_k}_{\mathbf{z},M,N} \tag{20}$$

of the approximants computed for all tried values of the regularization parameter λ.

It is clear that the best $L_{2,\rho_T}-$space approximation of the target regression function f_q by linear combinations $f_{\mathbf{z}}$ corresponds to the vector $\bar{c} = (c_1, c_2, \ldots, c_l)$ of ideal coefficients in (20) that solves the linear system $G\bar{c} = \bar{g}$ with the Gram matrix $G = \left(\left\langle f^{\lambda_k}_{\mathbf{z},M,N}, f^{\lambda_u}_{\mathbf{z},M,N}\right\rangle_{L_{2,\rho_T}}\right)^{l}_{k,u=1}$ and the right-hand side vector $\bar{g} = \left(\left\langle f_q, f^{\lambda_k}_{\mathbf{z},M,N}\right\rangle_{L_{2,\rho_T}}\right)^{l}_{k=1}$. But, of course, neither Gram matrix G nor the vector $\bar{g}$ is accessible, because there is no access to the target measure ρ_T.

To overcome this obstacle we first observe that the norms $\left\|f^{\lambda_k}_{\mathbf{z},M,N}\right\|_{\mathcal{H}_K}$ are under our control, such that we can put a threshold $\gamma_l > 0$ and consider only $\lambda_k, k = 1, 2, \ldots, l$, for which $\left\|f^{\lambda_k}_{\mathbf{z},M,N}\right\|_{\mathcal{H}_K} \leq \gamma_l$.

Then the following lemma is helpful.

Lemma 5. *Assume that conditions of Theorems 1 and 2 hold. Then for $\lambda_1, \lambda_2, \ldots, \lambda_l$ such that $\left\|f^{\lambda_k}_{\mathbf{z},M,N}\right\|_{\mathcal{H}_K} \leq \gamma_l,\ k = 1, 2, \ldots, l$, with probability $1 - \delta$ we have*

$$\begin{aligned}
\left\langle f_q, f^{\lambda_k}_{\mathbf{z},M,N}\right\rangle_{L_{2,\rho_T}} =& \frac{1}{n}\sum_{i=1}^{n} \bar{\beta}_{M,N}(x_i) y_i f^{\lambda_k}_{\mathbf{z},M,N}(x_i) \\
&+ O\left(\left(\log\frac{1}{\delta}\right)\left(m^{-\frac{1}{2}} + n^{-\frac{1}{2}} + \phi(\theta_\phi^{-1}(M^{-\frac{1}{2}} + N^{-\frac{1}{2}}))\right)\right) \\
\left\langle f^{\lambda_k}_{\mathbf{z},M,N}, f^{\lambda_u}_{\mathbf{z},M,N}\right\rangle_{L_{2,\rho_T}} =& \frac{1}{m}\sum_{j=1}^{m} f^{\lambda_k}_{\mathbf{z},M,N}(x'_j) f^{\lambda_u}_{\mathbf{z},M,N}(x'_j) + O\left((\log\frac{1}{\delta}) m^{-\frac{1}{2}}\right),
\end{aligned}$$

where the coefficients implicit in $O-$symbols may depend on a chosen threshold γ_l, but do not depend on m, n, M, N.

Proof. We prove only the first statement since the proof of the second follows analogously.

Keeping in mind that $f_q, f^{\lambda_k}_{\mathbf{z},M,N} \in \mathcal{H}_K$ we have

$$\begin{aligned}\left\langle f_q, f^{\lambda_k}_{\mathbf{z},M,N}\right\rangle_{L_{2,\rho_T}} &= \left\langle J_T f_q, J_T f^{\lambda_k}_{\mathbf{z},M,N}\right\rangle_{L_{2,\rho_T}} = \left\langle J_T^* J_T f_q, f^{\lambda_k}_{\mathbf{z},M,N}\right\rangle_{\mathcal{H}_K} \\ &= \left\langle S^*_{X_S} B_{M,N}\bar{y}, f^{\lambda_k}_{\mathbf{z},M,N}\right\rangle_{\mathcal{H}_K} + \left\langle J_T^* J_T f_q - S^*_{X_S} B_{M,N}\bar{y}, f^{\lambda_k}_{\mathbf{z},M,N}\right\rangle_{\mathcal{H}_K} \\ &= \left\langle B_{M,N}\bar{y}, S_{X_S} f^{\lambda_k}_{\mathbf{z},M,N}\right\rangle_{\mathbb{R}^n} + \left\langle J_T^* J_T f_q - S^*_{X_S} B_{M,N}\bar{y}, f^{\lambda_k}_{\mathbf{z},M,N}\right\rangle_{\mathcal{H}_K} \\ &= \frac{1}{n}\sum_{i=1}^{n} \tilde{\beta}_{M,N}(x_i) y_i f^{\lambda_k}_{\mathbf{z},M,N}(x_i) + \left\langle J_T^* J_T f_q - S^*_{X_S} B_{M,N}\bar{y}, f^{\lambda_k}_{\mathbf{z},M,N}\right\rangle_{\mathcal{H}_K}\end{aligned} \tag{21}$$

Moreover, from (19) with probability $1-\delta$ we have that

$$\begin{aligned}&\left|\langle J_T^* J_T f_q - S^*_{X_S} B_{M,N}\bar{y}, f^{\lambda_k}_{\mathbf{z},M,N}\rangle_{\mathcal{H}_K}\right| \\ &\qquad \le c\left\|f^{\lambda_k}_{\mathbf{z},M,N}\right\|_{\mathcal{H}_K}\left(\log\frac{1}{\delta}\right)\left(m^{-\frac{1}{2}} + n^{-\frac{1}{2}} + \phi(\theta_\phi^{-1}(M^{-\frac{1}{2}} + N^{-\frac{1}{2}}))\right)\end{aligned} \tag{22}$$

Now, the required statement follows from (21) and (22). □

Remark 3. *If in line with Remark 2, M and N are taken so large that one can assume the inequality $m^{-\frac{1}{2}} + n^{-\frac{1}{2}} \ge \phi(\theta_\phi^{-1}(M^{-\frac{1}{2}} + N^{-\frac{1}{2}}))$, then it is natural to consider only $\lambda_k, k = 1, 2, \ldots, l$, bounded from below by $m^{-\frac{1}{2}} + n^{-\frac{1}{2}}$, because one expects λ_k to be of order $\theta^{-1}\left(m^{-\frac{1}{2}} + n^{-\frac{1}{2}} + \phi(\theta_\phi^{-1}(M^{-\frac{1}{2}} + N^{-\frac{1}{2}}))\right) > m^{-\frac{1}{2}} + n^{-\frac{1}{2}}$. Using (5) and (18) we can conclude that for such λ_k with probability $1-\delta$ it holds*

$$\begin{aligned}\left\|f^{\lambda_k}_{\mathbf{z},M,N}\right\|_{\mathcal{H}_K} &\le \left\|g_{\lambda_k}(S^*_{X_S} B_{M,N} S_{X_S}) S^*_{X_S} B_{M,N} S_{X_S} f_q\right\|_{\mathcal{H}_K} \\ &\quad + \left\|g_{\lambda_k}(S^*_{X_S} B_{M,N} S_{X_S})(S^*_{X_S} B_{M,N}\bar{y} - S^*_{X_S} B_{M,N} S_{X_S} f_q)\right\|_{\mathcal{H}_K} \\ &\le \|f_q\|_{\mathcal{H}_K} \sup_t |g_{\lambda_k}(t)t| + \left\|S^*_{X_S} B_{M,N}\bar{y} - S^*_{X_S} B_{M,N} S_{X_S} f_q\right\|_{\mathcal{H}_K} \sup_t |g_{\lambda_k}(t)| \\ &\le (\gamma_0 + 1)\|f_q\|_{\mathcal{H}_K} + c\frac{\gamma_{-1}}{\lambda_k}\left(\log\frac{1}{\delta}\right)\left(m^{-\frac{1}{2}} + n^{-\frac{1}{2}} + \phi(\theta_\phi^{-1}(M^{-\frac{1}{2}} + N^{-\frac{1}{2}}))\right) \\ &\le c\log\frac{1}{\delta}.\end{aligned}$$

This means that for big enough M and N, the threshold value γ_l in Lemma 5 can be taken as $\gamma_l = O(\log\frac{1}{\delta})$.

At this point we would like to stress that in practice the number l of the elements in the set $\{f^{\lambda_k}_{\mathbf{z},M,N}\}_{k=1}^{l}$ can be assumed to be negligible compared to the

cardinalities m, n, M, N of the available data samples (usually not more than 10 - 15 approximants are computed for different values of the regularization parameters). Therefore, l-dependent coefficients do not affect the orders $O(m^{-\frac{1}{2}} + n^{-\frac{1}{2}} + \phi(\theta_\phi^{-1}(M^{-\frac{1}{2}} + N^{-\frac{1}{2}})))$ or $O(m^{-\frac{1}{2}})$. Then Lemma 5 suggests to approximate the inaccessible Gram matrix G and the vector $\bar{g}$ by respectively

$$\tilde{G} = \left(\frac{1}{m} \sum_{j=1}^{m} f_{\mathbf{z},M,N}^{\lambda_k}(x_j') f_{\mathbf{z},M,N}^{\lambda_u}(x_j') \right)_{k,u=1}^{l}, \ \tilde{g} = \left(\frac{1}{n} \sum_{i=1}^{n} \tilde{\beta}_{M,N}(x_i) y_i f_{\mathbf{z},M,N}^{\lambda_k}(x_i) \right)_{k=1}^{l},$$

which can be effectively computed from data samples. Moreover, Lemma 5 tells us that with probability $1 - \delta$ it holds

$$\begin{aligned} \|\bar{g} - \tilde{g}\|_{\mathbb{R}^l} &= O\left((\log \frac{1}{\delta}) \left(m^{-\frac{1}{2}} + n^{-\frac{1}{2}} + \phi(\theta_\phi^{-1}(M^{-\frac{1}{2}} + N^{-\frac{1}{2}})) \right) \right), \\ \left\| G - \tilde{G} \right\|_{\mathbb{R}^l \to \mathbb{R}^l} &= O\left((\log \frac{1}{\delta}) m^{-\frac{1}{2}} \right) \end{aligned} \tag{23}$$

With the matrix $\tilde{G}$ at hand one can easily check whether or not it is well-conditioned and $\tilde{G}^{-1}$ exists. If it is not the case, then some of approximants $f_{\mathbf{z},M,N}^{\lambda_k}$ are (almost) linearly dependent on others (a reason for that can be that some of λ_k are too close to each other). Such approximants cannot (essentially) influence the quality of the best approximation by linear combinations (20); they can be detected (by using, for example, the condition number of $\tilde{G}$ as a detection tool) and withdrawn from the consideration.

Thus, we assume that $\tilde{G}^{-1}$ exists. In view of Lemma 5 it is then natural to assume that m is so large that with probability $1 - \delta$ we have

$$\left\| G - \tilde{G} \right\|_{\mathbb{R}^l \to \mathbb{R}^l} < \frac{1}{\left\| \tilde{G}^{-1} \right\|_{\mathbb{R}^l \to \mathbb{R}^l}}. \tag{24}$$

This in its turn allows the application of the well known Banach theorem on inverse operators, which tells that

$$\left\| G^{-1} \right\|_{\mathbb{R}^l \to \mathbb{R}^l} \leq \frac{\left\| \tilde{G}^{-1} \right\|_{\mathbb{R}^l \to \mathbb{R}^l}}{1 - \left\| \tilde{G}^{-1} \right\|_{\mathbb{R}^l \to \mathbb{R}^l} \left\| G - \tilde{G} \right\|_{\mathbb{R}^l \to \mathbb{R}^l}} = O(1). \tag{25}$$

Theorem 4. *Consider* $\tilde{f}_{\mathbf{z}} = \sum_{k=1}^{l} \tilde{c}_k f_{\mathbf{z},M,N}^{\lambda_k}$, *where* $\tilde{c} = (\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_l) = \tilde{G}^{-1}\tilde{g}$, *and assume the conditions of Lemma 5 and (24). Then with probability* $1 - \delta$ *it holds*

$$\begin{aligned} \left\| f_q - \tilde{f}_{\mathbf{z}} \right\|_{L_{2,\rho_T}} \leq & \min_{c_k} \left\| f_q - \sum_{k=1}^{l} c_k f_{\mathbf{z},M,N}^{\lambda_k} \right\|_{L_{2,\rho_T}} \\ & + O\left((\log \frac{1}{\delta}) \left(m^{-\frac{1}{2}} + n^{-\frac{1}{2}} + \phi(\theta_\phi^{-1}(M^{-\frac{1}{2}} + N^{-\frac{1}{2}})) \right) \right), \end{aligned} \tag{26}$$

where a coefficient implicit in $O-$symbol may depend on l, but does not depend on m, n, M, N.

Proof. Note that the minimum in the right hand side of (26) is attained on $\bar{c} = (c_1, c_2, \ldots, c_l) = G^{-1}\bar{g}$. Let $\bar{f}_{\mathbf{z}}$ be a linear combination (20) defined by the vector of coefficients $\bar{c}$. Then from (23)-(25) with probability $1-\delta$ we have

$$\begin{aligned}\|\bar{c}-\tilde{c}\|_{\mathbb{R}^l} &\le \left\|\tilde{G}^{-1}\right\|_{\mathbb{R}^l\to\mathbb{R}^l}\left(\|\bar{g}-\tilde{g}\|_{\mathbb{R}^l}+\left\|G-\tilde{G}\right\|_{\mathbb{R}^l\to\mathbb{R}^l}\|\bar{c}\|_{\mathbb{R}^l}\right)\\ &= O\left((\log\frac{1}{\delta})\left(m^{-\frac{1}{2}}+n^{-\frac{1}{2}}+\phi(\theta_\phi^{-1}(M^{-\frac{1}{2}}+N^{-\frac{1}{2}}))\right)\right) \qquad (27)\end{aligned}$$

Moreover,

$$\begin{aligned}\left\|f_q-\tilde{f}_{\mathbf{z}}\right\|_{L_{2,\rho_T}} &\le \left\|f_q-\bar{f}_{\mathbf{z}}\right\|_{L_{2,\rho_T}}+\left\|\bar{f}_{\mathbf{z}}-\tilde{f}_{\mathbf{z}}\right\|_{L_{2,\rho_T}}\\ &\le \left\|f_q-\bar{f}_{\mathbf{z}}\right\|_{L_{2,\rho_T}}+\sqrt{l}\,\|\bar{c}-\tilde{c}\|_{\mathbb{R}^l}\max_k\left\|f^{\lambda_k}_{\mathbf{z},M,N}\right\|_{L_{2,\rho_T}}\\ &\le \left\|f_q-\bar{f}_{\mathbf{z}}\right\|_{L_{2,\rho_T}}+\sqrt{l}\gamma_l\,\|\bar{c}-\tilde{c}\|_{\mathbb{R}^l}\,, \qquad (28)\end{aligned}$$

and the statement of the theorem follows now from (27), (28). □

Remark 4. *Assume that the sequence $\lambda_1, \lambda_2, \ldots, \lambda_l$ of the tried values of the regularization parameter λ is so tight, and one of the values, say $\lambda = \lambda_\mu$, is so close to the value λ_δ indicated in Theorem 3, that the corresponding approximant $f^{\lambda_\mu}_{\mathbf{z},M,N}$ provides an accuracy of the order guaranteed by that theorem. Then under conditions of Theorem 4 the aggregate approximation $\tilde{f}_{\mathbf{z}}$ also guarantees an accuracy of the same order, but does not require any knowledge of the index functions φ, ϕ describing the smoothness of f_q and $\frac{d\rho_T}{d\rho_S}$. This follows from the fact that the second term of the right-hand side of (26) is negligible compared to the error bounds given by Theorem 3, and from the obvious inequality*

$$\min_{c_k}\left\|f_q-\sum_{k=1}^{l}c_k f^{\lambda_k}_{\mathbf{z},M,N}\right\|_{L_{2,\rho_T}} \le \left\|f_q-f^{\lambda_\mu}_{\mathbf{z},M,N}\right\|_{L_{2,\rho_T}}.$$

5 Numerical illustrations

5.1 Academic examples

Our first two illustrations are on toy data, and they are intended mainly to demonstrate potential advantages of the use of the general regularization scheme (7), (16) over a widely used importance weighted regularized least squares (IWRLS) corresponding to Tikhonov regularization and indexed by $g_\lambda(t) = (\lambda + t)^{-1}$.

In our first example, we simulate inputs $X_T = (x'_1, x'_2, \ldots, x'_m)$ in the target domain to be sampled from the continuous uniform distribution $\rho_T \sim U(0,1)$ over

$[0,1]$, while inputs $X_S = (x_1, x_2, \ldots, x_n)$ in the source domain are sampled from the beta distribution $\rho_S \sim B(\frac{1}{2}, 1)$. in this case, the Radon-Nikodym derivative $\beta = \frac{d\rho_T}{d\rho_S}$ is known to be $\beta = 2\sqrt{x}$.

The outputs are simulated as values of the function $f^*(x) = 1 + e^{-\frac{x^2}{2}}$ observed in Gaussian white noise, such that $y_i = f^*(x_i) + \epsilon_i$, $i = 1, 2, \ldots, n$, where ϵ_i are zero-mean Gaussian random variables with standard deviation δ.

In algorithms described in Section 3, we choose the kernel as

$$K(x, x') = 1 + \sqrt{xx'} + e^{-\frac{(x-x')^2}{2}},$$

such that, in the considered case, the corresponding space $\mathcal{H}_K$ contains both functions $\beta(x)$ and $f^*(x)$.

In our numerical experiments we employ a particular case of the general regularization scheme (16), where

$$g_\lambda(t) = g_{\lambda,k}(t) = \left(1 - \frac{\lambda^k}{(\lambda + t)^k}\right) t^{-1}, \tag{29}$$

and k is a natural number. As we already mentioned in Section 2.2, IWRLS-approach corresponds to (16), (29) with $k = 1$, while for $k = 2, 3, \ldots$, the approximation (16), (29) can be computed by applying IWRLS-approach iteratively k times, that corresponds to the so-called iterated Tikhonov regularization.

More precisely, for $g_\lambda(t) = g_{\lambda,k}(t)$ the approximant $f^\lambda_{\mathbf{z},M,N} = f^{\lambda,k}_{\mathbf{z},M,N}$ given by (16), (29) is the $k-$th term of the sequence

$$f^{\lambda,l}_{\mathbf{z},M,N}(x) = \sum_{i=1}^{n} c_i^l K(x, x_i), \quad l = 1, 2, \ldots, k, \tag{30}$$

where the coefficient vector $\bar{c}_l = (c_1^l, c_2^l, \ldots, c_n^l)$ can be calculated as

$$\bar{c}_l = (n\lambda \mathbf{I} + B_{M,N,\alpha}\mathbf{K})^{-1} \left(n\lambda \bar{c}_{l-1} + B_{M,N,\alpha}\bar{y}\right), \tag{31}$$

$$l = 1, 2, \ldots, k, \quad \bar{c}_0 = (0, 0, \ldots, 0);$$

here $\mathbf{I}$ is n by n identity matrix, $\mathbf{K} = (K(x_i, x_j))_{i,j=1}^n$,

$$B_{M,N,\alpha} = diag(\beta^\alpha_{M,N}(x_1), \beta^\alpha_{M,N}(x_2), \ldots, \beta^\alpha_{M,N}(x_n)),$$

and $\beta^\alpha_{M,N}(x)$ is defined by (15) with $g_\lambda(t) = (\lambda + t)^{-1}, \lambda = \alpha$.

Note that, in principle, the general regularization scheme can also be employed for computing the values of the approximate Radon-Nikodym derivative $\beta^\alpha_{M,N}(x_i)$, but for simplicity here we restrict ourselves to the use of KuLSIF-approach proposed in [29]. Moreover, we put $M = m$, $N = n$, such that $B_{M,N,\alpha} = diag(\beta^\alpha_1, \beta^\alpha_2, \ldots, \beta^\alpha_n)$, where $\beta^\alpha_i = \beta^\alpha_{m,n}(x_i)$ and the vector of diagonal elements $\bar{\beta}_\alpha = (\beta^\alpha_1, \beta^\alpha_2, \ldots, \beta^\alpha_n)$ can be calculated as

$$\bar{\beta}_\alpha = (n\alpha \mathbf{I} + \mathbf{K})^{-1} \bar{F}, \tag{32}$$

where $\bar{F} = (F_i)_{i=1}^n, F_i = \frac{n}{m}\sum_{j=1}^m K(x_i, x'_j)$.

The noisy inputs $y_i = f^*(x_i) + \epsilon_i, \epsilon_i \sim N(0, \delta^2)$, have been simulated with $\delta = 0.3$. Then the algorithm (30) - (32) has been implemented with $M = m = N = n = 10$, $\lambda = 0.1$, $\alpha = 0.5$, and $k = \{1, 2, 5, 10\}$. The performance of each implementation has been measure in terms of the root-mean-square deviation (RMSD).

$$RMSD = \left(m^{-1} \sum_{j=1}^{m} \left(f^*(x'_j) - f^{\lambda,k}_{\mathbf{z},M,N}(x'_j) \right)^2 \right)^{\frac{1}{2}}$$

A summary of the performance over 20 simulations of $(x_i)_{i=1}^n$, $(x'_j)_{j=1}^m$, $(y_i)_{i=1}^n$ is presented in the form of notched box plots in Figure 1. It is clear that in our first example the considered realization of the general regularization scheme outperforms the usual IWRLS ($k = 1$).

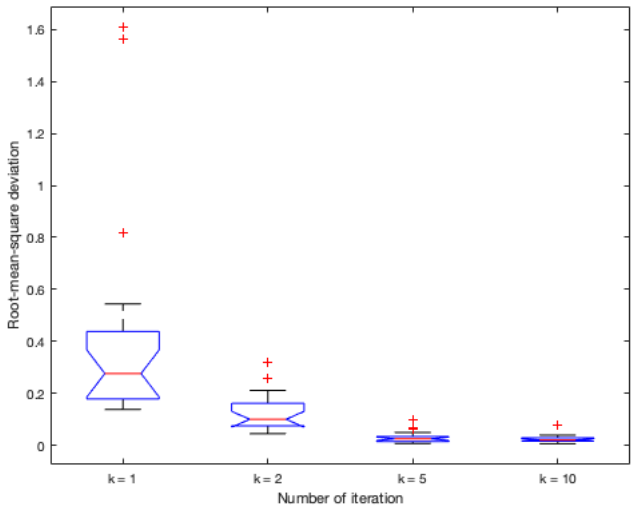

Figure 1: Root-mean-square deviations in the first example.

The performance of the algorithm (30) - (32) for a particular simulation is displayed in Figure 2. In this figure, the stars denote the exact values $f^*(x'_j)$, and the diamonds denote the values $f^{\lambda,5}_{\mathbf{z},M,N}(x'_j)$. Moreover, the triangles mark the values $f^{\lambda,5}_{\mathbf{z},M,N}(x'_j)$ resulting from the algorithm (30) - (32), where the matrix $B_{M,N,\alpha}$ is substituted by the matrix $B = diag(\beta(x_1), \beta(x_2), \ldots, \beta(x_n))$ containing the exact values of the Radon-Nikodym derivative, which in the considered case is $\beta(x) = 2\sqrt{x}$. As it can be seen from Figure 2, the performance is not essentially changed when the approximate matrix $B_{M,N,\alpha}$ is substituted by the ideal one. This

means, that in the considered case, KuLSIF-approach provides a reliable estimation of the Radon-Nikodym derivative.

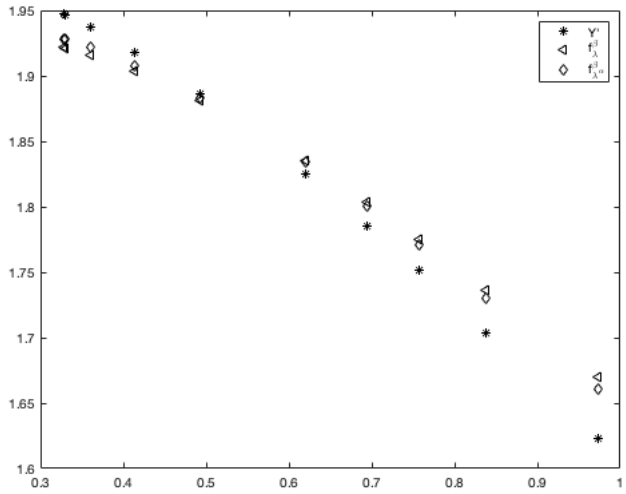

Figure 2: The performance of the algorithm in the first example for a particular simulation.

Recall that following [2], in our theoretical analysis we rely on the assumption that $\beta(x) = \frac{d\rho_T(x)}{d\rho_S(x)}$ is uniformly bounded for any $x \in \mathbf{X}$. Our second example shows how the algorithm (30) - (32) may perform in the situation where this assumption is violated. We keep the same setup as in the first example, but swap the source and the target distributions, such that now $\rho_T \sim B(\frac{1}{2}, 1)$, $\rho_S \sim U(0, 1)$, and $\beta(x) = \frac{1}{2\sqrt{x}}$ is an unbounded function on $\mathbf{X} = [0, 1]$. A summary of the performance over 20 simulations is presented in Figure 3. The figure shows that in the second example, the performance of the usual IWRLS ($k = 1$) becomes essentially worse as compared to the first example (Figure 1), while the considered realization of the general regularization scheme still performs reliably. This hints that the area of applications of algorithm (30) - (32) is wider than the one for the usual IWRLS.

5.2 Detection of vertebral artery stenosis based on diagnoses of carotid artery stenosis

In this section, we demonstrate an application of the aggregation approach presented in Section 4 to the problem of automatic stenosis detection from lumen diameters.

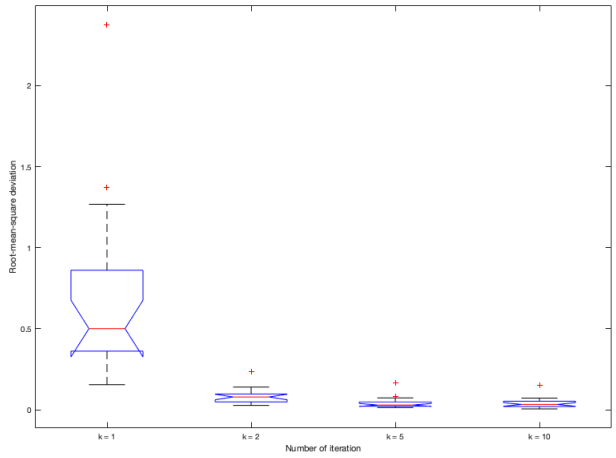

Figure 3: Root-mean-square deviations in the second example.

Stenosis is an abnormal narrowing in a blood vessel caused by lesion that reduces the space of lumen. In cervical arteries, such as internal carotid arteries (ICA) and vertebral arteries (VA), a stenosis can reduce, or even block, blood flow to the brain and significantly increases the risk of developing a stroke. Therefore, automatic stenosis detection is an important neuroradiological problem.

Such a detection problem can appear at the final or quantification stage of computerized tomography (CT) or magnetic resonance imaging (MRI) angiography, when vessel lumen segmentation and centerline extraction have already been performed. Then the above detection culminates the efforts spent in those previous stages, and for this reason it deserves a special consideration.

After the segmentation of CT/MRI scans, the existing software allows for estimating the diameters $d_s, s = 1, 2, \ldots$, of the vessel cross-sections at about 500 positions t_s along the vessel centerlines. Since the positions t_s and their total numbers may vary from patient to patient, it is reasonable to structure the above data in terms of functions $x(t)$, e.g. cubic interpolation splines with knots at t_s, describing the change in the vessel diameter and taking the values $x(t_s) = d_s, s = 1, 2, \ldots$. In this way, clinical data can be presented as a training sample $\mathbf{z}$ of functional inputs $x_i = x_i(t), i = 1, 2, \ldots, n$, labeled by outputs y_i taking the value $y_i = 0$ for the diagnosis no stenosis, and the value $y_i = 0.25, 0.5, 0.75, 1.0$, if the diagnosis is, respectively, light, medium, moderate or high stenosis. Then such data can be used to construct a predictor that automatically detects presence or absence of stenosis by assigning a label $y = 1$ or $y = 0$ to a new input function $x = x(t)$ extracted from CT/MRI scan.

At this point we would like to note that the prediction from functional inputs

$x = x(t)$ is not directly covered by our theoretical analysis. However, the algorithm (30) - (32) can still be used for such inputs if we employ there a kernel $K(x, x')$, which is natural extension to functions of a positive definite radial basis kernel. Examples of such kernels are given, for instance, in Table 1 of [19], and, in our further calculations, we have used inverse multiquadratic kernel

$$K(x, x') = 1 + \left(1 + \frac{\|x(\cdot) - x'(\cdot)\|^2}{2\epsilon}\right)^{-\frac{1}{2}}, \tag{33}$$

where

$$\|x(\cdot) - x'(\cdot)\|^2 = \int_0^b (x(t) - x'(t))^2 \, dt. \tag{34}$$

Recall that in the present context, $x(t)$, $x'(t)$ are functions of a position t along vessel centerlines. Since the lengths of such centerlines are patient-dependent, in (34) we need to consider $x(t)$, $x'(t)$ restricted to some interval $(0, b)$ of minimum length observed in the available clinical data, which in the considered case is $b = 140(mm)$.

We have permission for research-driven secondary use of anonymized clinical data collected at the Department of Neuroradiology and Department of Neurology, Medical University of Innsbruck, within the ReSect-study [34]. For the illustration below, the data of $n = 38$ ICA and $m = 35$ VA have been selected.

Source inputs $x_i = x_i(t), i = 1, 2, \ldots, n$, reflect the changes in diameters in selected ICA, of which 8 are affected by stenosis. These inputs are labeled by $y_i \in \{0, 0.25, 0.5, 0.75, 1.0\}$ depending on severity of stenosis.

Target inputs $x'_j = x'_j(t), j = 1, 2, \ldots, m$, reflect the changes in diameters in selected VA; 4 of these arteries are affected by stenosis, but all target inputs are used as unlabeled ones.

Then the training data $\mathbf{z} = \{(x_i, y_i)\}$ and unlabeled inputs $\{x'_j\}$ have been processed by the algorithm (30) - (34) to construct the potential VA stenosis classifiers $f^{\lambda,l}_{\mathbf{z},m,n}(x); l = 1$, i.e. this time we use a combination of IWRLS- and KuLSIF-approaches. The classifiers $f^{\lambda,1}_{\mathbf{z},m,n}$ are constructed for $\lambda = \lambda_k = 10^{-k}, k = 1, 2, 3, 4$. Concerning the value of the regularization parameter α in (32), we keep it the same as in the previous section, i.e. $\alpha = 0.5$. Moreover, in (33) we choose $\epsilon = 100$.

Then we aggregate $f^{\lambda_k}_{\mathbf{z},m,n} = f^{\lambda_k,1}_{\mathbf{z},m,n}, k = 1, 2, \ldots, l, l = 4$, into the proposed classifier $\tilde{f}_{\mathbf{z}}$ in the way described in Theorem 4, where in the definition of $\tilde{g}$ we use $\tilde{\beta}_{M,N}(x_i) = \beta^{\alpha}_i, i = 1, 2, \ldots, n$, given by (32).

Following [35] we use three metrics to evaluate the performance of stenosis detecting algorithms: the sensitivity $SE = \frac{TP}{TP+FN}$, the specificity $SP = \frac{TN}{TN+FP}$, and the positive predictive value $PPV = \frac{TP}{TP+FP}$, where TP is the number of cases when at least one stenosis in the considered artery has been detected by both the reference standard and the algorithm, regardless of the severity (because even mild narrowing of cervical artery calls for preventive measures); TN counts the cases when no stenosis in the considered artery has been detected by the reference

standard and by the algorithm; FN, FP are respectively the numbers of cases, when the absence or presence of stenosis has been wrongly detected by the algorithm.

Note that both SE and PPV give different information, and if one of them excels more than the other, the so-called $F1$-score can be a better metric compared to SE and PPV, because it is defined as their evenly weighted harmonic mean, i.e. $F1 = \frac{2SE \times PPV}{SE+PPV}$.

Note also that when a continuous-valued predictor $f(x)$ is used as binary classifier, its diagnostic ability depends on the so-called discrimination threshold c, such that a particular artery corresponding to an input, say x'_j, is assumed to be affected by stenosis if $f(x'_j) > c$.

The potential or optimal diagnostic ability of a particular classifier $f(x)$ can be assessed in terms of the receiver/relative operating characteristic (ROC) curve that visualizes the diagnostic ability of $f(x)$ as its discrimination thresholds c are varied. Recall that ROC curve is created by plotting the value SE against the value $1-SP$ for various threshold settings.

The ROC analysis provides tools to select a possibly optimal discrimination threshold c. One of them is the so-called Youden's method suggesting such $c = c_{opt}$ for which the point $(1 - SP, SE)$ has the minimal distance to the point $(0, 1)$.

Results of ROC analysis can be summarized as a single metric by computing the area under the ROC curve abbreviated as AUC, which ranges from near 0.5 for randomly assigned diagnoses to 1.0 for perfect diagnosing (classification). Below we report the performance of the considered classifiers on the target inputs $x'_j = x'_j(t), j = 1, 2, \ldots, m$, in all the above metrics under the assumption that the classifiers are equipped with their optimal (Youden's) discrimination thresholds $c = c_{opt}$.

The classifiers $f^{\lambda_k,1}_{\mathbf{z},m,n}, k = 1, 2, 3, 4$, produced by the algorithm (30) - (34) and their aggregation $\tilde{f}_{\mathbf{z}}$ have been applied to unlabeled target inputs $x'_j = x'_j(t), j = 1, 2, \ldots, m$, reflecting the change in diameters in selected vertebral arteries (VA). Then the detected presence or absence of stenosis indicated by the classifiers has been compared with known diagnoses. The results are reported in the first rows of Table 1 and are in agreement with our theoretical analysis, which predicts that the aggregation can perform better than the aggregated classifiers.

Algorithm	AUC	SE	SP	PPV	F1
$f^{\lambda_1,1}_{\mathbf{z},m,n}$ with $\lambda_1 = 10^{-1}$	0.209	0	0.882	0	NaN
$f^{\lambda_2,1}_{\mathbf{z},m,n}$ with $\lambda_2 = 10^{-2}$	0.621	0	0.882	0	NaN
$f^{\lambda_3,1}_{\mathbf{z},m,n}$ with $\lambda_3 = 10^{-3}$	0.669	1	0.912	0.25	0.4
$f^{\lambda_4,1}_{\mathbf{z},m,n}$ with $\lambda_4 = 10^{-4}$	0.677	1	0.912	0.25	0.4
Aggregation	0.976	1	0.969	0.75	0.857
Shahzad et.al.	0.924	0.857	0.818	0.5	0.631

Table 1: Performance of the compared classifiers on the target inputs

We may also compare the performance of the aggregation with some other approaches to automatic stenosis detection. The survey [35] gives a profound overview of the algorithms detecting stenosis from the diameters of the vessel cross-sections, i.e. from the same inputs as the ones considered above. The algorithms discussed in [35] have been developed for coronary artery stenosis detection, but in principle, they can also be used for diagnosing stenoses of other types of arteries, such as VA. Among the method overviewed in [35], the algorithm [36] outperforms the others and can be considered here as state-of-the-art algorithm. As it can be seen from the last row of Table 1, on the considered VA-dataset, the algorithm [36] equipped with ROC-optimized threshold performs even better than it was reported in [35] (there the reported values were $SE = 0.55, PPV = 0.27$). At the same time, Table 1 demonstrates that the proposed aggregation approach still gives a superior stenosis prediction performance in comparison to the method [36].

Acknowledgements

The research reported in this paper has been funded by the Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK), the Federal Ministry for Digital and Economic Affairs (BMDW), and the Province of Upper Austria in the frame of the COMET–Competence Centers for Excellent Technologies Programme and the COMET Module S3AI managed by the Austrian Research Promotion Agency FFG.

Sergiy Pereverzyev Jr. gratefully acknowledges the support of the Austrian Science Fund (FWF): project P 29514-N32.

The data used in Section 5.2 was acquired through a study performed at the Medical University of Innsbruck called "ReSect-Study". The ReSect-study is funded by the OeNB Anniversary Fund (#15644).

References

[1] H. Shimodaira, Improving predictive inference under covariate shift by weighting the log-likelihood function, Journal of Statistical Planning and Inference 90(2) (2000) 227–244. doi:https://doi.org/10.1016/S0378-3758(00)00115-4.

[2] J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, Correcting sample selection bias by unlabeled data, Advances in neural information processing systems 19 (2006) 601–608. doi:https://doi.org/10.5555/2976456.2976532.

[3] T. Poggio, S. Smale, The Mathematics of Learning: Dealing with Data *, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 1–19. doi:https://doi.org/10.1007/11362197_1.

[4] M. Sugiyama, K.-R. Müller, Input-dependent estimation of generalization error under covariate shift 23 (2005) 249–279. doi:https://doi.org/10.1524/stnd.2005.23.4.249.

[5] T. Kanamori, S. Hido, M. Sugiyama, A least-squares approach to direct importance estimation, Journal of Machine Learning Research 10 (2009) 1391–1445.

[6] F. Bauer, S. Pereverzev, L. Rosasco, On regularization algorithms in learning theory, Journal of Complexity 23 (2007) 52–72. doi:https://doi.org/10.1016/j.jco.2006.07.001.

[7] L. L. Gerfo, L. Rosasco, F. Odone, E. D. Vito, A. Verri, Spectral algorithms for supervised learning, Neural Computation 20(7) (2008) 1873–1879. doi:https://doi.org/10.1162/neco.2008.05-07-517.

[8] S. Lu, P. Mathe, S. Pereverzyev, Balancing principle in supervised learning for a general regularization scheme, Applied and Computational Harmonic Analysis 48 (2020) 123–148. doi:https://doi.org/10.1016/j.acha.2018.03.001.

[9] I. Schuster, M. Mollenhauer, S. Klus, K. Muandet, Kernel conditional density operators, Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics 108 (2020) 993–1004.

[10] G. Wilson, D. J. Cook, A survey of unsupervised deep domain adaptation, ACM Transactions on Intelligent Systems and Technology 11 (2020) 5–51. doi:https://doi.org/10.1145/3400066.

[11] G. Kriukova, O. Panasiuk, S. V. Pereverzyev, P. Tkachenko, A linear functional strategy for regularized ranking, Neural Networks 73 (2016) 26–35. doi:https://doi.org/10.1016/j.neunet.2015.08.012.

[12] S. V. Pereverzyev, P. Tkachenko, Regularization by the linear functional strategy with multiple kernels, Frontiers in Applied Mathematics and Statistics 3 (2017) 1. doi:https://doi.org/10.3389/fams.2017.00001.

[13] P. Tkachenko, G. Kriukova, M. Aleksandrova, O. Chertov, E. Renard, S. V. Pereverzyev, Prediction of nocturnal hypoglycemia by an aggregation of previously known prediction approaches: proof of concept for clinical application, Computer Methods and Programs in Biomedicine 134 (2016) 179–186. doi:https://doi.org/10.1016/j.cmpb.2016.07.003.

[14] S. Sampath, P. Tkachenko, E. Renard, S. V. Pereverzev, Glycemic control indices and theiraggregation in the prediction of nocturnal hypoglycemia from intermittent blood glucose measurements, Journal of Diabetes Science and Technology 10(6) (2016) 1245–1250. doi:https://doi.org/10.1177/1932296816670400.

[15] J. Chen, S. Pereverzyev Jr, Y. Xu, Aggregation of regularized solutions from multiple observation models, Inverse Problems 31(7) (2015) 075005. doi:https://doi.org/10.1088/0266-5611/31/7/075005.

[16] G. Montavon, M. L. Braun, K.-R. Müller, Kernel analysis of deep networks, Journal of Machine Learning Research 12 (2011) 2563–2581.

[17] C. A. Micchelli, Y. Xu, H. Zhang, Universal kernels, Journal of Machine Learning Research 7 (2006) 2651–2667. doi:https://doi.org/10.5555/1248547.1248642.

[18] A. Caponnetto, Y. Yao, Cross-validation based adaptation for regularization operators in learning, Analysis and Applications 8 (2010) 161–183. doi:https://doi.org/10.1142/S0219530510001564.

[19] Z. Szabó, B. K. Sriperumbudur, B. Póczos, A. Gretton, Learning theory of distribution regression, Journal of Machine Learning Research 17 (2016) 1–40. doi:https://doi.org/10.5555/2946645.3053434.

[20] G. Blanchard, N. Krämer, Convergence rates of kernel conjugate gradient for random design regression, Analysis and Applications 14 (2016) 763–794. doi:https://doi.org/10.1142/S0219530516400017.

[21] I. Pinelis, An approach to inequalities for the distributions of infinite-dimensional martingales, Probability in Banach Spaces 8 (1992) 128–134. doi:https://doi.org/10.1007/978-1-4612-0367-4_9.

[22] L. Rosasco, E. D. Vito, M. Belkin, On learning with integral operators, Journal of Machine Learning Research 11 (2010) 905–934.

[23] T. Evgeniou, M. Pontil, T. A. Poggio, Regularization networks and support vector machines, Advances in Computational Mathematics 13(1) (2000) 1–50. doi:https://doi.org/10.1023/A:1018946025316.

[24] S. Smale, D.-X. Zhou, Learning theory estimates via integral operators and their approximations, Constructive Approximation 26 (2007) 153–172. doi:https://doi.org/10.1007/s00365-006-0659-y.

[25] S. Lu, S. V. Pereverzyev, Regularization theory for ill-posed problems - selected topics, De Gruyter 58 (2013) . doi:https://doi.org/10.1515/9783110286496.

[26] P. Mathe, B. Hofmann, How general are general source conditions?, Inverse Problems 24 (2008) 015009. doi:https://doi.org/10.1088/0266-5611/24/1/015009.

[27] E. D. Vito, L. Rosasco, A. Caponnetto, U. D. Giovannini, F. Odone, Learning from examples as an inverse problem, Journal of Machine Learning Research 6 (2005) 883–904.

[28] E. D. Vito, L. Rosasco, A. Caponnetto, Discretization error analysis for Tikhonov regularization in learning theory, Analysis and Applications 4 (2006) 81–99. doi:https://doi.org/10.1142/S0219530506000711.

[29] T. Kanamori, T. Suzuki, M. Sugiyama, Statistical analysis of kernel-based least-squares density-ratio estimation, Machine Learning 86 (2012) 335–367. doi:https://doi.org/10.1007/s10994-011-5266-3.

[30] L. Oneto, S. Ridella, D. Anguita, Tikhonov, Ivanov and Morozov regularization for support vector machine learning, Machine Learning 103 (2016) 103–136. doi:https://doi.org/10.1007/s10994-015-5540-x.

[31] S. Page, S. Grünewälder, Ivanov-regularised least-squares estimators over large RKHSs and their interpolation spaces, The Journal of Machine Learning Research 20 (2019) 1–49.

[32] Q. Que, M. Belkin, Inverse density as an inverse problem: The Fredholm equation approach, Advances in Neural Information Processing Systems 26 (2013) .

[33] E. D. Vito, S. Pereverzyev, L. Rosasco, Adaptive kernel methods using the balancing principle, Foundations of Computational Mathematics 10 (2010) 455–479. doi:https://doi.org/10.1007/s10208-010-9064-2.

[34] L. Mayer, C. Boehme, T. Toell, B. Dejakum, J. Willeit, C. Schmidauer, K. Berek, C. Siedentopf, E. R. Gizewski, G. Ratzinger, S. Kiechl, M. Knoflach, Local signs and symptoms in spontaneous cervical artery dissection: A single centre cohort study, Journal Stroke 21(1) (2019) 112–115. doi:https://doi.org/10.5853/jos.2018.03055.

[35] H. Kirişli, M. Schaap, C. Metz, A. Dharampal, W. Meijboom, S. Papadopoulou, A. Dedic, K. Nieman, M. de Graaf, M. Meijs, M. Cramer, A. Broersen, S. Cetin, A. Eslami, L. Flórez-Valencia, B. M. K.L. Lor, I. Melki, B. Mohr, I. Öksüz, R. Shahzad, C. Wang, P. Kitslaar, G. Unal, A. Katouzian, M. Orkisz, C. Chen, F. Precioso, L. Najman, S. Masood, D. Ünay, L. van Vliet, R. Moreno, R. Goldenberg, E. Vuçini, G. Krestin, W. Niessen, T. van Walsum, Standardized evaluation framework for evaluating coronary artery stenosis detection, stenosis quantification and lumen segmentation algorithms in computed tomography angiography, Medical Image Analysis 17(8) (2013) 859–876. doi:https://doi.org/10.1016/j.media.2013.05.007.

[36] R. Shahzad, T. van Walsum, H. Kirisli, H. Tang, C. Metz, M. Schaap, L. van Vliet, W. Niessen, Automatic stenoses detection, quantification and lumen segmentation of the coronary arteries using a two point centerline extraction scheme, Proceedings of MICCAI Workshop 3D Cardiovascular Imaging: A MICCAI Segmentation Challengedoi:https://doi.org/10.13140/2.1.4409.2486.