# Adaptive Learning of Density Ratios in RKHS

**Werner Zellinger**                                        WERNER.ZELLINGER@OEAW.AC.AT
*Johann Radon Institute for Computational and Applied Mathematics*
*Austrian Academy of Sciences*
*Linz, Austria*


**Stefan Kindermann**                              KINDERMANN@INDMATH.UNI-LINZ.AC.AT
*Industrial Mathematics Institute*
*Johannes Kepler University Linz*
*Linz, Austria*


**Sergei V. Pereverzyev**                                  SERGEI.PEREVERZYEV@OEAW.AC.AT
*Johann Radon Institute for Computational and Applied Mathematics*
*Austrian Academy of Sciences*
*Linz, Austria*

**Editor:** Mahdi Soltanolkotabi

## Abstract

Estimating the ratio of two probability densities from finitely many observations of the densities is a central problem in machine learning and statistics with applications in two-sample testing, divergence estimation, generative modeling, covariate shift adaptation, conditional density estimation, and novelty detection. In this work, we analyze a large class of density ratio estimation methods that minimize a regularized Bregman divergence between the true density ratio and a model in a reproducing kernel Hilbert space (RKHS). We derive new finite-sample error bounds, and we propose a Lepskii type parameter choice principle that minimizes the bounds without knowledge of the regularity of the density ratio. In the special case of square loss, our method adaptively achieves a minimax optimal error rate. A numerical illustration is provided.

**Keywords:** density ratio estimation, reproducing kernel Hilbert space, learning theory, regularization, inverse problem

## 1. Introduction

### 1.1 Problem

Let $P$ and $Q$ be two probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$ with compact $\mathcal{X} \subset \mathbb{R}^d$ and $\sigma$-algebra $\mathcal{A}$ of its measurable subsets. Assume that $P$ is absolutely continuous with respect to $Q$. Then, the Radon-Nikodým theorem grants us a $Q$-integrable function

$\beta : \mathcal{X} \to [0, \infty)$ satisfying, for any $U \in \mathcal{A}$,

$$P(U) = \int_U \beta(x) \, \mathrm{d}Q(x). \tag{1}$$

If both measures, $P$ and $Q$, have densities with respect to another reference measure, e.g., the Lebesgue measure, then $\beta$ equals the *ratio of their densities*.

The problem of learning $\beta$ from samples $\mathbf{x} := (x_i)_{i=1}^m$ and $\mathbf{x}' := (x_i')_{i=1}^n$, drawn independently and identically (i.i.d.) from $P$ and $Q$, respectively, plays a central role in statistics and machine learning (Sugiyama et al., 2012a); its applications include anomaly detection (Smola et al., 2009; Hido et al., 2011), two-sample testing (Keziou and Leoni-Aubin, 2005; Kanamori et al., 2011), divergence estimation (Nguyen et al., 2007, 2010), covariate shift adaptation (Shimodaira, 2000; Gizewski et al., 2022; Dinu et al., 2023), generative modeling (Mohamed and Lakshminarayanan, 2016), conditional density estimation (Schuster et al., 2020), and classification from positive and unlabeled data (Kato et al., 2019). Note, that this problem is different from supervised learning (Cucker and Smale, 2001) as the given data do not include point evaluations $\beta(x)$ from the desired target $\beta$.

It has been first observed by Sugiyama, Suzuki, and Kanamori (2012b) that many algorithms for density ratio estimation compute finite-sample approximations of the objective $\beta$ by discretizing the minimization problem

$$\min_{f \in \mathcal{H}} B_F(\beta, g(f)) + \frac{\lambda}{2} \|f\|^2, \tag{2}$$

where $g(f)$ is a model of the Radon-Nikodým derivative $\beta$ with $f \in \mathcal{H}$ in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, $B_F(\beta, \widetilde{\beta}) := F(\beta) - F(\widetilde{\beta}) - \nabla F(\widetilde{\beta})[\beta - \widetilde{\beta}]$ is a Bregman divergence generated by some convex functional $F : L^1(Q) \to \mathbb{R}$, $\|.\|$ is a norm on $\mathcal{H}$, and $\lambda \in \mathbb{R}$ is a regularization parameter; see Example 1. For the exact form of the finite-sample approximations applied in practice, we ask the reader for patience until Eq. (11) is stated.

**Example 1**

- *As can be seen from (Sugiyama et al., 2012b, Section 3.2.1), the kernel unconstrained least-squares importance fitting procedure (KuLSIF) in (Kanamori et al., 2009) uses Eq. (2) with $F(h) = \int_{\mathcal{X}} (h(x) - 1)^2 / 2 \, \mathrm{d}Q(x)$ and $g = \mathrm{id}$, where $\mathrm{id}(f) = f$ is the identity (cf. (Kanamori et al., 2012, Eq. (4))). This leads to a discretization of the minimization problem*

$$\min_{f \in \mathcal{H}} \frac{1}{2} \|\beta - f\|_{L^2(Q)}^2 + \frac{\lambda}{2} \|f\|^2. \tag{3}$$

- *The logistic regression approach (LR) employed in (Bickel et al., 2009, Section 7) can be obtained from Eq. (2) using $F(h) = \int_{\mathcal{X}} h(x) \log(h(x)) - (1 + h(x)) \log(1 + h(x)) \, \mathrm{d}Q(x)$ and $g(f) = e^f$ with $f(x) = \theta^T \varphi(x)$ for some $\varphi : \mathcal{X} \to \mathbb{R}^d$; see (Sugiyama et al., 2012b, Section 3.2.3).*

2

- *The exponential function approach (Exp) introduced in (Menon and Ong, 2016, Section 7) can be obtained using $F(h) = \int_{\mathcal{X}} h(x)^{-3/2} \, dQ(x)$ and $g(f) = e^{2f}$; see (Menon and Ong, 2016, Table 1).*

- *The square loss (SQ) is used in (Menon and Ong, 2016, Table 1) with $F(h) = \int_X 1/(2h(x) + 2) \, dQ(x)$ and $g(f) = \frac{-1+2f}{2-2f}$.*

To the best of our knowledge, no error bounds are known for estimating Eq. (2) for the functionals in Example 1 that take into account the regularity of the density ratio $\beta$, except (Gizewski et al., 2022, Section 3.1) and (Nguyen et al., 2023) which hold only for KuLSIF. Moreover, except for KuLSIF, no method for choosing the regularization parameter $\lambda$ in Eq. (2) is known that adapts to the regularity of $\beta$, i.e., a method with error bounds provably decreasing for increasing regularity.

## 1.2 Results

Following Menon and Ong (2016), one can relate the minimization of the data fidelity term $B(f) := B_F(\beta, g(f))$ from Eq. (2), to the minimization of the expected loss with a loss function $\ell : \{-1, 1\} \times \mathbb{R} \to \mathbb{R}$ measuring by $\ell(-1, f(x))$ and $\ell(1, f(x))$, respectively, the error of a classifier $f(x)$ which predicts whether $x$ is drawn from $P$ or $Q$.

Our first result (Example 3) shows that all loss functions $\ell$ corresponding to Example 1, satisfy a property called *generalized self-concordance* (Marteau-Ferey et al., 2019). This allows us to extend results from supervised learning in RKHSs to density ratio estimation.

Our second result (Proposition 2) provides, in a well specified setting, for large enough sample sizes and with high probability, an error bound

$$B(f_{\mathbf{x},\mathbf{x}'}^{\lambda}) \leq O\left((m+n)^{-\frac{2r\alpha+\alpha}{2r\alpha+\alpha+1}}\right) \tag{4}$$

for a minimizer $f_{\mathbf{x},\mathbf{x}'}^{\lambda}$ of the empirical estimation Eq. (11) of Eq. (2) with appropriately chosen regularization parameter $\lambda$. The regularity index $r$ and the capacity index $\alpha$ in Eq. (4) originate from source conditions as used in inverse problems and learning theory. In the special case of square loss (SQ), the error rate in Eq. (4) is minimax optimal in the same sense as it is minimax optimal in supervised learning, see (Caponnetto and De Vito, 2007; Blanchard and Mücke, 2018).

Our third result (Eq. (30)) is a new method for choosing the regularization parameter $\lambda$ that does not need access to the regularity index $r$ but nevertheless achieves the same rate as in Eq. (4) (see Theorem 2). Our method is based on a Lepskii-type principle balancing bias and variance terms of a local quadratic approximation of the (generally not quadratic) data fidelity $B(f)$. Our parameter choice strategy achieves the error rate in Eq. (4) and is thus optimal in the case of square loss. A numerical example illustrates our theoretical findings.

## 1.3 Related Work

The representation of density ratio estimation methods as minimizers of empirical estimations of Eq. (2) has been first observed by Sugiyama et al. (2012b) and later characterized

 

in more detail by Menon and Ong (2016). Our work extends (Sugiyama et al., 2012b; Menon and Ong, 2016) by taking into account the effect of the regularization term $\frac{\lambda}{2}\left\|f\right\|^2$ in Eq. (2).

In general, to the best of our knowledge, there are no error bounds similar to Eq. (4) for the class of discretized minimization problems Eq. (2). For the special case of KuLSIF in Eq. (3), error rates have been discovered in (Kanamori et al., 2012; Gizewski et al., 2022; Nguyen et al., 2023). In particular, the error rate discovered in (Kanamori et al., 2012), for RKHS with *bracketing entropy* $\gamma \in (0,2)$, is

$$B(f_{\mathbf{z}}^{\lambda}) = O\left(\min\{m,n\}^{-\frac{2}{2+\gamma}}\right). \tag{5}$$

This rate is worse than the one in Eq. (4), whenever $\alpha(2r+1) > \frac{2}{\gamma}$. A similar result for KuLSIF and the special case of Gaussian kernel, is presented in Que and Belkin (2013), see Type I setting there. We can also mention (Gizewski et al., 2022, Remark 3) which provides the error bound

$$B(f_{\mathbf{z}}^{\lambda}) = O\left(\left(m^{-\frac{1}{2}} + n^{-\frac{1}{2}}\right)^{\frac{4r'+2}{2r'+2}}\right) \tag{6}$$

with $r'$ originating from a source condition similar to our Assumption 5 but for the covariance operator $T = \int_{\mathcal{X}} k(z,\cdot) \otimes k(z,\cdot)\, \mathrm{d}Q(z)$ of $Q$ instead of $\rho$, where $k$ denotes the kernel of $\mathcal{H}$ and $(u \otimes v)[f] = \langle u,f\rangle v$ the rank-1 operator. Our result in Eq. (4) improves Eq. (6) for $\alpha \in (1,\infty)$ and $0 < r' \le r \le \frac{1}{2}$. In the recent work (Nguyen et al., 2023), Eq. (6) is improved by additionally taking into account the capacity of the space $\mathcal{H}$ in terms of a new pointwise source condition on the reproducing kernel of $\mathcal{H}$ that allows to bound the regularized Christoffel function. Such condition can be used as alternative for the effective dimension as defined in our Assumption 6. As a result, an error bound similar to our Eq. (4) is obtained in (Nguyen et al., 2023). However, the error is measured in the $\mathcal{H}$-norm and the bound holds only for the special case of the KuLSIF.

To the best of our knowledge, the balancing principle introduced in this work is the first parameter choice method for density ratio estimation with discovered error rate. We can mention variants of cross-validation (Kanamori et al., 2012), the quasi-optimality criterion (Nguyen et al., 2023), and hints towards the use of balancing principles (Gizewski et al., 2022), but, even in the special case of KuLSIF, we are not aware of any finite-sample error bounds related to these approaches.

Some of our proof techniques rely on (Marteau-Ferey et al., 2019), where self-concordant loss functions in the context of kernel learning were studied first. Our analysis extends (Marteau-Ferey et al., 2019) to density ratio estimation and adaptive parameter choice.

## 2. Notation and Auxiliaries

### 2.1 Estimation of Bregman Divergence

Menon and Ong (2016) study properties of $F$ and $g$, which allow an intuitive estimation of Eq. (2) by discriminating i.i.d. observations $\mathbf{x} = (x_i)_{i=1}^m$ from $P$ and $\mathbf{x}' = (x_i')_{i=1}^n$ from $Q$. In the following, we will summarize some of these properties.

Let us start by introducing the underlying model: a probability measure $\rho$ on $\mathcal{X} \times \mathcal{Y}$, $\mathcal{Y} := \{-1, 1\}$, with conditional probability measures[1] defined by $\rho(x|y = 1) := P(x)$ and $\rho(x|y = -1) := Q(x)$ and a marginal (w.r.t. $y$) probability measure $\rho_{\mathcal{Y}}$ defined as Bernoulli measure with probability $\frac{1}{2}$ for both events $y = 1$ and $y = -1$. We make the following assumption on the data generation process.

**Assumption 1 (iid data)** *The data*

$$\mathbf{z} := (x_i, 1)_{i=1}^m \cup (x_i', -1)_{i=1}^n \in \mathcal{X} \times \mathcal{Y} \tag{7}$$

*is an i.i.d. sample of an $\mathcal{X} \times \mathcal{Y}$-valued random variable $Z$ with measure $\rho$.*

Assumption 1 allows us to introduce a binary classification problem for the accessible data $\mathbf{z}$. Recall the goal of the binary classification problem is to find, based on the data $\mathbf{z}$, a classifier $f : \mathcal{X} \to \mathbb{R}$ with a small *expected risk*

$$\mathcal{R}(f) := \mathbb{E}_{(x,y) \sim \rho}[\ell(y, f(x))] = \int_{\mathcal{X} \times \mathcal{Y}} \ell(y, f(x)) \, \mathrm{d}\rho(x, y) \tag{8}$$

for some loss function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$. In this work, we focus on *strictly proper composite* loss functions with twice differentiable Bayes risk, which are characterized by the following assumption.

**Assumption 2** *The function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ has an associated invertible link function $\Psi : [0, 1] \to \mathbb{R}$ such that*

1. *the Bayes classifier $f^* := \arg\min_{f:\mathcal{X} \to \mathbb{R}} \mathcal{R}(f)$ exists and $f^*(x) = \Psi \circ \rho(y = 1|x)$,*

2. *the Bayes risk $G(u) := u\ell(1, \Psi(u)) + (1 - u)\ell(-1, \Psi(u))$ is twice differentiable.*

The main result of Menon and Ong (2016) is the following representation of the data fidelity in Eq. (2), in terms of a loss function satisfying Assumption 2.

**Lemma 1 (Menon and Ong (2016), Proposition 3)** *Let the Bregman generator $F$ and the density ratio model $g(f)$ be defined by*

$$F(h) := -\int_{\mathcal{X}} (1 + h(x)) \cdot G\left(\frac{h(x)}{1 + h(x)}\right) \mathrm{d}Q(x), \quad g(f) := \frac{\Psi^{-1} \circ f}{1 - \Psi^{-1} \circ f}, \tag{9}$$

*for some link function $\Psi$ and the Bayes risk $G$ of a loss $\ell$ satisfying Assumption 2. Then*

$$B_F(\beta, g(f)) = 2\left(\mathcal{R}(f) - \mathcal{R}(f^*)\right). \tag{10}$$

---

1. Existence follows from $\mathcal{X} \times \mathcal{Y}$ being Polish (cf. (Dudley, 2002, Theorem 10.2.1), the product case).

Indeed, for all approaches in Example 1, we can find $F$ and $g$ as required by Lemma 1, see (Menon and Ong, 2016, Table 1). We refer to (Menon and Ong, 2016) for a recipe to construct new losses for Eq. (2) satisfying the requirements of Lemma 1.

**Example 2**

- *KuLSIF (Kanamori et al., 2009) uses $\ell(-1, v) = \frac{1}{2}v^2, \ell(1, v) = -v$ and $\Psi(v) = \frac{v}{1-v}$.*

- *LR (Bickel et al., 2009) uses $\ell(-1, v) = \log(1 + e^v), \ell(1, v) = \log(1 + e^{-v})$ and the link function $\Psi(v) = -\log(\frac{1}{v} - 1)$.*

- *Exp (Menon and Ong, 2016) uses $\ell(-1, v) = e^v, \ell(1, v) = e^{-v}$ and the link function $\Psi(v) = -\frac{1}{2}\log(\frac{1}{v} - 1)$.*

- *SQ (Menon and Ong, 2016) uses $\ell(-1, v) = (1 + v)^2, \ell(1, v) = (1 - v)^2$ and the link function $\Psi(v) = \frac{v+1}{2}$.*

According to Lemma 1, Eq. (2) can be expressed as regularized risk minimization objective, given $g$ and $F$ are constructed as specified in Lemma 1. This motivates the estimation of Eq. (2) by an empirical risk minimization

$$\min_{f \in \mathcal{H}} \frac{1}{m + n} \sum_{(x,y) \in \mathbf{z}} \ell(y, f(x)) + \frac{\lambda}{2} \|f\|^2. \tag{11}$$

Throughout this work, we will denote a minimizer of Eq. (11) by $f_{\mathbf{z}}^\lambda$. The goal of this work is to analyze the error $B(f_{\mathbf{z}}^\lambda) = B_F(\beta, g(f_{\mathbf{z}}^\lambda))$ of estimating $\beta$ by $g(f_{\mathbf{z}}^\lambda)$.

## 2.2 Learning in RKHS with Convex Losses

Lemma 1 identifies the methods in Example 1 as risk minimization problems with loss functions as detailed in Example 2. We will see in Section 3 that these loss functions satisfy the following property when applied to functions $f \in \mathcal{H}$ from an RKHS $\mathcal{H}$.

**Assumption 3 (generalized self-concordance (Marteau-Ferey et al., 2019))** *For any $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$, the function $\ell_z : \mathcal{H} \to \mathbb{R}, \ell_z(f) := \ell(y, f(x)), f \in \mathcal{H}$, is convex, three times differentiable, and it exists a set $\varphi(z) \subset \mathcal{H}$ such that*

$$\forall f \in \mathcal{H}, \forall h, k \in \mathcal{H} : |\nabla^3 \ell_z(f)[k, h, h]| \leq \nabla^2 \ell_z(f)[h, h] \cdot \sup_{p \in \varphi(z)} |k \cdot p|. \tag{12}$$

The definition of self-concordance originates from the analysis of Newton's method (Nesterov and Nemirovskii, 1994) and has been generalized in (Bach, 2010; Marteau-Ferey et al., 2019). Before we review error bounds, we follow Marteau-Ferey et al. (2019) and make the following technical assumptions.

**Assumption 4 (technical assumptions)** *For self-concordant $\ell_z(\cdot)$ with the set $\varphi(z) \subset \mathcal{H}$ in Eq. (12), it holds:*

1. *There exists $R \geq 0$ such that for any realization $z$ of the random variable $Z$, $\sup_{g \in \varphi(z)} \|g\| \leq R$ almost surely.*

2. *The following quantities are almost surely bounded for any realization $z$ of the random variable $Z$: $|\ell_z(0)|, \|\nabla\ell_z(0)\|, \text{Tr}(\nabla^2\ell_z(0))$.*

3. *There exists a minimizer $f_{\mathcal{H}} := \arg\min_{f\in\mathcal{H}} \mathcal{R}(f)$.*

In the sequel, the symbols $\varphi(Z)$ and $\ell_Z$ mean the varieties of sets $\varphi(z)$ and functions $\ell_z$ corresponding to the realizations $z$ of the random variable $Z$. The first point of Assumption 4 is standard in learning theory (Caponnetto and De Vito, 2007; Steinwart et al., 2009) and, together with point 2, assures that $\mathcal{R}(f)$ is well defined for all $f\in\mathcal{H}$. Point 3 imposes that our model is *well specified* in the sense that the function class $\mathcal{H}$ contains the minimizer of $\mathcal{R}$. Under these assumptions, Theorem 3 in (Marteau-Ferey et al., 2019) holds, which we state in our notation as follows:

**Lemma 2 (Marteau-Ferey et al. (2019), Theorem 3)** *Let Assumptions 1, 3, 4 be satisfied, $\delta\in(0,\frac{1}{2}]$, and define $B_1$ and $B_2$ as*

$$B_1 := \sup_{\|f\|\leq\|f_{\mathcal{H}}\|} \sup_{z\in\text{supp}(\rho)} \|\nabla\ell_z(f)\|, \quad B_2 := \sup_{\|f\|\leq\|f_{\mathcal{H}}\|} \sup_{z\in\text{supp}(\rho)} \text{Tr}(\nabla^2\ell_z(f)).$$

*Then, whenever $\lambda \leq B_2$ and $m+n$ is larger than*

$$\max\left\{512\,\|f_{\mathcal{H}}\|^2\,R^2\log\left(\frac{2}{\delta}\right), 512\log\left(\frac{2}{\delta}\right), 24\frac{B_2}{\lambda}\log\left(\frac{8B_2}{\lambda\delta}\right), 256\frac{R^2B_1^2}{\lambda^2}\log\left(\frac{2}{\delta}\right)\right\}, \quad (13)$$

*a minimizer $f_{\mathbf{z}}^{\lambda}$ of Eq. (11) satisfies, with probability at least $1-2\delta$,*

$$\mathcal{R}(f_{\mathbf{z}}^{\lambda}) - \mathcal{R}(f_{\mathcal{H}}) \leq \frac{84B_1^2}{\lambda(m+n)}\log\left(\frac{2}{\delta}\right) + 2\lambda\,\|f_{\mathcal{H}}\|^2. \quad (14)$$

Lemma 2 is a general result, which does not take into account cases where a regularity of $f_{\mathcal{H}}$ is assumed beyond $f_{\mathcal{H}}\in\mathcal{H}$ in Assumption 4. Such regularity can be encoded by so-called *source conditions* as used in the regularization theory for inverse problems (Engl et al., 1996; Bauer et al., 2007; Caponnetto and De Vito, 2007; Steinwart et al., 2009; Rudi and Rosasco, 2017; Blanchard and Mücke, 2018).

**Assumption 5 (source condition)** *There exist $r\in(0,\frac{1}{2}], v\in\mathcal{H}$ such that $f_{\mathcal{H}} = \mathbf{H}(f_{\mathcal{H}})^r v$ with the expected Hessian $\mathbf{H}(f) := \mathbb{E}[\nabla^2\ell_Z(f)]$.*

In the case of square loss, Assumption 5 recovers the polynomial source condition $f_{\mathcal{H}} = T^r v$ with covariance operator $T = \int_{\mathcal{X}} k(z,\cdot)\otimes k(z,\cdot)\,\mathrm{d}\rho(z)$, for the kernel $k$ of $\mathcal{H}$ and the rank-1 operator $(u\otimes v)[f] = \langle u, f\rangle v$. Starting from $r=0$ (assuming only $f_{\mathcal{H}}\in\mathcal{H}$), the regularity of $f_{\mathcal{H}}$ increases with increasing $r$, essentially reducing the number of eigenvectors of $T$ needed to well approximate $f_{\mathcal{H}}$.

It is well known in learning theory (Zhang, 2002; Caponnetto et al., 2005; Caponnetto and De Vito, 2007) that the rate of convergence of regularized learning algorithms is not only influenced by the regularity of $f_{\mathcal{H}}$, but also by the *capacity* of the space $\mathcal{H}$ measured by the so-called *effective dimension*. We therefore make the following assumption of (Marteau-Ferey et al., 2019, Assumption 7), which generalizes (Caponnetto and De Vito, 2007, Definition 1).

**Assumption 6 (capacity condition)** *There exist $\alpha \geq 1, Q_0 \geq 0$ such that $\mathrm{df}_\lambda \leq Q_0 \lambda^{-\frac{1}{\alpha}}$ with the degrees of freedom*

$$\mathrm{df}_\lambda := \mathbb{E}\left[\left\|\mathbf{H}_\lambda(f_\mathcal{H})^{-\frac{1}{2}}\nabla \ell_Z(f_\mathcal{H})\right\|^2\right] \tag{15}$$

*and $\mathbf{H}_\lambda(f) := \mathbf{H}(f) + \lambda I$.*

The *degrees of freedom* term $\mathrm{df}_\lambda$ appears as *Fisher information* in the asymptotic analysis of M-estimation (Van der Vaart, 2000; Lehmann and Casella, 2006), appears similarly in spline smoothing (Wahba, 1990), and reduces to the *effective dimension* (Caponnetto et al., 2005; Caponnetto and De Vito, 2007) $\mathrm{df}_\lambda = \mathrm{Tr}(TT_\lambda^{-1})$ with $T_\lambda := T + \lambda I$ for the square loss. In the latter case, Assumption 6 is equivalent to the condition $\sigma_j = O(j^{-\alpha})$ for a sequence $\sigma_1, \sigma_2, \ldots$ of non-increasing positive eigenvalues of $T$ (Guo et al., 2022, Theorem 5). That is, a bigger $\alpha$ implies that fewer eigenvectors are needed to approximate elements from the image space of $T$ with a given capacity. On the other hand, from a regularization theory perspective, Assumption 6 can be interpreted as specifying the *ill-posedness* of the problem of inverting $T$ (cf. (Blanchard and Mücke, 2018)).

Under Assumptions 5 and 6, the following result is stated in (Marteau-Ferey et al., 2019) and rephrased here in our notation. See Subsection 6.4 for more details on the derivation.

**Lemma 3 (Marteau-Ferey et al. (2019), Theorem 38)** *Let Assumptions 1, 3–6 be satisfied, $\delta \in (0, \frac{1}{2}]$, and define $B_1$ and $B_2$ by*

$$B_1^* := \sup_{z \in \mathrm{supp}(\rho)} \mathrm{Tr}(\nabla \ell_z(f_\mathcal{H})), \quad B_2^* := \sup_{z \in \mathrm{supp}(\rho)} \mathrm{Tr}(\nabla^2 \ell_z(f_\mathcal{H})), \quad L := \|f_\mathcal{H}\|_{\mathbf{H}^{-2r}(f_\mathcal{H})},$$

*where $\|f\|_{\mathbf{A}} := \left\|\mathbf{A}^{\frac{1}{2}}f\right\|$. Whenever $0 < \lambda \leq \min\{B_2^*, (2LR\log\frac{2}{\delta})^{-1/r}, Q_0^{2\alpha}(B_2^*)^\alpha(B_1^*)^{-2\alpha}\}$ and $m + n$ is larger than*

$$\max\left\{5184\frac{B_2^*}{\lambda}\log\left(\frac{8 \cdot 414^2 B_2^*}{\lambda\delta}\right), \frac{1296Q_0^2}{L^2\lambda^{1+2r+1/\alpha}}\right\}, \tag{16}$$

*then a minimizer $f_{\mathbf{z}}^\lambda$ of Eq. (11) satisfies, with probability at least $1 - 2\delta$,*

$$\mathcal{R}(f_{\mathbf{z}}^\lambda) - \mathcal{R}(f_\mathcal{H}) \leq \left\|f_{\mathbf{z}}^\lambda - f_\mathcal{H}\right\|_{\mathbf{H}(f_\mathcal{H})}^2 \leq \left\|f_{\mathbf{z}}^\lambda - f_\mathcal{H}\right\|_{\mathbf{H}_\lambda(f_\mathcal{H})}^2$$

$$\leq 414\frac{Q_0^2}{(m+n)\lambda^{1/\alpha}}\log\left(\frac{2}{\delta}\right) + 414L^2\lambda^{1+2r}. \tag{17}$$

In the special case of square loss (SQ), it is possible to find $\lambda$ in Lemma 3 that recovers the minimax optimal error rate discovered in (Caponnetto and De Vito, 2007); see (Marteau-Ferey et al., 2019, Corollary 39). However, (Marteau-Ferey et al., 2019, Corollary 39) uses an a priori choice of $\lambda$ that depends on the regularity level $r$, which cannot be computed from the available samples.

## 3. Error Rates under Regularity and Capacity

Let us start by Assumption 1. The goal of this section is to analyze the error of a density ratio model $g(f_{\mathbf{z}}^\lambda)$ with a minimizer $f_{\mathbf{z}}^\lambda$ of Eq. (11), in terms of the data fidelity $B(f) := B_F(\beta, g(f))$ as specified in Eq. (9). We note that this setting is general, in the sense that all methods in Example 1 are included.

Our analysis is based on the observation that all loss functions in Example 2 are self-concordant as defined in Assumption 3.

**Example 3**

- *KuLSIF (Kanamori et al., 2009) uses the losses $\ell_{(x,-1)}(f) = \frac{1}{2}f(x)^2, \ell_{(x,1)}(f) = -f(x)$ which are three times differentiable and Eq. (12) holds with $\varphi((x,y)) = \{0\} \subset \mathcal{H}$.*

- *LR (Bickel et al., 2009) is self-concordant with $\varphi((x,y)) = \{yK_x\}$, see (Marteau-Ferey et al., 2019, Example 2) and (Bach, 2010).*

- *Exp (Menon and Ong, 2016) uses $\ell_{(x,y)}(f) = e^{yf(x)}$ which is self-concordant with $\varphi((x,y)) = \{yK_x\}$.*

- *SQ (Menon and Ong, 2016) uses losses which are three times differentiable and Eq. (12) holds with $\varphi((x,y)) = \{0\} \subset \mathcal{H}$.*

Before discussing faster error rates with more advanced regularity concepts, let us discuss the special role of the parameter $\lambda$ when chosen to balance the two terms in Eq. (14). As an intermediate result, we will provide first error rates for regularized losses beyond the KuLSIF approach in Eq. (3).

### 3.1 Slow Error Rate with A Priori Parameter Choice

We have observed above that many methods for density ratio estimation are based on self-concordant loss functions. From Lemma 2 we know that, for large enough sample size $m+n$, with probability larger than $1-2\delta$, a minimizer $f_{\mathbf{z}}^\lambda$ of Eq. (11) satisfies the following bound:

$$B(f_{\mathbf{z}}^\lambda) - B(f_{\mathcal{H}}) \leq S(m+n, \delta, \lambda) + A(\lambda), \tag{18}$$

where

$$S(m+n, \delta, \lambda) := 168 \frac{B_1^2}{\lambda(m+n)} \log\left(\frac{2}{\delta}\right) \quad \text{and} \quad A(\lambda) := 4\lambda \left\| f_{\mathcal{H}} \right\|^2.$$

The question is: *how to choose $\lambda$?*

One strategy in learning theory is to use the value $\lambda^*$ which *balances* (Lepskii, 1991; Goldenshluger and Pereverzev, 2000; Birgé, 2001; Mathé, 2006; De Vito et al., 2010; Mücke, 2018; Blanchard et al., 2019; Lu et al., 2020; Zellinger et al., 2021)

$$\eta \cdot S(m+n, \delta, \lambda^*) = A(\lambda^*) \tag{19}$$

for some $\eta \geq 1$ ensuring that $\lambda^*$ is in the admissible set of values as, e.g., specified by Eq. (13). The value $\lambda^*$ in Eq. (19) is achieved, due to the monotonicity of $S(m+n, \delta, \lambda)$ (decreasing) and $A(\lambda)$ (increasing, starting from $A(0) = 0$).

The parameter choice $\lambda^*$ is particularly interesting, as it minimizes the right-hand side of Eq. (18) up to a constant factor and, as a consequence, allows to achieve error rates of optimal order; see Subsection 6.1.

**Proposition 1** *Let Assumptions 1–4 be satisfied, and let $\delta \in (0, \frac{1}{2}]$. Let further $\lambda^*$ be the solution of Eq. (19) for $\eta := \frac{256R^2}{42} \|f_{\mathcal{H}}\|^2$, which we assume to be larger than 1. The there exists a quantity $C > 0$ not depending on $m, n, \delta$, such that the minimizer $f_{\mathbf{z}}^{\lambda^*}$ of Eq. (11) satisfies*

$$B(f_{\mathbf{z}}^{\lambda^*}) - B(f_{\mathcal{H}}) \leq C(m+n)^{-\frac{1}{2}} \log^{\frac{1}{2}} \left( \frac{2}{\delta} \right) \tag{20}$$

*with probability at least $1 - 2\delta$, for large enough sample size $m + n$, which depends on $\delta$.*

Explicit bounds on the sample size $m + n$ and the quantity $C$ are given in Eq. (36) and Eq. (38), respectively. The value for $\lambda^*$ specified in Proposition 1 and Eq. (35) can be computed *a priori*. That is, all information needed to achieve the error rate $(m + n)^{-\frac{1}{2}}$ is available. However, this situation changes in Subsection 3.2 when the regularity of $f_{\mathcal{H}}$ is taken into account.

**Remark 4** *Together with Example 3, our Proposition 1 provides (to the best of our knowledge) the first error rate for regularized density ratio estimation methods beyond KuLSIF. At the same time, for $B(f_{\mathbf{z}}^{\lambda}) = \|\beta - f_{\mathbf{z}}^{\lambda}\|_{L^2(Q)}^2$ as used in KuLSIF, the error has been analyzed in (Kanamori et al., 2012) in terms of the so-called bracketing entropy $\gamma$ of $\mathcal{H}$. In our notation, this result tells us that, if $\gamma \in (0, 2)$ and $\beta \in \mathcal{H}$ (so that $B(f_{\mathcal{H}}) = 0$), then, for $\epsilon$ satisfying $1 - \frac{2}{2+\gamma} < \epsilon < 1$ and $\lambda^{-1} = O(\min\{m, n\}^{1-\epsilon})$, with probability $1 - \delta$, it holds that*

$$B(f_{\mathbf{z}}^{\lambda}) = O(\min\{m, n\}^{\epsilon - 1}). \tag{21}$$

*The rate in Eq. (21) is faster for smaller $\epsilon$ and tends, for RKHSs with $\gamma \to 2$, to its optimum $\min\{m, n\}^{-\frac{1}{2}}$, which is the same rate as given by our Proposition 1. For $m \approx n$ as motivated by Assumption 1, our result therefore recovers the state of the art (Kanamori et al., 2009) for KuLSIF and RKHS with bracketing entropy $\gamma \to 2$. In the same case, the rate has been recently improved in (Gizewski et al., 2022) for regular $\beta$, which we discuss in the next Section 3.2.*

### 3.2 Fast Error Rate Requiring A Posteriori Parameter Choice

The following result is based on Lemma 3 and it takes into account the unknown regularity of $f_{\mathcal{H}}$ by Assumption 5.

**Proposition 2** *Let Assumptions 1–6 be satisfied for some $\alpha \leq \frac{1}{1-2r}$, and let $\delta \in [\frac{2}{e^{1296}}, \frac{1}{2}]$. Let further $\lambda^*$ be a solution of Eq. (19) with $\eta := 1296 \log^{-1}\left(\frac{2}{\delta}\right)$. Then there exists a quantity $C > 0$ not depending on $m, n, \delta$, such that the minimizer $f_{\mathbf{z}}^{\lambda^*}$ of Eq. (11) satisfies*

$$B(f_{\mathbf{z}}^{\lambda^*}) - B(f_{\mathcal{H}}) \leq C(m+n)^{-\frac{2r\alpha+\alpha}{2r\alpha+\alpha+1}} \tag{22}$$

*with probability at least $1 - 2\delta$, for large enough sample size $m + n$, which depends on $\delta$.*

Explicit bounds on $m+n$ and $C$ are given in Eq. (41) and Eq. (43), respectively. The value $\lambda^*$ in Proposition 2, as specified precisely in Eq. (40), depends on the regularity index $r$ and can therefore not be computed *a priori*. A new method is required.

**Remark 5** *Together with Example 3, Proposition 2 provides (the to the best of our knowledge first) error rates result for regularized density ratio estimation methods that takes into account both, the regularity of the density ratio $\beta$ and the capacity of the space $\mathcal{H}$. As a result, even in the special case of KuLSIF with $B(f_{\mathbf{z}}^\lambda) = \left\| \beta - f_{\mathbf{z}}^\lambda \right\|_{L^2(Q)}^2$, Proposition 2 allows to refine existing results. To see this, consider the fastest error rate of (Kanamori et al., 2012) discovered for RKHS with bracketing entropy $\gamma \in (0,2)$ (cf. Remark 1):*

$$B(f_{\mathbf{z}}^\lambda) = O\left( \min\{m,n\}^{-\frac{2}{2+\gamma}} \right). \tag{23}$$

*This rate is improved by $O\left( (m+n)^{-\frac{2r\alpha+\alpha}{2r\alpha+\alpha+1}} \right)$ in Eq. (22) whenever $\alpha(2r+1) > \frac{2}{\gamma}$. We can also mention (Gizewski et al., 2022, Remark 3) which leads to the error bound*

$$B(f_{\mathbf{z}}^\lambda) - B(f_{\mathcal{H}}) = O\left( \left( m^{-\frac{1}{2}} + n^{-\frac{1}{2}} \right)^{\frac{4r'+2}{2r'+2}} \right) \tag{24}$$

*with $r'$ originating from a source condition similar to our Assumption 5 but for the covariance operator $T = \int_{\mathcal{X}} k(x,\cdot) \otimes k(x,\cdot)\, \mathrm{d}Q(x)$ of $Q$ instead of $\rho$. The rate in Eq. (24) is slower than the one in Eq. (22) for $\alpha \in (1,\infty)$ and $0 < r' \leq r \leq \frac{1}{2}$. In general, for the square loss (SQ), the error rate in Eq. (22) is optimal in a minimax sense, see (Caponnetto and De Vito, 2007; Steinwart et al., 2009; Blanchard and Mücke, 2018).*

## 4. Balancing Principle using Approximation by Norm

### 4.1 Parameter Choice when the Norm is Known

The value $\lambda^*$ in Proposition 2 cannot be computed directly as it depends on the regularity of $\beta$ encoded in $r$; see Eq. (40). One solution to this problem is to estimate $\lambda^*$ with Lepskii's *balancing principle* (Lepskii, 1991; Goldenshluger and Pereverzev, 2000; Birgé, 2001; Mathé, 2006; De Vito et al., 2010; Lu et al., 2020; Zellinger et al., 2021). However, in general, no representation of $B(f_{\mathbf{z}}^\lambda) - B(f_{\mathcal{H}})$ as squared norm is available. This prevents us from using the above referenced approaches directly, e.g., by using the norm estimates (De Vito et al., 2010, Proposition 1) or (Lu et al., 2020, Proposition 4.1).

In the following, we propose a novel form of the balancing principle based on the norm $\|\cdot\|_{\mathbf{H}_\lambda(f_{\mathcal{H}})}$ originating from the self-concordance-based upper bound in Eq. (17). Since the estimation of $\|\cdot\|_{\mathbf{H}_\lambda(f_{\mathcal{H}})}$ requires some care, we postpone it to the next Subsection 4.2.

We start by defining a suitable discretization of increasing parameter candidates $(\lambda_i)_{i=1}^l \in \mathbb{R}$ such that $\lambda^* \in [\lambda_1, \lambda_l]$ is in the range of admissible values (ensured by multiplication of $\eta$ in Eq. (19)). The balancing principle is then based on checking a necessary condition. Namely, whenever $\lambda_j \leq \lambda_i \leq \lambda^*$, Lemma 3 and the monotonicity of

$$S(m+n, \delta, \lambda) := \frac{414 Q_0^2}{(m+n)\lambda^{1/\alpha}} \log\left(\frac{2}{\delta}\right) \quad \text{and} \quad A(\lambda) := 414 L^2 \lambda^{1+2r},$$

grant us the inequality

$$
\begin{aligned}
\left\| f_{\mathbf{z}}^{\lambda_i} - f_{\mathbf{z}}^{\lambda_j} \right\|_{\mathbf{H}_{\lambda_j}(f_{\mathcal{H}})}^2 &\leq 2 \left\| f_{\mathbf{z}}^{\lambda_j} - f_{\mathcal{H}} \right\|_{\mathbf{H}_{\lambda_j}(f_{\mathcal{H}})}^2 + 2 \left\| f_{\mathbf{z}}^{\lambda_i} - f_{\mathcal{H}} \right\|_{\mathbf{H}_{\lambda_j}(f_{\mathcal{H}})}^2 \\
&\leq 2 \left\| f_{\mathbf{z}}^{\lambda_j} - f_{\mathcal{H}} \right\|_{\mathbf{H}_{\lambda_j}(f_{\mathcal{H}})}^2 + 2 \left\| f_{\mathbf{z}}^{\lambda_i} - f_{\mathcal{H}} \right\|_{\mathbf{H}_{\lambda_i}(f_{\mathcal{H}})}^2 \\
&\leq 2S(m+n,\delta,\lambda_j) + 2A(\lambda_j) + 2S(m+n,\delta,\lambda_i) + 2A(\lambda_i) \\
&\leq 4S(m+n,\delta,\lambda_j) + 4S(m+n,\delta,\lambda_i) \\
&\leq 8S(m+n,\delta,\lambda_j) \\
&\leq 8\eta \cdot S(m+n,\delta,\lambda_j),
\end{aligned}
\tag{25}
$$

where we have used that $\mathbf{H}_{\lambda_j}(f) = \mathbf{H}(f) + \lambda_j I \leq \mathbf{H}(f) + \lambda_i I = \mathbf{H}_{\lambda_i}(f)$ and $\eta$ as defined in Proposition 2. Following the above reasoning, if the error bound in Lemma 3 is sharp, then the maximum $\lambda_i \in (\lambda_i)_{i=1}^l$ satisfying the above condition for all smaller $\lambda_j, j \in \{1, \ldots, i-1\}$, can be expected to be close to $\lambda^*$. This leads to the following estimate

$$
\lambda_+ := \max \left\{ \lambda_i : \left\| f_{\mathbf{z}}^{\lambda_i} - f_{\mathbf{z}}^{\lambda_j} \right\|_{\mathbf{H}_{\lambda_j}(f_{\mathcal{H}})}^2 \leq 8\eta \cdot S(m+n,\delta,\lambda_j), j \in \{1, \ldots, i-1\} \right\}.
\tag{26}
$$

Importantly, Eq. (25) does not depend on the unknown regularity index $r \in (0, \frac{1}{2}]$ but only on the quantities $m, n, Q_0, \alpha$ which are either given or an upper bound can be easily estimated from data. We may state the following preparatory result.

**Theorem 1** *Let Assumptions 1–6 be satisfied and let*

$$
\lambda_i := \lambda_0 \cdot \xi^i, i \in \{1, \ldots, l\}
$$

*with $\xi > 1$, $\lambda_0 \leq \xi^{-l} \min \left\{ B_2^*, \left( 2LR \log \frac{2}{\delta} \right)^{-\frac{1}{r}}, Q_0^{2\alpha}(B_2^*)^\alpha(B_1^*)^{-2\alpha} \right\}$, $l < \frac{e^{1296}}{4} - 2$ and $\delta \in [\frac{2}{e^{1296}}, \frac{1}{4+2l}]$. Then there exists a quantity $C > 0$ not depending on $m, n, \delta$, such that the minimizer $f_{\mathbf{z}}^{\lambda_+}$ of Eq. (11) satisfies*

$$
B(f_{\mathbf{z}}^{\lambda_+}) - B(f_{\mathcal{H}}) \leq C(m+n)^{-\frac{2r\alpha+\alpha}{2r\alpha+\alpha+1}}
\tag{27}
$$

*with probability at least $1 - (4 + 2l)\delta$, for large enough sample size $m + n$ depending on $\delta$.*

Bounds on the sample size $m + n$ and the quantity $C$ are given in Eq. (44) and Eq. (48), respectively. We note that Theorem 1 achieves the same error rate as Proposition 2 but uses the parameter choice $\lambda_+$ which is a posteriori computable if we have access to the norm values $\left\| f_{\mathbf{z}}^{\lambda_i} - f_{\mathbf{z}}^{\lambda_j} \right\|_{\mathbf{H}_{\lambda_j}(f_{\mathcal{H}})}^2$. However, these values are not accessible, as we don't have access to the measure $\rho$. This issue is solved in the next Subsection 4.2.

## 4.2 Parameter Choice with Estimated Norm

The following concentration result is proven in Subsection 6.4 and essentially combines analytical arguments for self-concordant losses (Marteau-Ferey et al., 2019) with concentration inequalities for Hermitian operators (Rudi and Rosasco, 2017).

**Lemma 6** *Let the assumptions of Lemma 3 be satisfied, let $\delta \in (0, \frac{1}{2}]$, denote the empirical Hessian by $\widehat{\mathbf{H}}(f) := \frac{1}{m+n} \sum_{z \in \mathbf{z}} \nabla^2 \ell_z(f)$, and let us write $\mathbf{B} \preccurlyeq \mathbf{A}$ iff $\mathbf{A} - \mathbf{B}$ is positive semi-definite. Whenever $\|\mathbf{H}(f_{\mathbf{z}}^\lambda)\| > 0$ and*

$$m + n \geq \frac{16 B_2^*}{\|\mathbf{H}(f_{\mathbf{z}}^\lambda)\|} \log\left(\frac{2}{\delta}\right), \tag{28}$$

*then, with probability at least $1 - 4\delta$,*

$$\widehat{\mathbf{H}}_\lambda(f_{\mathbf{z}}^\lambda) \preccurlyeq 6\mathbf{H}_\lambda(f_{\mathcal{H}}) \preccurlyeq 48\widehat{\mathbf{H}}_\lambda(f_{\mathbf{z}}^\lambda). \tag{29}$$

Let us denote by $(\lambda_i)_{i=1}^l \in \mathbb{R}$ a geometric sequence with $\lambda_i := \lambda_0 \cdot \xi^i, i \in \{1, \dots, l\}$ as defined in Theorem 1. Then, Lemma 6 motivates the following assumption

**Assumption 7** *For the sequence $(\lambda_i)_{i=1}^l$ there exists $b^* > 0$ with $b^* \leq \min_{i \in \{1,\dots,l\}} \|\mathbf{H}(f_{\mathbf{z}}^{\lambda_i})\|$.*

Note that for strictly convex loss functions, the Hessian is positive definite, and thus Assumption 7 is necessarily satisfied.

The same reasoning as done for Eq. (25) motivates, under Assumption 7, the estimate

$$\lambda_{\mathrm{BP}} := \max\left\{\lambda_i : \left\|f_{\mathbf{z}}^{\lambda_i} - f_{\mathbf{z}}^{\lambda_j}\right\|_{\widehat{\mathbf{H}}_{\lambda_j}(f_{\mathbf{z}}^{\lambda_j})}^2 \leq 48\eta \cdot S(m+n, \delta, \lambda_j), j \in \{1, \dots, i-1\}\right\} \tag{30}$$

with $\eta := 1296 \log^{-1}(\frac{2}{\delta})$ as defined in Proposition 2. Note that Eq. (30) uses only given data or information for which upper bounds are easily estimable. Moreover, the following error rates result holds.

**Theorem 2** *Let all assumptions of Theorem 1 be satisfied for some $l < \frac{e^{1296}}{12} - 1$ and $\delta \in [\frac{2}{e^{1296}}, \frac{1}{6+6l}]$. Let further $(\lambda_i)_{i=1}^l$ satisfy Assumption 7. Then there exists a quantity $C > 0$ not depending on $m, n, \delta$, such that the minimizer $f_{\mathbf{z}}^{\lambda_{\mathrm{BP}}}$ of Eq. (11) satisfies*

$$B(f_{\mathbf{z}}^{\lambda_{\mathrm{BP}}}) - B(f_{\mathcal{H}}) \leq C(m+n)^{-\frac{2r\alpha+\alpha}{2r\alpha+\alpha+1}} \tag{31}$$

*with probability at least $1 - (6 + 6l)\delta$, for large enough sample size $m + n$ depending on $\delta$.*

Exact bounds on the sample size $m+n$ and the quantity $C$ are given in Eq. (57) and Eq. (61), respectively. Theorem 2 is the main result of this work. For large enough sample sizes, it provides an a posteriori parameter choice $\lambda_{\mathrm{BP}}$ that allows the minimizer $f_{\mathbf{z}}^{\lambda_{\mathrm{BP}}}$ of Eq. (11) to achieve an error rate that improves with the regularity of the true solution.

**Remark 7** *To the best of our knowledge, Theorem 2 gives the first proof of the adaptivity to the unknown target regularity, for a parameter choice strategy in density ratio estimation. Even if similar adaptivity results hold for cross-validation in supervised learning (Caponnetto and Yao, 2010) and the aggregation method in covariate shift domain adaptation (Gizewski et al., 2022), their extension to density ratio estimation is not straight forward and an open future problem. In the special case of square loss (SQ), the order of the error rate is optimal (see Remark 5), and Eq. (30) thus provides, in this sense, adaptivity at optimal rate.*
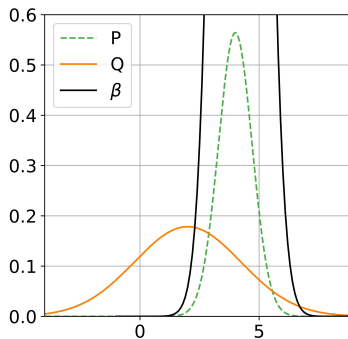
Figure 1: The goal is to learn the density ratio $\beta = \frac{\mathrm{d}P}{\mathrm{d}Q}$ (black, solid).

## 5. Numerical Example

To illustrate the behaviour of our approach, we rely on an example of (Shimodaira, 2000; Sugiyama et al., 2007; Dinu et al., 2023; Nguyen et al., 2023) with data from two Gaussian distributions $P, Q$ having means $4, 2$ and standard deviations $1/\sqrt{2}, \sqrt{5}$, respectively, see Figure 1.

We use two different density ratio estimation methods in Eq. (11), the Exp approach, and, the KuLSIF approach as described in Example 1. In particular, we approximate a minimizer of Eq. (11) by the conjugate gradient method applied to the parameters $\alpha_1, \ldots, \alpha_{m+n}$ of Eq. (11) when evaluated at a model ansatz $f_{\mathbf{z}}^\lambda = \sum_{j=1}^{m+n} \alpha_j k(x_j, \cdot)$ with reproducing kernel $k(x, x') := 1 + e^{-\frac{(x-x')^2}{2}}$ of $\mathcal{H}$, as suggested by the representer theorem (Schölkopf et al., 2001).

Following our goal of choosing $\lambda$, we fix a geometric sequence $(\lambda_i)_{i=1}^l \in \mathbb{R}$ with $\lambda_i := 10^i, i \in \{-3, \ldots, 1\}$ and compute

$$\lambda^\circ := \max\left\{\lambda_i : \left\|f_{\mathbf{z}}^{\lambda_i} - f_{\mathbf{z}}^{\lambda_j}\right\|_{\widehat{\mathbf{H}}_{\lambda_j}(f_{\mathbf{z}}^{\lambda_j})}^2 \leq \frac{M}{\lambda_j(m+n)}, j \in \{1, \ldots, i-1\}\right\} \tag{32}$$

for some $M \in \mathbb{R}$ and the pessimistic choice $\alpha = 1$ in Assumption 6. The empirical norm in Eq. (32) can be computed, for two minimizers $f_{\mathbf{z}}^{\lambda_s} = \sum_{i=1}^{m+n} \alpha_i k(x_i, \cdot)$ and $f_{\mathbf{z}}^{\lambda_t} = \sum_{i=1}^{m+n} \beta_i k(x_i, \cdot)$ of Eq. (11), by

$$\left\|f_{\mathbf{z}}^{\lambda_s} - f_{\mathbf{z}}^{\lambda_t}\right\|_{\widehat{\mathbf{H}}_{\lambda_t}(f_{\mathbf{z}}^{\lambda_t})}^2 = \frac{1}{m+n}(\boldsymbol{\alpha} - \boldsymbol{\beta})^{\mathrm{T}} \mathbf{K}\mathbf{E}\mathbf{K}(\boldsymbol{\alpha} - \boldsymbol{\beta}) + \lambda_t(\boldsymbol{\alpha} - \boldsymbol{\beta})^{\mathrm{T}}\mathbf{K}(\boldsymbol{\alpha} - \boldsymbol{\beta}),$$

where $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_{m+n})$, $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_{m+n})$, $\mathbf{K} := (k(x_i, x_j))_{i,j=1}^{m+n}$ and $\mathbf{E} := \mathrm{diag}(e_1, \ldots, e_{m+n})$ is a diagonal matrix with elements $e_i := e^{-y_i \cdot \sum_{j=1}^{m+n} \beta_j k(x_i, x_j)}$, in case of the Exp method, and $e_i := \frac{1-y_i}{2}$ in the case of KuLSIF, see Subsection 6.5.

To obtain theoretical error rate guarantees, Theorem 2 suggest to choose $M := C$ as given by Eq. (61). However, as already noted in (Lu et al., 2020, Remark 6.1), this choice is too rough for the low number of samples considered in practical applications. We therefore follow (Lu et al., 2020, Remark 6.1) and choose $M = M_j := \mathrm{Tr}(\widehat{\mathbf{H}}_{\lambda_j}(f_{\mathbf{z}}^{\lambda_j}))^{-2}$.
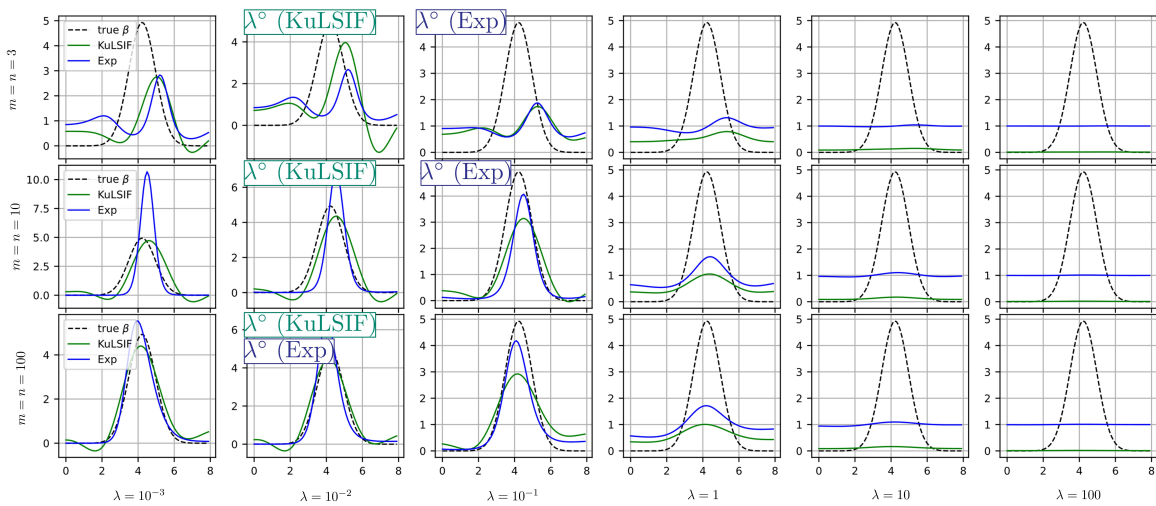
Figure 2: Results for approximating the density ratio $\beta$ (black, dashed) by the KuLSIF (green, solid) and Exp method (blue, solid). Rows: sample sizes $m = n \in \{3, 10, 100\}$. Columns: Parameter choices $\lambda = 10^i, i \in \{-3, \ldots, 2\}$. Choices of the proposed Lepskii type principle are marked by boxes in the upper left.

The results of our experiment can be found in Figure 2. The choice $\lambda^\circ$ in Eq. (32) corresponds to $\lambda = 10^{-2}$ for KuLSIF in all three cases of dataset size $m = n \in \{3, 10, 100\}$. In case of the Exp method, it chooses $\lambda = 10^{-1}$ for $m = n \in \{3, 10\}$ and $\lambda = 10^{-2}$ for $m = n = 100$. Although the choices are only optimal in two cases, w.r.t. the mean squared error evaluated at a dense grid of values, the parameter choice principle in Eq. (32) chooses always one of the two best values in the sequence.

## 6. Verification of Details

### 6.1 A priori parameter choice

The value $\lambda^*$ satisfying Eq. (19) minimizes Eq. (18) up to a constant factor (cf. Lu and Pereverzev (2013), Section 4.3). To see this, let us denote a minimizer by $\lambda_1 := \arg\min_{\lambda \geq 0} \eta S(m + n, \delta, \lambda) + A(\lambda)$. Then, either $\lambda^* < \lambda_1$, in which case

$$\eta S(m + n, \delta, \lambda^*) = A(\lambda^*) \leq A(\lambda_1) \leq \eta \inf_{\lambda \geq 0} \left[S(m + n, \delta, \lambda) + A(\lambda)\right],$$

or $\lambda_1 \leq \lambda^*$, and

$$A(\lambda^*) = \eta S(m + n, \delta, \lambda^*) \leq \eta S(m + n, \delta, \lambda_1) \leq \eta \inf_{\lambda \geq 0} \left[S(m + n, \delta, \lambda) + A(\lambda)\right],$$

which gives, for large enough sample sizes and with high probability,

$$B(f_{\mathbf{z}}^{\lambda^*}) - B(f_{\mathcal{H}}) \leq \eta S(m + n, \delta, \lambda^*) + A(\lambda^*) \leq 2\eta \inf_{\lambda \geq 0} \left[S(m + n, \delta, \lambda) + A(\lambda)\right]. \tag{33}$$

We will need the following technical lemma.

**Lemma 8** *If $A, B > 0$ and $k \geq 2A \log(2AB)$, then $A \log(Bk) \leq k$.*

**Proof** [Lemma 8] It holds that

$$A \log(Bk) = A \log\left(Bk \frac{2AB}{2AB}\right) = A \left(\log \frac{Bk}{2AB} + \log(2AB)\right) = A \left(\log \frac{k}{2A} + \log(2AB)\right)$$

$$\leq A \left(\frac{k}{2A} + \log(2AB)\right) = \frac{k}{2} + A \log(2AB),$$

which gives the desired result whenever $A \log(2AB) \leq k/2$. ∎

The following proof of Proposition 1 follows the essential steps of (Marteau-Ferey et al., 2019, Corollary 4) applied to Eq. (2) with explicit treatment of the balancing property Eq. (19) of $\lambda^*$.

**Proof** [Proposition 1] Using Assumptions 1 and 2, we obtain from Lemma 1

$$B(f_{\mathbf{z}}^\lambda) - B(f_{\mathcal{H}}) = 2 \left(\mathcal{R}(f_{\mathbf{z}}^\lambda) - \mathcal{R}(f^*)\right) - 2 \left(\mathcal{R}(f_{\mathcal{H}}) - \mathcal{R}(f^*)\right)$$

$$= 2 \left(\mathcal{R}(f_{\mathbf{z}}^\lambda) - \mathcal{R}(f_{\mathcal{H}})\right),$$

where $\mathcal{R}(f_{\mathbf{z}}^\lambda)$ is an expectation w.r.t. the loss function $\ell$ inducing $B(f_{\mathbf{z}}^\lambda) = B_F(\beta, g(f_{\mathbf{z}}^\lambda))$ as defined in Lemma 1. Let $B_1, B_2$ be as defined in Lemma 2. Then, under Assumptions 3 and 4, Lemma 2 gives, for large enough sample sizes $m + n$ (as specified later), with probability larger than $1 - 2\delta$,

$$B(f_{\mathbf{z}}^\lambda) - B(f_{\mathcal{H}}) \leq S(m + n, \delta, \lambda) + A(\lambda) \tag{34}$$

for $S(m + n, \delta, \lambda) = 168 \frac{B_1^2}{\lambda(m+n)} \log(\frac{2}{\delta})$ and $A(\lambda) = 4\lambda \|f_{\mathcal{H}}\|^2$. Solving

$$\eta \cdot S(m + n, \delta, \lambda^*) = A(\lambda^*)$$

for $\lambda^*$ with $\eta = \frac{16^2 R^2}{42} \|f_{\mathcal{H}}\|^2 \geq 1$ yields

$$\lambda^* = \frac{16 B_1 R}{\sqrt{m+n}} \log^{\frac{1}{2}}\left(\frac{2}{\delta}\right). \tag{35}$$

Let us now assume that $m + n$ is larger that

$$\max\left\{512 \|f_{\mathcal{H}}\|^2 R^2 \log\left(\frac{2}{\delta}\right), 512 \log\left(\frac{2}{\delta}\right), \frac{256 B_1^2 R^2}{B_2^2} \log\left(\frac{2}{\delta}\right), \frac{9 B_2^2 \log\left(\frac{6 B_2^2}{4\delta B_1^2 R^2} \log^{-1}\left(\frac{2}{\delta}\right)\right)}{B_1^2 R^2 \log\left(\frac{2}{\delta}\right)}\right\}. \tag{36}$$

In this case, $\lambda^*$ is an admissible value as specified by the conditions of Lemma 2:

1. The condition $\lambda^* \leq B_2$ is satisfied since $m + n \geq \frac{256 B_1^2 R^2}{B_2^2} \log\left(\frac{2}{\delta}\right)$.

2. The third element of the set from which the maximum is taken in Eq. (13) requires

$$m + n \geq 24\frac{B_2}{\lambda^*} \log\left(\frac{8B_2}{\lambda^*\delta}\right) = 3c_1(m+n)^{\frac{1}{2}} \log\left(\frac{1}{\delta}c_1(m+n)^{\frac{1}{2}}\right) \tag{37}$$

with $c_1 := \frac{8B_2}{\lambda^*(m+n)^{\frac{1}{2}}} = \frac{B_2}{2B_1R}\log^{-\frac{1}{2}}\left(\frac{2}{\delta}\right)$. Eq. (37) is guaranteed by Lemma 8 (with $k := (m+n)^{\frac{1}{2}}$, $A := 3c_1$ and $B := \frac{1}{\delta}c$), whenever $2A\log(2AB) \leq k$. That is, whenever $m+n$ is larger than

$$36c_1^2 \log^2\left(\frac{6c_1^2}{\delta}\right) = \frac{9B_2^2}{B_1^2R^2}\log^{-1}\left(\frac{2}{\delta}\right)\log\left(\frac{6B_2^2}{4\delta B_1^2R^2}\log^{-1}\left(\frac{2}{\delta}\right)\right).$$

3. The last element from the set from which the maximum is taken in Eq. (13) requires $m + n \geq 256\frac{R^2B_1^2}{(\lambda^*)^2}\log\left(\frac{2}{\delta}\right)$, which is satisfied.

Note that $\eta \cdot S(m+n,\delta,\lambda^*) + A(\lambda^*)$ upper bounds Eq. (34) evaluated at $\lambda^*$. The desired result follows with

$$C := 128\left\|f_{\mathcal{H}}\right\|^2 B_1 R. \tag{38}$$

∎

## 6.2 Error Rate Requiring A Posteriori Parameter Choice

The following proof of Proposition 2 uses Lemma 1 and follows the proof of (Marteau-Ferey et al., 2019, Corollary 9) applied to Eq. (2) (noting that Lemma 1 can be applied) with explicit treatment of the balancing property Eq. (19) of $\lambda^*$.

**Proof** [Proposition 2] Let us denote by $L := \|\mathbf{H}(f_{\mathcal{H}})^{-r}f_{\mathcal{H}}\|$. Using Assumptions 1 and 2, we obtain, from Lemma 1,

$$B(f_{\mathbf{z}}^{\lambda}) - B(f_{\mathcal{H}}) = 2\left(\mathcal{R}(f_{\mathbf{z}}^{\lambda}) - \mathcal{R}(f^*)\right) - 2\left(\mathcal{R}(f_{\mathcal{H}}) - \mathcal{R}(f^*)\right)$$
$$= 2\left(\mathcal{R}(f_{\mathbf{z}}^{\lambda}) - \mathcal{R}(f_{\mathcal{H}})\right),$$

where $\mathcal{R}(f_{\mathbf{z}}^{\lambda})$ is an expectation w.r.t. the loss function $\ell$ inducing $B(f_{\mathbf{z}}^{\lambda}) = B_F(\beta, g(f_{\mathbf{z}}^{\lambda}))$ as defined in Lemma 1. Let $B_1^*, B_2^*$ be as defined in Lemma 3. Then, under Assumptions 3–6, Lemma 3 gives, for large enough sample sizes $m+n$ (as specified below in Eq. (41)), with probability larger than $1 - 2\delta$,

$$B(f_{\mathbf{z}}^{\lambda}) - B(f_{\mathcal{H}}) \leq 2S(m+n,\delta,\lambda) + 2A(\lambda) \tag{39}$$

for $S(m+n,\delta,\lambda) = 414\frac{Q_0^2}{\lambda^{1/\alpha}(m+n)}\log\left(\frac{2}{\delta}\right)$ and $A(\lambda) = 414L^2\lambda^{1+2r}$. Solving Eq. (19) for $\lambda^*$ and $\eta = 1296\log^{-1}\left(\frac{2}{\delta}\right)$ (notably being larger than 1 since $\delta \geq 2e^{-1296}$), gives

$$\lambda^* = \left(\frac{1296Q_0^2}{(m+n)L^2}\right)^{\frac{\alpha}{1+2r\alpha+\alpha}}. \tag{40}$$

In what follows, we will show that $\lambda^*$ satisfies all assumptions from Lemma 3 whenever $m + n$ is larger than

$$1296 \cdot \max\left\{ \frac{Q_0^2}{L^2(B_2^*)^{1+2r+\frac{1}{\alpha}}}, \frac{Q_0^2}{L^2}\left(2LR\log\frac{2}{\delta}\right)^{\frac{\alpha+2\alpha r+1}{\alpha r}}, \frac{1296(B_1^*)^{2+4\alpha r+2\alpha}}{L^2Q_0^{4\alpha r+2\alpha}(B_2^*)^{1+2\alpha r+\alpha}}, \right.$$
$$\left. \frac{L^2\left(16\cdot648 B_2^*\right)^{1+2r+\frac{1}{\alpha}}}{1296^2 Q_0^2}\log^{1+2r+\frac{1}{\alpha}}\left(\frac{48(B_2^*)^2 1296\cdot414^2}{\delta}\left(\frac{1296Q_0^2}{L^2}\right)^{-\frac{2\alpha}{1+2r\alpha+\alpha}}\right)\right\}. \tag{41}$$

1. The requirement $\lambda^* \leq B_2^*$ is equivalent to

$$m + n \geq \frac{1296Q_0^2}{L^2(B_2^*)^{1+2r+\frac{1}{\alpha}}}.$$

2. The requirement $\lambda^* \leq (2LR\log\frac{2}{\delta})^{-1/r}$ is equivalent to

$$m + n \geq \frac{1296Q_0^2}{L^2}\left(2LR\log\frac{2}{\delta}\right)^{\frac{\alpha+2\alpha r+1}{\alpha r}}.$$

3. The requirement $\lambda^* \leq Q_0^{2\alpha}(B_2^*)^{\alpha}(B_1^*)^{-2\alpha}$ is equivalent to

$$m + n \geq \frac{1296(B_1^*)^{2+4\alpha r+2\alpha}}{L^2Q_0^{4\alpha r+2\alpha}(B_2^*)^{1+2\alpha r+\alpha}}.$$

4. The requirement $m + n \geq \frac{1296Q_0^2}{L^2(\lambda^*)^{1+2r+1/\alpha}}$ follows directly from the definition of $\lambda^*$.

5. The last requirement is

$$m + n \geq 5184\frac{B_2^*}{\lambda^*}\log\left(\frac{8\cdot414^2 B_2^*}{\lambda^*\delta}\right)$$
$$= 648(m+n)^{\frac{\alpha}{1+2r\alpha+\alpha}}c_1\log\left(\frac{414^2}{\delta}c_1(m+n)^{\frac{\alpha}{1+2r\alpha+\alpha}}\right) \tag{42}$$

with $c_1 := 8B_2^*\left(\frac{1296Q_0^2}{L^2}\right)^{-\frac{\alpha}{1+2r\alpha+\alpha}}$. Eq. (42) follows from Lemma 8 ($k := (m+n)^{\frac{\alpha}{1+2r\alpha+\alpha}}$, $A := 648c_1$ and $B := \frac{414^2}{\delta}c_1$), whenever

$$m + n \geq \frac{L^2\left(16\cdot648 B_2^*\right)^{1+2r+\frac{1}{\alpha}}}{1296Q_0^2}\log^{1+2r+\frac{1}{\alpha}}\left(\frac{64(B_2^*)^2 1296\cdot414^2}{\delta}\left(\frac{1296Q_0^2}{L^2}\right)^{-\frac{2\alpha}{1+2r\alpha+\alpha}}\right),$$

since $(m+n)^{\frac{\alpha}{1+2r\alpha+\alpha}} \leq (m+n)^{\frac{1}{2}}$. The latter follows from $\alpha \leq \frac{1}{1-2r}$.

18

Evaluating $2\eta \cdot S(m+n, \delta, \lambda^*) + 2A(\lambda^*)$, which upper bounds Eq. (39) evaluated at $\lambda^*$, gives the desired result for

$$C := 2 \cdot 828 L^2 \left( \frac{1296 Q_0^2}{L^2} \right)^{\frac{\alpha + 2r\alpha}{1 + 2\alpha r + \alpha}}. \tag{43}$$

∎

## 6.3 Parameter Choice with Known Norm

We first prove the result assuming access to the norm $\|\cdot\|_{\mathbf{H}_\lambda(f_\mathcal{H})}$.

**Proof** [Theorem 1] Using Assumptions 1 and 2, we obtain, from Lemma 1,

$$
\begin{aligned}
B(f_\mathbf{z}^{\lambda+}) - B(f_\mathcal{H}) &= 2\left(\mathcal{R}(f_\mathbf{z}^{\lambda+}) - \mathcal{R}(f^*)\right) - 2\left(\mathcal{R}(f_\mathcal{H}) - \mathcal{R}(f^*)\right) \\
&= 2\left(\mathcal{R}(f_\mathbf{z}^{\lambda+}) - \mathcal{R}(f_\mathcal{H})\right),
\end{aligned}
$$

where $\mathcal{R}(f_\mathbf{z}^{\lambda+})$ is an expectation w.r.t. the loss function $\ell$ inducing $B(f_\mathbf{z}^{\lambda+}) = B_F(\beta, f_\mathbf{z}^{\lambda+})$ as defined in Lemma 1. From Lemma 3 we further know that, with probability at least $1 - 2\delta$,

$$B(f_\mathbf{z}^{\lambda+}) - B(f_\mathcal{H}) \leq 2 \left\| f_\mathbf{z}^{\lambda+} - f_\mathcal{H} \right\|_{\mathbf{H}(f_\mathcal{H})}^2,$$

whenever

$$m + n \geq \max\left\{ 5184 \frac{B_2^*}{\lambda_0} \log\left( \frac{8 \cdot 414^2 B_2^*}{\lambda_0 \delta} \right), \frac{1296 Q_0^2}{L^2 \lambda_0^{1+2r+1/\alpha}} \right\}. \tag{44}$$

We now define

$$\overline{\lambda} := \max\{\lambda_i : A(\lambda_i) \leq \eta \cdot S(m+n, \delta, \lambda_i)\}$$

with $S(m+n, \delta, \lambda) := \frac{414 Q_0^2}{(m+n)\lambda^{1/\alpha}} \log\left(\frac{2}{\delta}\right)$, $A(\lambda) := 414 L^2 \lambda^{1+2r}$ as used in Lemma 3 and $\eta := 1296 \log^{-1}\left(\frac{2}{\delta}\right)$ as used in Proposition 2. Using $(a+b)^2 \leq 2a^2 + 2b^2$ and the triangle inequality, we obtain

$$
\begin{aligned}
B(f_\mathbf{z}^{\lambda+}) - B(f_\mathcal{H}) &\leq 4 \left\| f_\mathbf{z}^{\lambda+} - f_\mathbf{z}^{\overline{\lambda}} \right\|_{\mathbf{H}(f_\mathcal{H})}^2 + 4 \left\| f_\mathbf{z}^{\overline{\lambda}} - f_\mathcal{H} \right\|_{\mathbf{H}(f_\mathcal{H})}^2 \\
&\leq 4 \left\| f_\mathbf{z}^{\lambda+} - f_\mathbf{z}^{\overline{\lambda}} \right\|_{\mathbf{H}_{\overline{\lambda}}(f_\mathcal{H})}^2 + 4 \left\| f_\mathbf{z}^{\overline{\lambda}} - f_\mathcal{H} \right\|_{\mathbf{H}_{\overline{\lambda}}(f_\mathcal{H})}^2,
\end{aligned} \tag{45}
$$

where the last inequality follows from the fact that $\mathbf{H}_\lambda(f) = \mathbf{H}(f) + \lambda I \succcurlyeq \mathbf{H}(f)$ and it holds with probability at least $1 - 2\delta$, for large enough sample size.

We will now prove bounds on the two terms in Eq. (45) separately. To bound the first term, we note that $\overline{\lambda} \leq \lambda^*$ since $A(\lambda)$ is increasing and $S(m+n, \delta, \lambda)$ is decreasing in $\lambda$. Eq. (25) implies, for any $\lambda_j \leq \overline{\lambda} \leq \lambda^*$,

$$\left\| f_{\mathbf{z}}^{\overline{\lambda}} - f_{\mathbf{z}}^{\lambda_j} \right\|_{\mathbf{H}_{\lambda_j}(f_{\mathcal{H}})}^2 \leq 8\eta \cdot S(m+n, \delta, \lambda_j),$$

which holds with probability at least $1 - 2l\delta$ for all $j$ at the same time. As a consequence, we get $\overline{\lambda} \leq \lambda_+$ from the maximizing property of $\lambda_+$ in Eq. (26). Another consequence of $\overline{\lambda} \leq \lambda_+$ and Eq. (26) is that

$$\left\| f_{\mathbf{z}}^{\lambda_+} - f_{\mathbf{z}}^{\overline{\lambda}} \right\|_{\mathbf{H}_{\overline{\lambda}}(f_{\mathcal{H}})}^2 \leq 8\eta \cdot S(m+n, \delta, \overline{\lambda}). \tag{46}$$

The bound

$$\left\| f_{\mathbf{z}}^{\overline{\lambda}} - f_{\mathcal{H}} \right\|_{\mathbf{H}_{\overline{\lambda}}(f_{\mathcal{H}})}^2 \leq S(m+n, \delta, \overline{\lambda}) + A(\overline{\lambda}) \leq 2\eta \cdot S(m+n, \delta, \overline{\lambda}) \tag{47}$$

follows from Lemma 3 and the definition of $\overline{\lambda}$. Eq. (47) holds with probability at least $1 - 2\delta$. Inserting the bounds Eq. (46) and Eq. (47) into Eq. (45) gives

$$B(f_{\mathbf{z}}^{\lambda_+}) - B(f_{\mathcal{H}}) \leq 40\eta \cdot S(m+n, \delta, \overline{\lambda}),$$

which holds with probability at least $1 - (4+2l)\delta$. Without loss of generality assume that $\lambda_0 \cdot \xi^s = \overline{\lambda} \leq \lambda^* = \lambda_0 \cdot \xi^{s+1}$. It follows that $\lambda^* \leq \overline{\lambda} \cdot \xi$ and $\xi^{1/\alpha} S(m+n, \delta, \lambda^*) \geq S(m+n, \delta, \overline{\lambda})$. Consequently,

$$B(f_{\mathbf{z}}^{\lambda_+}) - B(f_{\mathcal{H}}) \leq 40\xi^{1/\alpha}\eta \cdot S(m+n, \delta, \lambda^*) \leq C(m+n)^{-\frac{2r\alpha+\alpha}{2r\alpha+\alpha+1}}$$

holds with probability at least $1 - (4+2l)\delta$ and with

$$C := 16560\xi^{1/\alpha}L^2 \left( \frac{1296 Q_0^2}{L^2} \right)^{\frac{\alpha+2r\alpha}{1+2\alpha r+\alpha}}. \tag{48}$$

∎

## 6.4 Parameter Choice with Empirical Norm

In the following, let $\mathcal{R}_\lambda(f) := \mathcal{R}(f) + \frac{\lambda}{2}\|f\|^2$, $f^\lambda := \arg\min_{f \in \mathcal{H}} \mathcal{R}_\lambda(f)$ and $\mathbf{t}(f) := \sup_{z \in \text{supp}(\rho)} \sup_{g \in \varphi(z)} |f \cdot g|$ with $\ell_z, Z, \rho, \varphi$ as in Assumptions 1, 4 and 3. Recall also the definitions

$$\mathbf{H}_\lambda(f) = \mathbf{H}(f) + \lambda I = \mathbb{E}[\nabla^2 \ell_Z(f)] + \lambda I \quad \text{and} \quad \widehat{\mathbf{H}}(f) = \frac{1}{m+n} \sum_{z \in \mathbf{z}} \nabla^2 \ell_z(f).$$

We need the following three lemmas.

**Lemma 9 (Marteau-Ferey et al. (2019), Proposition 15)** *Let Assumptions 1, 4 and 3 be satisfied, $\lambda \geq 0$ and $f_1, f_2 \in \mathcal{H}$. Then, we have*

$$\mathbf{H}_\lambda(f_1) \preccurlyeq e^{\mathbf{t}(f_1 - f_2)} \mathbf{H}_\lambda(f_2) \tag{49}$$

$$\mathcal{R}_\lambda(f_1) - \mathcal{R}_\lambda(f_0) - \nabla \mathcal{R}_\lambda(f_0)(f_1 - f_0) \leq \psi(\mathbf{t}(f_1 - f_0)) \|f_1 - f_0\|^2_{\mathbf{H}_\lambda(f_0)} \tag{50}$$

*with $\psi(t) = (e^t - t - 1)/t^2$.*

**Lemma 10 (Marteau-Ferey et al. (2019))** *Under the conditions of Lemma 3, we have, with probability at least $1 - 2\delta$,*

$$\mathbf{t}(f_\mathcal{H} - f^\lambda) \leq \log(2), \quad \mathbf{t}(f^\lambda - f^\lambda_{\mathbf{z}}) \leq \log(2). \tag{51}$$

**Proof** Lemma 33 in (Marteau-Ferey et al., 2019) shows that (noting that $\max\{\mathrm{df}_\lambda, (Q_0^*)^2\} \leq (B_1)^2/\lambda$ from their Proposition 18 and $1/r_\lambda(\theta^*) \leq R/\sqrt{\lambda}$ by the remark after their Assumption 6) the conditions on $n$ in Lemma 3 allow to apply their Proposition 16, which gives the desired result. ∎

**Lemma 11 (Marteau-Ferey et al. (2019), Rudi and Rosasco (2017))** *Let the conditions of Lemma 3 be satisfied. Then, it holds, with probability at least $1 - \delta$,*

$$\mathbf{H}_\lambda(f) \preccurlyeq 2\widehat{\mathbf{H}}_\lambda(f). \tag{52}$$

*If, in addition, $0 < \|\mathbf{H}(f)\|$, then it holds, for all*

$$m + n \geq \frac{16 B_2^*}{\|\mathbf{H}(f)\|} \log\left(\frac{2}{\delta}\right), \tag{53}$$

*with probability at least $1 - \delta$,*

$$\widehat{\mathbf{H}}_\lambda(f) \preccurlyeq \frac{3}{2} \mathbf{H}_\lambda(f). \tag{54}$$

**Proof** Eq. (52) is proven in (Marteau-Ferey et al., 2019, Lemma 30). Eq. (54) follows from (Marteau-Ferey et al., 2019, Proposition 47) (attributed to (Rudi and Rosasco, 2017), Proposition 6 and Proposition 8) and (Marteau-Ferey et al., 2019, Remark 48). ∎

We are now ready to provide more precise steps on the derivation of Lemma 3, which is essentially a re-phrasing of the proof of (Marteau-Ferey et al., 2019, Theorem 38) in our notation.

**Proof** [Lemma 3] From Eq. (50) and the fact that $\mathbf{H}(f) \preccurlyeq \mathbf{H}_\lambda(f)$, we obtain

$$\mathcal{R}(f^\lambda_{\mathbf{z}}) - \mathcal{R}(f_\mathcal{H}) \leq \psi(\mathbf{t}(f^\lambda_{\mathbf{z}} - f_\mathcal{H})) \left\| f^\lambda_{\mathbf{z}} - f_\mathcal{H} \right\|^2_{\mathbf{H}(f_\mathcal{H})}$$

$$\leq \psi(\mathbf{t}(f^\lambda_{\mathbf{z}} - f_\mathcal{H})) \left\| f^\lambda_{\mathbf{z}} - f_\mathcal{H} \right\|^2_{\mathbf{H}_\lambda(f_\mathcal{H})}, \tag{55}$$

which is exactly how the proof of (Marteau-Ferey et al., 2019, Theorem 38) works. Indeed, they proceed by showing how Eq. (55) can be upper bounded by $414\frac{Q_0^2}{(m+n)\lambda^{1/\alpha}}\log\left(\frac{2}{\delta}\right) + 414L^2\lambda^{1+2r}$. To obtain Lemma 3, we just remark that $\mathbf{t}(f_\mathbf{z}^\lambda - f_\mathcal{H}) \leq \mathbf{t}(f_\mathbf{z}^\lambda - f^\lambda) + \mathbf{t}(f_\mathbf{z}^\lambda - f^\lambda)) \leq 2\log(2)$ follows from Lemma 10. We further get $\psi(\mathbf{t}(f_\mathbf{z}^\lambda - f_\mathcal{H})) \leq \psi(2\log(2)) \leq 1$ since $\psi(t)$ increases with $t$. ∎

Lemma 6 can be obtained as follows.

**Proof** [Lemma 6] From Lemma 11 we get, with probability at least $1 - \delta$,

$$\widehat{\mathbf{H}}_\lambda(f_\mathbf{z}^\lambda) \preccurlyeq \frac{3}{2}\mathbf{H}_\lambda(f_\mathcal{H}).$$

From Lemma 9, we further obtain

$$\frac{3}{2}\mathbf{H}_\lambda(f_\mathcal{H}) \preccurlyeq \frac{3}{2}e^{\mathbf{t}(f_\mathbf{z}^\lambda - f_\mathcal{H})}\mathbf{H}_\lambda(f_\mathcal{H}) \preccurlyeq \frac{3}{2}e^{\mathbf{t}(f_\mathbf{z}^\lambda - f^\lambda)+\mathbf{t}(f^\lambda - f_\mathcal{H})}\mathbf{H}_\lambda(f_\mathcal{H}) \preccurlyeq 6\mathbf{H}_\lambda(f_\mathcal{H}),$$

where the last inequality uses also Lemma 10. Proceeding with Eq. (52) in Lemma 11 gives

$$6\mathbf{H}_\lambda(f_\mathcal{H}) \preccurlyeq 12\widehat{\mathbf{H}}_\lambda(f_\mathcal{H}) \preccurlyeq 12e^{\mathbf{t}(f_\mathbf{z}^\lambda - f_\mathcal{H})}\widehat{\mathbf{H}}_\lambda(f_\mathbf{z}^\lambda) \preccurlyeq 48\widehat{\mathbf{H}}_\lambda(f_\mathbf{z}^\lambda).$$

The conclusion follows whenever the above inequalities are satisfied at the same time, which happens with probability at least $1 - 4\delta$. ∎

We are now ready to prove our main result.

**Proof** [Theorem 2] Following the proof of Theorem 1 line by line up to Eq. (45) and replacing $\lambda_+$ by $\lambda_{\text{BP}}$ gives, with probability at least $1 - 2\delta$,

$$B(f_\mathbf{z}^{\lambda_{\text{BP}}}) - B(f_\mathcal{H}) \leq 4\left\|f_\mathbf{z}^{\lambda_{\text{BP}}} - f_\mathbf{z}^{\overline{\lambda}}\right\|_{\mathbf{H}_{\overline{\lambda}}(f_\mathcal{H})}^2 + 4\left\|f_\mathbf{z}^{\overline{\lambda}} - f_\mathcal{H}\right\|_{\mathbf{H}_{\overline{\lambda}}(f_\mathcal{H})}^2, \tag{56}$$

whenever

$$m + n \geq \max\left\{5184\frac{B_2^*}{\lambda_0}\log\left(\frac{8 \cdot 414^2 B_2^*}{\lambda_0\delta}\right), \frac{1296Q_0^2}{L^2\lambda_0^{2+1/\alpha}}, \frac{16B_2^*}{1296b^*}\log\left(\frac{2}{\delta}\right)\right\}, \tag{57}$$

where

$$\overline{\lambda} := \max\{\lambda_i : A(\lambda_i) \leq \eta \cdot S(m+n, \delta, \lambda_i)\},$$

$S(m+n, \delta, \lambda) := \frac{414Q_0^2}{(m+n)\lambda^{1/\alpha}}\log\left(\frac{2}{\delta}\right)$ and $A(\lambda) := 414L^2\lambda^{1+2r}$ are defined as in Lemma 3, and $\eta := 1296\log^{-1}\left(\frac{2}{\delta}\right)$ as defined in Proposition 2. Note that we used the requirement $\frac{1296Q_0^2}{L^2\lambda_0^{2+1/\alpha}} \geq \frac{1296Q_0^2}{L^2\lambda_0^{1+2r+1/\alpha}}$ in Eq. (57) that is independent of $r$, and, we added a requirement

based on $b^*$ from Assumption 7 to satisfy all conditions needed to apply Lemma 6. From Lemma 3 and the definition of $\overline{\lambda}$ we obtain a bound for the second term of Eq. (56)

$$\left\| f_{\mathbf{z}}^{\overline{\lambda}} - f_{\mathcal{H}} \right\|_{\mathbf{H}_{\overline{\lambda}}(f_{\mathcal{H}})}^2 \leq S(m+n,\delta,\overline{\lambda}) + A(\overline{\lambda}) \leq 2\eta \cdot S(m+n,\delta,\overline{\lambda}), \tag{58}$$

with holds again with probability at least $1 - 2\delta$.

To bound the first term in Eq. (56), we note that Lemma 6 implies that

$$\left\| f_{\mathbf{z}}^{\lambda_{\mathrm{BP}}} - f_{\mathbf{z}}^{\overline{\lambda}} \right\|_{\mathbf{H}_{\overline{\lambda}}(f_{\mathcal{H}})}^2 \leq 8 \left\| f_{\mathbf{z}}^{\lambda_{\mathrm{BP}}} - f_{\mathbf{z}}^{\overline{\lambda}} \right\|_{\widehat{\mathbf{H}}_{\overline{\lambda}}(f_{\mathbf{z}}^{\overline{\lambda}})}^2, \tag{59}$$

which holds with probability at least $1 - 4l\delta$ for all possibilities $\lambda_{\mathrm{BP}}, \overline{\lambda} \in \{\lambda_0, \ldots, \lambda_l\}$. Let us now focus on all $j$ which satisfy $\lambda_j \leq \overline{\lambda}$ and note that $\overline{\lambda} \leq \lambda^*$ by the monotonicity of $S(m+n,\delta,\lambda)$ and $A(\lambda)$. For the above mentioned $j$, we apply Lemma 6, which grants us, with probability at least $1 - 2l\delta$ for each $j$ at the same time,

$$\left\| f_{\mathbf{z}}^{\overline{\lambda}} - f_{\mathbf{z}}^{\lambda_j} \right\|_{\widehat{\mathbf{H}}_{\lambda_j}(f_{\mathbf{z}}^{\lambda_j})}^2 \leq 6 \left\| f_{\mathbf{z}}^{\overline{\lambda}} - f_{\mathbf{z}}^{\lambda_j} \right\|_{\mathbf{H}_{\lambda_j}(f_{\mathcal{H}})}^2.$$

Now, the same reasoning as for Eq. (25) can be applied, to obtain, for any $\lambda_j \leq \overline{\lambda} \leq \lambda^*$,

$$\left\| f_{\mathbf{z}}^{\overline{\lambda}} - f_{\mathbf{z}}^{\lambda_j} \right\|_{\widehat{\mathbf{H}}_{\lambda_j}(f_{\mathbf{z}}^{\lambda_j})}^2 \leq 48\eta \cdot S(m+n,\delta,\lambda_j),$$

which holds with probability at least $1 - (2+2l)\delta$ resulting from the joint application of Lemma 3 and Lemma 6. As a consequence, we get $\overline{\lambda} \leq \lambda_{\mathrm{BP}}$ from the maximizing property of $\lambda_{\mathrm{BP}}$ in Eq. (30). Since $\overline{\lambda} \leq \lambda_{\mathrm{BP}}$ and Eq. (30) holds, we get

$$\left\| f_{\mathbf{z}}^{\lambda_{\mathrm{BP}}} - f_{\mathbf{z}}^{\overline{\lambda}} \right\|_{\widehat{\mathbf{H}}_{\overline{\lambda}}(f_{\mathbf{z}}^{\overline{\lambda}})}^2 \leq 48\eta \cdot S(m+n,\delta,\overline{\lambda}), \tag{60}$$

which realizes with probability at least $1-(2+2l)\delta$. Combining the results above gives

$$B(f_{\mathbf{z}}^{\lambda_{\mathrm{BP}}}) - B(f_{\mathcal{H}}) \leq 32 \cdot 48\eta \cdot S(m+n,\delta,\overline{\lambda}) + 2 \cdot 4\eta \cdot S(m+n,\delta,\overline{\lambda}) = 1544\eta \cdot S(m+n,\delta,\overline{\lambda}),$$

which holds with probability at least $(6+6l)\delta$. With the same arguments as used in the proof of Theorem 1, it holds that $S(m+n,\delta,\overline{\lambda}) \leq \xi^{1/\alpha} S(m+n,\delta,\lambda^*)$ and further

$$B(f_{\mathbf{z}}^{\lambda_{\mathrm{BP}}}) - B(f_{\mathcal{H}}) \leq 1544\eta \cdot \xi^{1/\alpha} S(m+n,\delta,\lambda^*) \leq C(m+n)^{-\frac{2r\alpha+\alpha}{2r\alpha+\alpha+1}}$$

with the constant

$$C := 1296 \cdot 1544\xi^{1/\alpha} \cdot 828L^2 \left( \frac{1296Q_0^2}{L^2} \right)^{\frac{\alpha+2r\alpha}{1+2\alpha r+\alpha}}. \tag{61}$$

$\blacksquare$

### 6.5 Derivation of Empirical Norms

**Exp method**  Let $\ell_z(f) = e^{-yf(x)}$ for $z = (x, y)$. Then, the reproducing property of the RKHS $\mathcal{H}$ with kernel $k$ gives

$$\nabla^2 \ell_z(f) = \nabla^2 e^{-y \cdot f(x)} = \nabla^2 e^{-y \cdot \langle k(x, \cdot), f \rangle} = e^{-y \cdot \langle k(x, \cdot), f \rangle} k(x, \cdot) k(x, \cdot).$$

For the minimizer $f_{\mathbf{z}}^{\lambda} = \sum_{j=1}^{m+n} \alpha_j k(x_j, \cdot)$ of Eq. (11) as determined by the representer theorem, see, e.g. (Schölkopf et al., 2001), it follows that

$$\widehat{\mathbf{H}}_{\lambda}(f_{\mathbf{z}}^{\lambda}) = \frac{1}{m+n} \sum_{i=1}^{m+n} e^{-y_i \cdot \sum_{j=1}^{m+n} \alpha_j k(x_i, x_j)} k(x_i, \cdot) k(x_i, \cdot) + \lambda I.$$

For $f_{\mathbf{z}}^{\lambda_s} = \sum_{j=1}^{m+n} \alpha_j k(x_j, \cdot)$ and $f_{\mathbf{z}}^{\lambda_t} = \sum_{j=1}^{m+n} \beta_j k(x_j, \cdot)$ we obtain

$$\left\| f_{\mathbf{z}}^{\lambda_s} - f_{\mathbf{z}}^{\lambda_t} \right\|_{\widehat{\mathbf{H}}_{\lambda_t}(f_{\mathbf{z}}^{\lambda_t})}^2 = \left\langle \widehat{\mathbf{H}}_{\lambda_t}^{\frac{1}{2}}(f_{\mathbf{z}}^{\lambda_t})(f_{\mathbf{z}}^{\lambda_s} - f_{\mathbf{z}}^{\lambda_t}), \widehat{\mathbf{H}}_{\lambda_t}^{\frac{1}{2}}(f_{\mathbf{z}}^{\lambda_t})(f_{\mathbf{z}}^{\lambda_s} - f_{\mathbf{z}}^{\lambda_t}) \right\rangle$$

$$= \left\langle \widehat{\mathbf{H}}_{\lambda_t}(f_{\mathbf{z}}^{\lambda_t})(f_{\mathbf{z}}^{\lambda_s} - f_{\mathbf{z}}^{\lambda_t}), f_{\mathbf{z}}^{\lambda_s} - f_{\mathbf{z}}^{\lambda_t} \right\rangle$$

$$= \left\langle \widehat{\mathbf{H}}_{\lambda_t}(f_{\mathbf{z}}^{\lambda_t}) \sum_{j=1}^{m+n} (\alpha_j - \beta_j) k(x_j, \cdot), \sum_{j=1}^{m+n} (\alpha_j - \beta_j) k(x_j, \cdot) \right\rangle$$

$$= \frac{1}{m+n} \sum_{i=1}^{m+n} \left\langle \sum_{t=1}^{m+n} (\alpha_t - \beta_t) k(x_i, x_t) e_i k(x_i, \cdot), \sum_{j=1}^{m+n} (\alpha_j - \beta_j) k(x_j, \cdot) \right\rangle$$

$$\quad + \lambda_t (\boldsymbol{\alpha} - \boldsymbol{\beta})^{\mathrm{T}} \mathbf{K} (\boldsymbol{\alpha} - \boldsymbol{\beta})$$

$$= \frac{1}{m+n} (\boldsymbol{\alpha} - \boldsymbol{\beta})^{\mathrm{T}} \mathbf{K} \mathbf{E} \mathbf{K} (\boldsymbol{\alpha} - \boldsymbol{\beta}) + \lambda_t (\boldsymbol{\alpha} - \boldsymbol{\beta})^{\mathrm{T}} \mathbf{K} (\boldsymbol{\alpha} - \boldsymbol{\beta}),$$

where $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_{m+n})$, $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_{m+n})$, diagonal matrix $\mathbf{E} := \mathrm{diag}(e_1, \ldots, e_{m+n})$, $e_i := e^{-y_i \cdot \sum_{j=1}^{m+n} \beta_j k(x_i, x_j)}$ and $\mathbf{K} := (k(x_i, x_j))_{i,j=1}^{m+n}$.

**KuLSIF**  Let $\ell_{(x,-1)}(f) = \frac{1}{2} f(x)^2$ and $\ell_{(x,1)}(f) = -f(x)$. Then, the reproducing property of the RKHS $\mathcal{H}$ with kernel $k$ gives

$$\nabla^2 \ell_{(x,-1)}(f) = k(x, \cdot) k(x, \cdot), \nabla^2 \ell_{(x,1)}(f) = 0.$$

For the minimizer $f_{\mathbf{z}}^{\lambda} = \sum_{j=1}^{m+n} \alpha_j k(x_j, \cdot)$ of Eq. (11), it follows that

$$\widehat{\mathbf{H}}_{\lambda}(f_{\mathbf{z}}^{\lambda}) = \frac{1}{m+n} \sum_{i=1}^{m+n} \frac{1 - y_i}{2} k(x_i, \cdot) k(x_i, \cdot) + \lambda I.$$

For $f_{\mathbf{z}}^{\lambda_s} = \sum_{j=1}^{m+n} \alpha_j k(x_j, \cdot)$ and $f_{\mathbf{z}}^{\lambda_t} = \sum_{j=1}^{m+n} \beta_j k(x_j, \cdot)$ we get, analogously to the derivation for the Exp method,

$$\left\| f_{\mathbf{z}}^{\lambda_s} - f_{\mathbf{z}}^{\lambda_t} \right\|_{\widehat{\mathbf{H}}_{\lambda_t}(f_{\mathbf{z}}^{\lambda_t})}^2 = \frac{1}{m+n} (\boldsymbol{\alpha} - \boldsymbol{\beta})^{\mathrm{T}} \mathbf{K} \mathbf{E} \mathbf{K} (\boldsymbol{\alpha} - \boldsymbol{\beta}) + \lambda_t (\boldsymbol{\alpha} - \boldsymbol{\beta})^{\mathrm{T}} \mathbf{K} (\boldsymbol{\alpha} - \boldsymbol{\beta}),$$

where $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_{m+n})$, $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_{m+n})$, diagonal matrix $\mathbf{E} := \mathrm{diag}(e_1, \ldots, e_{m+n})$, $e_i := \frac{1 - y_i}{2}$ and $\mathbf{K} := (k(x_i, x_j))_{i,j=1}^{m+n}$.

## 7. Conclusion and Future Work

We proposed error bounds for a large class of kernel methods for density ratio estimation for which no error bounds were known before (to the best of our knowledge). The class consists of methods which minimize a regularized Bregman divergence over a reproducing kernel Hilbert space. We also proposed a Lepskii type principle for choosing the regularization parameter. The principle minimizes the error bounds without knowledge of the unknown regularity of the density ratio. As a consequence it achieves a minimax optimal error rate in the special case of square loss. The key proof technique for both, the error rates and the adaptive parameter choice, is a local quadratic approximation of the Bregman divergence, which is equivalent to an expected risk for a proper composite self-concordant loss function. The minimax optimality of our error bounds for non-square losses is an open problem.

## Acknowledgements

## References

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4(none):384 – 414, 2010.

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, 2007.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(9), 2009.

Lucien Birgé. An alternative point of view on lepski's method. *Lecture Notes-Monograph Series*, pages 113–133, 2001.

Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18:971–1013, 2018.

Gilles Blanchard, Peter Mathé, and Nicole Mücke. Lepskii principle in supervised learning. *arXiv preprint arXiv:1905.10764*, 2019.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Andrea Caponnetto and Yuan Yao. Cross-validation based adaptation for regularization operators in learning theory. *Analysis and Applications*, 8(02):161–183, 2010.

Andrea Caponnetto, Lorenzo Rosasco, Ernesto De Vito, and Alessandro Verri. Empirical effective dimension and optimal rates for regularized least squares algorithm. Technical report, Computer Science and Artificial Intelligence Laboratory (CSAIL), MIT, 2005.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2001.

Ernesto De Vito, Sergei V Pereverzyev, and Lorenzo Rosasco. Adaptive kernel methods using the balancing principle. *Foundations of Computational Mathematics*, 10(4):455–479, 2010.

M.-C. Dinu, M. Holzleitner, M. Beck, D. H. Nguyen, A. Huber, H. Eghbal-zadeh, B. A. Moser, S. V. Pereverzyev, S. Hochreiter, and W. Zellinger. Addressing parameter choice issues in unsupervised domain adaptation by aggregation. *International Conference on Learning Representations*, 2023.

Richard M Dudley. *Cambridge studies in advanced mathematics: Real analysis and probability.* 74. Cambridge University Press, 2nd edition, 2002.

Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

Elke R Gizewski, Lukas Mayer, Bernhard A Moser, Duc Hoan Nguyen, Sergiy Pereverzyev Jr, Sergei V Pereverzyev, Natalia Shepeleva, and Werner Zellinger. On a regularization of unsupervised domain adaptation in RKHS. *Applied and Computational Harmonic Analysis*, 57:201–227, 2022.

Alexander Goldenshluger and Sergei V Pereverzev. Adaptive estimation of linear functionals in Hilbert scales from indirect white noise observations. *Probability Theory and Related Fields*, 118(2):169–186, 2000.

Xin Guo, Zheng-Chu Guo, and Lei Shi. Capacity dependent analysis for functional online learning algorithms. *arXiv preprint arXiv:2209.12198*, 2022.

Shohei Hido, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, and Takafumi Kanamori. Statistical outlier detection using direct density ratio estimation. *Knowledge and Information Systems*, 26:309–336, 2011.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445, 2009.

Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. $f$-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE Transactions on Information Theory*, 58(2):708–720, 2011.

Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86(3):335–367, 2012.

Masahiro Kato, Takeshi Teshima, and Junya Honda. Learning from positive and unlabeled data with a selection bias. In *International Conference on Learning Representations*, 2019.

Amor Keziou and Samuela Leoni-Aubin. Test of homogeneity in semiparametric two-sample density ratio models. *Comptes Rendus Mathématique*, 340(12):905–910, 2005.

Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

OV Lepskii. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466, 1991.

Shuai Lu and Sergei V Pereverzev. Regularization theory for ill-posed problems. In *Regularization Theory for Ill-posed Problems*. de Gruyter, 2013.

Shuai Lu, Peter Mathé, and Sergei V Pereverzev. Balancing principle in supervised learning for a general regularization scheme. *Applied and Computational Harmonic Analysis*, 48 (1):123–148, 2020.

Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis Bach, and Alessandro Rudi. Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. In *Conference on Learning Theory*, pages 2294–2340. PMLR, 2019.

Peter Mathé. The Lepskii principle revisited. *Inverse problems*, 22(3):L11, 2006.

Aditya Menon and Cheng Soon Ong. Linking losses for density ratio and class-probability estimation. In *International Conference on Machine Learning*, pages 304–313. PMLR, 2016.

Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

Nicole Mücke. Adaptivity for regularized kernel methods by Lepskii's principle. *arXiv preprint arXiv:1804.05433*, 2018.

Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

DH Nguyen, W Zellinger, and S Pereverzyev. On regularized Radon-Nikodym differentiation. *RICAM-Report 2023-13*, 2023.

XuanLong Nguyen, Martin J Wainwright, and Michael Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. *Advances in Neural Information Processing Systems*, 20, 2007.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Qichao Que and Mikhail Belkin. Inverse density as an inverse problem: The Fredholm equation approach. *Advances in Neural Information Processing Systems*, 26, 2013.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. *Advances in Neural Information Processing Systems*, 30, 2017.

Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International Conference on Computational Learning Theory*, pages 416–426. Springer, 2001.

Ingmar Schuster, Mattes Mollenhauer, Stefan Klus, and Krikamol Muandet. Kernel conditional density operators. In *International Conference on Artificial Intelligence and Statistics*, pages 993–1004. PMLR, 2020.

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

Alex Smola, Le Song, and Choon Hui Teo. Relative novelty detection. In *Artificial Intelligence and Statistics*, pages 536–543. PMLR, 2009.

Ingo Steinwart, Don R Hush, Clint Scovel, et al. Optimal rates for regularized least squares regression. In *Conference on Learning Theory*, pages 79–93, 2009.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8 (5), 2007.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012a.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044, 2012b.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Grace Wahba. *Spline models for observational data*. SIAM, 1990.

Werner Zellinger, Natalia Shepeleva, Marius-Constantin Dinu, Hamid Eghbal-zadeh, Hoan Duc Nguyen, Bernhard Nessler, Sergei Pereverzyev, and Bernhard A Moser. The balancing principle for parameter choice in distance-regularized domain adaptation. *Advances in Neural Information Processing Systems*, 34, 2021.

Tong Zhang. Effective dimension and generalization of kernel learning. *Advances in Neural Information Processing Systems*, 15, 2002.