UNIVERSITY *of York*

This is a repository copy of *The Goldenshluger-Lepski Method for Constrained Least-Squares Estimators over RKHSs*.

The Goldenshluger–Lepski method for constrained least-squares estimators over RKHSs

STEPHEN PAGE[1] and STEFFEN GRÜNEWÄLDER[2]

[1]*STOR-i, Lancaster University, Lancaster, LA1 4YF, United Kingdom. E-mail: s.page@lancaster.ac.uk*
[2]*Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, United Kingdom.*
E-mail: s.grunewalder@lancaster.ac.uk

We study an adaptive estimation procedure called the Goldenshluger–Lepski method in the context of reproducing kernel Hilbert space (RKHS) regression. Adaptive estimation provides a way of selecting tuning parameters for statistical estimators using only the available data. This allows us to perform estimation without making strong assumptions about the estimand. In contrast to procedures such as training and validation, the Goldenshluger–Lepski method uses all of the data to produce non-adaptive estimators for a range of values of the tuning parameters. An adaptive estimator is selected by performing pairwise comparisons between these non-adaptive estimators. Applying the Goldenshluger–Lepski method is non-trivial as it requires a simultaneous high-probability bound on all of the pairwise comparisons. In the RKHS regression context, we choose our non-adaptive estimators to be clipped least-squares estimators constrained to lie in a ball in an RKHS. Applying the Goldenshluger–Lepski method in this context is made more complicated by the fact that we cannot use the L^2 norm for performing the pairwise comparisons as it is unknown. We use the method to address two regression problems. In the first problem the RKHS is fixed, while in the second problem we adapt over a collection of RKHSs.

Keywords: Adaptive estimation; Goldenshluger–Lepski method; RKHS regression

1. Introduction

In nonparametric statistics, it is assumed that the estimand belongs to a very large parameter space in order to avoid model misspecification. Such misspecification can lead to large approximation errors and poor estimator performance. However, it is often challenging to produce estimators which are robust against such large parameter spaces. An important tool which allows us to achieve this aim is adaptive estimation. Adaptive estimators behave as if they know the true model from a collection of models, despite being a function of the data. In particular, adaptive estimators can often achieve the same optimal rates of convergence as the best estimators when the true model is known.

In this paper, we study an adaptive estimation procedure called the Goldenshluger–Lepski method in the context of reproducing kernel Hilbert space (RKHS) regression. The Goldenshluger–Lepski method works by performing pairwise comparisons between non-adaptive estimators with a range of values for the tuning parameters. As far as we are aware, this is the first time that this method has been applied in the context of RKHS regression. The Goldenshluger–Lepski method, introduced in the series of papers [16–18,20], is an extension of Lepski's method. While Lepski's method focusses on adaptation over a single parameter, the Goldenshluger–Lepski method can be used to perform adaptation over multiple parameters.

The Goldenshluger–Lepski method operates by selecting an estimator which minimises the sum of a proxy for the unknown bias and an inflated variance term. The proxy for the bias is calculated by performing pairwise comparisons between the estimator in question and all estimators which are in

some sense less smooth than this estimator. A key challenge in applying the Goldenshluger–Lepski method is proving a high-probability bound on all of these pairwise comparisons simultaneously. This bound is called a majorant.

We now describe the RKHS regression problem studied in this paper. We denote our covariate space by S and we assume that the regression function lies in an interpolation space between $C(S)$, the space of continuous functions equipped with the supremum norm, and an RKHS. Depending on the setting, this RKHS may be fixed or we may perform adaptation over a collection of RKHSs. The non-adaptive estimators we use in this context are clipped versions of least-squares estimators which are constrained to lie in a ball of predefined radius in an RKHS. These estimators are discussed in detail in [32]. Constraining an estimator to lie in a ball of predefined radius is a form of Ivanov regularisation (see [31]). One advantage of these estimators is that there is a clear way of producing a majorant for them, especially when the RKHS is fixed. This is because we can control the estimator constrained to lie in a ball of radius r by bounding quantities of the form rZ for some random variable Z which does not depend on r.

When the RKHS is fixed, the only tuning parameter to be selected is the radius of the ball in which the least-squares estimator is constrained to lie. Estimators for which the radius is larger are considered to be less smooth. In order to provide a majorant for the Goldenshluger–Lepski method, we must prove regression results which control these estimators for all radii simultaneously. When we perform adaptation over a collection of RKHSs, we must prove regression results which control the same estimators for all RKHSs and all ball radii in these RKHSs simultaneously. We demonstrate this approach for a collection of RKHSs with Gaussian kernels. Estimators for which both the width parameter of the Gaussian kernel is smaller and the radius of the ball in the RKHS is larger are considered to be less smooth. These results extend those of [32].

Our main results are Theorems 4 (page 2248) and 9 (page 2253). These show that a fixed quantile of the squared $L^2(P)$ error of a clipped version of the estimator produced by the Goldenshluger–Lepski method is of order $n^{-\beta/(1+\beta)}$. Here, n is the number of data points and β parametrises the interpolation space between $C(S)$ and the RKHS containing the regression function. We use $C(S)$ when interpolating so that we have direct control over approximation errors in the $L^2(P_n)$ norm. Theorem 4 addresses the case in which the RKHS is fixed and Theorem 9 addresses the case in which we perform adaptation over a collection of RKHSs with Gaussian kernels. The order $n^{-\beta/(1+\beta)}$ for the squared $L^2(P)$ error of the adaptive estimators matches the order of the smallest bounds obtained in [32] for the squared $L^2(P)$ error of the non-adaptive estimators. In the sense discussed in [32], this order is the optimal power of n if we make the slightly weaker assumption that the regression function is an element of the interpolation space between $L^2(P)$ and the RKHS parametrised by β. In particular, for a fixed $\beta \in (0, 1)$, Steinwart et al. [38] show that there is an instance of our problem such that the following holds. For all estimators $\hat{f}$ of g, for some $\varepsilon > 0$, we have

$$\|\hat{f} - g\|^2_{L^2(P)} \geq C_{\alpha,\varepsilon} n^{-\alpha}$$

with probability at least ε for all $n \geq 1$, for some constant $C_{\alpha,\varepsilon} > 0$, for all $\alpha > \beta/(1+\beta)$. This implies that for all estimators $\hat{f}$ of g, we have

$$\mathbb{E}\big(\|\hat{f} - g\|^2_{L^2(P)}\big) \geq C_{\alpha,\varepsilon} \varepsilon n^{-\alpha}$$

for all $n \geq 1$, for all $\alpha > \beta/(1+\beta)$. In this sense, our expectation bound in this setting is optimal because it attains the order $n^{-\beta/(1+\beta)}$, the smallest possible power of n.

Under further assumptions on the effective dimension of the RKHS faster rates of convergence can be obtained. For instance, in Steinwart et al. [38] optimal rates of convergence under assumptions on the effective dimension are attained and our rates correspond to their worst-case rates. Also see [10,28] for results on estimating the effective dimension from data.

1.1. Literature review

Lepski's method was introduced in the series of papers [24–26] as a method for adaptation over a single parameter. It has since been studied in, for example, [6] and [15]. Lepski's method selects the smoothest non-adaptive estimator from a collection, subject to a bound on a series of pairwise comparisons involving all estimators at most as smooth as the resulting estimator. The method can only adapt to one parameter because of the need for an ordering of the collection of non-adaptive estimators.

Lepski's method has been applied to RKHS regression under the name of the balancing principle. However, as far as we are aware, Lepski's method has not been used to target the true regression function, but instead an RKHS element which approximates the true regression function. In [13], the authors note the difficulty in using Lespki's method to control the squared $L^2(P)$ error of an adaptive version of a support vector machine (SVM). This difficulty arises because Lepski's method generally requires the norm we are interested in controlling to be known in order to perform the pairwise comparisons. However, P is unknown in this situation. The authors of [13] get around the problem that P is unknown as follows. Lepski's method is used to control the known squared $L^2(P_n)$ error and squared RKHS error of two different adaptive SVMs. The results of these procedures are combined to produce an adaptive SVM whose squared $L^2(P)$ error is bounded. The above alteration is also noted in [27]. Furthermore, the authors show that it is possible to greatly reduce the number of pairwise comparisons which must be performed to produce an adaptive estimator. This is done by only comparing each estimator to the estimator which is next less smooth.

Lepski's method is common in the inverse problems literature and is there also known as the balancing principle. In this context, it is not assumed that the target function lies in a particular RKHS but within a scale of Hilbert spaces. Scales of Hilbert spaces have been developed in [23] and have been introduced to the inverse problems literature in [29] and [30]. These scales of Hilbert spaces are closely linked to interpolation spaces [23]. Adaptation over these scales is typically achieved by combining Tikhonov regularisation with Lepski's method. This approach achieves in various settings close to optimal rates of convergence (e.g. [19,35]).

The Goldenshluger–Lepski method extends Lepski's method in order to perform adaptation over multiple parameters. The method is introduced in the series of papers [16–18,20]. The first two papers concentrate on function estimation in the presence of white noise. The first paper considers the problem of pointwise estimation, while the second paper examines estimation in the L^p norm for $p \in [1, \infty]$. The third paper produces adaptive bandwidth estimators for kernel density estimation and the fourth paper considers general methodology for selecting a linear estimator from a collection.

Another popular approach to model selection is to use a training and validation set. The training set is used to produce a collection of non-adaptive estimators for a range of different values for the tuning parameters and the validation set is used to select the best estimator from this collection. This selection is performed by calculating a proxy for the cost function that we wish to minimise. An example of using training and validation to perform adaptation over a Gaussian kernel parameter for an SVM can be found in [14]. The procedure produces an adaptive estimator of a bounded regression function from a range of Sobolev spaces. This estimator is analysed using union bounding, as opposed to the chaining techniques used to analyse the Goldenshluger–Lepski method in this paper. The training and validation approach is studied in a more general context in [11] using inverse problem techniques. One important advantage of the Goldenshluger–Lepski method in comparison to training and validation is that it uses all of the data to calculate the non-adaptive estimators. This is because it does not require data for calculating a proxy cost function. However, the Goldenshluger–Lepski method does require us to calculate a majorant, as discussed above, which is often a challenging task.

There are various other approaches to gain adaptive estimators. For instance, in [2] regression under random design is studied. The estimators are chosen adaptively from a family of finite dimensional

subspaces of $L^2(\nu)$, where ν is a known measure on the covariate space. The estimation approach selects a member of the family of subspaces by minimising a linear combination of the best least-squares error that can be achieved within the class and a penalty term which penalises model complexity. In contrast to Lepski's method the model class is selected based solely on this value and no pairwise comparison between model classes is performed. Model selection based on penalised least-squares values is a well-studied area, in particular, in the Gaussian noise case. To mention but a few references, in [8] a versatile framework for model selection under a penalised least-squares criterion is developed and in [9] the role of the penalty term in this framework is analysed in depth. Closely related are earlier works [3,7] where penalty terms are used in conjunction with sieves to produce adaptive estimators.

We perform in this paper adaptation over a family of kernel functions. This approach touches on the topic of multiple kernel learning (MKL) where a family of kernels is given and a regressor is chosen from the corresponding family of RKHSs. Early MKL papers like [1,12] focus on the optimisation aspect of the problem, i.e. how to efficiently find a suitable kernel within a large family of kernels to gain small empirical errors. In [21] a broad overview of MKL techniques up to the year 2011 is given. A more recent overview is contained in [34]. Much of the MKL literature focuses solely on the optimisation problem and not on controlling the risk of estimators. To the best of our knowledge, the risk has not been studied in the MKL literature under our assumption that the unknown regression function can lie outside of all the involved RKHSs.

2. Problem definition

We give a formal definition of the RKHS regression problem. First, recall that an RKHS H on S is a Hilbert space of real-valued functions on S such that, for all $x \in S$, there is some $k_x \in H$ such that $h(x) = \langle h, k_x \rangle_H$ for all $h \in H$. The function $k(x_1, x_2) = \langle k_{x_1}, k_{x_2} \rangle_H$ for $x_1, x_2 \in S$ is known as the kernel and is symmetric and positive-definite.

For a topological space T, let $\mathcal{B}(T)$ be its Borel σ-algebra. Let $(S, \mathcal{S})$ be a measurable space and (X_i, Y_i) for $1 \le i \le n$ be i.i.d. $(S \times \mathbb{R}, \mathcal{S} \otimes \mathcal{B}(\mathbb{R}))$-valued random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We assume $X_i \sim P$ and $\mathbb{E}(Y_i^2) < \infty$, where $\mathbb{E}$ denotes integration with respect to $\mathbb{P}$. We have $\mathbb{E}(Y_i | X_i) = g(X_i)$ almost surely for some function g which is measurable on $(S, \mathcal{S})$ (Section A3.2 of [40]). Our goal is to estimate g from the data (X_i, Y_i), $1 \le i \le n$. Since $\mathbb{E}(Y_i^2) < \infty$, it follows that $g \in L^2(P)$ by Jensen's inequality. We consider the quadratic loss with the corresponding risk function for an estimate $\hat{h}$ given by

$$\mathbb{E}(\hat{h}(X) - Y)^2,$$

where the pair (X, Y) has the same distribution as (X_1, Y_1) and is independent of (X_i, Y_i), $1 \le i \le n$. Furthermore, $\hat{h}$ is assumed measurable with respect to (X_i, Y_i), $1 \le i \le n$, and to attain values in an RKHS or in a collection of RKHSs.

We assume throughout that

$(g1)$ $\|g\|_\infty \le C$ for $C > 0$.

We also need to make an assumption on the behaviour of the errors of the response variables Y_i for $1 \le i \le n$. Let U and V be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. We say U is σ^2-sub-Gaussian if

$$\mathbb{E}(\exp(tU)) \le \exp(\sigma^2 t^2 / 2)$$

for all $t \in \mathbb{R}$. We say U is σ^2-sub-Gaussian given V if

$$\mathbb{E}(\exp(tU) | V) \le \exp(\sigma^2 t^2 / 2)$$

almost surely for all $t \in \mathbb{R}$. We assume

(Y) $Y_i - g(X_i)$ is σ^2-sub-Gaussian given X_i for $1 \le i \le n$.

3. Regression for a fixed RKHS

We continue by providing simultaneous bounds on our collection of non-adaptive estimators for a fixed RKHS. Our results in this section depend on how well the regression function g can be approximated by elements of an RKHS H with kernel k. We make the following assumptions.

(H) The RKHS H with kernel k has the following properties:

- The RKHS H is separable.
- The kernel k is bounded.
- The kernel k is a measurable function on $(S \times S, \mathcal{S} \otimes \mathcal{S})$.

We define

$$\|k\|_{\text{diag}} = \sup_{x \in S} k(x, x) < \infty.$$

We can guarantee that H is separable by, for example, assuming that k is continuous and S is a separable topological space (Lemma 4.33 of [37]). The fact that H has a kernel k which is measurable on $(S \times S, \mathcal{S} \otimes \mathcal{S})$ guarantees that all functions in H are measurable on $(S, \mathcal{S})$ (Lemma 4.24 of [37]).

Let B_H be the closed unit ball of H and $r > 0$. We define the estimator

$$\hat{h}_r = \arg\min_{f \in r B_H} \frac{1}{n} \sum_{i=1}^{n} \big(f(X_i) - Y_i\big)^2$$

of the regression function g. We make this definition unique by demanding that $\hat{h}_r \in \text{sp}\{k_{X_i} : 1 \le i \le n\}$ (see Lemma 3 of [32]). We also define $\hat{h}_0 = 0$. Lemma 11 on page 11 states that this estimator is well defined.

Since we assume $(g1)$, that g is bounded in $[-C, C]$, we can make $\hat{h}_r$ closer to g by constraining it to lie in the same interval. As in [32], we define the projection $V : \mathbb{R} \to [-C, C]$ by

$$V(t) = \begin{cases} -C & \text{if } t < -C, \\ t & \text{if } |t| \le C, \\ C & \text{if } t > C, \end{cases}$$

for $t \in \mathbb{R}$. For a function $f : S \to \mathbb{R}$ we write Vf for the composition of V and f.

We now prove a series of result which allow us to control $\hat{h}_r$ for $r \ge 0$ simultaneously, extending the results of [32] while using similar proof techniques. This is crucial in order to apply the Goldenshluger–Lepski method to these estimators. The results assign probabilities to events which occur for all $r \ge 0$ and all $h_r \in r B_H$. These events are measurable due to the separability of $[0, \infty)$ and $r B_H$, as well as the continuity in r of the quantities in question, including $\hat{h}_r$ by Lemma 11. By Lemma 2 of [32], we have

$$\|\hat{h}_r - h_r\|_{L^2(P_n)}^2 \le \frac{4}{n} \sum_{i=1}^{n} \big(Y_i - g(X_i)\big)\big(\hat{h}_r(X_i) - h_r(X_i)\big) + 4\|h_r - g\|_{L^2(P_n)}^2$$

for all $r > 0$ and all $h_r \in rB_H$. We can remove $\hat{h}_r$ in the first term on the right-hand side by taking a supremum over rB_H. After applying the reproducing kernel property and the Cauchy-Schwarz inequality, we obtain a quadratic form of sub-Gaussians which can be controlled using Lemma 36 of [32]. See Lemma 22 in [33] on page 1 for details.

It is useful to be able to transfer a bound on the squared $L^2(P_n)$ error of an estimator, including the result above, to a bound on the squared $L^2(P)$ error of the estimator. By using Talagrand's inequality, we can obtain a high-probability bound on

$$\sup_{r>0} \sup_{f_1,f_2 \in rB_H} \frac{1}{r} \big| \|Vf_1 - Vf_2\|_{L^2(P_n)}^2 - \|Vf_1 - Vf_2\|_{L^2(P)}^2 \big|$$

by proving an expectation bound on the same quantity. By using symmetrisation (Lemma 2.3.1 of [39]) and the contraction principle for Rademacher processes (Theorem 3.2.1 of [15]), we again obtain a quadratic form of sub-Gaussians, which in this case are Rademacher random variables. The result is stated in Lemma 24 in [33] on page 3.

To capture how well g can be approximated by elements of H, we define

$$I_\infty(g, r) = \inf\big\{ \|h_r - g\|_\infty^2 : h_r \in rB_H \big\}$$

for $r \geq 0$. We use this measure of approximation as it is compatible with the use of the bound

$$\|h_r - g\|_{L^2(P_n)}^2 \leq \|h_r - g\|_\infty^2$$

in the proof of Lemma 22. It is easy to show that $I_\infty(g, r)$ is continuous in r under Assumption (H).

We obtain a bound on the squared $L^2(P)$ error of $V\hat{h}_r$ by combining Lemmas 22 and 24. To state this result we need the following events

$$A_{1,t} = \left\{ \|\hat{h}_r - h_r\|_{L^2(P_n)}^2 \leq \frac{20\|k\|_{\text{diag}}^{1/2}\sigma r t^{1/2}}{n^{1/2}} + 4\|h_r - g\|_\infty^2, \text{ for all } r \geq 0 \right\},$$

$$A_{2,t} = \left\{ \sup_{f_1,f_2 \in rB_H} \big| \|Vf_1 - Vf_2\|_{L^2(P_n)}^2 - \|Vf_1 - Vf_2\|_{L^2(P)}^2 \big| \right.$$

$$\left. \leq \frac{97\|k\|_{\text{diag}}^{1/2}Crt^{1/2}}{n^{1/2}} + \frac{8\|k\|_{\text{diag}}^{1/2}Crt}{3n}, \text{ for all } r \geq 0 \right\}.$$

Theorem 1. *Assume* $(g1)$, (Y) *and* (H) *and let* $t \geq 1$. *On the set* $A_{1,t} \cap A_{2,t} \in \mathcal{F}$, *for which* $\mathbb{P}(A_{1,t} \cap A_{2,t}) \geq 1 - 2e^{-t}$, *we have*

$$\|V\hat{h}_r - g\|_{L^2(P)}^2 \leq \frac{2\|k\|_{\text{diag}}^{1/2}(97C + 20\sigma)rt^{1/2}}{n^{1/2}} + \frac{16\|k\|_{\text{diag}}^{1/2}Crt}{3n} + 10I_\infty(g, r)$$

simultaneously for all $r \geq 0$.

3.1. The Goldenshluger–Lepski method for a fixed RKHS

We now produce bounds on our adaptive estimator for a fixed RKHS. Lemma 12 (p. 2255), which is a simple consequence of Lemma 22, can be used to define the majorant of the non-adaptive estimators.

This motivates the following definition of the adaptive estimator used in the Goldenshluger–Lepski method.

Let $R \subseteq [0, \infty)$ be closed and non-empty. The Goldenshluger–Lepski method defines an adaptive estimator using

$$\hat{r} = \arg\min_{r \in R} \left(\sup_{s \in R, s \geq r} \left(\|\hat{h}_r - \hat{h}_s\|^2_{L^2(P_n)} - \frac{\tau(r+s)}{n^{1/2}} \right) + \frac{2(1+\nu)\tau r}{n^{1/2}} \right) \tag{1}$$

for tuning parameters $\tau, \nu > 0$. The supremum of pairwise comparisons can be viewed as a proxy for the unknown bias, while the other term is an inflated variance term. Note that the supremum is at least the value at r, so

$$\sup_{s \in R, s \geq r} \left(\|\hat{h}_r - \hat{h}_s\|^2_{L^2(P_n)} - \frac{\tau(r+s)}{n^{1/2}} \right) + \frac{2(1+\nu)\tau r}{n^{1/2}} \geq \frac{2\nu\tau r}{n^{1/2}}. \tag{2}$$

The role of the tuning parameter ν is simply to control this bound. In fact, the parameter ν only affects the constants and not the rate of convergence in Theorem 4 and Theorem 9. The parameter τ controls the probability with which our bound on the squared $L^2(P)$ error of $V\hat{h}_{\hat{r}}$ holds. Lemma 13 on page 2255 shows that the estimator $\hat{r}$ is well-defined by Equation (1).

It may be that $\hat{r}$ is not a random variable on $(\Omega, \mathcal{F})$ in some cases, but we assume

$(\hat{r})$ $\hat{r}$ is a well-defined random variable on $(\Omega, \mathcal{F})$

throughout. Later, we assume that R is finite, in which case $\hat{r}$ is certainly a random variable on $(\Omega, \mathcal{F})$. If $\hat{r}$ is a random variable on $(\Omega, \mathcal{F})$, then $\hat{h}_{\hat{r}}$ is a $(H, \mathcal{B}(H))$-valued measurable function on $(\Omega, \mathcal{F})$ by Lemma 11.

By Lemma 12, the supremum in the definition of $\hat{r}$ is at most $40 I_\infty(g, r)$ for an appropriate value of t. The definition of $\hat{r}$ then gives us control over the squared $L^2(P_n)$ norm of $\hat{h}_{\hat{r}} - \hat{h}_r$ when $\hat{r} \leq r$. When $\hat{r} \geq r$, we can control the squared $L^2(P_n)$ norm of $\hat{h}_{\hat{r}} - \hat{h}_r$ using Lemma 12. However, we must control a term of order $\hat{r}/n^{1/2}$ using (2) and the definition of $\hat{r}$. In both cases, this gives a bound on the squared $L^2(P_n)$ norm of $V\hat{h}_{\hat{r}} - V\hat{h}_r$. Extra terms appear when moving to a bound on the squared $L^2(P)$ norm of $V\hat{h}_{\hat{r}} - V\hat{h}_r$ using Lemma 24. However, these terms are very similar to the inflated variance term, and can be controlled in the same way. Applying

$$\|V\hat{h}_{\hat{r}} - g\|^2_{L^2(P)} \leq 2\|V\hat{h}_{\hat{r}} - V\hat{h}_r\|^2_{L^2(P)} + 2\|V\hat{h}_r - g\|^2_{L^2(P)}$$

leads to Theorem 14 (p. 2256). Combining it with Theorem 1 gives us the following main result.

Theorem 2. *Assume (g1), (Y), (H) and (κ). Let* $\tau \geq 80\|k\|^{1/2}_{\mathrm{diag}}\sigma$, $\nu > 0$ *and*

$$t = \left(\frac{\tau}{80\|k\|^{1/2}_{\mathrm{diag}}\sigma} \right)^2 \geq 1.$$

On the set $A_{1,t} \cap A_{2,t} \in \mathcal{F}$, *for which* $\mathbb{P}(A_{1,t} \cap A_{2,t}) \geq 1 - 2e^{-t}$, *we have*

$$\|V\hat{h}_{\hat{r}} - g\|^2_{L^2(P)} \leq \inf_{r \in R} \left((1 + D_1\tau n^{-1/2})(D_2\tau r n^{-1/2} + D_3 I_\infty(g, r)) \right)$$

for constants $D_1, D_2, D_3 > 0$ *not depending on* τ, r *or* n.

We can obtain rates of convergence for our estimator $V\hat{h}_{\hat{r}}$ if we make an assumption about how well g can be approximated by elements of H. Such assumptions are typically stated in terms of interpolation spaces between a Banach space $(Z, \|\cdot\|_Z)$ and a dense subspace $(V, \|\cdot\|_V)$ (see [5]). The K-functional of (Z, V) is

$$K(z, t) = \inf_{v \in V} \left(\|z - v\|_Z + t \|v\|_V \right)$$

for $z \in Z$ and $t > 0$. We define

$$\|z\|_{\beta, q} = \left(\int_0^\infty \left(t^{-\beta} K(z, t) \right)^q t^{-1} \, dt \right)^{1/q} \quad \text{and} \quad \|z\|_{\beta, \infty} = \sup_{t > 0} \left(t^{-\beta} K(z, t) \right)$$

for $z \in Z$, $\beta \in (0, 1)$ and $1 \leq q < \infty$. We then define the interpolation space $[Z, V]_{\beta, q}$ to be the set of $z \in Z$ such that $\|z\|_{\beta, q} < \infty$. The size of $[Z, V]_{\beta, q}$ decreases as β increases. The following result is Lemma 1 of [32], which is essentially Theorem 3.1 of [36].

Lemma 3. *Let $(Z, \|\cdot\|_Z)$ be a Banach space, $(V, \|\cdot\|_V)$ be a dense subspace of Z and $z \in [Z, V]_{\beta, \infty}$. We have*

$$\inf \left\{ \|v - z\|_Z : v \in V, \|v\|_V \leq r \right\} \leq \frac{\|z\|_{\beta, \infty}^{1/(1-\beta)}}{r^{\beta/(1-\beta)}}.$$

Results like this have a long history in inverse problem theory [4,22]. From the above, when H is dense in $C(S)$, we can define the interpolation spaces $[C(S), H]_{\beta, q}$. We set $q = \infty$ and work with the largest space of functions for a fixed $\beta \in (0, 1)$. We are then able to apply the approximation result in Lemma 3.

Let us assume

$(g2)$ $g \in [C(S), H]_{\beta, \infty}$ with norm at most B for $\beta \in (0, 1)$ and $B > 0$.

The assumption $(g2)$, together with Lemma 3, give

$$I_\infty(g, r) \leq \frac{B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}} \tag{3}$$

for $r > 0$. In order for us to apply Theorem 2 to this setting, we need to make an assumption on R. We assume either

$(R1)$ $R = [0, \infty)$

or

$(R2)$ $R = \{bi : 0 \leq i \leq I - 1\} \cup \{an^{1/2}\}$ and $\rho = an^{1/2}$ for $a, b > 0$ and $I = \lceil an^{1/2}/b \rceil$.

The assumption $(R1)$ is mainly of theoretical interest and would make it difficult to calculate $\hat{r}$ in practice. The estimator $\hat{r}$ can be computed under the assumption $(R2)$, since in this case R is finite. We obtain a high-probability bound on a fixed quantile of the squared $L^2(P)$ error of $V\hat{h}_{\hat{r}}$ of order $t^{1/2} n^{-\beta/(1+\beta)}$ with probability at least $1 - e^{-t}$ when τ is an appropriate multiple of $t^{1/2}$.

Corollary 4. *Assume $(g1)$, $(g2)$, (Y) and (H). Let $\tau \geq 80\|k\|_{\mathrm{diag}}^{1/2}\sigma$, $\nu > 0$ and*

$$t = \left(\frac{\tau}{80\|k\|_{\mathrm{diag}}^{1/2}\sigma} \right)^2 \geq 1.$$

Also assume $(R1)$ and $(\varkappa)$, or $(R2)$. On the set $A_{1,t} \cap A_{2,t} \in \mathcal{F}$, for which $\mathbb{P}(A_{1,t} \cap A_{2,t}) \geq 1 - 2e^{-t}$, we have

$$\|V\hat{h}_{\hat{r}} - g\|^2_{L^2(P)} \leq D_1 \tau n^{-\beta/(1+\beta)} + D_2 \tau^2 n^{-(1+3\beta)/(2(1+\beta))}$$

for constants $D_1, D_2 > 0$ not depending on n or τ.

From the high probability bound we obtain a bound on the expected squared $L^2(P)$ error.

Corollary 5. *Assume* $(g1), (g2), (Y), (H)$ *and either* $(R1)$ *and* $(\varkappa)$, *or* $(R2)$. *We have*

$$\mathbb{E}\big(\|V\hat{h}_{\hat{r}} - g\|^2_{L^2(P)}\big) \leq Dn^{-\beta/(1+\beta)}$$

for constant D not depending on n.

4. Regression for a collection of RKHSs

In this section, we again provide simultaneous bounds on our collection of non-adaptive estimators. Our results still depend on how well the regression function g can be approximated by elements of an RKHS. However, this RKHS now comes from a collection instead of being fixed. Let $\mathcal{K}$ be a set of kernels on $S \times S$. We make the following assumptions.

$(\mathcal{K}1)$ The covariate set S and the set of kernels $\mathcal{K}$ have the following properties:

- The covariate set S is a separable topological space.
- The set of kernels $(\mathcal{K}, \|\cdot\|_\infty)$ is separable.
- The kernel k is bounded for all $k \in \mathcal{K}$.
- The kernel k is continuous for all $k \in \mathcal{K}$.

Since $(\mathcal{K}, \|\cdot\|_\infty)$ is a separable set of kernels, we have that $\mathcal{K}$ has a countable dense subset $\mathcal{K}_0$. For all $\varepsilon > 0$ and all $k \in \mathcal{K}$, there exists $k_0 \in \mathcal{K}_0$ such that

$$\|k_0 - k\|_\infty = \sup_{x_1, x_2 \in S} \big|k_0(x_1, x_2) - k(x_1, x_2)\big| < \varepsilon.$$

Let H_k be the RKHS with kernel k for $k \in \mathcal{K}$. Since k is continuous and S is a separable topological space, we have that H_k is separable by Lemma 4.33 of [37]. Hence, the assumption (H) holds for H_k. We use the notation $\|\cdot\|_k$ and $\langle \cdot, \cdot \rangle_k$ for the norm and inner product of H_k.

Let B_k be the closed unit ball of H_k for $k \in \mathcal{K}$ and $r > 0$. We define the estimator

$$\hat{h}_{k,r} = \arg\min_{f \in rB_k} \frac{1}{n} \sum_{i=1}^n \big(f(X_i) - Y_i\big)^2$$

of the regression function g. We make this definition unique by demanding that $\hat{h}_{k,r} \in \text{sp}\{k_{X_i} : 1 \leq i \leq n\}$ (see Lemma 3 of [32]). We also define $\hat{h}_{k,0} = 0$. Since we assume $(g1)$, that g is bounded in $[-C, C]$, we can make $\hat{h}_{k,r}$ closer to g by clipping it to obtain $V\hat{h}_{k,r}$. Lemma 26 ([33], p. 6) shows that $\hat{h}_{k,r}$ is a well-defined estimator.

Let

$$\mathcal{L} = \big\{k/\|k\|_{\text{diag}} : k \in \mathcal{K}\big\} \cup \{0\}$$

and

$$D = \sup_{f_1, f_2 \in \mathcal{L}} \|f_1 - f_2\|_\infty \leq 2.$$

We include 0 in the definition of $\mathcal{L}$ so that, when analysing stochastic processes over $\mathcal{L}$ using chaining, we can start all chains at 0. Note that $(\mathcal{L}, \|\cdot\|_\infty)$ is separable since $\mathcal{L} \setminus \{0\}$ is the image of a continuous function on $(\mathcal{K}, \|\cdot\|_\infty)$, which is itself separable. Let $N(a, M, d)$ be the minimum size of an $a > 0$ cover of a metric space (M, d). The following quantity is key to quantifying the complexity of our collection of kernels,

$$J = \left(162 \int_0^{D/2} \log\big(2N\big(a, \mathcal{L}, \|\cdot\|_\infty\big)\big) \, da + 1 \right)^{1/2}.$$

Using J we upper bound the empirical L^2-distance between the estimator $\hat{h}_{k,r}$ and $h_{k,r}$ in Lemma 29 ([33], p. 9). Instead of one quadratic form of sub-Gaussians as in Lemma 22, we obtain a supremum over $\mathcal{K}$ of quadratic forms of sub-Gaussians. This can be controlled by chaining using Lemma 28 of [33].

It is again useful to be able to transfer a bound on the squared $L^2(P_n)$ error of an estimator to a bound on the squared $L^2(P)$ error of the estimator. Lemma 31 ([33], p. 11) allows us to do such a transfer. The result is proved using the same method as Lemma 24, although we obtain a supremum of quadratic forms of sub-Gaussians which are controlled using chaining. The event in the result is measurable by Lemma 30 of [33].

To capture how well g can be approximated by elements of H_k, we define

$$I_\infty(g, k, r) = \inf\big\{\|h_{k,r} - g\|_\infty^2 : h_{k,r} \in r B_k\big\}$$

for $k \in \mathcal{K}$ and $r \geq 0$. We obtain a bound on the squared $L^2(P)$ error of $V\hat{h}_{k,r}$ by combining Lemmas 29 and 31. In the following, let $A_{3,t} \in \mathcal{F}$ be an event such that $\mathbb{P}(A_{3,t}) \geq 1 - e^{-t}$ and on which

$$\|\hat{h}_{k,r} - h_{k,r}\|_{L^2(P_n)}^2 \leq \frac{21 J \|k\|_{\text{diag}}^{1/2} \sigma r t^{1/2}}{n^{1/2}} + 4\|h_{k,r} - g\|_\infty^2$$

simultaneously for all $k \in \mathcal{K}$, all $r \geq 0$ and all $h_{k,r} \in r B_k$. By Lemma 29 in [33] such an $A_{3,t}$ exists. Furthermore, let

$$A_{4,t} = \Bigg\{ \sup_{f_1, f_2 \in r B_k} \big| \|Vf_1 - Vf_2\|_{L^2(P_n)}^2 - \|Vf_1 - Vf_2\|_{L^2(P)}^2 \big|$$

$$\leq \frac{151 J \|k\|_{\text{diag}}^{1/2} C r t^{1/2}}{n^{1/2}} + \frac{8\|k\|_{\text{diag}}^{1/2} C r t}{3n}, \text{ for all } k \in \mathcal{K} \text{ and all } r \geq 0 \Bigg\}.$$

With these definition in place the following holds.

Theorem 6. *Assume* $(g1)$, (Y) *and* $(\mathcal{K}1)$. *Let* $t \geq 1$ *and recall the definitions of* $A_{3,t}$ *and* $A_{4,t}$ *from Lemmas 29 and 31. On the set* $A_{3,t} \cap A_{4,t} \in \mathcal{F}$, *for which* $\mathbb{P}(A_{3,t} \cap A_{4,t}) \geq 1 - 2e^{-t}$, *we have*

$$\|V\hat{h}_{k,r} - g\|_{L^2(P)}^2 \leq \frac{2 J \|k\|_{\text{diag}}^{1/2} (151C + 21\sigma) r t^{1/2}}{n^{1/2}} + \frac{16\|k\|_{\text{diag}}^{1/2} C r t}{3n} + 10 I_\infty(g, k, r)$$

simultaneously for all $k \in \mathcal{K}$ *and all* $r \geq 0$.

4.1. The Goldenshluger–Lepski method for a collection of RKHSs with Gaussian kernels

We now apply the Goldenshluger–Lepski method again in the context of RKHS regression. However, we now produce an estimator which adapts over a collection of RKHSs with Gaussian kernels. We make the following assumptions on S and $\mathcal{K}$.

($\mathcal{K}2$) The covariate set S and the set of kernels $\mathcal{K}$ have the following properties:

- The covariate set $S \subseteq \mathbb{R}^d$ for $d \geq 1$.
- The set of kernels
$$\mathcal{K} = \left\{ k_\gamma(x_1, x_2) = \gamma^{-d} \exp\left(-\|x_1 - x_2\|_2^2/\gamma^2\right) : \gamma \in \Gamma \text{ and } x_1, x_2 \in S \right\}$$
for $\Gamma \subseteq [u, v]$ non-empty for $v \geq u > 0$.

Recalling the definitions from the previous section, we have
$$\mathcal{L} = \left\{ f_\gamma(x_1, x_2) = \exp\left(-\|x_1 - x_2\|_2^2/\gamma^2\right) : \gamma \in \Gamma \text{ and } x_1, x_2 \in S \right\} \cup \{0\}.$$

The assumption ($\mathcal{K}2$) implies the assumption ($\mathcal{K}1$). This is because Lemma 33 of [33] shows that $(\mathcal{L}, \|\cdot\|_\infty)$, and hence $(\mathcal{K}, \|\cdot\|_\infty)$, is separable. We change notation slightly. Let H_γ be the RKHS with kernel k_γ for $\gamma \in \Gamma$, let $\|\cdot\|_\gamma$ and $\langle \cdot, \cdot \rangle_\gamma$ be the norm and inner product of H_γ, and let B_γ be the closed unit ball of H_γ. Furthermore, we write $\hat{h}_{\gamma,r}$ in place of $\hat{h}_{k_\gamma,r}$ and $I_\infty(g, \gamma, r)$ in place of $I_\infty(g, k_\gamma, r)$.

The scaling of the kernels is selected so that the following lemma holds. The result is immediate from Proposition 4.46 of [37] and the way that the norm of an RKHS scales with its kernel (Theorem 4.21 of [37]).

Lemma 7. *Assume* ($\mathcal{K}2$). *Let* $\gamma, \eta \in \Gamma$ *with* $\gamma \geq \eta$. *We have* $B_\gamma \subseteq B_\eta$.

By Lemma 33 of [33], the function $F : \Gamma \to \mathcal{L} \setminus \{0\}$ by $F(\gamma) = f_\gamma$ is continuous. Hence, the function $G : \Gamma \to \mathcal{K}$ by $G(\gamma) = k_\gamma$ is continuous. Lemma 34 ([33], p. 15), which follows directly from Lemma 26, now states that the estimator is well-defined.

Recall the definition of J from the previous section. Lemma 36 ([33], p. 15) provides us with a bound on J. Furthermore, Lemma 15 on page 2259 can be used to define the majorant of the non-adaptive estimators and is a simple consequence of Lemma 29. This motivates the definition of the adaptive estimator used in the Goldenshluger–Lepski method.

Let $R \subseteq [0, \infty)$ be non-empty. The Goldenshluger–Lepski method creates an adaptive estimator by defining $(\hat{\gamma}, \hat{r})$ to be the minimiser of

$$\sup_{\eta \in \Gamma, \eta \leq \gamma} \sup_{s \in R, s \geq r} \left(\|\hat{h}_{\gamma,r} - \hat{h}_{\eta,s}\|_{L_2(P_n)}^2 - \frac{\tau(\gamma^{-d/2}r + \eta^{-d/2}s)}{n^{1/2}} \right) + \frac{2(1+v)\tau\gamma^{-d/2}r}{n^{1/2}} \quad (4)$$

over $(\gamma, r) \in \Gamma \times R$ for tuning parameters $\tau, v > 0$. Again, the supremum of pairwise comparisons can be viewed as a proxy for the unknown bias, while the other term is an inflated variance term. Note that the supremum is at least the value at (γ, r), which means the above is at least

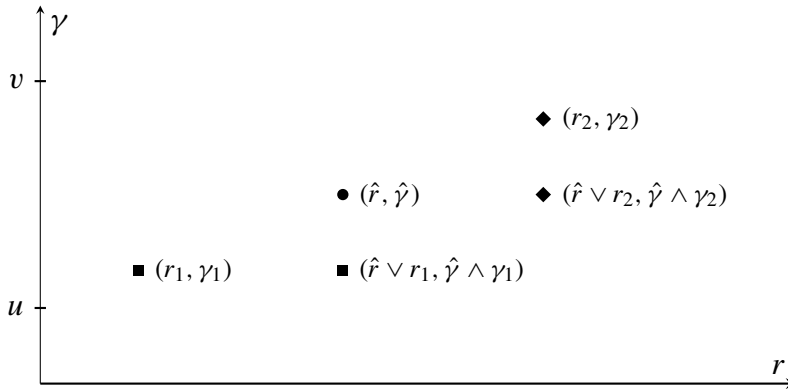$$\frac{2v\tau\gamma^{-d/2}r}{n^{1/2}}. \quad (5)$$

Figure 1. A demonstration of the parameter comparisons made in the proof of Theorem 16.

Again, the role of the tuning parameter v is simply to control this bound. The parameter τ controls the probability with which our bound on the squared $L^2(P)$ error of $V\hat{h}_{\hat{\gamma},\hat{r}}$ holds. It may be that $\hat{\gamma}$ is not a well-defined random variable on $(\Omega, \mathcal{F})$ in some cases, but we assume

   ($\hat{\gamma}$) $\hat{\gamma}$ is a well-defined random variable on $(\Omega, \mathcal{F})$

throughout. Later, we assume that R and Γ are finite, in which case $\hat{\gamma}$ and $\hat{r}$ are certainly well-defined random variables on $(\Omega, \mathcal{F})$. If $\hat{\gamma}$ and $\hat{r}$ are well-defined random variables on $(\Omega, \mathcal{F})$, then $\hat{h}_{\hat{\gamma},\hat{r}}$ is an $(C(S), \mathcal{B}(C(S)))$-valued measurable function on $(\Omega, \mathcal{F})$ by Lemma 34.

By Lemma 15, the supremum in the definition of $(\hat{\gamma}, \hat{r})$ is at most $40I_\infty(g, \gamma, r)$ for an appropriate value of t. The definition of $(\hat{\gamma}, \hat{r})$ then gives us control over the squared $L^2(P_n)$ norm of $\hat{h}_{\hat{\gamma},\hat{r}} - \hat{h}_{\hat{\gamma} \wedge \gamma, \hat{r} \vee r}$. We can control the squared $L^2(P_n)$ norm of $\hat{h}_{\hat{\gamma} \wedge \gamma, \hat{r} \vee r} - h_{\gamma,r}$ using Lemma 15. In both cases, we use the boundedness of Γ when controlling the squared $L^2(P_n)$ norm before clipping the estimators using V. Extra terms appear when moving from bounds on the squared $L^2(P_n)$ norm to bounds on the squared $L^2(P)$ norm using Lemma 31. We must then control terms of order $\hat{\gamma}^{-d/2}\hat{r}/n^{1/2}$ using (5) and the definition of $(\hat{\gamma}, \hat{r})$. Combining the bounds gives a bound on the squared $L^2(P)$ norm of $Vh_{\hat{\gamma},\hat{r}} - Vh_{\gamma,r}$. Applying

$$\|V\hat{h}_{\hat{r}} - g\|^2_{L^2(P)} \leq 2\|Vh_{\hat{\gamma},\hat{r}} - Vh_{\gamma,r}\|^2_{L^2(P)} + 2\|Vh_{\gamma,r} - g\|^2_{L^2(P)}$$

leads to Theorem 16 on page 2260. Comparisons between $(\hat{r}, \hat{\gamma})$, (r, γ) and $(\hat{\gamma} \wedge \gamma, \hat{r} \vee r)$ are demonstrated in Figure 1 for two different values of (r, γ).

We now combine Theorems 6 and 16.

Theorem 8. *Assume (g1), (Y) and (K2). Let $\tau \geq 84J\sigma$, $v > 0$ and*

$$t = \left(\frac{\tau}{84J\sigma}\right)^2 \geq 1.$$

On the set $A_{3,t} \cap A_{4,t} \in \mathcal{F}$, for which $\mathbb{P}(A_{3,t} \cap A_{4,t}) \geq 1 - 2e^{-t}$, we have

$$\|V\hat{h}_{\gamma,\hat{r}} - g\|^2_{L^2(P)} \leq \inf_{\gamma \in \Gamma} \inf_{r \in R}\left((1 + D_1\tau n^{-1/2})\left(D_2\tau\gamma^{-d/2}rn^{-1/2} + D_3I_\infty(g, \gamma, r)\right)\right)$$

for constants $D_1, D_2, D_3 > 0$ not depending on τ, γ, r or n.

We can obtain rates of convergence for our estimator $V\hat{h}_{\hat{\gamma},\hat{r}}$ if we make an assumption about how well g can be approximated by elements of H_α for $\alpha \in [u, v]$. Let us assume

(g3) $g \in [C(S), H_\alpha]_{\beta,\infty}$ with norm at most B for $\alpha \in [u, v]$, $\beta \in (0, 1)$ and $B > 0$.

The assumption (g3), together with Lemma 3, give

$$I_\infty(g, \alpha, r) \leq \frac{B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}} \tag{6}$$

for $r > 0$. In order for us to apply Theorem 8 to this setting, we need to make assumptions on Γ and R. We assume either (R1) and

(Γ1) $\Gamma = [u, v]$,

or (R2) and

(Γ2) $\Gamma = \{uc^i : 0 \leq i \leq L - 1\} \cup \{v\}$ for $c > 1$ and $L = \lceil \log(v/u)/\log(c) \rceil$.

The assumptions (R1) and (Γ1) are mainly of theoretical interest and would make it difficult to calculate $(\hat{\gamma}, \hat{r})$ in practice. The estimator $(\hat{\gamma}, \hat{r})$ can be computed under the assumptions (R2) and (Γ2), since in this case R and Γ are finite. We obtain a high-probability bound on a fixed quantile of the squared $L^2(P)$ error of $V\hat{h}_{\hat{r},\hat{\gamma}}$ of order $t^{1/2}n^{-\beta/(1+\beta)}$ with probability at least $1 - e^{-t}$ when τ is an appropriate multiple of $t^{1/2}$.

Corollary 9. *Assume* (g1)*,* (g3)*,* (Y) *and* (K2)*. Let* $\tau \geq 84J\sigma$*,* $v > 0$ *and*

$$t = \left(\frac{\tau}{84J\sigma}\right)^2 \geq 1.$$

Also assume (R1)*,* (Γ1)*,* (κ) *and* (ϰ)*, or* (R2) *and* (Γ2)*. On the set* $A_{3,t} \cap A_{4,t} \in \mathcal{F}$*, for which* $\mathbb{P}(A_{3,t} \cap A_{4,t}) \geq 1 - 2e^{-t}$*, we have*

$$\|V\hat{h}_{\hat{\gamma},\hat{r}} - g\|_{L^2(P)}^2 \leq D_1\tau n^{-\beta/(1+\beta)} + D_2\tau^2 n^{-(1+3\beta)/(2(1+\beta))}$$

for constants $D_1, D_2 > 0$ *not depending on* n *or* τ*.*

From the high probability bound we obtain a bound on the expected squared $L^2(P)$ error similarly to Corollary 5.

Corollary 10. *Assume* (g1)*,* (g3)*,* (Y)*,* (K2) *and either* (R1)*,* (Γ1)*,* (κ) *and* (ϰ)*, or* (R2) *and* (Γ2)*. We have*

$$\mathbb{E}\big(\|V\hat{h}_{\hat{\gamma},\hat{r}} - g\|_{L^2(P)}^2\big) \leq Dn^{-\beta/(1+\beta)}$$

for constant D *not depending on* n*.*

5. Discussion

In this paper, we show how the Goldenshluger–Lepski method can be applied when performing regression over an RKHS H, which is separable with a bounded and measurable kernel k, or a collection of such RKHSs. We produce an adaptive estimator from a collection of clipped versions of least-squares

estimators which are constrained to lie in a ball of predefined radius in H. Since the $L^2(P)$ norm is unknown, we use the $L^2(P_n)$ norm when calculating the pairwise comparisons for the proxy for the unknown bias of this collection of non-adaptive estimators. When H is fixed, our estimator need only adapt to the radius of the ball in H. However, when H comes from a collection of RKHSs with Gaussian kernels, the estimator must also adapt to the width parameter of the kernel. As far as we are aware, this is the first time that the Goldenshluger–Lepski method has been applied in the context of RKHS regression. In order to apply the Goldenshluger–Lepski method in this context, we must provide a majorant by controlling all of the non-adaptive estimators simultaneously, extending the results of [32].

By assuming that the regression function lies in an interpolation space between $C(S)$ and H parametrised by β, we obtain a bound on a fixed quantile of the squared $L^2(P)$ error of our adaptive estimator of order $n^{-\beta/(1+\beta)}$. This is true for both the case in which H is fixed and the case in which H comes from a collection of RKHSs with Gaussian kernels. The order $n^{-\beta/(1+\beta)}$ for the squared $L^2(P)$ error of the adaptive estimators matches the order of the smallest bounds obtained in [32] for the squared $L^2(P)$ error of the non-adaptive estimators. In the sense discussed in [32], this order is the optimal power of n if we make the slightly weaker assumption that the regression function is an element of the interpolation space between $L^2(P)$ and H parametrised by β.

For the case in which H comes from a collection of RKHSs with Gaussian kernels, our current results rely on the boundedness of the set Γ of width parameters of the kernels. This is somewhat limiting as allowing the width parameter to tend to 0 as n tends to infinity would allow us to estimate a greater collection of functions. We hope that in the future the analysis in the proof of Theorem 16 can be extended to allow for such flexibility.

The results in this paper warrant the investigation of whether it is possible to extend the use of the Goldenshluger–Lepski method from the case in which H comes from a collection of RKHSs with Gaussian kernels to other cases. The analysis in this paper relies on the fact that the closed unit ball of the RKHS generated by a Gaussian kernel increases as the width of the kernel decreases. It may be possible to apply a similar analysis to other situations in which H belongs to a collection of RKHSs which also exhibit this nestedness property. If the RKHSs did not exhibit this property, then a new form of analysis would be necessary to apply the Goldenshluger–Lepski method. In particular, we would need a new criterion for deciding on the smoothness of the non-adaptive estimators when performing the pairwise comparisons.

Appendix A: Proof of the Goldenshluger–Lepski method for a fixed RKHS

This section is composed of two subsections: in Section A.1 we derive a number of auxiliary results and in Section A.2 we provide proofs for our main results.

A.1. Auxiliary results

The following combines parts of Lemmas 3 and 32 of [32].

Lemma 11. *Assume* (H). *We have that* $\hat{h}_r$ *is a* $(H, \mathcal{B}(H))$-*valued measurable function on* $(\Omega \times [0, \infty), \mathcal{F} \otimes \mathcal{B}([0, \infty)))$, *where* r *varies in* $[0, \infty)$. *Furthermore,* $\|\hat{h}_r - \hat{h}_s\|_H^2 \leq |r^2 - s^2|$ *for* $r, s \in [0, \infty)$.

We bound the distance between $\hat{h}_r$ and $\hat{h}_s$ in the $L^2(P_n)$ norm for $s \geq r \geq 0$ to prove the following Lemma.

Lemma 12. *Assume* (Y) *and* (H). *Let* $t \geq 1$ *and recall the definition of* $A_{1,t}$. *On the set* $A_{1,t} \in \mathcal{F}$, *for which* $\mathbb{P}(A_{1,t}) \geq 1 - e^{-t}$, *we have*

$$\|\hat{h}_r - \hat{h}_s\|^2_{L^2(P_n)} \leq \frac{80\|k\|^{1/2}_{\text{diag}}\sigma(r+s)t^{1/2}}{n^{1/2}} + 40I_\infty(g,r)$$

simultaneously for all $s \geq r \geq 0$.

Proof. By Lemma 22, we have

$$\|\hat{h}_r - \hat{h}_s\|^2_{L^2(P_n)} \leq 4\|\hat{h}_r - h_r\|^2_{L^2(P_n)} + 4\|h_r - g\|^2_{L^2(P_n)}$$

$$+ 4\|g - h_s\|^2_{L^2(P_n)} + 4\|h_s - \hat{h}_s\|^2_{L^2(P_n)}$$

$$\leq \frac{80\|k\|^{1/2}_{\text{diag}}\sigma(r+s)t^{1/2}}{n^{1/2}} + 20\|h_r - g\|^2_\infty + 20\|h_s - g\|^2_\infty$$

for all $r, s \geq 0$ and all $h_r \in rB_H$, $h_s \in sB_H$. Taking an infimum over $h_r \in rB_H$ and $h_s \in sB_H$ gives

$$\|\hat{h}_r - \hat{h}_s\|^2_{L^2(P_n)} \leq \frac{80\|k\|^{1/2}_{\text{diag}}\sigma(r+s)t^{1/2}}{n^{1/2}} + 20I_\infty(g,r) + 20I_\infty(g,s).$$

The result follows. $\square$

Next, we show that the estimator $\hat{r}$ is well-defined.

Lemma 13. *Let* $\hat{r}$ *be the infimum of all points attaining the minimum in* (1). *Then* $\hat{r}$ *is well-defined.*

Proof. Let K be the $n \times n$ symmetric matrix with $K_{i,j} = k(X_i, X_j)$. By Lemma 3 of [32], we have that K is an $(\mathbb{R}^{n\times n}, \mathcal{B}(\mathbb{R}^{n\times n}))$-valued measurable matrix on $(\Omega, \mathcal{F})$ and that there exist an orthogonal matrix A and a diagonal matrix D which are both $(\mathbb{R}^{n\times n}, \mathcal{B}(\mathbb{R}^{n\times n}))$-valued measurable matrices on $(\Omega, \mathcal{F})$ such that $K = ADA^\mathsf{T}$. Furthermore, we can demand that the diagonal entries of D are non-negative and non-increasing. Let $m = \text{rk}\,K$ and

$$\rho = \left(\sum_{i=1}^m D_{i,i}^{-1}\left(A^\mathsf{T}Y\right)_i^2\right)^{1/2},$$

which are random variables on $(\Omega, \mathcal{F})$. By Lemma 3 of [32], we have that $\hat{h}_r$ is constant in r for $r \geq \rho$. Hence,

$$\inf_{r \in R}\left(\sup_{s \in R, s \geq r}\left(\|\hat{h}_r - \hat{h}_s\|^2_{L^2(P_n)} - \frac{\tau(r+s)}{n^{1/2}}\right) + \frac{2(1+v)\tau r}{n^{1/2}}\right)$$

$$= \inf_{r \in R \cap [0,\rho]}\left(\sup_{s \in R, s \geq r}\left(\|\hat{h}_r - \hat{h}_s\|^2_{L^2(P_n)} - \frac{\tau(r+s)}{n^{1/2}}\right) + \frac{2(1+v)\tau r}{n^{1/2}}\right). \quad (7)$$

By Lemma 11, we have

$$\|\hat{h}_r - \hat{h}_s\|^2_{L^2(P_n)} - \frac{\tau(r+s)}{n^{1/2}}$$

is continuous in r for all $s \in R$ such that $s \geq r$. The supremum of a collection of lower semi-continuous functions is lower semi-continuous. Therefore,

$$\sup_{s \in R, s \geq r} \left(\|\hat{h}_r - \hat{h}_s\|^2_{L^2(P_n)} - \frac{\tau(r+s)}{n^{1/2}} \right) + \frac{2(1+v)\tau r}{n^{1/2}}$$

is lower semi-continuous in r. Hence, the infimum (7) is attained as it is the infimum of a lower semi-continuous function on a compact set. By lower semi-continuity, $\hat{r}$ also attains the infimum and is well-defined.   □

A.2. Main results

We use the Goldenshluger–Lepski method to prove the following theorem.

Theorem 14. *Assume* (Y), (H) *and* (κ). *Let* $\tau \geq 80\|k\|^{1/2}_{\text{diag}}\sigma$, $v > 0$ *and*

$$t = \left(\frac{\tau}{80\|k\|^{1/2}_{\text{diag}}\sigma} \right)^2 \geq 1.$$

Recall the definitions of $A_{1,t}$ *and* $A_{2,t}$. *On the set* $A_{1,t} \cap A_{2,t} \in \mathcal{F}$, *for which* $\mathbb{P}(A_{1,t} \cap A_{2,t}) \geq 1 - 2e^{-t}$, *we have*

$$\|V\hat{h}_{\hat{r}} - g\|^2_{L^2(P)}$$

is at most

$$\inf_{r \in R} \left(\max\left\{ \frac{2\tau r}{n^{1/2}} + \left(\frac{1}{v} + \frac{97C}{80\sigma v} + \frac{C\tau}{2400\|k\|^{1/2}_{\text{diag}}\sigma^2 v n^{1/2}} \right) \left(40I_\infty(g,r) + \frac{2(1+v)\tau r}{n^{1/2}} \right), \right.$$

$$\left. \frac{4(2+v)\tau r}{n^{1/2}} + \frac{97C\tau r}{40\sigma n^{1/2}} + \frac{C\tau^2 r}{1200\|k\|^{1/2}_{\text{diag}}\sigma^2 n} \right\} + 80I_\infty(g,r) + 2\|V\hat{h}_r - g\|^2_{L^2(P)} \right).$$

Proof. Since we assume (Y) and (H), we find that Lemma 22 holds, which implies that Lemma 12 holds. By our choice of t, we have

$$\|\hat{h}_r - \hat{h}_s\|^2_{L^2(P_n)} \leq \frac{\tau(r+s)}{n^{1/2}} + 40I_\infty(g,r) \tag{8}$$

simultaneously for all $s, r \in R$ such that $s \geq r \geq 0$. Fix $r \in R$ and suppose that $\hat{r} \leq r$. By the definition of $\hat{r}$ in (1) and (8), we have

$$\|\hat{h}_{\hat{r}} - \hat{h}_r\|^2_{L^2(P_n)} = \|\hat{h}_{\hat{r}} - \hat{h}_r\|^2_{L^2(P_n)} - \frac{\tau(\hat{r}+r)}{n^{1/2}} + \frac{\tau(\hat{r}+r)}{n^{1/2}}$$

$$\leq \sup_{s \in R, s \geq \hat{r}} \left(\|\hat{h}_{\hat{r}} - \hat{h}_s\|^2_{L^2(P_n)} - \frac{\tau(\hat{r}+s)}{n^{1/2}} \right) + \frac{2\tau r}{n^{1/2}}$$

$$\leq \sup_{s \in R, s \geq r} \left(\|\hat{h}_r - \hat{h}_s\|^2_{L^2(P_n)} - \frac{\tau(r+s)}{n^{1/2}} \right) + \frac{2(2+\nu)\tau r}{n^{1/2}} - \frac{2(1+\nu)\tau \hat{r}}{n^{1/2}}$$

$$\leq 40 I_\infty(g, r) + \frac{2(2+\nu)\tau r}{n^{1/2}}.$$

This shows

$$\|V\hat{h}_{\hat{r}} - V\hat{h}_r\|^2_{L^2(P_n)} \leq 40 I_\infty(g, r) + \frac{2(2+\nu)\tau r}{n^{1/2}},$$

and it follows from Lemma 24 and our choice of t that

$$\|V\hat{h}_{\hat{r}} - V\hat{h}_r\|^2_{L^2(P)} \leq 40 I_\infty(g, r) + \frac{2(2+\nu)\tau r}{n^{1/2}} + \frac{97 C \tau r}{80 \sigma n^{1/2}} + \frac{C \tau^2 r}{2400 \|k\|^{1/2}_{\text{diag}} \sigma^2 n}.$$

Hence,

$$\|V\hat{h}_{\hat{r}} - g\|^2_{L^2(P)}$$

$$\leq 2\|V\hat{h}_{\hat{r}} - V\hat{h}_r\|^2_{L^2(P)} + 2\|V\hat{h}_r - g\|^2_{L^2(P)}$$

$$\leq 80 I_\infty(g, r) + \frac{4(2+\nu)\tau r}{n^{1/2}} + \frac{97 C \tau r}{40 \sigma n^{1/2}} + \frac{C \tau^2 r}{1200 \|k\|^{1/2}_{\text{diag}} \sigma^2 n} + 2\|V\hat{h}_r - g\|^2_{L^2(P)}.$$

Now suppose instead that $\hat{r} \geq r$. Since (8) holds simultaneously for all $s, r \in R$ such that $s \geq r \geq 0$, we have

$$\|\hat{h}_{\hat{r}} - \hat{h}_r\|^2_{L^2(P_n)} \leq \frac{\tau(r + \hat{r})}{n^{1/2}} + 40 I_\infty(g, r).$$

This shows

$$\|V\hat{h}_{\hat{r}} - V\hat{h}_r\|^2_{L^2(P_n)} \leq \frac{\tau r}{n^{1/2}} + 40 I_\infty(g, r) + \frac{\tau \hat{r}}{n^{1/2}},$$

and it follows from Lemma 24 that

$$\|V\hat{h}_{\hat{r}} - V\hat{h}_r\|^2_{L^2(P)}$$

$$\leq \frac{\tau r}{n^{1/2}} + 40 I_\infty(g, r) + \frac{\tau \hat{r}}{n^{1/2}} + \frac{97 C \tau \hat{r}}{80 \sigma n^{1/2}} + \frac{C \tau^2 \hat{r}}{2400 \|k\|^{1/2}_{\text{diag}} \sigma^2 n}$$

$$= \frac{\tau r}{n^{1/2}} + 40 I_\infty(g, r) + \left(\frac{1}{2\nu} + \frac{97 C}{160 \sigma \nu} + \frac{C \tau}{4800 \|k\|^{1/2}_{\text{diag}} \sigma^2 \nu n^{1/2}} \right) \frac{2\nu \tau \hat{r}}{n^{1/2}}.$$

By (2), the definition of $\hat{r}$ in (1) and (8), we have

$$\frac{2\nu \tau \hat{r}}{n^{1/2}} \leq \sup_{s \in R, s \geq \hat{r}} \left(\|\hat{h}_{\hat{r}} - \hat{h}_s\|^2_{L^2(P_n)} - \frac{\tau(\hat{r} + s)}{n^{1/2}} \right) + \frac{2(1+\nu)\tau \hat{r}}{n^{1/2}}$$

$$\leq \sup_{s \in R, s \geq r} \left(\|\hat{h}_r - \hat{h}_s\|^2_{L^2(P_n)} - \frac{\tau(r + s)}{n^{1/2}} \right) + \frac{2(1+\nu)\tau r}{n^{1/2}}$$

$$\leq 40 I_\infty(g, r) + \frac{2(1+\nu)\tau r}{n^{1/2}}.$$

Hence,

$$\|V\hat{h}_{\hat{r}} - g\|_{L^2(P)}^2$$

$$\leq 2\|V\hat{h}_{\hat{r}} - V\hat{h}_r\|_{L^2(P)}^2 + 2\|V\hat{h}_r - g\|_{L^2(P)}^2$$

$$\leq \frac{2\tau r}{n^{1/2}} + 80I_\infty(g,r) + \left(\frac{1}{\nu} + \frac{97C}{80\sigma\nu} + \frac{C\tau}{2400\|k\|_{\text{diag}}^{1/2}\sigma^2\nu n^{1/2}}\right)\left(40I_\infty(g,r) + \frac{2(1+\nu)\tau r}{n^{1/2}}\right)$$

$$+ 2\|V\hat{h}_r - g\|_{L^2(P)}^2.$$

The result follows. $\qquad\square$

We assume $(g1)$ to bound the distance between $V\hat{h}_{\hat{r}}$ and g in the $L^2(P)$ norm and prove Theorem 2.

Proof of Theorem 2. By Theorem 14, we have

$$\|V\hat{h}_{\hat{r}} - g\|_{L^2(P)}^2 \leq \inf_{r\in R}\left((1 + D_4\tau n^{-1/2})(D_5\tau rn^{-1/2} + D_6I_\infty(g,r)) + 2\|V\hat{h}_r - g\|_{L^2(P)}^2\right)$$

for some constants $D_4, D_5, D_6 > 0$ not depending on τ, r or n. By Theorem 1, we have

$$\|V\hat{h}_r - g\|_{L^2(P)}^2 \leq \frac{(97C + 20\sigma)\tau r}{40\sigma n^{1/2}} + \frac{C\tau^2 r}{1200\|k\|_{\text{diag}}^{1/2}\sigma^2 n} + 10I_\infty(g,r)$$

$$\leq D_7\tau rn^{-1/2} + D_8\tau^2 rn^{-1} + 10I_\infty(g,r).$$

for all $r \in R$, for some constants $D_7, D_8 > 0$ not depending on τ, r or n. This gives

$$\|V\hat{h}_{\hat{r}} - g\|_{L^2(P)}^2 \leq \inf_{r\in R}\left((1 + D_4\tau n^{-1/2})(D_5\tau rn^{-1/2} + D_6I_\infty(g,r))\right.$$

$$\left. + 2D_7\tau rn^{-1/2} + 2D_8\tau^2 rn^{-1} + 20I_\infty(g,r)\right).$$

Hence, the result follows with

$$D_1 = \frac{D_4 D_5 + 2D_8}{D_5 + 2D_7}, \qquad D_2 = D_5 + 2D_7, \qquad D_3 = D_6 + 20. \qquad\square$$

We assume $(g2)$ to prove Theorem 4.

Proof of Corollary 4. If we assume $(R1)$, then $r = an^{(1-\beta)/(2(1+\beta))} \in R$ and

$$\|V\hat{h}_{\hat{r}} - g\|_{L^2(P)}^2 \leq (1 + D_3\tau n^{-1/2})(D_4\tau rn^{-1/2} + D_5I_\infty(g,r))$$

$$\leq (1 + D_3\tau n^{-1/2})\left(D_4\tau an^{-\beta/(1+\beta)} + \frac{D_5 B^{2/(1-\beta)}}{a^{2\beta/(1-\beta)}n^{\beta/(1+\beta)}}\right)$$

for some constants $D_3, D_4, D_5 > 0$ not depending on n or τ by Theorem 2 and (3). If we assume $(R2)$, then there is a unique $r \in R$ such that

$$an^{(1-\beta)/(2(1+\beta))} \leq r < an^{(1-\beta)/(2(1+\beta))} + b$$

and

$$\|V\hat{h}_{\hat{r}} - g\|_{L^2(P)}^2$$
$$\leq (1 + D_3\tau n^{-1/2})(D_4\tau rn^{-1/2} + D_5 I_\infty(g,r))$$
$$\leq (1 + D_3\tau n^{-1/2})\left(D_4\tau(an^{(1-\beta)/(2(1+\beta))} + b)n^{-1/2} + \frac{D_5 B^{2/(1-\beta)}}{a^{2\beta/(1-\beta)}n^{\beta/(1+\beta)}}\right)$$

by Theorem 2 and (3). In either case,

$$\|V\hat{h}_{\hat{r}} - g\|_{L^2(P)}^2 \leq D_1\tau n^{-\beta/(1+\beta)} + D_2\tau^2 n^{-(1+3\beta)/(2(1+\beta))}$$

for some constants $D_1, D_2 > 0$ not depending on n or τ. □

Proof of Corollary 5. Under the stated assumptions, we can apply Theorem 2. Let $D_1' = 80\|k\|_{\text{diag}}^{1/2} \times \sigma D_1 n^{-\beta/(1+\beta)}$ and $D_2' = 6400\|k\|_{\text{diag}} D_2 \sigma^2 n^{-(1+3\beta)/(2(1+\beta))}$ with constants D_1, D_2 from Theorem 2. Furthermore, let $\phi(t) = D_1' t^{1/2} + D_2' t$ then

$$\mathbb{E}(\|V\hat{h}_{\hat{r}} - g\|_{L^2(P)}^2)$$
$$= \int_0^\infty \mathbb{P}(\|V\hat{h}_{\hat{r}} - g\|_{L^2(P)}^2 \geq a)\,da$$
$$= \int_0^\infty \mathbb{P}(\|V\hat{h}_{\hat{r}} - g\|_{L^2(P)}^2 \geq \phi(t))(D_1' t^{1/2} + D_2')\,dt$$
$$\leq (\pi^{1/2} + 1)D_1' + 3D_2',$$

using the bound on $\mathbb{P}(\|V\hat{h}_{\hat{r}} - g\|_{L^2(P)}^2 \geq \phi(t))$ for $t \geq 1$. The result follows by letting $D = 80(\pi^{1/2} + 1)\|k\|_{\text{diag}}^{1/2}\sigma D_1 + 19200\|k\|_{\text{diag}} D_2 \sigma^2$. □

Appendix B: Proof of the Goldenshluger–Lepski method for a collection of RKHSs with Gaussian kernels

This section is composed of two subsections, one containing auxiliary results (Section B.1) and one containing the proofs of our main results (Section B.2).

B.1. Auxiliary results

We bound the distance between $\hat{h}_{\gamma,r}$ and $\hat{h}_{\eta,s}$ in the $L^2(P_n)$ norm for $\gamma, \eta \in \Gamma$ with $\eta \leq \gamma$ and $s \geq r \geq 0$ to prove the following lemma.

Lemma 15. *Assume (Y) and (K2). Let $t \geq 1$ and recall the definition of $A_{3,t}$. On the set $A_{3,t} \in \mathcal{F}$, for which $\mathbb{P}(A_{3,t}) \geq 1 - e^{-t}$, we have*

$$\|\hat{h}_{\gamma,r} - \hat{h}_{\eta,s}\|_{L^2(P_n)}^2 \leq \frac{84 J\sigma(\gamma^{-d/2}r + \eta^{-d/2}s)t^{1/2}}{n^{1/2}} + 40 I_\infty(g,\gamma,r)$$

simultaneously for all $\gamma, \eta \in \Gamma$ such that $\eta \leq \gamma$ and all $s \geq r \geq 0$.

Proof. By Lemma 29, we have

$$\|\hat{h}_{\gamma,r} - \hat{h}_{\eta,s}\|^2_{L^2(P_n)} \le 4\|\hat{h}_{\gamma,r} - h_{\gamma,r}\|^2_{L^2(P_n)} + 4\|h_{\gamma,r} - g\|^2_{L^2(P_n)}$$

$$+ 4\|g - h_{\eta,s}\|^2_{L^2(P_n)} + 4\|h_{\eta,s} - \hat{h}_{\eta,s}\|^2_{L^2(P_n)}$$

$$\le \frac{84J\sigma(\gamma^{-d/2}r + \eta^{-d/2}s)t^{1/2}}{n^{1/2}} + 20\|h_{\gamma,r} - g\|^2_\infty + 20\|h_{\eta,s} - g\|^2_\infty$$

for all $\gamma, \eta \in \Gamma$, all $r, s \ge 0$ and all $h_{\gamma,r} \in rB_\gamma$, $h_{\eta,s} \in sB_\eta$. Taking an infimum over $h_{\gamma,r} \in rB_\gamma$ and $h_{\eta,s} \in sB_\eta$ gives

$$\|\hat{h}_{\gamma,r} - \hat{h}_{\eta,s}\|^2_{L^2(P_n)} \le \frac{84J\sigma(\gamma^{-d/2}r + \eta^{-d/2}s)t^{1/2}}{n^{1/2}} + 20I_\infty(g, \gamma, r) + 20I_\infty(g, \eta, s).$$

The result follows from Lemma 7. $\qquad\square$

B.2. Main results

We use the Goldenshluger–Lepski method to prove the following theorem.

Theorem 16. *Assume* (Y) *and* $(\mathcal{K}2)$. *Let* $\tau \ge 84J\sigma$, $v > 0$ *and*

$$t = \left(\frac{\tau}{84J\sigma}\right)^2 \ge 1.$$

Recall the definitions of $A_{3,t}$ *and* $A_{4,t}$. *On the set* $A_{3,t} \cap A_{4,t} \in \mathcal{F}$, *for which* $\mathbb{P}(A_{3,t} \cap A_{4,t}) \ge 1 - 2e^{-t}$, *we have*

$$\|V\hat{h}_{\hat{\gamma},\hat{r}} - g\|^2_{L^2(P)}$$

is at most

$$\inf_{\gamma \in \Gamma} \inf_{r \in R} \left(320I_\infty(g, \gamma, r) + \frac{4v^{d/2}(5 + 2v)\tau\gamma^{-d/2}r}{u^{d/2}n^{1/2}} \right.$$

$$+ \frac{302Cv^{d/2}\tau\gamma^{-d/2}r}{21u^{d/2}\sigma n^{1/2}} + \frac{4Cv^{d/2}\tau^2\gamma^{-d/2}r}{1323J^2u^{d/2}\sigma^2 n}$$

$$+ \left(\frac{12v^{d/2}}{u^{d/2}v} + \frac{302Cv^{d/2}}{21u^{d/2}\sigma v} + \frac{4Cv^{d/2}\tau}{1323J^2u^{d/2}\sigma^2 vn^{1/2}} \right) \left(20I_\infty(g, \gamma, r) + \frac{(1 + v)\tau\gamma^{-d/2}r}{n^{1/2}} \right)$$

$$\left. + 2\|V\hat{h}_{\gamma,r} - g\|^2_{L_2(P)} \right).$$

Proof. Since we assume (Y) and $(\mathcal{K}2)$, which implies $(\mathcal{K}1)$, we find that Lemma 29 holds, which implies that Lemma 15 holds. By our choice of t, we have

$$\|\hat{h}_{\gamma,r} - \hat{h}_{\eta,s}\|^2_{L^2(P_n)} \le \frac{\tau(\gamma^{-d/2}r + \eta^{-d/2}s)}{n^{1/2}} + 40I_\infty(g, \gamma, r) \qquad (9)$$

simultaneously for all $\gamma, \eta \in \Gamma$ and all $r, s \in R$ such that $\eta \leq \gamma$ and $s \geq r$. Fix $\gamma \in \Gamma$ and $r \in R$. Then

$$\|V\hat{h}_{\hat{\gamma},\hat{r}} - V\hat{h}_{\gamma,r}\|^2_{L_2(P)} \leq 2\|V\hat{h}_{\hat{\gamma},\hat{r}} - V\hat{h}_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r}\|^2_{L_2(P)} + 2\|Vh_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r} - V\hat{h}_{\gamma,r}\|^2_{L_2(P)}.$$

We now bound the right-hand side. By $\Gamma \subseteq [u, v]$, the definition of $(\hat{\gamma}, \hat{r})$ in (4) and (9), we have

$$\|\hat{h}_{\hat{\gamma},\hat{r}} - \hat{h}_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r}\|^2_{L_2(P_n)}$$

$$= \|\hat{h}_{\hat{\gamma},\hat{r}} - \hat{h}_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r}\|^2_{L_2(P_n)} - \frac{\tau(\hat{\gamma}^{-d/2}\hat{r} + (\hat{\gamma} \wedge \gamma)^{-d/2}(\hat{r} \vee r))}{n^{1/2}}$$

$$+ \frac{\tau(\hat{\gamma}^{-d/2}\hat{r} + (\hat{\gamma} \wedge \gamma)^{-d/2}(\hat{r} + r))}{n^{1/2}}$$

$$\leq \sup_{\eta\in\Gamma,\eta\leq\hat{\gamma}} \sup_{s\in R,s\geq\hat{r}} \left(\|\hat{h}_{\hat{\gamma},\hat{r}} - \hat{h}_{\eta,s}\|^2_{L_2(P_n)} - \frac{\tau(\hat{\gamma}^{-d/2}\hat{r} + \eta^{-d/2}s)}{n^{1/2}} \right)$$

$$+ \frac{\tau(\hat{\gamma}^{-d/2}\hat{r} + (v/u)^{d/2}(\hat{\gamma}^{-d/2}\hat{r} + \gamma^{-d/2}r))}{n^{1/2}}$$

$$\leq \sup_{\eta\in\Gamma,\eta\leq\gamma} \sup_{s\in R,s\geq r} \left(\|\hat{h}_{\gamma,r} - \hat{h}_{\eta,s}\|^2_{L_2(P_n)} - \frac{\tau(\gamma^{-d/2}r + \eta^{-d/2}s)}{n^{1/2}} \right)$$

$$+ \frac{2(1+v)\tau\gamma^{-d/2}r}{n^{1/2}} - \frac{2(1+v)\tau\hat{\gamma}^{-d/2}\hat{r}}{n^{1/2}} + \frac{v^{d/2}\tau(2\hat{\gamma}^{-d/2}\hat{r} + \gamma^{-d/2}r)}{u^{d/2}n^{1/2}}$$

$$\leq 40I_\infty(g,\gamma,r) + \frac{v^{d/2}(3+2v)\tau\gamma^{-d/2}r}{u^{d/2}n^{1/2}} + \frac{2v^{d/2}\tau\hat{\gamma}^{-d/2}\hat{r}}{u^{d/2}n^{1/2}}.$$

This shows

$$\|V\hat{h}_{\hat{\gamma},\hat{r}} - V\hat{h}_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r}\|^2_{L_2(P_n)} \leq 40I_\infty(g,\gamma,r) + \frac{v^{d/2}(3+2v)\tau\gamma^{-d/2}r}{u^{d/2}n^{1/2}} + \frac{2v^{d/2}\tau\hat{\gamma}^{-d/2}\hat{r}}{u^{d/2}n^{1/2}},$$

and it follows from Lemma 31, our choice of t and $\Gamma \subseteq [u, v]$ that

$$\|V\hat{h}_{\hat{\gamma},\hat{r}} - V\hat{h}_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r}\|^2_{L_2(P)}$$

$$\leq 40I_\infty(g,\gamma,r) + \frac{v^{d/2}(3+2v)\tau\gamma^{-d/2}r}{u^{d/2}n^{1/2}} + \frac{2v^{d/2}\tau\hat{\gamma}^{-d/2}\hat{r}}{u^{d/2}n^{1/2}}$$

$$+ \left(\frac{151C\tau}{84\sigma n^{1/2}} + \frac{C\tau^2}{2646J^2\sigma^2 n} \right)(\hat{\gamma} \wedge \gamma)^{-d/2}(\hat{r} \vee r)$$

$$\leq 40I_\infty(g,\gamma,r) + \frac{v^{d/2}(3+2v)\tau\gamma^{-d/2}r}{u^{d/2}n^{1/2}} + \frac{2v^{d/2}\tau\hat{\gamma}^{-d/2}\hat{r}}{u^{d/2}n^{1/2}}$$

$$+ \left(\frac{151C\tau}{84\sigma n^{1/2}} + \frac{C\tau^2}{2646J^2\sigma^2 n} \right)(v/u)^{d/2}(\hat{\gamma}^{-d/2}\hat{r} + \gamma^{-d/2}r)$$

$$= 40I_\infty(g,\gamma,r) + \frac{v^{d/2}(3+2v)\tau\gamma^{-d/2}r}{u^{d/2}n^{1/2}} + \frac{151Cv^{d/2}\tau\gamma^{-d/2}r}{84u^{d/2}\sigma n^{1/2}} + \frac{Cv^{d/2}\tau^2\gamma^{-d/2}r}{2646J^2u^{d/2}\sigma^2 n}$$

$$+ \left(\frac{v^{d/2}}{u^{d/2}v} + \frac{151Cv^{d/2}}{168u^{d/2}\sigma v} + \frac{Cv^{d/2}\tau}{5292J^2u^{d/2}\sigma^2 vn^{1/2}} \right)\frac{2v\tau\hat{\gamma}^{-d/2}\hat{r}}{n^{1/2}}.$$

By (5), the definition of $(\hat{\gamma}, \hat{r})$ in (4) and (9), we have

$$\frac{2\nu\tau\hat{\gamma}^{-d/2}\hat{r}}{n^{1/2}}$$

$$\leq \sup_{\eta\in\Gamma,\eta\leq\hat{\gamma}}\sup_{s\in R,s\geq\hat{r}}\left(\|\hat{h}_{\hat{\gamma},\hat{r}} - \hat{h}_{\eta,s}\|^2_{L_2(P_n)} - \frac{\tau(\hat{\gamma}^{-d/2}\hat{r} + \eta^{-d/2}s)}{n^{1/2}}\right) + \frac{2(1+\nu)\tau\hat{\gamma}^{-d/2}\hat{r}}{n^{1/2}}$$

$$\leq \sup_{\eta\in\Gamma,\eta\leq\gamma}\sup_{s\in R,s\geq r}\left(\|\hat{h}_{\gamma,r} - \hat{h}_{\eta,s}\|^2_{L_2(P_n)} - \frac{\tau(\gamma^{-d/2}r + \eta^{-d/2}s)}{n^{1/2}}\right) + \frac{2(1+\nu)\tau\gamma^{-d/2}r}{n^{1/2}}$$

$$\leq 40I_\infty(g,\gamma,r) + \frac{2(1+\nu)\tau\gamma^{-d/2}r}{n^{1/2}}. \tag{10}$$

Hence,

$$\|V\hat{h}_{\hat{\gamma},\hat{r}} - V\hat{h}_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r}\|^2_{L_2(P)}$$

$$\leq 40I_\infty(g,\gamma,r) + \frac{v^{d/2}(3+2\nu)\tau\gamma^{-d/2}r}{u^{d/2}n^{1/2}} + \frac{151Cv^{d/2}\tau\gamma^{-d/2}r}{84u^{d/2}\sigma n^{1/2}} + \frac{Cv^{d/2}\tau^2\gamma^{-d/2}r}{2646J^2u^{d/2}\sigma^2 n}$$

$$+ \left(\frac{2v^{d/2}}{u^{d/2}\nu} + \frac{151Cv^{d/2}}{84u^{d/2}\sigma\nu} + \frac{Cv^{d/2}\tau}{2646J^2u^{d/2}\sigma^2\nu n^{1/2}}\right)\left(20I_\infty(g,\gamma,r) + \frac{(1+\nu)\tau\gamma^{-d/2}r}{n^{1/2}}\right).$$

Since (9) holds simultaneously for all $\gamma, \eta \in \Gamma$ and all $r, s \in R$ such that $\eta \leq \gamma$ and $s \geq r$, we have

$$\|\hat{h}_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r} - \hat{h}_{\gamma,r}\|^2_{L_2(P_n)} \leq 40I_\infty(g,\gamma,r) + \frac{\tau(\gamma^{-d/2}r + (\hat{\gamma}\wedge\gamma)^{-d/2}(\hat{r}\vee r))}{n^{1/2}}.$$

This shows

$$\|V\hat{h}_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r} - V\hat{h}_{\gamma,r}\|^2_{L_2(P_n)} \leq 40I_\infty(g,\gamma,r) + \frac{\tau(\gamma^{-d/2}r + (\hat{\gamma}\wedge\gamma)^{-d/2}(\hat{r}\vee r))}{n^{1/2}},$$

and it follows from Lemma 31, our choice of t and (10) that

$$\|V\hat{h}_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r} - V\hat{h}_{\gamma,r}\|^2_{L_2(P)}$$

$$\leq 40I_\infty(g,\gamma,r) + \frac{\tau(\gamma^{-d/2}r + (\hat{\gamma}\wedge\gamma)^{-d/2}(\hat{r}\vee r))}{n^{1/2}}$$

$$+ \left(\frac{151C\tau}{84\sigma n^{1/2}} + \frac{C\tau^2}{2646J^2\sigma^2 n}\right)(\hat{\gamma}\wedge\gamma)^{-d/2}(\hat{r}\vee r)$$

$$= 40I_\infty(g,\gamma,r) + \frac{\tau\gamma^{-d/2}r}{n^{1/2}} + \left(\frac{\tau}{n^{1/2}} + \frac{151C\tau}{84\sigma n^{1/2}} + \frac{C\tau^2}{2646J^2\sigma^2 n}\right)(\hat{\gamma}\wedge\gamma)^{-d/2}(\hat{r}\vee r)$$

$$\leq 40I_\infty(g,\gamma,r) + \frac{\tau\gamma^{-d/2}r}{n^{1/2}}$$

$$+ \left(\frac{\tau}{n^{1/2}} + \frac{151C\tau}{84\sigma n^{1/2}} + \frac{C\tau^2}{2646J^2\sigma^2 n}\right)(v/u)^{d/2}(\hat{\gamma}^{-d/2}\hat{r} + \gamma^{-d/2}r)$$

$$\leq 40I_\infty(g,\gamma,r) + \frac{2v^{d/2}\tau\gamma^{-d/2}r}{u^{d/2}n^{1/2}} + \frac{151Cv^{d/2}\tau\gamma^{-d/2}r}{84u^{d/2}\sigma n^{1/2}} + \frac{Cv^{d/2}\tau^2\gamma^{-d/2}r}{2646J^2u^{d/2}\sigma^2 n}$$

$$+ \left(\frac{v^{d/2}}{2u^{d/2}v} + \frac{151Cv^{d/2}}{168u^{d/2}\sigma v} + \frac{Cv^{d/2}\tau}{5292J^2u^{d/2}\sigma^2 vn^{1/2}} \right) \frac{2v\tau\hat{\gamma}^{-d/2}\hat{r}}{n^{1/2}}$$

$$\leq 40I_\infty(g,\gamma,r) + \frac{2v^{d/2}\tau\gamma^{-d/2}r}{u^{d/2}n^{1/2}} + \frac{151Cv^{d/2}\tau\gamma^{-d/2}r}{84u^{d/2}\sigma n^{1/2}} + \frac{Cv^{d/2}\tau^2\gamma^{-d/2}r}{2646J^2u^{d/2}\sigma^2 n}$$

$$+ \left(\frac{v^{d/2}}{u^{d/2}v} + \frac{151Cv^{d/2}}{84u^{d/2}\sigma v} + \frac{Cv^{d/2}\tau}{2646J^2u^{d/2}\sigma^2 vn^{1/2}} \right) \left(20I_\infty(g,\gamma,r) + \frac{(1+v)\tau\gamma^{-d/2}r}{n^{1/2}} \right).$$

Hence,

$$\|V\hat{h}_{\hat{\gamma},\hat{r}} - V\hat{h}_{\gamma,r}\|^2_{L_2(P)}$$

$$\leq 2\|V\hat{h}_{\hat{\gamma},\hat{r}} - V\hat{h}_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r}\|^2_{L_2(P)} + 2\|Vh_{\hat{\gamma}\wedge\gamma,\hat{r}\vee r} - V\hat{h}_{\gamma,r}\|^2_{L_2(P)}$$

$$\leq 160I_\infty(g,\gamma,r) + \frac{2v^{d/2}(5+2v)\tau\gamma^{-d/2}r}{u^{d/2}n^{1/2}} + \frac{151Cv^{d/2}\tau\gamma^{-d/2}r}{21u^{d/2}\sigma n^{1/2}} + \frac{2Cv^{d/2}\tau^2\gamma^{-d/2}r}{1323J^2u^{d/2}\sigma^2 n}$$

$$+ \left(\frac{6v^{d/2}}{u^{d/2}v} + \frac{151Cv^{d/2}}{21u^{d/2}\sigma v} + \frac{2Cv^{d/2}\tau}{1323J^2u^{d/2}\sigma^2 vn^{1/2}} \right) \left(20I_\infty(g,\gamma,r) + \frac{(1+v)\tau\gamma^{-d/2}r}{n^{1/2}} \right).$$

We have

$$\|V\hat{h}_{\hat{\gamma},\hat{r}} - g\|^2_{L_2(P)} \leq 2\|V\hat{h}_{\hat{\gamma},\hat{r}} - V\hat{h}_{\gamma,r}\|^2_{L_2(P)} + 2\|V\hat{h}_{\gamma,r} - g\|^2_{L_2(P)}$$

and the result follows. $\qquad\square$

We assume (g1) to bound the distance between $V\hat{h}_{\hat{\gamma},\hat{r}}$ and g in the $L^2(P)$ norm and prove Theorem 8.

Proof of Theorem 8. By Theorem 16, we have

$$\|V\hat{h}_{\hat{\gamma},\hat{r}} - g\|^2_{L^2(P)}$$

$$\leq \inf_{\gamma\in\Gamma} \inf_{r\in R} \left((1 + D_4\tau n^{-1/2})(D_5\tau\gamma^{-d/2}rn^{-1/2} + D_6 I_\infty(g,\gamma,r)) + 2\|V\hat{h}_{\gamma,r} - g\|^2_{L^2(P)} \right)$$

for some constants $D_4, D_5, D_6 > 0$ not depending on τ, γ, r or n. By Theorem 6, we have

$$\|V\hat{h}_{\gamma,r} - g\|^2_{L^2(P)} \leq \frac{(151C + 21\sigma)\tau\gamma^{-d/2}r}{42\sigma n^{1/2}} + \frac{C\tau^2\gamma^{-d/2}r}{1323J^2\sigma^2 n} + 10I_\infty(g,\gamma,r)$$

$$\leq D_7\tau\gamma^{-d/2}rn^{-1/2} + D_8\tau^2\gamma^{-d/2}rn^{-1} + 10I_\infty(g,\gamma,r)$$

for all $\gamma \in \Gamma$ and all $r \in R$, for some constants $D_7, D_8 > 0$ not depending on τ, γ, r or n. This gives

$$\|V\hat{h}_{\hat{\gamma},\hat{r}} - g\|^2_{L^2(P)} \leq \inf_{\gamma\in\Gamma} \inf_{r\in R} \left((1 + D_4\tau n^{-1/2})(D_5\tau\gamma^{-d/2}rn^{-1/2} + D_6 I_\infty(g,\gamma,r)) \right.$$

$$\left. + 2D_7\tau\gamma^{-d/2}rn^{-1/2} + 2D_8\tau^2\gamma^{-d/2}rn^{-1} + 20I_\infty(g,\gamma,r) \right).$$

Hence, the result follows with

$$D_1 = \frac{D_4 D_5 + 2D_8}{D_5 + 2D_7}, \qquad D_2 = D_5 + 2D_7, \qquad D_3 = D_6 + 20. \qquad\square$$

We assume $(g3)$ to prove Theorem 9.

Proof of Corollary 9. If we assume $(R1)$ and $(\Gamma1)$, then $\alpha \in \Gamma$ and $r = an^{(1-\beta)/(2(1+\beta))} \in R$, so

$$\|V\hat{h}_{\hat{\gamma},\hat{r}} - g\|^2_{L^2(P)} \leq \left(1 + D_3\tau n^{-1/2}\right)\left(D_4\tau\alpha^{-d/2}rn^{-1/2} + D_5I_\infty(g,\alpha,r)\right)$$

$$\leq \left(1 + D_3\tau n^{-1/2}\right)\left(D_4\tau\alpha^{-d/2}an^{-\beta/(1+\beta)} + \frac{D_5B^{2/(1-\beta)}}{a^{2\beta/(1-\beta)}n^{\beta/(1+\beta)}}\right)$$

for some constants $D_3, D_4, D_5 > 0$ not depending on n or τ by Theorem 8 and (6). If we assume $(R2)$ and $(\Gamma2)$, then there is a unique $\gamma \in \Gamma$ such that $\alpha/c < \gamma \leq \alpha$ and a unique $r \in R$ such that

$$an^{(1-\beta)/(2(1+\beta))} \leq r < an^{(1-\beta)/(2(1+\beta))} + b.$$

By Theorem 8, Lemma 7 and (6), we have

$$\|V\hat{h}_{\hat{\gamma},\hat{r}} - g\|^2_{L^2(P)}$$

$$\leq \left(1 + D_3\tau n^{-1/2}\right)\left(D_4\tau\gamma^{-d/2}rn^{-1/2} + D_5I_\infty(g,\gamma,r)\right)$$

$$\leq \left(1 + D_3\tau n^{-1/2}\right)\left(D_4\tau c^{d/2}\alpha^{-d/2}\left(an^{(1-\beta)/(2(1+\beta))} + b\right)n^{-1/2} + \frac{D_5B^{2/(1-\beta)}}{a^{2\beta/(1-\beta)}n^{\beta/(1+\beta)}}\right).$$

In either case,

$$\|V\hat{h}_{\hat{\gamma},\hat{r}} - g\|^2_{L^2(P)} \leq D_1\tau n^{-\beta/(1+\beta)} + D_2\tau^2 n^{-(1+3\beta)/(2(1+\beta))}$$

for some constants $D_1, D_2 > 0$ not depending on n or τ. $\qquad\square$

Proof of Corollary 10. The proof is identical to the proof of Corollary 5. $\qquad\square$

Supplementary Material

Regression results (DOI: 10.3150/20-BEJ1307SUPP; .pdf). We provide the proofs of the results for the two regression problems, including the majorants, along with some technical results.

References

[1] Bach, F.R., Lanckriet, G.R.G. and Jordan, M.I. (2004). Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the Twenty-First International Conference on Machine Learning* 6.

[2] Baraud, Y. (2002). Model selection for regression on a random design. *ESAIM Probab. Stat.* **6** 127–146. MR1918295 https://doi.org/10.1051/ps:2002007

[3] Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413. MR1679028 https://doi.org/10.1007/s004400050210

[4] Baumeister, J. (1987). *Stable Solution of Inverse Problems. Advanced Lectures in Mathematics*. Braunschweig: Friedr. Vieweg & Sohn. MR0889048 https://doi.org/10.1007/978-3-322-83967-1

[5] Bergh, J. and Löfström, J. (1976). *Interpolation Spaces. An Introduction*. Berlin: Springer. MR0482275

[6] Birgé, L. (2001). An alternative point of view on Lepski's method. In *State of the Art in Probability and Statistics* (*Leiden*, 1999). *Institute of Mathematical Statistics Lecture Notes – Monograph Series* **36** 113–133. Beachwood, OH: IMS. MR1836557 https://doi.org/10.1214/lnms/1215090065

[7] Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam* 55–87. New York: Springer. MR1462939

[8] Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* (*JEMS*) **3** 203–268. MR1848946 https://doi.org/10.1007/s100970100031

[9] Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields* **138** 33–73. MR2288064 https://doi.org/10.1007/s00440-006-0011-8

[10] Blanchard, G., Mathé, P. and Mücke, N. (2019). Lepskii principle in supervised learning. Preprint. Available at arXiv:1905.10764.

[11] Caponnetto, A. and Yao, Y. (2010). Cross-validation based adaptation for regularization operators in learning theory. *Anal. Appl.* (*Singap.*) **8** 161–183. MR2646909 https://doi.org/10.1142/S0219530510001564

[12] Chapelle, O., Vapnik, V., Bousquet, O. and Mukherjee, S. (2002). Choosing multiple parameters for support vector machines. *Mach. Learn.* **46**.

[13] De Vito, E., Pereverzyev, S. and Rosasco, L. (2010). Adaptive kernel methods using the balancing principle. *Found. Comput. Math.* **10** 455–479. MR2657949 https://doi.org/10.1007/s10208-010-9064-2

[14] Eberts, M. and Steinwart, I. (2013). Optimal regression rates for SVMs using Gaussian kernels. *Electron. J. Stat.* **7** 1–42. MR3020412 https://doi.org/10.1214/12-EJS760

[15] Giné, E. and Nickl, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. New York: Cambridge Univ. Press. MR3588285 https://doi.org/10.1017/CBO9781107337862

[16] Goldenshluger, A. and Lepski, O. (2008). Universal pointwise selection rule in multivariate function estimation. *Bernoulli* **14** 1150–1190. MR2543590 https://doi.org/10.3150/08-BEJ144

[17] Goldenshluger, A. and Lepski, O. (2009). Structural adaptation via $\mathbb{L}_p$-norm oracle inequalities. *Probab. Theory Related Fields* **143** 41–71. MR2449122 https://doi.org/10.1007/s00440-007-0119-5

[18] Goldenshluger, A. and Lepski, O. (2011). Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. *Ann. Statist.* **39** 1608–1632. MR2850214 https://doi.org/10.1214/11-AOS883

[19] Goldenshluger, A. and Pereverzev, S.V. (2003). On adaptive inverse estimation of linear functionals in Hilbert scales. *Bernoulli* **9** 783–807. MR2047686 https://doi.org/10.3150/bj/1066418878

[20] Goldenshluger, A.V. and Lepski, O.V. (2013). General selection rule from a family of linear estimators. *Theory Probab. Appl.* **57** 209–226. MR3201652 https://doi.org/10.1137/S0040585X97985923

[21] Gönen, M. and Alpaydın, E. (2011). Multiple kernel learning algorithms. *J. Mach. Learn. Res.* **12** 2211–2268. MR2825425 https://doi.org/10.1016/j.patcog.2012.09.002

[22] Hofmann, B. (2005). Approximate source conditions in regularization and an application to multiplication operators. In *Proceedings*: 10*th International Conference "Mathematical Modelling and Analysis* 2005*" and* 2*nd International Conference "Computational Methods in Applied Mathematics"* 29–34. Vilnius: Technika. MR2194416

[23] Kreĭn, S.G. and Petunin, J.I. (1966). Scales of Banach spaces. *Uspekhi Mat. Nauk* **21** 89–168. MR0193499

[24] Lepski, O. (1991). On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 454–466.

[25] Lepskiĭ, O.V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatn. Primen.* **36** 645–659. MR1147167 https://doi.org/10.1137/1136085

[26] Lepskiĭ, O.V. (1992). Asymptotically minimax adaptive estimation. II. Schemes without optimal adaptation. Adaptive estimates. *Teor. Veroyatn. Primen.* **37** 468–481. MR1214353 https://doi.org/10.1137/1137095

[27] Lu, S., Mathé, P. and Pereverzev, S.V. (2020). Balancing principle in supervised learning for a general regularization scheme. *Appl. Comput. Harmon. Anal.* **48** 123–148. MR4016987 https://doi.org/10.1016/j.acha.2018.03.001

[28] Lu, S., Mathé, P. and Pereverzev, S. Jr. (2019). Analysis of regularized Nyström subsampling for regression functions of low smoothness. *Anal. Appl.* (*Singap.*) **17** 931–946. MR4009938 https://doi.org/10.1142/S0219530519500039

[29] Mair, B.A. and Ruymgaart, F.H. (1996). Statistical inverse estimation in Hilbert scales. *SIAM J. Appl. Math.* **56** 1424–1444. MR1409127 https://doi.org/10.1137/S0036139994264476

[30] Mathé, P. and Pereverzev, S.V. (2001). Optimal discretization of inverse problems in Hilbert scales. Regularization and self-regularization of projection methods. *SIAM J. Numer. Anal.* **38** 1999–2021. MR1856240 https://doi.org/10.1137/S003614299936175X

[31] Oneto, L., Ridella, S. and Anguita, D. (2016). Tikhonov, Ivanov and Morozov regularization for support vector machine learning. *Mach*. *Learn*. **103** 103–136. MR3472725 https://doi.org/10.1007/s10994-015-5540-x

[32] Page, S. and Grünewälder, S. (2019). Ivanov-regularised least-squares estimators over large RKHSs and their interpolation spaces. *J*. *Mach*. *Learn*. *Res*. **20** 120. MR4002874

[33] Page, S. and Grünewälder, S. (2021). Supplement to "The Goldenshluger–Lepski method for constrained least-squares estimators over RKHSs." https://doi.org/10.3150/20-BEJ1307SUPP

[34] Pereverzyev, S. and Tkachenko, P. (2017). Regularization by the linear functional strategy with multiple kernels. *Front*. *Appl*. *Math*. *Stat*. **3**.

[35] Pricop-Jeckstadt, M. (2019). Nonlinear Tikhonov regularization in Hilbert scales with balancing principle tuning parameter in statistical inverse problems. *Inverse Probl*. *Sci*. *Eng*. **27** 205–236. MR3883351 https://doi.org/10.1080/17415977.2018.1454918

[36] Smale, S. and Zhou, D.-X. (2003). Estimating the approximation error in learning theory. *Anal*. *Appl*. (*Singap*.) **1** 17–41. MR1959283 https://doi.org/10.1142/S0219530503000089

[37] Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. New York: Springer. MR2450103

[38] Steinwart, I., Hush, D.R. and Scovel, C. (2009). Optimal rates for regularized least squares regression. In *The* 22*nd Conference on Learning Theory*.

[39] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*. New York: Springer. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

[40] Williams, D. (1991). *Probability with Martingales*. Cambridge: Cambridge Univ. Press. MR1155402 https://doi.org/10.1017/CBO9780511813658