Journal of Machine Learning Research 20 (2019) 1-49 Submitted 11/17; Revised 6/19; Published 9/19

Ivanov-Regularised Least-Squares Estimators over Large RKHSs and Their Interpolation Spaces

Stephen Page s.page@lancaster.ac.uk
STOR-i
Lancaster University
Lancaster, LA1 4YF, United Kingdom

Steffen Grünewälder s.grunewalder@lancaster.ac.uk
Department of Mathematics and Statistics
Lancaster University
Lancaster, LA1 4YF, United Kingdom

Editor: Ingo Steinwart

Abstract

We study kernel least-squares estimation under a norm constraint. This form of regularisation is known as Ivanov regularisation and it provides better control of the norm of the estimator than the well-established Tikhonov regularisation. Ivanov regularisation can be studied under minimal assumptions. In particular, we assume only that the RKHS is separable with a bounded and measurable kernel. We provide rates of convergence for the expected squared L^2 error of our estimator under the weak assumption that the variance of the response variables is bounded and the unknown regression function lies in an interpolation space between L^2 and the RKHS. We then obtain faster rates of convergence when the regression function is bounded by clipping the estimator. In fact, we attain the optimal rate of convergence. Furthermore, we provide a high-probability bound under the stronger assumption that the response variables have subgaussian errors and that the regression function lies in an interpolation space between L^∞ and the RKHS. Finally, we derive adaptive results for the settings in which the regression function is bounded.

Keywords: Interpolation Space, Ivanov Regularisation, Regression, RKHS, Training and Validation

1. Introduction

One of the key problems to overcome in nonparametric regression is overfitting, due to estimators coming from large hypothesis classes. To avoid this phenomenon, it is common to ensure that both the empirical risk and some regularisation function are small when defining an estimator. There are three natural ways to achieve this goal. We can minimise the empirical risk subject to a constraint on the regularisation function, minimise the regularisation function subject to a constraint on the empirical risk or minimise a linear combination of the two. These techniques are known as Ivanov regularisation, Morozov regularisation and Tikhonov regularisation respectively (Oneto, Ridella, and Anguita, 2016). Ivanov and

Morozov regularisation can be viewed as dual problems, while Tikhonov regularisation can be viewed as the Lagrangian relaxation of either.

Tikhonov regularisation has gained popularity as it provides a closed-form estimator in many situations. In particular, Tikhonov regularisation in which the estimator is selected from a reproducing-kernel Hilbert space (RKHS) has been extensively studied (Smale and Zhou, 2007; Caponnetto and de Vito, 2007; Steinwart and Christmann, 2008; Mendelson and Neeman, 2010; Steinwart, Hush, and Scovel, 2009). Although Tikhonov regularisation produces an estimator in closed form, it is Ivanov regularisation which provides the greatest control over the hypothesis class, and hence over the estimator it produces. For example, if the regularisation function is the norm of the RKHS, then the bound on this function forces the estimator to lie in a ball of predefined radius inside the RKHS. An RKHS norm measures the smoothness of a function, so the norm constraint bounds the smoothness of the estimator. By contrast, Tikhonov regularisation provides no direct control over the smoothness of the estimator.

The control we have over the Ivanov-regularised estimator is useful in many settings. The most obvious use of Ivanov regularisation is when the regression function lies in a ball of known radius inside the RKHS. In this case, Ivanov regularisation can be used to constrain the estimator to lie in the same ball. Suppose, for example, that we are interested in estimating the trajectory of a particle from noisy observations over time. Assume that the velocity or acceleration of the particle is constrained by certain physical conditions. Constraints of this nature can be imposed by bounding the norm of the trajectory in a Sobolev space. Certain Sobolev spaces are RKHSs, so it is possible to use Ivanov regularisation to enforce physical conditions on an estimator of the trajectory which match those of the trajectory itself. Ivanov regularisation can also be used within larger inference methods. It is compatible with validation, allowing us to control an estimator selected from an uncountable collection. This is because the Ivanov-regularised estimator is continuous in the size of the ball containing it (see Lemma 32), so the estimators parametrised by an interval of ball sizes can be controlled simultaneously using chaining.

In addition to the other useful properties of the Ivanov-regularised estimator, Ivanov regularisation can be performed almost as quickly as Tikhonov regularisation. The Ivanov-regularised estimator is a support vector machine (SVM) with regularisation parameter selected to match the norm constraint (see Lemma 3). This parameter can be selected to within a tolerance ε using interval bisection with order $\log(1/\varepsilon)$ iterations. In general, Ivanov regularisation requires the calculation of order $\log(1/\varepsilon)$ SVMs.

In this paper, we study the behaviour of the Ivanov-regularised least-squares estimator with regularisation function equal to the norm of the RKHS. We derive a number of novel results concerning the rate of convergence of the estimator in various settings and under various assumptions. Our analysis is performed by controlling empirical processes over balls in the RKHS. By contrast, the analysis of Tikhonov-regularised estimators usually relies on the spectral decomposition of the kernel operator T on $L^2(P)$. Here, P is the covariate distribution.

We first prove an expectation bound on the squared $L^2(P)$ error of our estimator of order $n^{-\beta/2}$, under the weak assumption that the response variables have bounded variance. Here, n is the number of data points, and β parametrises the interpolation space between $L^2(P)$ and H containing the regression function. As far as we are aware, the analysis of an estimator in this setting has not previously been considered. The definition of an interpolation space is given in Section 2. The expected squared $L^2(P)$ error can be viewed as the expected squared error of our estimator at a new independent covariate, with the same distribution P. If we also assume that the regression function is bounded, then it makes sense to clip our estimator so that it takes values in the same interval as the regression function. This further assumption allows us to achieve an expectation bound on the squared $L^2(P)$ error of the clipped estimator of order $n^{-\beta/(1+\beta)}$.

We then move away from the average behaviour of the error towards its behaviour in the worst case. We obtain high-probability bounds of the same order, under the stronger assumption that the response variables have subgaussian errors and the interpolation space is between L^∞ and H. The second assumption is quite natural as we already assume that the regression function is bounded, and H can be continuously embedded in L^∞ since it has a bounded kernel k. Note that this assumption means that the set of possible regression functions is independent of the covariate distribution.

When the regression function is bounded, we also analyse an adaptive version of our estimator, which does not require us to know which interpolation space contains the regression function. This adaptive estimator obtains bounds of the same order as the non-adaptive one. The adaptive estimator is created using training and validation. The bounds do not contain higher-order terms because we treat the validation step as a form of Ivanov regularisation, as opposed to, for example, using a union bound.

Our expectation bound of order $n^{-\beta/(1+\beta)}$, when the regression function is bounded, improves on the high-probability bound of Smale and Zhou (2007) of order $n^{-\beta/2}$. Their bound is attained under the stronger assumption that the regression function lies in the image of a power of the kernel operator, instead of an interpolation space (see Steinwart and Scovel, 2012). The authors also assume that the response variables are bounded. Furthermore, for a fixed $\beta \in (0,1)$, Steinwart et al. (2009) show that there is an instance of our problem with a bounded regression function such that the following holds. For all estimators $\hat{f}$ of g, for some $\varepsilon > 0$, we have

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

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

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

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. Our expectation bound on the adaptive version of our estimator is also optimal, because the bound is of the same order as in the easier non-adaptive setting.

The high-probability bound of Steinwart et al. (2009) is optimal in a similar sense, although the authors achieve faster rates by assuming a fixed rate of decay of the eigenvalues of the kernel operator T, as discussed in Section 3. Since there is an additional parameter for the decay of the eigenvalues, the collection of problem instances for a fixed set of parameters is smaller in their paper. This means that our optimal rates are the slowest of the optimal rates in Steinwart et al. (2009).

Note that all of the bounds in this paper are on the squared $L^2(P)$ error of our estimators. We do not consider errors based on other norms, such as the interpolation space norms, as it is unclear how the techniques used in this paper could be adapted to such errors. It seems likely that a completely different approach would be necessary to bound these other errors.

2. RKHSs and Their Interpolation Spaces

A Hilbert space H of real-valued functions on S is an RKHS if the evaluation functional $L_x : H \to \mathbb{R}$, $L_x h = h(x)$, is bounded for all $x \in S$. In this case, $L_x \in H^*$ the dual of H and the Riesz representation theorem tells us that there is some $k_x \in H$ such that $h(x) = \langle h, k_x \rangle_H$ for all $h \in H$. The kernel is then given by $k(x_1, x_2) = \langle k_{x_1}, k_{x_2} \rangle_H$ for $x_1, x_2 \in S$, and is symmetric and positive-definite.

Now suppose that $(S, \mathcal{S})$ is a measurable space on which P is a probability measure. We can define a range of interpolation spaces between $L^2(P)$ and H (Bergh and Löfström, 1976). Let $(Z, \|\cdot\|_Z)$ be a Banach space and $(V, \|\cdot\|_V)$ be a dense subspace of Z. The K-functional of (Z, V) is

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

for $z \in Z$ and $t > 0$. For $\beta \in (0, 1)$ and $1 \leq q < \infty$, we define

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

for $z \in Z$. The interpolation space $[Z, V]_{\beta,q}$ is defined to be the set of $z \in Z$ such that $\|z\|_{\beta,q} < \infty$. Smaller values of β give larger spaces. The space $[Z, V]_{\beta,q}$ is not much larger than V when β is close to 1, but we obtain spaces which get closer to Z as β decreases. The following result is essentially Theorem 3.1 of Smale and Zhou (2003). The authors only consider the case in which $\|v\|_Z \leq \|v\|_V$ for all $v \in V$, however the result holds by the same proof even without this condition.

Lemma 1 *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*

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

When H is dense in $L^2(P)$, we can define the interpolation spaces $[L^2(P), H]_{\beta,q}$, where $L^2(P)$ is the space of measurable functions f on $(S, \mathcal{S})$ such that f^2 is integrable with respect to P. We work with $q = \infty$, which gives the largest space of functions for a fixed $\beta \in (0,1)$. We can then use the approximation result in Lemma 1. When H is dense in L^∞, we also require $[L^\infty, H]_{\beta,q}$, where L^∞ is the space of bounded measurable functions on $(S, \mathcal{S})$.

3. Literature Review

Early research on RKHS regression does not make assumptions on the rate of decay of the eigenvalues of the kernel operator. For example, Smale and Zhou (2007) assume that the response variables are bounded and the regression function is of the form $g = T^{\beta/2} f$ for $\beta \in (0,1]$ and $f \in L^2(P)$. Here, $T : L^2(P) \to L^2(P)$ is the kernel operator and P is the covariate distribution. The authors achieve a squared $L^2(P)$ error of order $n^{-\beta/2}$ with high probability by using SVMs.

Initial research which does make assumptions on the rate of decay of the eigenvalues of the kernel operator, such as that of Caponnetto and de Vito (2007), assumes that the regression function is at least as smooth as an element of H. However, their paper still allows for regression functions of varying smoothness by letting $g \in T^{(\beta-1)/2}(H)$ for $\beta \in [1,2]$. By assuming that the ith eigenvalue of T is of order $i^{-1/p}$ for $p \in (0,1]$, the authors achieve a squared $L^2(P)$ error of order $n^{-\beta/(\beta+p)}$ with high probability by using SVMs. This squared $L^2(P)$ error is shown to be of optimal order for $\beta \in (1,2]$.

Later research focuses on the case in which the regression function is at most as smooth as an element of H. Often, this research demands that the response variables are bounded. For example, Mendelson and Neeman (2010) assume that $g \in T^{\beta/2}(L^2(P))$ for $\beta \in (0,1)$ to obtain a squared $L^2(P)$ error of order $n^{-\beta/(1+p)}$ with high probability by using Tikhonov-regularised least-squares estimators. The authors also show that if the eigenfunctions of the kernel operator T are uniformly bounded in L^∞, then the order can be improved to $n^{-\beta/(\beta+p)}$. Steinwart et al. (2009) relax the condition on the eigenfunctions to the condition

$$\|h\|_\infty \le C_p \|h\|_H^p \|h\|_{L^2(P)}^{1-p}$$

for all $h \in H$ and some constant $C_p > 0$. The same rate is attained by using clipped Tikhonov-regularised least-squares estimators, including clipped SVMs, and is shown to be optimal. The authors assume that g is in an interpolation space between $L^2(P)$ and H, which is slightly more general than the assumption of Mendelson and Neeman (2010). A detailed discussion about the image of $L^2(P)$ under powers of T and interpolation spaces between $L^2(P)$ and H is given by Steinwart and Scovel (2012).

Lately, the assumption that the response variables must be bounded has been relaxed to allow for subexponential errors. However, the assumption that the regression function is bounded has been maintained. For example, Fischer and Steinwart (2017) assume that $g \in T^{\beta/2}(L^2(P))$ for $\beta \in (0,2]$ and that g is bounded. The authors also assume that $T^{\alpha/2}(L^2(P))$ is continuously embedded in L^∞, with respect to an appropriate norm on $T^{\alpha/2}(L^2(P))$, for some $\alpha < \beta$. This gives the same squared $L^2(P)$ error of order $n^{-\beta/(\beta+p)}$ with high probability by using SVMs.

4. Contribution

In this paper, we provide bounds on the squared $L^2(P)$ error of our Ivanov-regularised least-squares estimator when the regression function comes from an interpolation space between $L^2(P)$ and an RKHS H, which is separable with a bounded and measurable kernel k. We use the norm of the RKHS as our regularisation function. Under the weak assumption that the response variables have bounded variance, we prove a bound on the expected squared $L^2(P)$ error of order $n^{-\beta/2}$ (Theorem 5 on page 12). As far as we are aware, the analysis of an estimator in this setting has not previously been considered. If we assume that the regression function is bounded, then we can clip the estimator and achieve an expected squared $L^2(P)$ error of order $n^{-\beta/(1+\beta)}$ (Theorem 7 on page 13).

Under the stronger assumption that the response variables have subgaussian errors and the regression function comes from an interpolation space between L^∞ and H, we show that the squared $L^2(P)$ error is of order $n^{-\beta/(1+\beta)}$ with high probability (Theorem 11 on page 17). For the settings in which the regression function is bounded, we use training and validation on the data in order to select the size of the constraint on the norm of our estimator. This gives us an adaptive estimation result which does not require us to know which interpolation space contains the regression function. We obtain a squared $L^2(P)$ error of order $n^{-\beta/(1+\beta)}$ in expectation and with high probability, depending on the setting (Theorems 9 and 13 on pages 15 and 18). In order to perform training and validation, the response variables in the validation set must have subgaussian errors. The expectation results for bounded regression functions are of optimal order in the sense discussed at the end of Section 1.

Regression Function	$L^2(P)$ Interpolation	L^∞ Bound	L^∞ Interpolation
Response Variables	Bounded Variance	Bounded Variance	Subgaussian Errors
Bound Type	Expectation	Expectation	High Probability
Bound Order	$n^{-\beta/2}$	$n^{-\beta/(1+\beta)}$ $(*)$	$n^{-\beta/(1+\beta)}$

Table 1: Orders of bounds on squared $L^2(P)$ error

The validation results are summarised in Table 2. Again, the columns for which there is an L^∞ bound on the regression function also make the $L^2(P)$ interpolation assumption. The assumptions on the response variables relate to those in the validation set, which has $\tilde{n}$ data points. We assume that $\tilde{n}$ is equal to some multiple of n. Again, orders of bounds marked with $(*)$ are known to be optimal.

Regression Function	L^∞ Bound	L^∞ Interpolation
Response Variables	Subgaussian Errors	Subgaussian Errors
Bound Type	Expectation	High Probability
Bound Order	$n^{-\beta/(1+\beta)}$ $(*)$	$n^{-\beta/(1+\beta)}$

Table 2: Orders of validation bounds on squared $L^2(P)$ error

The validation results are summarised in Table 2. Again, the columns for which there is an L^∞ bound on the regression function also make the $L^2(P)$ interpolation assumption. The

assumptions on the response variables relate to those in the validation set, which has $\tilde{n}$ data points. We assume that $\tilde{n}$ is equal to some multiple of n. Again, orders of bounds marked with $(*)$ are known to be optimal.

5. Problem Definition

We now formally define our regression problem. For a topological space T, let $\mathcal{B}(T)$ be the Borel σ-algebra of T. Let $(S, \mathcal{S})$ be a measurable space. Assume that (X_i, Y_i) for $1 \leq i \leq n$ are $(S \times \mathbb{R}, \mathcal{S} \otimes \mathcal{B}(\mathbb{R}))$-valued random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which are i.i.d. with $X_i \sim P$ and $\mathbb{E}(Y_i^2) < \infty$, where $\mathbb{E}$ denotes integration with respect to $\mathbb{P}$. Since any version of $\mathbb{E}(Y_i | X_i)$ is $\sigma(X_i)$-measurable, where $\sigma(X_i)$ is the σ-algebra generated by X_i, we have that $\mathbb{E}(Y_i | X_i) = g(X_i)$ almost surely for some function g which is measurable on $(S, \mathcal{S})$ (Section A3.2 of Williams, 1991). From the definition of conditional expectation and the identical distribution of the (X_i, Y_i), it is clear that we can choose g to be the same for all $1 \leq i \leq n$. The conditional expectation used is that of Kolmogorov, defined using the Radon–Nikodym derivative. Its definition is unique almost surely. Since $\mathbb{E}(Y_i^2) < \infty$, it follows that $g \in L^2(P)$ by Jensen's inequality. To summarise, $\mathbb{E}(Y_i | X_i) = g(X_i)$ almost surely for $1 \leq i \leq n$ with $g \in L^2(P)$. We assume throughout that

$$(Y1) \qquad \operatorname{var}(Y_i | X_i) \leq \sigma^2 \text{ almost surely for } 1 \leq i \leq n.$$

Our results depend on how well 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

$$\|k\|_\infty = \sup_{x \in S} k(x, x)^{1/2} < \infty.$$

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 Steinwart and Christmann, 2008). 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 Steinwart and Christmann, 2008).

6. Ivanov Regularisation

We now consider Ivanov regularisation for least-squares estimators. Let P_n be the empirical distribution of the X_i for $1 \leq i \leq n$. The definition of Ivanov regularisation provides us with the following result.

Lemma 2 *Let $A \subseteq L^2(P)$. It may be that A is a function of $\omega \in \Omega$ and does not contain g. Let*

$$\hat{f} \in \arg\min_{f \in A} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2.$$

Then, for all $f \in A$ and all $\omega \in \Omega$, we have

$$\|\hat{f} - f\|_{L^2(P_n)}^2 \leq \frac{4}{n} \sum_{i=1}^{n} (Y_i - g(X_i))(\hat{f}(X_i) - f(X_i)) + 4\|f - g\|_{L^2(P_n)}^2.$$

The function g in Lemma 2 need not be the regression function, however this is the case of interest. In general, the first term of the right-hand side of the inequality must be controlled by bounding it with

$$\sup_{f_1, f_2 \in A} \frac{4}{n} \sum_{i=1}^{n} (Y_i - g(X_i))(f_1(X_i) - f_2(X_i)). \tag{1}$$

This is usually not measurable. However, if A is a fixed subset of a separable RKHS, then A is separable and the function which evaluates $f \in A$ at X_i is continuous for $1 \leq i \leq n$. This means that the supremum can be replaced with a countable supremum, so the quantity is a random variable on $(\Omega, \mathcal{F})$. Clearly, this term increases as A gets larger. However, if A gets larger, then we may select $f \in A$ closer to g. Hence, we can make the second term of the right-hand side of the inequality in Lemma 2 smaller. This demonstrates the trade-off in selecting the size of A for the Ivanov-regularised least-squares estimator constrained to lie in A.

The next step in analysing $\hat{f}$ is to move to a bound on

$$\|\hat{f} - f\|_{L^2(P)}^2 \leq \|\hat{f} - f\|_{L^2(P_n)}^2 + \sup_{f_1, f_2 \in A} \left| \|f_1 - f_2\|_{L^2(P_n)}^2 - \|f_1 - f_2\|_{L^2(P)}^2 \right|. \tag{2}$$

The second term on the right-hand side of this inequality is measurable when A is a fixed subset of a separable RKHS. It also increases with A. Finally, we obtain a bound on

$$\|\hat{f} - g\|_{L^2(P)}^2 \leq 2\|\hat{f} - f\|_{L^2(P)}^2 + 2\|f - g\|_{L^2(P)}^2.$$

This again demonstrates why $f \in A$ should be close to g.

7. Estimator Definition

In this section, we consider the Ivanov-regularised estimators of Section 6 in the context of RKHS regression. Recall the RKHS H from the definition of our regression problem in Section 5. Let B_H be the closed unit ball of H and $r > 0$. The Ivanov-regularised least-squares estimator constrained to lie in rB_H is

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

We also define $\hat{h}_0 = 0$. The next result shows that $\hat{h}_r$ is an SVM with regularisation parameter $\mu(r)$.

Lemma 3 *Assume (H). Let K be the $n \times n$ symmetric matrix with $K_{i,j} = k(X_i, X_j)$. Then K is an $(\mathbb{R}^{n \times n}, \mathcal{B}(\mathbb{R}^{n \times n}))$-valued measurable matrix on $(\Omega, \mathcal{F})$ and 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, the diagonal entries of D are non-negative and non-increasing. Let $m = \operatorname{rk} K$, which is a random variable on $(\Omega, \mathcal{F})$. For $r > 0$, if*

$$r^2 < \sum_{i=1}^{m} D_{i,i}^{-1} (A^{\mathsf{T}} Y)_i^2,$$

then define $\mu(r) > 0$ by

$$\sum_{i=1}^{m} \frac{D_{i,i}}{(D_{i,i} + n\mu(r))^2} (A^{\mathsf{T}} Y)_i^2 = r^2. \tag{3}$$

Otherwise, let $\mu(r) = 0$. We have that $\mu(r)$ is strictly decreasing when $\mu(r) > 0$, and $\mu(r)$ is measurable on $(\Omega \times (0, \infty), \mathcal{F} \otimes \mathcal{B}((0, \infty)))$, where r varies in $(0, \infty)$. Let $a \in \mathbb{R}^n$ be defined by

$$(A^{\mathsf{T}} a)_i = (D_{i,i} + n\mu(r))^{-1} (A^{\mathsf{T}} Y)_i$$

for $1 \leq i \leq m$ and $(A^{\mathsf{T}} a)_i = 0$ for $m + 1 \leq i \leq n$, noting that A^{T} has the inverse A since it is an orthogonal matrix. For $r \geq 0$, we can uniquely define $\hat{h}_r$ by demanding that $\hat{h}_r \in \operatorname{sp}\{k_{X_i} : 1 \leq i \leq n\}$. This gives

$$\hat{h}_r = \sum_{i=1}^{n} a_i k_{X_i}$$

for $r > 0$ and $\hat{h}_0 = 0$. 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 $[0, \infty)$.

Let $r > 0$. There are multiple methods for calculating $\mu(r)$ to within a given tolerance $\varepsilon > 0$. We call this value $\nu(r)$.

7.1. Diagonalising K

Firstly, $\mu(r) = 0$ if and only if

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

so in this case we set $\nu(r) = 0$. Otherwise, $\mu(r) > 0$ and

$$\begin{aligned} r^2 &= \sum_{i=1}^{m} \frac{D_{i,i}}{(D_{i,i} + n\mu(r))^2} (A^{\mathsf{T}} Y)_i^2 \\ &\leq n^{-2} \left(\sum_{i=1}^{m} D_{i,i} (A^{\mathsf{T}} Y)_i^2 \right) \mu(r)^{-2}. \end{aligned}$$

Hence,

$$\mu(r) \leq n^{-1} \left(\sum_{i=1}^{m} D_{i,i} (A^\mathsf{T} Y)_i^2 \right)^{1/2} r^{-1}. \tag{4}$$

The function

$$\sum_{i=1}^{m} \frac{D_{i,i}}{(D_{i,i} + n\mu)^2} (A^\mathsf{T} Y)_i^2$$

of $\mu \geq 0$ is continuous. Hence, we can calculate $\nu(r)$ using interval bisection on the interval with lower end point 0 and upper end point equal to the right-hand side of (4). We can then approximate a by replacing $\mu(r)$ with $\nu(r)$ in the calculation of a in Lemma 3.

7.2. Not Diagonalising K

We can calculate an alternative $\nu(r)$ without diagonalising K. Note that if $\mu(r) > 0$, then (3) can be written as

$$Y^\mathsf{T} (K + n\mu(r)I)^{-1} K (K + n\mu(r)I)^{-1} Y = r^2.$$

Since $\mu(r)$ is strictly decreasing for $\mu(r) > 0$, we have

$$r \geq \left(Y^\mathsf{T} (K + n\varepsilon I)^{-1} K (K + n\varepsilon I)^{-1} Y \right)^{1/2}$$

if and only if $\mu(r) \in [0, \varepsilon]$, so in this case we set $\nu(r) = \varepsilon$. Otherwise, $\mu(r) > \varepsilon$ and (4) can be written as

$$\mu(r) \leq n^{-1} (Y^\mathsf{T} K Y)^{1/2} r^{-1}. \tag{5}$$

The function

$$Y^\mathsf{T} (K + n\mu I)^{-1} K (K + n\mu I)^{-1} Y$$

of $\mu > 0$ is continuous. Hence, we can calculate $\nu(r)$ using interval bisection on the interval with lower end point ε and upper end point equal to the right-hand side of (5). When $\mu(r) > 0$ or K is invertible, we can also calculate a in Lemma 3 using $a = (K + n\mu(r)I)^{-1} Y$. Since $\nu(r) > 0$, we can approximate a by $(K + n\nu(r)I)^{-1} Y$.

If we have that K is invertible, then we can calculate the $\nu(r)$ in Subsection 7.1 while still not diagonalising K. We have $\mu(r) = 0$ if and only if $r \geq (Y^\mathsf{T} K^{-1} Y)^{1/2}$, so in this case we set $\nu(r) = 0$. Otherwise, $\mu(r) > 0$ and (4) can be written as

$$\mu(r) \leq n^{-1} (Y^\mathsf{T} K Y)^{1/2} r^{-1},$$

so we can again use interval bisection to calculate $\nu(r)$. We can still approximate a by $(K + n\nu(r)I)^{-1} Y$.

7.3. Approximating $\hat{h}_r$

Having discussed how to approximate $\mu(r)$ by $\nu(r)$ to within a given tolerance $\varepsilon > 0$, we now consider the estimator produced by this approximation. We find that this estimator is equal to $\hat{h}_s$ for some $s > 0$. We only have $\nu(r) = 0$ for the methods considered above when $\mu(r) = 0$, in which case we can let $s = r$ to obtain the approximate estimator $\hat{h}_s = \hat{h}_r$. Otherwise, $\nu(r) > 0$. Let

$$s = \left(\sum_{i=1}^{m} \frac{D_{i,i}}{(D_{i,i} + n\nu(r))^2} (A^\mathsf{T} Y)_i^2 \right)^{1/2} .$$

By (3), we have $\mu(s) = \nu(r)$ and the approximate estimator is equal to $\hat{h}_s$. Assume that r is bounded away from 0 as $n \to \infty$ and let $C > 0$ be some constant not depending on n. We can ensure that s is of the same order as r as $n \to \infty$ by demanding that s is within C of r. This is enough to ensure that the orders of convergence for $\hat{h}_r$ apply to $\hat{h}_s$. In order to attain this value of $\nu(r)$, interval bisection should terminate at $x \in \mathbb{R}$ such that

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

is within C of r. Note that this guarantees $\|\hat{h}_s - \hat{h}_r\|_H \leq C^{1/2}(r+s)^{1/2}$ by Lemma 32.

8. Expectation Bounds

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

$$I_2(g, r) = \inf \left\{ \|h_r - g\|_{L^2(P)}^2 : h_r \in rB_H \right\}$$

for $r > 0$. We consider the distance of g from rB_H because we constrain our estimator $\hat{h}_r$ to lie in this set. The supremum in (1) with $A = rB_H$ can be controlled using the reproducing kernel property and the Cauchy–Schwarz inequality to obtain

$$8r \left(\frac{1}{n^2} \sum_{i,j=1}^{n} (Y_i - g(X_i))(Y_j - g(X_j)) k(X_i, X_j) \right)^{1/2} .$$

The expectation of this quantity can be bounded using Jensen's inequality. Something very similar to this argument gives the first term of the bound in Theorem 4 below. The expectation of the supremum in (2) with $A = rB_H$ can be controlled using symmetrisation (Lemma 2.3.1 of van der Vaart and Wellner, 1996) to obtain

$$2\,\mathbb{E} \left(\sup_{f \in 2rB_H} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i)^2 \right| \right) ,$$

where the ε_i for $1 \le i \le n$ are i.i.d. Rademacher random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, independent of the X_i. Since $\|f\|_\infty \le 2\|k\|_\infty r$ for all $f \in 2rB_H$, we can remove the squares on the $f(X_i)$ by using the contraction principle for Rademacher processes (Theorem 3.2.1 of Giné and Nickl, 2016). This quantity can then be bounded in a similar way to the supremum in (1), giving the second term of the bound in Theorem 4 below.

Theorem 4 *Assume (Y1) and (H). Let $r > 0$. We have*

$$\mathbb{E}\left(\|\hat{h}_r - g\|^2_{L^2(P)}\right) \le \frac{8\|k\|_\infty \sigma r}{n^{1/2}} + \frac{64\|k\|^2_\infty r^2}{n^{1/2}} + 10 I_2(g, r).$$

We can obtain rates of convergence for our estimator $\hat{h}_r$ if we make an assumption about how well g can be approximated by elements of H. Let us assume

$$(g1) \qquad g \in [L^2(P), H]_{\beta,\infty} \text{ with norm at most } B \text{ for } \beta \in (0,1) \text{ and } B > 0.$$

The assumption $(g1)$, together with Lemma 1, give

$$I_2(g, r) \le \frac{B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}} \tag{6}$$

for $r > 0$. We obtain an expectation bound on the squared $L^2(P)$ error of our estimator $\hat{h}_r$ of order $n^{-\beta/2}$.

Theorem 5 *Assume (Y1), (H) and (g1). Let $r > 0$. We have*

$$\mathbb{E}\left(\|\hat{h}_r - g\|^2_{L^2(P)}\right) \le \frac{8\|k\|_\infty \sigma r}{n^{1/2}} + \frac{64\|k\|^2_\infty r^2}{n^{1/2}} + \frac{10 B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}}.$$

Let $D_1 > 0$. Setting

$$r = D_1 \|k\|_\infty^{-(1-\beta)} B n^{(1-\beta)/4}$$

gives

$$\mathbb{E}\left(\|\hat{h}_r - g\|^2_{L^2(P)}\right) \le D_2 \|k\|^{2\beta}_\infty B^2 n^{-\beta/2} + D_3 \|k\|^\beta_\infty B \sigma n^{-(1+\beta)/4}$$

for constants $D_2, D_3 > 0$ depending only on D_1 and β.

Since we must let $r \to \infty$ for the initial bound in Theorem 5 to tend to 0, the second term of the initial bound is asymptotically larger than the first. If we ignore the first term and minimise the second and third terms over $r > 0$, we get

$$r = \left(\frac{5\beta}{32(1-\beta)}\right)^{(1-\beta)/2} \|k\|_\infty^{-(1-\beta)} B n^{(1-\beta)/4}.$$

In particular, r is of the form in Theorem 5. This choice of r gives

$$D_2 = 64\left(\frac{5\beta}{32(1-\beta)}\right)^{1-\beta} + 10\left(\frac{32(1-\beta)}{5\beta}\right)^{\beta} \text{ and } D_3 = 8\left(\frac{5\beta}{32(1-\beta)}\right)^{(1-\beta)/2}.$$

The fact that the second term of the initial bound is larger than the first produces some interesting observations. Firstly, the choice of r above does not depend on σ^2. Secondly, we can decrease the bound if we can find a way to reduce the second term, without having to alter the other terms. The increased size of the second term is due to the fact that the bound on $f \in 2rB_H$ is given by $\|f\|_\infty \leq 2\|k\|_\infty r$ when applying the contraction principle for Rademacher processes. If we can use a bound which does not depend on r, then we can reduce the size of the second term.

We now also assume

$$\|g\|_\infty \leq C \text{ for } C > 0 \tag{$g2$}$$

and clip our estimator. Let $r > 0$. Since g is bounded in $[-C, C]$, we can make $\hat{h}_r$ closer to g by constraining it to lie in the same interval. Similarly to Chapter 7 of Steinwart and Christmann (2008) and Steinwart et al. (2009), we define the projection $V : \mathbb{R} \to [-C, C]$ by

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

for $t \in \mathbb{R}$. We can apply the inequality

$$\|V\hat{h}_r - Vh_r\|_{L^2(P_n)}^2 \leq \|\hat{h}_r - h_r\|_{L^2(P_n)}^2$$

for all $h_r \in rB_H$. We continue analysing $V\hat{h}_r$ by bounding

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

The expectation of the supremum can be bounded in the same way as before, with some adjustments. After symmetrisation, we can remove the squares on the $Vf_1(X_i) - Vf_2(X_i)$ for $f_1, f_2 \in rB_H$ and $1 \leq i \leq n$ by using the contraction principle for Rademacher processes with $\|Vf_1 - Vf_2\|_\infty \leq 2C$. We can then use the triangle inequality to remove $Vf_2(X_i)$, before applying the contraction principle again to remove V. The expectation bound on the squared $L^2(P)$ error of our estimator $V\hat{h}_r$ follows in the same way as before.

Theorem 6 *Assume (Y1), (H) and (g2). Let $r > 0$. We have*

$$\mathbb{E}\left(\|V\hat{h}_r - g\|_{L^2(P)}^2\right) \leq \frac{8\|k\|_\infty(16C + \sigma)r}{n^{1/2}} + 10I_2(g, r).$$

We can obtain rates of convergence for our estimator $V\hat{h}_r$ by again assuming ($g1$). We obtain an expectation bound on the squared $L^2(P)$ error of $V\hat{h}_r$ of order $n^{-\beta/(1+\beta)}$.

Theorem 7 *Assume (Y1), (H), (g1) and (g2). Let $r > 0$. We have*

$$\mathbb{E}\left(\|V\hat{h}_r - g\|_{L^2(P)}^2\right) \leq \frac{8\|k\|_\infty(16C + \sigma)r}{n^{1/2}} + \frac{10B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}}.$$

Let $D_1 > 0$. Setting

$$r = D_1\|k\|_\infty^{-(1-\beta)/(1+\beta)} B^{2/(1+\beta)} (16C + \sigma)^{-(1-\beta)/(1+\beta)} n^{(1-\beta)/(2(1+\beta))}$$

gives

$$\mathbb{E}\left(\|V\hat{h}_r - g\|^2_{L^2(P)}\right) \leq D_2\|k\|_\infty^{2\beta/(1+\beta)} B^{2/(1+\beta)}(16C+\sigma)^{2\beta/(1+\beta)} n^{-\beta/(1+\beta)}$$

for a constant $D_2 > 0$ depending only on D_1 and β.

If we minimise the initial bound in Theorem 7 over $r > 0$, we get

$$r = \left(\frac{5\beta}{2(1-\beta)}\right)^{(1-\beta)/(1+\beta)} \|k\|_\infty^{-(1-\beta)/(1+\beta)} B^{2/(1+\beta)}(16C+\sigma)^{-(1-\beta)/(1+\beta)} n^{(1-\beta)/(2(1+\beta))}.$$

In particular, r is of the form in Theorem 7. This choice of r gives

$$D_2 = 2 \cdot 5^{(1-\beta)/(1+\beta)} \cdot 4^{2\beta/(1+\beta)} \left(\left(\frac{2\beta}{1-\beta}\right)^{(1-\beta)/(1+\beta)} + \left(\frac{1-\beta}{2\beta}\right)^{2\beta/(1+\beta)}\right).$$

Although the second bound in Theorem 7 is of theoretical interest, it is in practice impossible to select r of the correct order in n for the bound to hold without knowing β. Since assuming that we know β is not realistic, we must use some other method for determining a good choice of r.

8.1. Validation

Suppose that we have an independent second data set $(\tilde{X}_i, \tilde{Y}_i)$ for $1 \leq i \leq \tilde{n}$ which are $(S \times \mathbb{R}, \mathcal{S} \otimes \mathcal{B}(\mathbb{R}))$-valued random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let the $(\tilde{X}_i, \tilde{Y}_i)$ be i.i.d. with $\tilde{X}_i \sim P$ and $\mathbb{E}(\tilde{Y}_i|\tilde{X}_i) = g(\tilde{X}_i)$ almost surely. Let $\rho \geq 0$ and $R \subseteq [0, \rho]$ be non-empty and compact. Furthermore, let $F = \{V\hat{h}_r : r \in R\}$. We estimate a value of r which makes the squared $L^2(P)$ error of $V\hat{h}_r$ small by

$$\hat{r} = \operatorname*{arg\,min}_{r \in R} \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} (V\hat{h}_r(\tilde{X}_i) - \tilde{Y}_i)^2.$$

The minimum is attained because Lemma 32 shows that it is the minimum of a continuous function over a compact set. In the event of ties, we may take $\hat{r}$ to be the infimum of all points attaining the minimum. Lemma 33 shows that the estimator $\hat{r}$ is a random variable on $(\Omega, \mathcal{F})$. Hence, by Lemma 3, $\hat{h}_{\hat{r}}$ is a $(H, \mathcal{B}(H))$-valued random variable on $(\Omega, \mathcal{F})$.

The definition of $\hat{r}$ means that we can analyse $V\hat{h}_{\hat{r}}$ using Lemma 2. The expectation of the supremum in (1) with $A = F$ can be bounded using chaining (Theorem 2.3.7 of Giné and Nickl, 2016). The diameter of $(F, \|\cdot\|_\infty)$ is $2C$, which is an important bound for the use of chaining. Hence, this form of analysis can only be performed under the assumption $(g2)$. After symmetrisation, the expectation of the supremum in (2) with $A = F$ can be bounded in the same way. In order to perform chaining, we need to make an assumption on the behaviour of the errors of the response variables $\tilde{Y}_i$ for $1 \leq i \leq \tilde{n}$. Let U and V be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. We say U is σ^2-subgaussian if

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

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

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

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

$(\tilde{Y})$ $$\tilde{Y}_i - g(\tilde{X}_i) \text{ is } \tilde{\sigma}^2\text{-subgaussian given } \tilde{X}_i \text{ for } 1 \leq i \leq \tilde{n}.$$

This is stronger than the equivalent of the assumption $(Y1)$, that $\operatorname{var}(\tilde{Y}_i|\tilde{X}_i) \leq \tilde{\sigma}^2$ almost surely.

Theorem 8 *Assume (H) and ($\tilde{Y}$). Let $r_0 \in R$. We have*

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

is at most

$$\frac{32C(4C+\tilde{\sigma})}{\tilde{n}^{1/2}}\left(\left(2\log\left(2+\frac{\|k\|^2_\infty \rho^2}{C^2}\right)\right)^{1/2} + \pi^{1/2}\right) + 10\,\mathbb{E}\left(\|V\hat{h}_{r_0} - g\|^2_{L^2(P)}\right).$$

In order for us to apply the validation result in Theorem 8 to the initial bound in Theorem 7, we need to make an assumption on R. We assume either

$(R1)$ $$R = [0, \rho] \text{ for } \rho = an^{1/2} \text{ and } a > 0$$

or

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

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 an expectation bound on the squared $L^2(P)$ error of $V\hat{h}_{\hat{r}}$ of order $n^{-\beta/(1+\beta)}$. This is the same order in n as the second bound in Theorem 7.

Theorem 9 *Assume (Y1), (H), (g1), (g2) and ($\tilde{Y}$). Also assume (R1) or (R2) and that $\tilde{n}$ increases at least linearly in n. We have*

$$\mathbb{E}\left(\|V\hat{h}_{\hat{r}} - g\|^2_{L^2(P)}\right) \leq D_1 n^{-\beta/(1+\beta)}$$

for a constant $D_1 > 0$ not depending on n or $\tilde{n}$.

9. High-Probability Bounds

In this section, we look at how to extend our expectation bounds on our estimators to high-probability bounds. In order to do this, we must control the second term of the bound in Lemma 2 with $A = rB_H$ for $r > 0$, which is

$$\|h_r - g\|^2_{L^2(P_n)} \tag{7}$$

for $h_r \in rB_H$. There is no way to bound (7) in high-probability without strict assumptions on g. In fact, the most natural assumption is ($g2$) that $\|g\|_\infty \leq C$ for $C > 0$, which we assume throughout this section. Bounding (7) also requires us to introduce a new measure of how well g can be approximated by elements of H. We define

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

for $r > 0$. Note that $I_\infty(g, r) \geq I_2(g, r)$. Using $I_\infty(g, r)$ instead of $I_2(g, r)$ means that we do not have to control (7) by relying on $\|h_r - g\|_\infty \leq \|k\|_\infty r + C$. Using Hoeffding's inequality, this would add a term of order $r^2 t^{1/2}/n^{1/2}$ for $t \geq 1$ to the bound in Theorem 10 below, which holds with probability $1 - 3e^{-t}$, substantially increasing its size.

It may be possible to avoid this problem by instead considering the Ivanov-regularised least-squares estimator

$$\hat{f}_r = \operatorname*{arg\,min}_{f \in V(rB_H)} \frac{1}{n} \sum_{i=1}^n (f(X_i) - Y_i)^2$$

for $r > 0$, where $V(rB_H) = \{V h_r : h_r \in rB_H\}$. The second term of the bound in Lemma 2 with $A = V(rB_H)$ is

$$\|V h_r - g\|_{L^2(P_n)}^2 \tag{8}$$

for $h_r \in rB_H$. Since $\|V h_r - g\|_\infty \leq 2C$, using Hoeffding's inequality to bound (8) would only add a term of order $C^2 t^{1/2}/n^{1/2}$ to the bound in Theorem 10 below, which would not alter its size. However, the calculation and analysis of the estimator $\hat{f}_r$ is outside the scope of this paper. This is because the calculation of $\hat{f}_r$ involves minimising a quadratic form subject to a series of linear constraints, and its analysis requires a bound on the supremum in (1) with $A = V(rB_H)$.

The rest of the analysis of $V\hat{h}_r$ is similar to that of the expectation bound. The supremum in (1) with $A = rB_H$ can again be bounded by

$$8r \left(\frac{1}{n^2} \sum_{i,j=1}^n (Y_i - g(X_i))(Y_j - g(X_j)) k(X_i, X_j)\right)^{1/2}.$$

The quadratic form can be bounded using Lemma 35, under an assumption on the behaviour of the errors of the response variables Y_i for $1 \leq i \leq n$. The proof of Theorem 10 below uses a very similar argument to this one. The supremum in (2) with $A = rB_H$ can be bounded using Talagrand's inequality (Theorem A.9.1 of Steinwart and Christmann, 2008). In order to use Lemma 35, we must assume

($Y2$) $\quad\quad Y_i - g(X_i)$ is σ^2-subgaussian given X_i for $1 \leq i \leq n$.

This assumption is stronger than ($Y1$). In particular, Theorem 6 still holds under the assumptions ($Y2$), (H) and ($g2$).

Theorem 10 *Assume ($Y2$), (H) and ($g2$). Let $r > 0$ and $t \geq 1$. With probability at least $1 - 3e^{-t}$, we have*

$$\|V\hat{h}_r - g\|_{L^2(P)}^2 \leq \left(D_1 + D_2 r^{1/2} + D_3 r\right) t^{1/2} n^{-1/2} + D_4 t n^{-1} + 10 I_\infty(g, r)$$

for some constants $D_1, D_2, D_3, D_4 > 0$ not depending on n or t.

We can obtain rates of convergence for our estimator $V\hat{h}_r$, but we must make a new assumption about how well g can be approximated by elements of H, instead of ($g1$). We now assume

($g3$) $$g \in [L^\infty, H]_{\beta,\infty} \text{ with norm at most } B \text{ for } \beta \in (0,1) \text{ and } B > 0,$$

instead of $g \in [L^2(P), H]_{\beta,\infty}$ with norm at most B. This assumption is stronger than ($g1$), as it implies that the norm of $g \in [L^2(P), H]_{\beta,\infty}$ is

$$\sup_{t>0}(t^{-\beta} \inf_{h\in H}(\|g-h\|_{L^2(P)} + t\|h\|_H)) \leq \sup_{t>0}(t^{-\beta} \inf_{h\in H}(\|g-h\|_{L^\infty} + t\|h\|_H)) \leq B.$$

In particular, Theorem 7 still holds under the assumptions ($Y1$), (H), ($g2$) and ($g3$) or ($Y2$), (H), ($g2$) and ($g3$). The assumption ($g3$), together with Lemma 1, give

$$I_\infty(g, r) \leq \frac{B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}}. \tag{9}$$

We obtain a high-probability bound on the squared $L^2(P)$ error of $V\hat{h}_r$ of order $t^{\beta/(1+\beta)} n^{-\beta/(1+\beta)}$ with probability at least $1 - e^{-t}$.

Theorem 11 *Assume ($Y2$), (H), ($g2$) and ($g3$). Let $r > 0$ and $t \geq 1$. With probability at least $1 - 3e^{-t}$, we have*

$$\|V\hat{h}_r - g\|^2_{L^2(P)} \leq \left(D_1 + D_2 r^{1/2} + D_3 r\right) t^{1/2} n^{-1/2} + D_4 t n^{-1} + D_5 r^{-2\beta/(1-\beta)}$$

for some constants $D_1, D_2, D_3, D_4, D_5 > 0$ not depending on n or t. Let $D_6 > 0$. Setting

$$r = D_6 \left(\frac{n}{t}\right)^{\frac{1-\beta}{2(1+\beta)}}$$

gives

$$\|V\hat{h}_r - g\|^2_{L^2(P)} \leq D_7 \left(\frac{t}{n}\right)^{\frac{\beta}{1+\beta}} + D_8 \left(\frac{t}{n}\right)^{\frac{1+3\beta}{4(1+\beta)}} + D_1 \left(\frac{t}{n}\right)^{1/2} + D_4 \left(\frac{t}{n}\right)$$

for constants $D_7, D_8 > 0$ not depending on n or t.

Since we must let $r \to \infty$ for the initial bound in Theorem 11 to tend to 0, the asymptotically largest terms in the bound are

$$D_3 r t^{1/2} n^{-1/2} + D_5 r^{-2\beta/(1-\beta)}.$$

If we minimise this over $r > 0$, we get r of the form in Theorem 11 with

$$D_6 = \left(\frac{2D_5\beta}{D_3(1-\beta)}\right)^{(1-\beta)/(1+\beta)}.$$

It is possible to avoid the dependence of r on t by setting

$$r = D_6 n^{(1-\beta)/(2(1+\beta))}$$

to obtain the bound

$$D_7 t^{1/2} n^{-\beta/(1+\beta)} + D_8 t^{1/2} n^{-(1+3\beta)/(4(1+\beta))} + D_1 t^{1/2} n^{-1/2} + D_4 t n^{-1}.$$

9.1. Validation

We now extend our expectation bound on $V\hat{h}_{\hat{r}}$ to a high-probability bound. The supremum in (1) with $A = F$ can be bounded using chaining (Exercise 1 of Section 2.3 of Giné and Nickl, 2016), while the supremum in (2) with $A = F$ can be bounded using Talagrand's inequality.

Theorem 12 *Assume (H) and ($\tilde{Y}$). Let $r_0 \in R$ and $t \geq 1$. With probability at least $1 - 3e^{-t}$, we have*

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

is at most

$$\frac{20C(C+\tilde{\sigma})t^{1/2}}{\tilde{n}^{1/2}}\left(1 + 32\left(\left(2\log\left(2 + \frac{\|k\|^2_\infty \rho^2}{C^2}\right)\right)^{1/2} + \pi^{1/2}\right)\right)$$
$$+ \frac{48C^2t^{1/2}}{\tilde{n}^{1/2}} + \frac{16C^2t}{3\tilde{n}} + 10\|V\hat{h}_{r_0} - g\|^2_{L^2(P)}.$$

We can apply the validation result in Theorem 12 to the initial bound in Theorem 11 by assuming either ($R1$) or ($R2$). We obtain a high-probability bound on 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 $1 - e^{-t}$. This is the same order in n as the second bound in Theorem 11.

Theorem 13 *Assume (Y2), (H), (g2), (g3) and ($\tilde{Y}$). Let $t \geq 1$. Also assume (R1) or (R2) and that $\tilde{n}$ increases at least linearly in n. With probability at least $1 - 6e^{-t}$, we have*

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

for constants $D_1, D_2 > 0$ not depending on n, $\tilde{n}$ or t.

10. Discussion

In this paper, we show how Ivanov regularisation can be used to produce smooth estimators which have a small squared $L^2(P)$ error. We first consider the case in which the regression function lies in an interpolation space between $L^2(P)$ and an RKHS H. We achieve bounds on the squared $L^2(P)$ error under the assumption that H is separable with a bounded and measurable kernel. Under the weak assumption that the response variables have bounded variance, we prove an expectation bound on the squared $L^2(P)$ error of our estimator of order $n^{-\beta/2}$. Here, β parametrises the interpolation space between $L^2(P)$ and H containing the regression function. As far as we are aware, the analysis of an estimator in this setting has not previously been considered.

If we assume that the regression function is bounded, then we can clip the estimator and show that the clipped estimator has an expected squared $L^2(P)$ error of order $n^{-\beta/(1+\beta)}$. Under the stronger assumption that the response variables have subgaussian errors and that the regression function comes from an interpolation space between L^∞ and H, we show that the squared $L^2(P)$ error is of order $n^{-\beta/(1+\beta)}$ with high probability. For the settings in which the regression function is bounded, we can use training and validation on the data set to obtain bounds of the same order of $n^{-\beta/(1+\beta)}$. This allows us to select the size of the norm constraint for our Ivanov regularisation without knowing which interpolation space contains the regression function. The response variables in the validation set must have subgaussian errors.

The expectation bounds of order $n^{-\beta/(1+\beta)}$ for bounded regression functions is optimal in the sense discussed at the end of Section 1. We use Ivanov regularisation instead of Tikhonov regularisation to control empirical processes over balls in the RKHS. By contrast, the analysis of Tikhonov-regularised estimators usually uses the spectral decomposition of the kernel operator (Mendelson and Neeman, 2010; Steinwart et al., 2009). Analysing the Ivanov-regularised estimator using this decomposition would give a more complete picture of the differences between Ivanov and Tikhonov regularisation for RKHS regression.

It would be useful to extend the lower bound of order $n^{-\beta/(1+\beta)}$, discussed at the end of Section 1, to the case in which the regression function lies in an interpolation space between L^∞ and the RKHS. This would show that our high-probability bounds are also of optimal order. However, it is possible that estimation can be performed with a high-probability bound on the squared $L^2(P)$ error of smaller order.

Appendix A. Proof of Expectation Bound for Unbounded Regression Function

The proofs of all of the bounds in this paper follow the outline in Section 6. We first prove Lemma 2.

Proof of Lemma 2 Since $f \in A$, the definition of $\hat{f}$ gives

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(X_i) - Y_i)^2 \leq \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2.$$

Expanding

$$(\hat{f}(X_i) - Y_i)^2 = \Big((\hat{f}(X_i) - f(X_i)) + (f(X_i) - Y_i)\Big)^2,$$

substituting into the above and rearranging gives

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(X_i) - f(X_i))^2 \leq \frac{2}{n}\sum_{i=1}^{n}(Y_i - f(X_i))(\hat{f}(X_i) - f(X_i)).$$

Substituting

$$Y_i - f(X_i) = (Y_i - g(X_i)) + (g(X_i) - f(X_i))$$

into the above and applying the Cauchy–Schwarz inequality to the second term gives

$$\begin{aligned}\|\hat{f} - f\|^2_{L^2(P_n)} \leq{}& \frac{2}{n}\sum_{i=1}^{n}(Y_i - g(X_i))(\hat{f}(X_i) - f(X_i)) \\ &+ 2\|g - f\|_{L^2(P_n)}\|\hat{f} - f\|_{L^2(P_n)}.\end{aligned}$$

For constants $a, b \in \mathbb{R}$ and a variable $x \in \mathbb{R}$, we have

$$x^2 \leq a + 2bx \implies x^2 \leq 2a + 4b^2$$

by completing the square and rearranging. Applying this result to the above inequality proves the lemma. ■

The following lemma is useful for bounding the expectation of both of the suprema in Section 6.

Lemma 14 *Assume (H). Let the ε_i be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{E}(\varepsilon_i|X) = 0$ almost surely and $\operatorname{var}(\varepsilon_i|X) \leq \sigma^2$ almost surely for $1 \leq i \leq n$ and $\operatorname{cov}(\varepsilon_i, \varepsilon_j|X) = 0$ almost surely for $1 \leq i, j \leq n$ with $i \neq j$. Then*

$$\mathbb{E}\left(\sup_{f \in rB_H} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right) \leq \frac{\|k\|_\infty \sigma r}{n^{1/2}}.$$

Proof This proof method is due to Remark 6.1 of Sriperumbudur (2016). By the reproducing kernel property and the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
\sup_{f \in rB_H} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right| &= \sup_{f \in rB_H} \left|\left\langle \frac{1}{n}\sum_{i=1}^n \varepsilon_i k_{X_i}, f \right\rangle_H\right| \\
&= r \left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i k_{X_i}\right\|_H \\
&= r \left(\frac{1}{n^2}\sum_{i,j=1}^n \varepsilon_i \varepsilon_j k(X_i, X_j)\right)^{1/2}.
\end{aligned}$$

By Jensen's inequality, we have

$$\begin{aligned}
\mathbb{E}\left(\sup_{f \in rB_H} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right| \,\middle|\, X\right) &\leq r \left(\frac{1}{n^2}\sum_{i,j=1}^n \operatorname{cov}(\varepsilon_i, \varepsilon_j|X) k(X_i, X_j)\right)^{1/2} \\
&\leq r \left(\frac{\sigma^2}{n^2}\sum_{i=1}^n k(X_i, X_i)\right)^{1/2}
\end{aligned}$$

almost surely and again, by Jensen's inequality, we have

$$\mathbb{E}\left(\sup_{f \in rB_H} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right) \leq r \left(\frac{\sigma^2 \|k\|_\infty^2}{n}\right)^{1/2}.$$

The result follows. ■

We bound the distance between $\hat{h}_r$ and h_r in the $L^2(P_n)$ norm for $r > 0$ and $h_r \in rB_H$.

Lemma 15 *Assume (Y1) and (H). Let $h_r \in rB_H$. We have*

$$\mathbb{E}\left(\|\hat{h}_r - h_r\|_{L^2(P_n)}^2\right) \leq \frac{4\|k\|_\infty \sigma r}{n^{1/2}} + 4\|h_r - g\|_{L^2(P)}^2.$$

Proof By Lemma 2 with $A = rB_H$, we have

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

We now bound the expectation of the right-hand side. We have

$$\mathbb{E}\left(\|h_r - g\|_{L^2(P_n)}^2\right) = \|h_r - g\|_{L^2(P)}^2.$$

Furthermore,

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n (Y_i - g(X_i))h_r(X_i)\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}(Y_i - g(X_i)|X_i)h_r(X_i)\right) = 0.$$

Finally, by Lemma 14 with $\varepsilon_i = Y_i - g(X_i)$, we have

$$\begin{aligned}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n (Y_i - g(X_i))\hat{h}_r(X_i)\right) &\leq \mathbb{E}\left(\sup_{f \in rB_H}\left|\frac{1}{n}\sum_{i=1}^n (Y_i - g(X_i))f(X_i)\right|\right) \\ &\leq \frac{\|k\|_\infty \sigma r}{n^{1/2}}.\end{aligned}$$

The result follows. ■

The following lemma is useful for moving the bound on the distance between $\hat{h}_r$ and h_r from the $L^2(P_n)$ norm to the $L^2(P)$ norm for $r > 0$ and $h_r \in rB_H$.

Lemma 16 *Assume (H). We have*

$$\mathbb{E}\left(\sup_{f \in rB_H}\left|\|f\|_{L^2(P_n)}^2 - \|f\|_{L^2(P)}^2\right|\right) \leq \frac{8\|k\|_\infty^2 r^2}{n^{1/2}}.$$

Proof Let the ε_i for $1 \leq i \leq n$ be i.i.d. Rademacher random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, independent of the X_i. Lemma 2.3.1 of van der Vaart and Wellner (1996) shows

$$\mathbb{E}\left(\sup_{f \in rB_H}\left|\frac{1}{n}\sum_{i=1}^n f(X_i)^2 - \int f^2 dP\right|\right) \leq 2\,\mathbb{E}\left(\sup_{f \in rB_H}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)^2\right|\right)$$

by symmetrisation. Since $|f(X_i)| \leq \|k\|_\infty r$ for all $f \in rB_H$, we find

$$\frac{f(X_i)^2}{2\|k\|_\infty r}$$

is a contraction vanishing at 0 as a function of $f(X_i)$ for all $1 \leq i \leq n$. By Theorem 3.2.1 of Giné and Nickl (2016), we have

$$\mathbb{E}\left(\sup_{f \in rB_H}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\frac{f(X_i)^2}{2\|k\|_\infty r}\right|\middle|X\right) \leq 2\,\mathbb{E}\left(\sup_{f \in rB_H}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\middle|X\right)$$

almost surely. By Lemma 14 with $\sigma^2 = 1$, we have

$$\mathbb{E}\left(\sup_{f \in rB_H}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right) \leq \frac{\|k\|_\infty r}{n^{1/2}}.$$

The result follows. ■

We move the bound on the distance between $\hat{h}_r$ and h_r from the $L^2(P_n)$ norm to the $L^2(P)$ norm for $r > 0$ and $h_r \in rB_H$.

Corollary 17 *Assume (Y1) and (H). Let $h_r \in rB_H$. We have*

$$\mathbb{E}\left(\|\hat{h}_r - h_r\|_{L^2(P)}^2\right) \leq \frac{4\|k\|_\infty \sigma r}{n^{1/2}} + \frac{32\|k\|_\infty^2 r^2}{n^{1/2}} + 4\|h_r - g\|_{L^2(P)}^2.$$

Proof By Lemma 15, we have

$$\mathbb{E}\left(\|\hat{h}_r - h_r\|_{L^2(P_n)}^2\right) \leq \frac{4\|k\|_\infty \sigma r}{n^{1/2}} + 4\|h_r - g\|_{L^2(P)}^2.$$

Since $\hat{h}_r - h_r \in 2rB_H$, by Lemma 16 we have

$$\begin{aligned}\mathbb{E}\left(\|\hat{h}_r - h_r\|_{L^2(P)}^2 - \|\hat{h}_r - h_r\|_{L^2(P_n)}^2\right) &\leq \mathbb{E}\left(\sup_{f \in 2rB_H}\left|\|f\|_{L^2(P_n)}^2 - \|f\|_{L^2(P)}^2\right|\right)\\ &\leq \frac{32\|k\|_\infty^2 r^2}{n^{1/2}}.\end{aligned}$$

The result follows. ■

We bound the distance between $\hat{h}_r$ and g in the $L^2(P)$ norm for $r > 0$ to prove Theorem 4.

Proof of Theorem 4 Fix $h_r \in rB_H$. We have

$$\begin{aligned}\|\hat{h}_r - g\|_{L^2(P)}^2 &\leq \left(\|\hat{h}_r - h_r\|_{L^2(P)}^2 + \|h_r - g\|_{L^2(P)}^2\right)^2\\ &\leq 2\|\hat{h}_r - h_r\|_{L^2(P)}^2 + 2\|h_r - g\|_{L^2(P)}^2.\end{aligned}$$

By Corollary 17, we have

$$\mathbb{E}\left(\|\hat{h}_r - h_r\|_{L^2(P)}^2\right) \leq \frac{4\|k\|_\infty \sigma r}{n^{1/2}} + \frac{32\|k\|_\infty^2 r^2}{n^{1/2}} + 4\|h_r - g\|_{L^2(P)}^2.$$

Hence,

$$\mathbb{E}\left(\|\hat{h}_r - g\|_{L^2(P)}^2\right) \leq \frac{8\|k\|_\infty \sigma r}{n^{1/2}} + \frac{64\|k\|_\infty^2 r^2}{n^{1/2}} + 10\|h_r - g\|_{L^2(P)}^2.$$

Taking an infimum over $h_r \in rB_H$ proves the result. ■

We assume $(g1)$ to prove Theorem 5.

Proof of Theorem 5 The initial bound follows from Theorem 4 and (6). Based on this bound, setting

$$r = D_1 \|k\|_\infty^{-(1-\beta)} B n^{(1-\beta)/4}$$

gives

$$\mathbb{E}\left(\|\hat{h}_r - g\|_{L^2(P)}^2\right) \leq \left(64D_1^2 + 10D_1^{-2\beta/(1-\beta)}\right) \|k\|_\infty^{2\beta} B^2 n^{-\beta/2} + 8D_1 \|k\|_\infty^\beta B\sigma n^{-(1+\beta)/4}.$$

Hence, the next bound follows with

$$D_2 = 64D_1^2 + 10D_1^{-2\beta/(1-\beta)} \text{ and } D_3 = 8D_1.$$

■

Appendix B. Proof of Expectation Bound for Bounded Regression Function

We can obtain a bound on the distance between $V\hat{h}_r$ and Vh_r in the $L^2(P_n)$ norm for $r > 0$ and $h_r \in rB_H$ from Lemma 15. The following lemma is useful for moving the bound on the distance between $V\hat{h}_r$ and Vh_r from the $L^2(P_n)$ norm to the $L^2(P)$ norm.

Lemma 18 *Assume (H). We have*

$$\mathbb{E}\left(\sup_{f_1,f_2 \in rB_H} \left| \|Vf_1 - Vf_2\|_{L^2(P_n)}^2 - \|Vf_2 - Vf_2\|_{L^2(P)}^2 \right|\right) \leq \frac{64\|k\|_\infty Cr}{n^{1/2}}.$$

Proof Let the ε_i for $1 \leq i \leq n$ be i.i.d. Rademacher random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, independent of the X_i. Lemma 2.3.1 of van der Vaart and Wellner (1996) shows

$$\mathbb{E}\left(\sup_{f_1,f_2 \in rB_H} \left| \frac{1}{n}\sum_{i=1}^n (Vf_1(X_i) - Vf_2(X_i))^2 - \int (Vf_1 - Vf_2)^2 dP \right|\right)$$

is at most

$$2\,\mathbb{E}\left(\sup_{f_1,f_2 \in rB_H} \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i (Vf_1(X_i) - Vf_2(X_i))^2 \right|\right)$$

by symmetrisation. Since $|Vf_1(X_i) - Vf_2(X_i)| \leq 2C$ for all $f_1, f_2 \in rB_H$, we find

$$\frac{(Vf_1(X_i) - Vf_2(X_i))^2}{4C}$$

is a contraction vanishing at 0 as a function of $Vf_1(X_i) - Vf_2(X_i)$ for all $1 \leq i \leq n$. By Theorem 3.2.1 of Giné and Nickl (2016), we have

$$\mathbb{E}\left(\sup_{f_1,f_2 \in rB_H} \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i \frac{(Vf_1(X_i) - Vf_2(X_i))^2}{4C} \right| \,\middle|\, X\right)$$

is at most

$$2\,\mathbb{E}\left(\sup_{f_1,f_2\in rB_H}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i(Vf_1(X_i)-Vf_2(X_i))\right|\,\middle|X\right)$$

almost surely. Therefore,

$$\mathbb{E}\left(\sup_{f_1,f_2\in rB_H}\left|\frac{1}{n}\sum_{i=1}^n(Vf_1(X_i)-Vf_2(X_i))^2-\int(Vf_1-Vf_2)^2dP\right|\right)$$

is at most

$$16C\,\mathbb{E}\left(\sup_{f_1,f_2\in rB_H}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i(Vf_1(X_i)-Vf_2(X_i))\right|\right)\le 32C\,\mathbb{E}\left(\sup_{f\in rB_H}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_iVf(X_i)\right|\right)$$

by the triangle inequality. Again, by Theorem 3.2.1 of Giné and Nickl (2016), we have

$$\mathbb{E}\left(\sup_{f_1,f_2\in rB_H}\left|\frac{1}{n}\sum_{i=1}^n(Vf_1(X_i)-Vf_2(X_i))^2-\int(Vf_1-Vf_2)^2dP\right|\right)$$

is at most

$$64C\,\mathbb{E}\left(\sup_{f\in rB_H}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_if(X_i)\right|\right)$$

since V is a contraction vanishing at 0. The result follows from Lemma 14 with $\sigma^2=1$. ■

We move the bound on the distance between $V\hat{h}_r$ and Vh_r from the $L^2(P_n)$ norm to the $L^2(P)$ norm for $r>0$ and $h_r\in rB_H$.

Corollary 19 *Assume (Y1) and (H). Let $h_r\in rB_H$. We have*

$$\mathbb{E}\left(\|V\hat{h}_r-Vh_r\|^2_{L^2(P)}\right)\le\frac{4\|k\|_\infty(16C+\sigma)r}{n^{1/2}}+4\|h_r-g\|^2_{L^2(P)}.$$

Proof By Lemma 15, we have

$$\mathbb{E}\left(\|\hat{h}_r-h_r\|^2_{L^2(P_n)}\right)\le\frac{4\|k\|_\infty\sigma r}{n^{1/2}}+4\|h_r-g\|^2_{L^2(P)},$$

so

$$\mathbb{E}\left(\|V\hat{h}_r-Vh_r\|^2_{L^2(P_n)}\right)\le\frac{4\|k\|_\infty\sigma r}{n^{1/2}}+4\|h_r-g\|^2_{L^2(P)}.$$

Since $\hat{h}_r,h_r\in rB_H$, by Lemma 18 we have

$$\begin{aligned}&\mathbb{E}\left(\|V\hat{h}_r-Vh_r\|^2_{L^2(P)}-\|V\hat{h}_r-Vh_r\|^2_{L^2(P_n)}\right)\\ \le\ &\mathbb{E}\left(\sup_{f_1,f_2\in rB_H}\left|\|Vf_1-Vf_2\|^2_{L^2(P_n)}-\|Vf_2-Vf_2\|^2_{L^2(P)}\right|\right)\\ \le\ &\frac{64\|k\|_\infty Cr}{n^{1/2}}.\end{aligned}$$

The result follows. ■

We assume $(g2)$ to bound the distance between $V\hat{h}_r$ and g in the $L^2(P)$ norm for $r > 0$ and prove Theorem 6.

Proof of Theorem 6 Fix $h_r \in rB_H$. We have

$$\begin{aligned}\|V\hat{h}_r - g\|_{L^2(P)}^2 &\leq \left(\|V\hat{h}_r - Vh_r\|_{L^2(P)}^2 + \|Vh_r - g\|_{L^2(P)}^2\right)^2 \\ &\leq 2\|V\hat{h}_r - Vh_r\|_{L^2(P)}^2 + 2\|Vh_r - g\|_{L^2(P)}^2 \\ &\leq 2\|V\hat{h}_r - Vh_r\|_{L^2(P)}^2 + 2\|h_r - g\|_{L^2(P)}^2.\end{aligned}$$

By Corollary 19, we have

$$\mathbb{E}\left(\|V\hat{h}_r - Vh_r\|_{L^2(P)}^2\right) \leq \frac{4\|k\|_\infty(16C + \sigma)r}{n^{1/2}} + 4\|h_r - g\|_{L^2(P)}^2.$$

Hence,

$$\mathbb{E}\left(\|V\hat{h}_r - g\|_{L^2(P)}^2\right) \leq \frac{8\|k\|_\infty(16C + \sigma)r}{n^{1/2}} + 10\|h_r - g\|_{L^2(P)}^2.$$

Taking an infimum over $h_r \in rB_H$ proves the result. ■

We assume $(g1)$ to prove Theorem 7.

Proof of Theorem 7 The initial bound follows from Theorem 6 and (6). Based on this bound, setting

$$r = D_1\|k\|_\infty^{-(1-\beta)/(1+\beta)} B^{2/(1+\beta)} (16C + \sigma)^{-(1-\beta)/(1+\beta)} n^{(1-\beta)/(2(1+\beta))}$$

gives

$$\mathbb{E}\left(\|V\hat{h}_r - g\|_{L^2(P)}^2\right)$$

is at most

$$\left(8D_1 + 10D_1^{-2\beta/(1-\beta)}\right)\|k\|_\infty^{2\beta/(1+\beta)} B^{2/(1+\beta)} (16C + \sigma)^{2\beta/(1+\beta)} n^{-\beta/(1+\beta)}.$$

Hence, the next bound follows with

$$D_2 = 8D_1 + 10D_1^{-2\beta/(1-\beta)}.$$

■

Appendix C. Proof of Expectation Bound for Validation

We need to introduce some definitions for stochastic processes. A stochastic process W on $(\Omega, \mathcal{F})$ indexed by a metric space (M, d) is d^2-subgaussian if it is centred and $W(s) - W(t)$ is $d(s,t)^2$-subgaussian for all $s, t \in M$. W is separable if there exists a countable set $M_0 \subseteq M$ such that the following holds for all $\omega \in \Omega_0$, where $\mathbb{P}(\Omega_0) = 1$. For all $s \in M$ and $\varepsilon > 0$, $W(s)$ is in the closure of $\{W(t) : t \in M_0, d(s,t) \leq \varepsilon\}$.

We also need to introduce the concept of covering numbers for the next result. The covering number $N(M, d, \varepsilon)$ is the minimum number of d-balls of size $\varepsilon > 0$ needed to cover M.

The following lemma is useful for bounding the expectation of both of the suprema in Section 6.

Lemma 20 *Assume (H). Let the ε_i be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $(\tilde{X}_i, \varepsilon_i)$ is i.i.d. for $1 \leq i \leq \tilde{n}$ and let ε_i be $\tilde{\sigma}^2$-subgaussian given $\tilde{X}_i$. Let $r_0 \in R$, $f_0 = V\hat{h}_{r_0}$ and*

$$W(f) = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \varepsilon_i (f(\tilde{X}_i) - f_0(\tilde{X}_i))$$

for $f \in F$. Then W is $\tilde{\sigma}^2 \|\cdot\|_\infty^2 / \tilde{n}$-subgaussian given $\tilde{X}$ and separable on $(F, \tilde{\sigma}\|\cdot\|_\infty / \tilde{n}^{1/2})$. Furthermore,

$$\mathbb{E}\left(\sup_{f \in F} |W(f)|\right) \leq \frac{4C\tilde{\sigma}}{\tilde{n}^{1/2}} \left(\left(2 \log\left(2 + \frac{\|k\|_\infty^2 \rho^2}{C^2}\right)\right)^{1/2} + \pi^{1/2}\right).$$

Proof Let $W_i(f) = \varepsilon_i(f(\tilde{X}_i) - f_0(\tilde{X}_i))$ for $1 \leq i \leq \tilde{n}$ and $f \in F$. Note that the W_i are independent and centred. Since $W_i(f_1) - W_i(f_2)$ is $\tilde{\sigma}^2 \|f_1 - f_2\|_\infty^2$-subgaussian given $\tilde{X}_i$ for $1 \leq i \leq \tilde{n}$ and $f_1, f_2 \in F$, the process W is $\tilde{\sigma}^2 \|\cdot\|_\infty^2 / \tilde{n}$-subgaussian given $\tilde{X}$. By Lemma 32, we have that $(F, \tilde{\sigma}\|\cdot\|_\infty / \tilde{n}^{1/2})$ is separable. Hence, W is separable on $(F, \tilde{\sigma}\|\cdot\|_\infty / \tilde{n}^{1/2})$ since it is continuous. The diameter of $(F, \tilde{\sigma}\|\cdot\|_\infty / \tilde{n}^{1/2})$ is

$$D = \sup_{f_1, f_2 \in F} \tilde{\sigma}\|f_1 - f_2\|_\infty / \tilde{n}^{1/2} \leq 2C\tilde{\sigma}/\tilde{n}^{1/2}.$$

We have

$$\begin{aligned}
\int_0^\infty (\log(N(F, \tilde{\sigma}\|\cdot\|_\infty / \tilde{n}^{1/2}, \varepsilon)))^{1/2} d\varepsilon &= \int_0^\infty (\log(N(F, \|\cdot\|_\infty, \tilde{n}^{1/2}\varepsilon/\tilde{\sigma})))^{1/2} d\varepsilon \\
&= \frac{\tilde{\sigma}}{\tilde{n}^{1/2}} \int_0^\infty (\log(N(F, \|\cdot\|_\infty, u)))^{1/2} du.
\end{aligned}$$

This is finite by Lemma 37. Hence, by Theorem 2.3.7 of Giné and Nickl (2016) and Lemma 37, we have

$$\mathbb{E}\left(\sup_{f \in F} |W(f)| \middle| \tilde{X}, X, Y\right)$$

is at most

$$\begin{aligned}
&\mathbb{E}(|W(f_0)|\,|\tilde{X},X,Y) + 2^{5/2}\int_0^{C\tilde{\sigma}/\tilde{n}^{1/2}} (\log(2N(F,\tilde{\sigma}\|\cdot\|_\infty/\tilde{n}^{1/2},\varepsilon)))^{1/2} d\varepsilon \\
&= 2^{5/2}\int_0^{C\tilde{\sigma}/\tilde{n}^{1/2}} (\log(2N(F,\|\cdot\|_\infty,\tilde{n}^{1/2}\varepsilon/\tilde{\sigma})))^{1/2} d\varepsilon \\
&= \frac{2^{5/2}\tilde{\sigma}}{\tilde{n}^{1/2}}\int_0^{C} (\log(2N(F,\|\cdot\|_\infty,u)))^{1/2} du \\
&\leq \frac{2^{5/2}\tilde{\sigma}}{\tilde{n}^{1/2}}\left(\left(\log\left(2+\frac{\|k\|_\infty^2\rho^2}{C^2}\right)\right)^{1/2} C + \left(\frac{\pi}{2}\right)^{1/2} C\right) \\
&= \frac{4C\tilde{\sigma}}{\tilde{n}^{1/2}}\left(\left(2\log\left(2+\frac{\|k\|_\infty^2\rho^2}{C^2}\right)\right)^{1/2} + \pi^{1/2}\right)
\end{aligned}$$

almost surely, noting $W(f_0) = 0$. The result follows. ■

We bound the distance between $V\hat{h}_{\hat{r}}$ and $V\hat{h}_{r_0}$ in the $L^2(\tilde{P}_{\tilde{n}})$ norm for $r_0 \in R$.

Lemma 21 *Assume (H) and ($\tilde{Y}$). Let $r_0 \in R$. We have*

$$\mathbb{E}\left(\|V\hat{h}_{\hat{r}} - V\hat{h}_{r_0}\|^2_{L^2(\tilde{P}_{\tilde{n}})}\right)$$

is at most

$$\frac{16C\tilde{\sigma}}{\tilde{n}^{1/2}}\left(\left(2\log\left(2+\frac{\|k\|_\infty^2\rho^2}{C^2}\right)\right)^{1/2} + \pi^{1/2}\right) + 4\,\mathbb{E}\left(\|V\hat{h}_{r_0} - g\|^2_{L^2(P)}\right).$$

Proof By Lemma 2 with $A = F$ and n, X, Y and P_n replaced by $\tilde{n}$, $\tilde{X}$, $\tilde{Y}$ and $\tilde{P}_{\tilde{n}}$, we have

$$\|V\hat{h}_{\hat{r}} - V\hat{h}_{r_0}\|^2_{L^2(\tilde{P}_{\tilde{n}})} \leq \frac{4}{\tilde{n}}\sum_{i=1}^{\tilde{n}}(\tilde{Y}_i - g(\tilde{X}_i))(V\hat{h}_{\hat{r}}(\tilde{X}_i) - V\hat{h}_{r_0}(\tilde{X}_i)) + 4\|V\hat{h}_{r_0} - g\|^2_{L^2(\tilde{P}_{\tilde{n}})}.$$

We now bound the expectation of the right-hand side. We have

$$\mathbb{E}\left(\|V\hat{h}_{r_0} - g\|^2_{L^2(\tilde{P}_{\tilde{n}})}\right) = \mathbb{E}\left(\|V\hat{h}_{r_0} - g\|^2_{L^2(P)}\right).$$

Let $f_0 = V\hat{h}_{r_0}$. By Lemma 20 with $\varepsilon_i = Y_i - g(X_i)$, we have

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}(\tilde{Y}_i - g(\tilde{X}_i))(V\hat{h}_{\hat{r}}(\tilde{X}_i) - V\hat{h}_{r_0}(\tilde{X}_i))\right) \\
&\leq \mathbb{E}\left(\sup_{f\in F}\left|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}(\tilde{Y}_i - g(\tilde{X}_i))(f(\tilde{X}_i) - f_0(\tilde{X}_i))\right|\right) \\
&\leq \frac{4C\tilde{\sigma}}{\tilde{n}^{1/2}}\left(\left(2\log\left(2+\frac{\|k\|_\infty^2\rho^2}{C^2}\right)\right)^{1/2} + \pi^{1/2}\right).
\end{aligned}$$

The result follows. ∎

The following lemma is useful for moving the bound on the distance between $V\hat{h}_{\hat{r}}$ and $V\hat{h}_{r_0}$ from the $L^2(\tilde{P}_{\tilde{n}})$ norm to the $L^2(P)$ norm for $r_0 \in R$.

Lemma 22 *Assume (H). Let $r_0 \in R$ and $f_0 = V\hat{h}_{r_0}$. We have*

$$\mathbb{E}\left(\sup_{f\in F}\left|\|f-f_0\|^2_{L^2(\tilde{P}_{\tilde{n}})} - \|f-f_0\|^2_{L^2(P)}\right|\right) \leq \frac{64C^2}{\tilde{n}^{1/2}}\left(\left(2\log\left(2+\frac{\|k\|^2_\infty \rho^2}{C^2}\right)\right)^{1/2} + \pi^{1/2}\right).$$

Proof Let the ε_i be i.i.d. Rademacher random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ for $1 \leq i \leq \tilde{n}$, independent of $\tilde{X}$, X and Y. Lemma 2.3.1 of van der Vaart and Wellner (1996) shows

$$\mathbb{E}\left(\sup_{f\in F}\left|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}(f(\tilde{X}_i)-f_0(\tilde{X}_i))^2 - \int (f-f_0)^2 dP\right| \,\middle|\, X, Y\right)$$

is at most

$$2\,\mathbb{E}\left(\sup_{f\in F}\left|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\varepsilon_i(f(\tilde{X}_i)-f_0(\tilde{X}_i))^2\right| \,\middle|\, X, Y\right)$$

almost surely by symmetrisation. Since $|f(\tilde{X}_i) - f_0(\tilde{X}_i)| \leq 2C$ for all $f \in F$, we find

$$\frac{(f(\tilde{X}_i)-f_0(\tilde{X}_i))^2}{4C}$$

is a contraction vanishing at 0 as a function of $f(\tilde{X}_i) - f_0(\tilde{X}_i)$ for all $1 \leq i \leq \tilde{n}$. By Theorem 3.2.1 of Giné and Nickl (2016), we have

$$\mathbb{E}\left(\sup_{f\in F}\left|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\varepsilon_i\frac{(f(\tilde{X}_i)-f_0(\tilde{X}_i))^2}{4C}\right| \,\middle|\, \tilde{X}, X, Y\right)$$

is at most

$$2\,\mathbb{E}\left(\sup_{f\in F}\left|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\varepsilon_i(f(\tilde{X}_i)-f_0(\tilde{X}_i))\right| \,\middle|\, \tilde{X}, X, Y\right)$$

almost surely. Therefore,

$$\mathbb{E}\left(\sup_{f\in F}\left|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}(f(\tilde{X}_i)-f_0(\tilde{X}_i))^2 - \int (f-f_0)^2 dP\right| \,\middle|\, X, Y\right)$$

is at most

$$16C\,\mathbb{E}\left(\sup_{f\in F}\left|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\varepsilon_i(f(\tilde{X}_i)-f_0(\tilde{X}_i))\right| \,\middle|\, X, Y\right)$$

almost surely. The result follows from Lemma 20 with $\tilde{\sigma}^2 = 1$. ∎

We move the bound on the distance between $V\hat{h}_{\hat{r}}$ and $V\hat{h}_{r_0}$ from the $L^2(\tilde{P}_{\tilde{n}})$ norm to the $L^2(P)$ norm for $r_0 \in R$.

Corollary 23 *Assume (H) and ($\tilde{Y}$). Let $r_0 \in R$. We have*

$$\mathbb{E}\left(\|V\hat{h}_{\hat{r}} - V\hat{h}_{r_0}\|_{L^2(P)}^2\right)$$

is at most

$$\frac{16C(4C+\tilde{\sigma})}{\tilde{n}^{1/2}}\left(\left(2\log\left(2+\frac{\|k\|_\infty^2\rho^2}{C^2}\right)\right)^{1/2}+\pi^{1/2}\right)+4\,\mathbb{E}\left(\|V\hat{h}_{r_0}-g\|_{L^2(P)}^2\right).$$

Proof By Lemma 21, we have

$$\mathbb{E}\left(\|V\hat{h}_{\hat{r}} - V\hat{h}_{r_0}\|_{L^2(\tilde{P}_{\tilde{n}})}^2\right)$$

is at most

$$\frac{16C\tilde{\sigma}}{\tilde{n}^{1/2}}\left(\left(2\log\left(2+\frac{\|k\|_\infty^2\rho^2}{C^2}\right)\right)^{1/2}+\pi^{1/2}\right)+4\,\mathbb{E}\left(\|V\hat{h}_{r_0}-g\|_{L^2(P)}^2\right).$$

Let $f_0 = V\hat{h}_{r_0}$. Since $\hat{h}_{\hat{r}} \in F$, by Lemma 22 we have

$$\begin{aligned}
&\mathbb{E}\left(\|V\hat{h}_{\hat{r}} - V\hat{h}_{r_0}\|_{L^2(P)}^2 - \|V\hat{h}_{\hat{r}} - V\hat{h}_{r_0}\|_{L^2(\tilde{P}_{\tilde{n}})}^2\right)\\
&\leq \mathbb{E}\left(\sup_{f\in F}\left|\|f-f_0\|_{L^2(\tilde{P}_{\tilde{n}})}^2 - \|f-f_0\|_{L^2(P)}^2\right|\right)\\
&\leq \frac{64C^2}{\tilde{n}^{1/2}}\left(\left(2\log\left(2+\frac{\|k\|_\infty^2\rho^2}{C^2}\right)\right)^{1/2}+\pi^{1/2}\right).
\end{aligned}$$

The result follows. ■

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

Proof of Theorem 8 We have

$$\begin{aligned}
\|V\hat{h}_{\hat{r}} - g\|_{L^2(P)}^2 &\leq \left(\|V\hat{h}_{\hat{r}} - Vh_{r_0}\|_{L^2(P)}^2 + \|Vh_{r_0} - g\|_{L^2(P)}^2\right)^2\\
&\leq 2\|V\hat{h}_{\hat{r}} - Vh_{r_0}\|_{L^2(P)}^2 + 2\|Vh_{r_0} - g\|_{L^2(P)}^2.
\end{aligned}$$

By Corollary 23, we have

$$\mathbb{E}\left(\|V\hat{h}_{\hat{r}} - V\hat{h}_{r_0}\|_{L^2(P)}^2\right)$$

is at most

$$\frac{16C(4C+\tilde{\sigma})}{\tilde{n}^{1/2}}\left(\left(2\log\left(2+\frac{\|k\|_\infty^2\rho^2}{C^2}\right)\right)^{1/2}+\pi^{1/2}\right)+4\,\mathbb{E}\left(\|V\hat{h}_{r_0}-g\|_{L^2(P)}^2\right).$$

Hence,

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

is at most

$$\frac{32C(4C+\tilde{\sigma})}{\tilde{n}^{1/2}}\left(\left(2\log\left(2+\frac{\|k\|_\infty^2\rho^2}{C^2}\right)\right)^{1/2}+\pi^{1/2}\right)+10\,\mathbb{E}\left(\|V\hat{h}_{r_0}-g\|_{L^2(P)}^2\right).$$

■

We assume the conditions of Theorem 7 to prove Theorem 9.

Proof of Theorem 9 If we assume $(R1)$, then $r_0 = an^{(1-\beta)/(2(1+\beta))} \in R$ and

$$\mathbb{E}\left(\|V\hat{h}_{r_0}-g\|_{L^2(P)}^2\right) \le \frac{8\|k\|_\infty(16C+\sigma)an^{(1-\beta)/(2(1+\beta))}}{n^{1/2}} + \frac{10B^{2/(1-\beta)}}{a^{2\beta/(1-\beta)}n^{\beta/(1+\beta)}}$$

by Theorem 7. If we assume $(R2)$, then there is at least one $r_0 \in R$ such that

$$an^{(1-\beta)/(2(1+\beta))} \le r_0 < an^{(1-\beta)/(2(1+\beta))} + b$$

and

$$\begin{aligned}\mathbb{E}\left(\|V\hat{h}_{r_0}-g\|_{L^2(P)}^2\right) &\le \frac{8\|k\|_\infty(16C+\sigma)r_0}{n^{1/2}} + \frac{10B^{2/(1-\beta)}}{r_0^{2\beta/(1-\beta)}}\\ &\le \frac{8\|k\|_\infty(16C+\sigma)\left(an^{(1-\beta)/(2(1+\beta))}+b\right)}{n^{1/2}} + \frac{10B^{2/(1-\beta)}}{a^{2\beta/(1-\beta)}n^{\beta/(1+\beta)}}\end{aligned}$$

by Theorem 7. In either case,

$$\mathbb{E}\left(\|V\hat{h}_{r_0}-g\|_{L^2(P)}^2\right) \le D_2 n^{-\beta/(1+\beta)}$$

for some constant $D_2 > 0$ not depending on n or $\tilde{n}$. By Theorem 8, we have

$$\mathbb{E}\left(\|V\hat{h}_{\hat{r}}-g\|_{L^2(P)}^2\right) \le D_3\log(n)^{1/2}\tilde{n}^{-1/2} + 10D_2 n^{-\beta/(1+\beta)}$$

for some constant $D_3 > 0$ not depending on n or $\tilde{n}$. Since $\tilde{n}$ increases at least linearly in n, there exists some constant $D_4 > 0$ such that $\tilde{n} \ge D_4 n$. We then have

$$\begin{aligned}\mathbb{E}\left(\|V\hat{h}_{\hat{r}}-g\|_{L^2(P)}^2\right) &\le D_4^{-1/2}D_3\log(n)^{1/2}n^{-1/2} + 10D_2 n^{-\beta/(1+\beta)}\\ &\le D_1 n^{-\beta/(1+\beta)}\end{aligned}$$

for some constant $D_1 > 0$ not depending on n or $\tilde{n}$. ■

Appendix D. Proof of High-Probability Bound for Bounded Regression Function

We bound the distance between $\hat{h}_r$ and h_r in the $L^2(P_n)$ norm for $r > 0$ and $h_r \in rB_H$.

Lemma 24 *Assume (Y2) and (H). Let $r > 0$, $h_r \in rB_H$ and $t \geq 1$. With probability at least $1 - 2e^{-t}$, we have*

$$\|\hat{h}_r - h_r\|_{L^2(P_n)}^2 \leq \frac{20\|k\|_\infty \sigma r t^{1/2}}{n^{1/2}} + 4\|h_r - g\|_\infty^2.$$

Proof By Lemma 2 with $A = rB_H$, we have

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

We now bound the right-hand side. We have

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

Furthermore,

$$-\frac{1}{n}\sum_{i=1}^n (Y_i - g(X_i))h_r(X_i)$$

is subgaussian given X with parameter

$$\frac{1}{n^2}\sum_{i=1}^n \sigma^2 h_r(X_i)^2 \leq \frac{\|k\|_\infty^2 \sigma^2 r^2}{n}.$$

So we have

$$-\frac{1}{n}\sum_{i=1}^n (Y_i - g(X_i))h_r(X_i) \leq \frac{\|k\|_\infty \sigma r (2t)^{1/2}}{n^{1/2}} \leq \frac{2\|k\|_\infty \sigma r t^{1/2}}{n^{1/2}}$$

with probability at least $1 - e^{-t}$ by Chernoff bounding. Finally, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n (Y_i - g(X_i))\hat{h}_r(X_i) &\leq \sup_{f \in rB_H} \left| \frac{1}{n}\sum_{i=1}^n (Y_i - g(X_i)) f(X_i) \right| \\
&= \sup_{f \in rB_H} \left| \left\langle \frac{1}{n}\sum_{i=1}^n (Y_i - g(X_i)) k_{X_i}, f \right\rangle_H \right| \\
&= r \left\| \frac{1}{n}\sum_{i=1}^n (Y_i - g(X_i)) k_{X_i} \right\|_H \\
&= r \left(\frac{1}{n^2} \sum_{i,j=1}^n (Y_i - g(X_i))(Y_j - g(X_j)) k(X_i, X_j) \right)^{1/2}
\end{aligned}$$

by the reproducing kernel property and the Cauchy–Schwarz inequality. Let K be the $n \times n$ matrix with $K_{i,j} = k(X_i, X_j)$ and let ε be the vector of the $Y_i - g(X_i)$. Then

$$\frac{1}{n^2} \sum_{i,j=1}^n (Y_i - g(X_i))(Y_j - g(X_j)) k(X_i, X_j) = \varepsilon^\mathsf{T} (n^{-2} K) \varepsilon.$$

Furthermore, since k is a measurable function on $(S \times S, \mathcal{S} \otimes \mathcal{S})$, we have that $n^{-2}K$ is an $(\mathbb{R}^{n\times n}, \mathcal{B}(\mathbb{R}^{n\times n}))$-valued measurable matrix on $(\Omega, \mathcal{F})$ and non-negative-definite. Let a_i for $1 \le i \le n$ be the eigenvalues of $n^{-2}K$. Then

$$\max_i a_i \le \operatorname{tr}(n^{-2}K) \le n^{-1}\|k\|_\infty^2$$

and

$$\operatorname{tr}((n^{-2}K)^2) = \|a\|_2^2 \le \|a\|_1^2 \le n^{-2}\|k\|_\infty^4.$$

Therefore, by Lemma 35 with $M = n^{-2}K$, we have

$$\varepsilon^{\mathsf{T}}(n^{-2}K)\varepsilon \le \|k\|_\infty^2 \sigma^2 n^{-1}(1 + 2t + 2(t^2 + t)^{1/2})$$

and

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - g(X_i))\hat{h}_r(X_i) \le \frac{3\|k\|_\infty \sigma r t^{1/2}}{n^{1/2}}$$

with probability at least $1 - e^{-t}$. The result follows. ■

The following lemma is useful for bounding the supremum in (2).

Lemma 25 *Let $D > 0$ and $A \subseteq L^\infty$ be separable with $\|f\|_\infty \le D$ for all $f \in A$. Let*

$$Z = \sup_{f\in A}\left|\|f\|_{L^2(P_n)}^2 - \|f\|_{L^2(P)}^2\right|.$$

Then, for $t > 0$, we have

$$Z \le \mathbb{E}(Z) + \left(\frac{2D^4 t}{n} + \frac{4D^2\,\mathbb{E}(Z) t}{n}\right)^{1/2} + \frac{2D^2 t}{3n}$$

with probability at least $1 - e^{-t}$.

Proof We have

$$Z = \sup_{f\in A}\left|\sum_{i=1}^{n} n^{-1}\left(f(X_i)^2 - \|f\|_{L^2(P)}^2\right)\right|$$

and

$$\mathbb{E}\left(n^{-1}\left(f(X_i)^2 - \|f\|_{L^2(P)}^2\right)\right) = 0,$$
$$n^{-1}\left|f(X_i)^2 - \|f\|_{L^2(P)}^2\right| \le \frac{D^2}{n},$$
$$\mathbb{E}\left(n^{-2}\left(f(X_i)^2 - \|f\|_{L^2(P)}^2\right)^2\right) \le \frac{D^4}{n^2}$$

for all $1 \le i \le n$ and $f \in A$. Furthermore, A is separable, so Z is a random variable on $(\Omega, \mathcal{F})$ and we can use Talagrand's inequality (Theorem A.9.1 of Steinwart and Christmann, 2008) to show

$$Z > \mathbb{E}(Z) + \left(2t\left(\frac{D^4}{n} + \frac{2D^2\,\mathbb{E}(Z)}{n}\right)\right)^{1/2} + \frac{2tD^2}{3n}$$

with probability at most e^{-t}. The result follows. ■

The following lemma is useful for moving the bound on the distance between $V\hat{h}_r$ and Vh_r from the $L^2(P_n)$ norm to the $L^2(P)$ norm for $r > 0$ and $h_r \in rB_H$.

Lemma 26 *Assume (H). Let $r > 0$ and $t \geq 1$. With probability at least $1 - e^{-t}$, we have*

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

is at most

$$\frac{8\left(C^2 + 4\|k\|_\infty^{1/2} C^{3/2} r^{1/2} + 8\|k\|_\infty Cr\right) t^{1/2}}{n^{1/2}} + \frac{8C^2 t}{3n}.$$

Proof Let $A = \{Vf_1 - Vf_2 : f_1, f_2 \in rB_H\}$ and

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

Then $A \subseteq L^\infty$ is separable because H is separable and has a bounded kernel k. Furthermore, $\|Vf_1 - Vf_2\|_\infty \leq 2C$ for all $f_1, f_2 \in rB_H$. By Lemma 25, we have

$$Z \leq \mathbb{E}(Z) + \left(\frac{32C^4 t}{n} + \frac{16C^2\,\mathbb{E}(Z)t}{n}\right)^{1/2} + \frac{8C^2 t}{3n}$$

with probability at least $1 - e^{-t}$. By Lemma 18, we have

$$\mathbb{E}(Z) \leq \frac{64\|k\|_\infty Cr}{n^{1/2}}.$$

The result follows. ■

We move the bound on the distance between $V\hat{h}_r$ and Vh_r from the $L^2(P_n)$ norm to the $L^2(P)$ norm for $r > 0$ and $h_r \in rB_H$.

Corollary 27 *Assume (Y2) and (H). Let $r > 0$, $h_r \in rB_H$ and $t \geq 1$. With probability at least $1 - 3e^{-t}$, we have*

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

is at most

$$\frac{4\left(2C^2 + 8\|k\|_\infty^{1/2} C^{3/2} r^{1/2} + \|k\|_\infty (16C + 5\sigma) r\right) t^{1/2}}{n^{1/2}} + \frac{8C^2 t}{3n} + 4\|h_r - g\|_\infty^2.$$

Proof By Lemma 24, we have

$$\|\hat{h}_r - h_r\|_{L^2(P_n)}^2 \leq \frac{20\|k\|_\infty \sigma r t^{1/2}}{n^{1/2}} + 4\|h_r - g\|_\infty^2.$$

with probability at least $1 - 2e^{-t}$, so

$$\|V\hat{h}_r - Vh_r\|_{L^2(P_n)}^2 \leq \frac{20\|k\|_\infty \sigma r t^{1/2}}{n^{1/2}} + 4\|h_r - g\|_\infty^2.$$

Since $\hat{h}_r, h_r \in rB_H$, by Lemma 26 we have

$$\begin{aligned}
&\|V\hat{h}_r - Vh_r\|^2_{L^2(P)} - \|V\hat{h}_r - Vh_r\|^2_{L^2(P_n)} \\
\leq &\sup_{f_1, f_2 \in rB_H} \left| \|Vf_1 - Vf_2\|^2_{L^2(P_n)} - \|Vf_2 - Vf_2\|^2_{L^2(P)} \right| \\
\leq &\frac{8\left(C^2 + 4\|k\|_\infty^{1/2} C^{3/2} r^{1/2} + 8\|k\|_\infty Cr\right) t^{1/2}}{n^{1/2}} + \frac{8C^2 t}{3n}
\end{aligned}$$

with probability at least $1 - e^{-t}$. The result follows. ∎

We assume $(g2)$ to bound the distance between $V\hat{h}_r$ and g in the $L^2(P)$ norm for $r > 0$ and prove Theorem 10.

Proof of Theorem 10 Fix $h_r \in rB_H$. We have

$$\begin{aligned}
\|V\hat{h}_r - g\|^2_{L^2(P)} &\leq \left(\|V\hat{h}_r - Vh_r\|^2_{L^2(P)} + \|Vh_r - g\|^2_{L^2(P)}\right)^2 \\
&\leq 2\|V\hat{h}_r - Vh_r\|^2_{L^2(P)} + 2\|Vh_r - g\|^2_{L^2(P)} \\
&\leq 2\|V\hat{h}_r - Vh_r\|^2_{L^2(P)} + 2\|h_r - g\|^2_{L^2(P)}.
\end{aligned}$$

By Corollary 27, we have

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

is at most

$$\frac{4\left(2C^2 + 8\|k\|_\infty^{1/2} C^{3/2} r^{1/2} + \|k\|_\infty (16C + 5\sigma) r\right) t^{1/2}}{n^{1/2}} + \frac{8C^2 t}{3n} + 4\|h_r - g\|^2_\infty.$$

with probability at least $1 - 3e^{-t}$. Hence,

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

is at most

$$\frac{8\left(2C^2 + 8\|k\|_\infty^{1/2} C^{3/2} r^{1/2} + \|k\|_\infty (16C + 5\sigma) r\right) t^{1/2}}{n^{1/2}} + \frac{16C^2 t}{3n} + 10\|h_r - g\|^2_\infty.$$

Take a sequence of $h_{r,n} \in rB_H$ for $n \geq 1$ with

$$\|h_{r,n} - g\|^2_\infty \downarrow I_\infty(g, r)$$

as $n \to \infty$ and let $A_n \in \mathcal{F}$ be the set with $\mathbb{P}(A_n) \geq 1 - 3e^{-t}$ on which the above inequality holds with $h_r = h_{r,n}$. Since the A_n are non-increasing sets, we have that

$$\mathbb{P}\left(\bigcap_{n=1}^\infty A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n) \geq 1 - 3e^{-t}.$$

The result follows with

$$D_1 = 16C^2,\ D_2 = 64\|k\|_\infty^{1/2} C^{3/2},\ D_3 = 8\|k\|_\infty (16C + 5\sigma) \text{ and } D_4 = 16C^2/3.$$

■

We assume $(g3)$ to prove Theorem 11.

Proof of Theorem 11 The initial bound follows from Theorem 10 and (9) with $D_5 = 10B^{2/(1-\beta)}$. Based on this bound, setting

$$r = D_6 t^{-(1-\beta)/(2(1+\beta))} n^{(1-\beta)/(2(1+\beta))}$$

gives

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

is at most

$$\left(D_3 D_6 + D_5 D_6^{-2\beta/(1-\beta)}\right) t^{\beta/(1+\beta)} n^{-\beta/(1+\beta)} + D_2 D_6^{1/2} t^{(1+3\beta)/(4(1+\beta))} n^{-(1+3\beta)/(4(1+\beta))}$$
$$+ D_1 t^{1/2} n^{-1/2} + D_4 t n^{-1}.$$

Hence, the next bound follows with

$$D_7 = D_3 D_6 + D_5 D_6^{-2\beta/(1-\beta)} \text{ and } D_8 = D_2 D_6^{1/2}.$$

■

Appendix E. Proof of High-Probability Bound for Validation

We need to introduce some new notation for the next result. Let U and V be random variables on $(\Omega, \mathcal{F})$. Then

$$\|U\|_{\psi_2} = \inf\{a \in (0,\infty) : \mathbb{E}\,\psi_2(|U|/a) \le 1\},$$
$$\|U|V\|_{\psi_2} = \inf\{a \in (0,\infty) : \mathbb{E}(\psi_2(|U|/a)|V) \le 1 \text{ almost surely}\},$$

where $\psi_2(x) = \exp(x^2) - 1$ for $x \in \mathbb{R}$. Note that these infima are attained by the monotone convergence theorem. Exercise 5 of Section 2.3 of Giné and Nickl (2016) shows that $\|U\|_{\psi_2}$ is a norm on the space of U such that $\|U\|_{\psi_2} < \infty$ and $\|U|V\|_{\psi_2}$ is a norm on the space of U such that $\|U|V\|_{\psi_2} < \infty$.

We bound the distance between $V\hat{h}_{\hat{r}}$ and $V\hat{h}_{r_0}$ in the $L^2(\tilde{P}_{\tilde{n}})$ norm for $r_0 \in R$.

Lemma 28 *Assume (H) and ($\tilde{Y}$). Let $r_0 \in R$ and $t \ge 1$. With probability at least $1 - 2e^{-t}$, we have*

$$\|V\hat{h}_{\hat{r}} - V\hat{h}_{r_0}\|_{L^2(\tilde{P}_{\tilde{n}})}^2$$

is at most

$$\frac{292 C\tilde{\sigma} t^{1/2}}{\tilde{n}^{1/2}} \left(\left(2\log\left(1 + \frac{\|k\|_\infty^2 \rho^2}{8C^2}\right)\right)^{1/2} + \pi^{1/2}\right) + \frac{24 C^2 t^{1/2}}{\tilde{n}^{1/2}} + 4\|V\hat{h}_{r_0} - g\|_{L^2(P)}^2.$$

Proof By Lemma 2 with $A = F$ and n, X, Y and P_n replaced by $\tilde{n}$, $\tilde{X}$, $\tilde{Y}$ and $\tilde{P}_{\tilde{n}}$, we have

$$\|V\hat{h}_{\hat{r}} - V\hat{h}_{r_0}\|^2_{L^2(\tilde{P}_{\tilde{n}})} \leq \frac{4}{\tilde{n}} \sum_{i=1}^{\tilde{n}} (\tilde{Y}_i - g(\tilde{X}_i))(V\hat{h}_{\hat{r}}(\tilde{X}_i) - V\hat{h}_{r_0}(\tilde{X}_i)) + 4\|V\hat{h}_{r_0} - g\|^2_{L^2(\tilde{P}_{\tilde{n}})}.$$

We now bound the right-hand side. We have

$$\|V\hat{h}_{r_0} - g\|^2_{L^2(\tilde{P}_{\tilde{n}})} = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \left((V\hat{h}_{r_0}(\tilde{X}_i) - g(\tilde{X}_i))^2 - \|V\hat{h}_{r_0} - g\|^2_{L^2(P)} \right) + \|V\hat{h}_{r_0} - g\|^2_{L^2(P)}.$$

Since

$$\left| (V\hat{h}_{r_0}(\tilde{X}_i) - g(\tilde{X}_i))^2 - \|V\hat{h}_{r_0} - g\|^2_{L^2(P)} \right| \leq 4C^2$$

for all $1 \leq i \leq \tilde{n}$, we find

$$\|V\hat{h}_{r_0} - g\|^2_{L^2(\tilde{P}_{\tilde{n}})} - \|V\hat{h}_{r_0} - g\|^2_{L^2(P)} > t$$

with probability at most

$$\exp\left(-\frac{\tilde{n}t^2}{32C^4} \right).$$

by Hoeffding's inequality. Therefore, we have

$$\|V\hat{h}_{r_0} - g\|^2_{L^2(\tilde{P}_{\tilde{n}})} - \|V\hat{h}_{r_0} - g\|^2_{L^2(P)} \leq \frac{32^{1/2}C^2t^{1/2}}{\tilde{n}^{1/2}} \leq \frac{6C^2t^{1/2}}{\tilde{n}^{1/2}}$$

with probability at least $1 - e^{-t}$. Now let $f_0 = V\hat{h}_{r_0}$ and

$$W(f) = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} (\tilde{Y}_i - g(\tilde{X}_i))(f(\tilde{X}_i) - f_0(\tilde{X}_i))$$

for $f \in F$. W is $\tilde{\sigma}^2\|\cdot\|^2_\infty/\tilde{n}$-subgaussian given $\tilde{X}$ and separable on $(F, \tilde{\sigma}\|\cdot\|_\infty/\tilde{n}^{1/2})$ by Lemma 20. The diameter of $(F, \tilde{\sigma}\|\cdot\|_\infty/\tilde{n}^{1/2})$ is

$$D = \sup_{f_1, f_2 \in F} \tilde{\sigma}\|f_1 - f_2\|_\infty/\tilde{n}^{1/2} \leq 2C\tilde{\sigma}/\tilde{n}^{1/2}.$$

From Lemma 37, we have

$$\begin{aligned} \int_0^\infty (\log(N(F, \tilde{\sigma}\|\cdot\|_\infty/\tilde{n}^{1/2}, \varepsilon)))^{1/2} d\varepsilon &= \int_0^\infty (\log(N(F, \|\cdot\|_\infty, \tilde{n}^{1/2}\varepsilon/\tilde{\sigma})))^{1/2} d\varepsilon \\ &= \frac{\tilde{\sigma}}{\tilde{n}^{1/2}} \int_0^\infty (\log(N(F, \|\cdot\|_\infty, u)))^{1/2} du \end{aligned}$$

is finite. Hence, by Exercise 1 of Section 2.3 of Giné and Nickl (2016) and Lemma 37, we have

$$\left\| \sup_{f \in F} |W(f)| \,\middle|\, \tilde{X}, X, Y \right\|_{\psi_2}$$

is at most

$$
\begin{aligned}
&\left\| W(f_0) \middle| \tilde{X}, X, Y \right\|_{\psi_2} + 1536^{1/2} \int_0^{2C\tilde{\sigma}/\tilde{n}^{1/2}} (\log N(F, \tilde{\sigma}\|\cdot\|_\infty/\tilde{n}^{1/2}, \varepsilon))^{1/2} d\varepsilon \\
&= 1536^{1/2} \int_0^{2C\tilde{\sigma}/\tilde{n}^{1/2}} (\log N(F, \|\cdot\|_\infty, \tilde{n}^{1/2}\varepsilon/\tilde{\sigma}))^{1/2} d\varepsilon \\
&= \frac{1536^{1/2}\tilde{\sigma}}{\tilde{n}^{1/2}} \int_0^{2C} (\log N(F, \|\cdot\|_\infty, u))^{1/2} du \\
&\leq \frac{1536^{1/2}\tilde{\sigma}}{\tilde{n}^{1/2}} \left(2 \left(\log \left(1 + \frac{\|k\|_\infty^2 \rho^2}{8C^2} \right) \right)^{1/2} C + (2\pi)^{1/2} C \right) \\
&= \frac{3072^{1/2} C\tilde{\sigma}}{\tilde{n}^{1/2}} \left(\left(2 \log \left(1 + \frac{\|k\|_\infty^2 \rho^2}{8C^2} \right) \right)^{1/2} + \pi^{1/2} \right),
\end{aligned}
$$

noting $W(f_0) = 0$. By Chernoff bounding, we have $\sup_{f \in F} |W(f)|$ is at most

$$
\begin{aligned}
&\frac{3072^{1/2} C\tilde{\sigma}(t + \log(2))^{1/2}}{\tilde{n}^{1/2}} \left(\left(2 \log \left(1 + \frac{\|k\|_\infty^2 \rho^2}{8C^2} \right) \right)^{1/2} + \pi^{1/2} \right) \\
&\leq \frac{73 C\tilde{\sigma} t^{1/2}}{\tilde{n}^{1/2}} \left(\left(2 \log \left(1 + \frac{\|k\|_\infty^2 \rho^2}{8C^2} \right) \right)^{1/2} + \pi^{1/2} \right)
\end{aligned}
$$

with probability at least $1 - e^{-t}$. In particular,

$$
\frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} (\tilde{Y}_i - g(\tilde{X}_i))(V\hat{h}_{\hat{r}}(\tilde{X}_i) - V\hat{h}_{r_0}(\tilde{X}_i))
$$

is at most

$$
\frac{73 C\tilde{\sigma} t^{1/2}}{\tilde{n}^{1/2}} \left(\left(2 \log \left(1 + \frac{\|k\|_\infty^2 \rho^2}{8C^2} \right) \right)^{1/2} + \pi^{1/2} \right)
$$

with probability at least $1 - e^{-t}$. The result follows. ■

The following lemma is useful for moving the bound on the distance between $V\hat{h}_{\hat{r}}$ and $V\hat{h}_{r_0}$ from the $L^2(\tilde{P}_{\tilde{n}})$ norm to the $L^2(P)$ norm for $r_0 \in R$.

Lemma 29 *Assume (H). Let $r_0 \in R$, $f_0 = V\hat{h}_{r_0}$ and $t \geq 1$. With probability at least $1 - e^{-t}$, we have*

$$
\sup_{f \in F} \left| \|f - f_0\|_{L^2(\tilde{P}_{\tilde{n}})}^2 - \|f - f_0\|_{L^2(P)}^2 \right|
$$

is at most

$$
\frac{10C^2 t^{1/2}}{\tilde{n}^{1/2}} \left(1 + 32 \left(\left(2 \log \left(2 + \frac{\|k\|_\infty^2 \rho^2}{C^2} \right) \right)^{1/2} + \pi^{1/2} \right) \right) + \frac{8C^2 t}{3\tilde{n}}.
$$

Proof Let $A = \{f - f_0 : f \in F\}$ and

$$Z = \sup_{f \in F} \left| \|f - f_0\|^2_{L^2(\tilde{P}_{\tilde{n}})} - \|f - f_0\|^2_{L^2(P)} \right|.$$

Then $A \subseteq L^\infty$ is separable by Lemma 32. Furthermore, $\|f - f_0\|_\infty \leq 2C$ for all $f \in F$. By Lemma 25 with n and P_n replaced by $\tilde{n}$ and $\tilde{P}_{\tilde{n}}$, we have

$$Z \leq \mathbb{E}(Z) + \left(\frac{32C^4 t}{\tilde{n}} + \frac{16C^2\,\mathbb{E}(Z)t}{\tilde{n}} \right)^{1/2} + \frac{8C^2 t}{3\tilde{n}}$$

with probability at least $1 - e^{-t}$. By Lemma 22, we have

$$\mathbb{E}(Z) \leq \frac{64C^2}{\tilde{n}^{1/2}} \left(\left(2 \log \left(2 + \frac{\|k\|^2_\infty \rho^2}{C^2} \right) \right)^{1/2} + \pi^{1/2} \right).$$

The result follows. ■

We move the bound on the distance between $V\hat{h}_{\hat{r}}$ and $V\hat{h}_{r_0}$ from the $L^2(\tilde{P}_{\tilde{n}})$ norm to the $L^2(P)$ norm for $r_0 \in R$.

Corollary 30 *Assume (H) and ($\tilde{Y}$). Let $r_0 \in R$ and $t \geq 1$. With probability at least $1 - 3e^{-t}$, we have*

$$\|V\hat{h}_{\hat{r}} - V\hat{h}_{r_0}\|^2_{L^2(P)}$$

is at most

$$\frac{10C(C + \tilde{\sigma})t^{1/2}}{\tilde{n}^{1/2}} \left(1 + 32 \left(\left(2 \log \left(2 + \frac{\|k\|^2_\infty \rho^2}{C^2} \right) \right)^{1/2} + \pi^{1/2} \right) \right)$$
$$+ \frac{24C^2 t^{1/2}}{\tilde{n}^{1/2}} + \frac{8C^2 t}{3\tilde{n}} + 4\|V\hat{h}_{r_0} - g\|^2_{L^2(P)}.$$

Proof By Lemma 28, we have

$$\|V\hat{h}_{\hat{r}} - V\hat{h}_{r_0}\|^2_{L^2(\tilde{P}_{\tilde{n}})}$$

is at most

$$\frac{292C\tilde{\sigma}t^{1/2}}{\tilde{n}^{1/2}} \left(\left(2 \log \left(1 + \frac{\|k\|^2_\infty \rho^2}{8C^2} \right) \right)^{1/2} + \pi^{1/2} \right) + \frac{24C^2 t^{1/2}}{\tilde{n}^{1/2}} + 4\|V\hat{h}_{r_0} - g\|^2_{L^2(P)}$$

with probability at least $1 - 2e^{-t}$. Let $f_0 = V\hat{h}_{r_0}$. Since $\hat{h}_{\hat{r}} \in F$, by Lemma 29 we have

$$\begin{aligned}
&\|V\hat{h}_{\hat{r}} - V\hat{h}_{r_0}\|^2_{L^2(P)} - \|V\hat{h}_{\hat{r}} - V\hat{h}_{r_0}\|^2_{L^2(\tilde{P}_{\tilde{n}})} \\
\leq\ &\sup_{f \in F} \left| \|f - f_0\|^2_{L^2(\tilde{P}_{\tilde{n}})} - \|f - f_0\|^2_{L^2(P)} \right| \\
\leq\ &\frac{10C^2 t^{1/2}}{\tilde{n}^{1/2}} \left(1 + 32 \left(\left(2 \log \left(2 + \frac{\|k\|^2_\infty \rho^2}{C^2} \right) \right)^{1/2} + \pi^{1/2} \right) \right) + \frac{8C^2 t}{3\tilde{n}}
\end{aligned}$$

with probability at least $1 - e^{-t}$. The result follows. ■

We bound the distance between $V\hat{h}_{\hat{r}}$ and g in the $L^2(P)$ norm to prove Theorem 12.

Proof of Theorem 12 We have

$$\begin{aligned}\|V\hat{h}_{\hat{r}} - g\|_{L^2(P)}^2 &\leq \left(\|V\hat{h}_{\hat{r}} - Vh_{r_0}\|_{L^2(P)}^2 + \|Vh_{r_0} - g\|_{L^2(P)}^2\right)^2 \\ &\leq 2\|V\hat{h}_{\hat{r}} - Vh_{r_0}\|_{L^2(P)}^2 + 2\|Vh_{r_0} - g\|_{L^2(P)}^2.\end{aligned}$$

By Corollary 30, we have

$$\|V\hat{h}_{\hat{r}} - V\hat{h}_{r_0}\|_{L^2(P)}^2$$

is at most

$$\begin{aligned}&\frac{10C(C+\tilde{\sigma})t^{1/2}}{\tilde{n}^{1/2}}\left(1 + 32\left(\left(2\log\left(2 + \frac{\|k\|_\infty^2\rho^2}{C^2}\right)\right)^{1/2} + \pi^{1/2}\right)\right) \\ &+ \frac{24C^2t^{1/2}}{\tilde{n}^{1/2}} + \frac{8C^2t}{3\tilde{n}} + 4\|V\hat{h}_{r_0} - g\|_{L^2(P)}^2.\end{aligned}$$

with probability at least $1 - 3e^{-t}$. Hence,

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

is at most

$$\begin{aligned}&\frac{20C(C+\tilde{\sigma})t^{1/2}}{\tilde{n}^{1/2}}\left(1 + 32\left(\left(2\log\left(2 + \frac{\|k\|_\infty^2\rho^2}{C^2}\right)\right)^{1/2} + \pi^{1/2}\right)\right) \\ &+ \frac{48C^2t^{1/2}}{\tilde{n}^{1/2}} + \frac{16C^2t}{3\tilde{n}} + 10\|V\hat{h}_{r_0} - g\|_{L^2(P)}^2.\end{aligned}$$

The result follows. ■

We assume the conditions of Theorem 11 to prove Theorem 13.

Proof of Theorem 13 If we assume $(R1)$, then $r_0 = an^{(1-\beta)/(2(1+\beta))} \in R$ and

$$\|V\hat{h}_{r_0} - g\|_{L^2(P)}^2$$

is at most

$$\begin{aligned}&\left(D_3 + D_4a^{1/2}n^{(1-\beta)/(4(1+\beta))} + D_5an^{(1-\beta)/(2(1+\beta))}\right)t^{1/2}n^{-1/2} \\ &+ D_6tn^{-1} + D_7a^{-2\beta/(1-\beta)}n^{-\beta/(1+\beta)}\end{aligned}$$

with probability at least $1 - 3e^{-t}$ for some constants $D_3, D_4, D_5, D_6, D_7 > 0$ not depending on n, $\tilde{n}$ or t by Theorem 11. If we assume $(R2)$, then there is at least one $r_0 \in R$ such that

$$an^{(1-\beta)/(2(1+\beta))} \leq r_0 < an^{(1-\beta)/(2(1+\beta))} + b$$

and

$$\|V\hat{h}_{r_0} - g\|_{L^2(P)}^2$$

is at most

$$\begin{aligned}&\left(D_3 + D_4 r_0^{1/2} + D_5 r_0\right) t^{1/2} n^{-1/2} + D_6 t n^{-1} + D_7 r_0^{-2\beta/(1-\beta)} \\ \leq\ &\left(D_3 + D_4\left(a^{1/2} n^{(1-\beta)/(4(1+\beta))} + b^{1/2}\right) + D_5\left(a n^{(1-\beta)/(2(1+\beta))} + b\right)\right) t^{1/2} n^{-1/2} \\ +\ &D_6 t n^{-1} + D_7 a^{-2\beta/(1-\beta)} n^{-\beta/(1+\beta)}\end{aligned}$$

with probability at least $1 - 3e^{-t}$ by Theorem 11. In either case,

$$\|V\hat{h}_{r_0} - g\|_{L^2(P)}^2 \leq D_8 t^{1/2} n^{-\beta/(1+\beta)} + D_9 t n^{-1}$$

for some constants $D_8, D_9 > 0$ not depending on n, $\tilde{n}$ or t. By Theorem 12, we have

$$\|V\hat{h}_{\hat{r}} - g\|_{L^2(P)}^2 \leq D_{10} t^{1/2} \log(n)^{1/2} \tilde{n}^{-1/2} + D_{11} t \tilde{n}^{-1} + 10 D_8 t^{1/2} n^{-\beta/(1+\beta)} + 10 D_9 t n^{-1}$$

with probability at least $1 - 6e^{-t}$ for some constants $D_{10}, D_{11} > 0$ not depending on n, $\tilde{n}$ or t. Since $\tilde{n}$ increases at least linearly in n, there exists some constant $D_{12} > 0$ such that $\tilde{n} \geq D_{12} n$. We then have

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

is at most

$$\begin{aligned}&D_{12}^{-1/2} D_{10} t^{1/2} \log(n)^{1/2} n^{-1/2} + D_{12}^{-1} D_{11} t n^{-1} + 10 D_8 t^{1/2} n^{-\beta/(1+\beta)} + 10 D_9 t n^{-1} \\ \leq\ &D_1 t^{1/2} n^{-\beta/(1+\beta)} + D_2 t n^{-1}\end{aligned}$$

for some constants $D_1, D_2 > 0$ not depending on n, $\tilde{n}$ or t. ■

Appendix F. Estimator Calculation and Measurability

The following result is essentially Theorem 2.1 from Quintana and Rodríguez (2014). The authors show that that a strictly-positive-definite matrix which is a $(\mathbb{C}^{n\times n}, \mathcal{B}(\mathbb{C}^{n\times n}))$-valued measurable matrix on $(\Omega, \mathcal{F})$ can be diagonalised by an unitary matrix and a diagonal matrix which are both $(\mathbb{C}^{n\times n}, \mathcal{B}(\mathbb{C}^{n\times n}))$-valued measurable matrices on $(\Omega, \mathcal{F})$. The result holds for non-negative-definite matrices by adding the identity matrix before diagonalisation and subtracting it afterwards. Furthermore, the construction of the unitary matrix produces a matrix with real entries, which is to say an orthogonal matrix, when the strictly-positive-definite matrix has real entries.

Lemma 31 *Let M be a non-negative-definite matrix which is an $(\mathbb{R}^{n\times n}, \mathcal{B}(\mathbb{R}^{n\times n}))$-valued measurable matrix on $(\Omega, \mathcal{F})$. 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 $M = ADA^{\mathsf{T}}$.*

We prove Lemma 3.

Proof of Lemma 3 Let $H_n = \mathrm{sp}\{k_{X_i} : 1 \leq i \leq n\}$. The subspace H_n is closed in H, so there is an orthogonal projection $Q : H \to H_n$. Since $f - Qf \in H_n^{\perp}$ for all $f \in H$, we have

$$f(X_i) - (Qf)(X_i) = \langle f - Qf, k_{X_i}\rangle = 0$$

for all $1 \leq i \leq n$. Hence,

$$\begin{aligned}\inf_{f \in rB_H} \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2 &= \inf_{f \in rB_H} \frac{1}{n}\sum_{i=1}^{n}((Qf)(X_i) - Y_i)^2 \\ &= \inf_{f \in (rB_H)\cap H_n} \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2.\end{aligned}$$

Let $f \in (rB_H) \cap H_n$ and write

$$f = \sum_{i=1}^{n} a_i k_{X_i}$$

for some $a \in \mathbb{R}^n$. Then

$$\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2 = n^{-1}(Ka - Y)^{\mathsf{T}}(Ka - Y)$$

and $\|f\|_H^2 = a^{\mathsf{T}} K a$, so we can write the norm constraint as $a^{\mathsf{T}} K a + s = r^2$, where $s \geq 0$ is a slack variable. The Lagrangian can be written as

$$\begin{aligned}L(a, s; \mu) &= n^{-1}(Ka - Y)^{\mathsf{T}}(Ka - Y) + \mu(a^{\mathsf{T}} K a + s - r^2) \\ &= a^{\mathsf{T}}(n^{-1}K^2 + \mu K)a - 2n^{-1}Y^{\mathsf{T}}Ka + \mu s + n^{-1}Y^{\mathsf{T}}Y - \mu r^2,\end{aligned}$$

where μ is the Lagrangian multiplier for the norm constraint. We seek to minimise the Lagrangian for a fixed value of μ. Note that we require $\mu \geq 0$ for the Lagrangian to have a finite minimum, due to the term in s. We have

$$\frac{\partial L}{\partial a} = 2(n^{-1}K^2 + \mu K)a - 2n^{-1}KY.$$

This being 0 is equivalent to $K((K + n\mu I)a - Y) = 0$.

Since the kernel k is a measurable function on $(S \times S, \mathcal{S} \otimes \mathcal{S})$ and the X_i are $(S, \mathcal{S})$-valued random variables on $(\Omega, \mathcal{F})$, we find that K is an $(\mathbb{R}^{n\times n}, \mathcal{B}(\mathbb{R}^{n\times n}))$-valued measurable matrix on $(\Omega, \mathcal{F})$. Furthermore, since the kernel k takes real values and is non-negative definite, K is non-negative definite with real entries. By Lemma 31, 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 $K = ADA^{\mathsf{T}}$. Note that the diagonal entries of D must be non-negative and we may assume that they are non-increasing. Inserting this diagonalisation into $K((K + n\mu I)a - Y) = 0$ gives

$$AD((D + n\mu I)A^{\mathsf{T}}a - A^{\mathsf{T}}Y) = 0.$$

Since A has the inverse A^{T}, this is equivalent to

$$D((D + n\mu I)A^{\mathsf{T}}a - A^{\mathsf{T}}Y) = 0.$$

This in turn is equivalent to

$$(A^{\mathsf{T}}a)_i = (D_{i,i} + n\mu)^{-1}(A^{\mathsf{T}}Y)_i$$

for $1 \leq i \leq m$. The same f is produced for all such a, because if w is the difference between two such a, then $(A^\mathsf{T} w)_i = 0$ for $1 \leq i \leq m$ and the squared H norm of

$$\sum_{i=1}^{n} w_i k_{X_i}$$

is $w^\mathsf{T} K w = w^\mathsf{T} A D A^\mathsf{T} w = 0$. Hence, we are free to set $(A^\mathsf{T} a)_i = 0$ for $m+1 \leq i \leq n$. This uniquely defines $A^\mathsf{T} a$, which in turn uniquely defines a, since A^T has the inverse A. Note that this definition of a is measurable on $(\Omega \times [0, \infty), \mathcal{F} \otimes \mathcal{B}([0, \infty)))$, where μ varies in $[0, \infty)$.

We now search for a value of μ such that a and s satisfy the norm constraint. We call this value $\mu(r)$. There are two cases. If

$$r^2 < \sum_{i=1}^{m} D_{i,i}^{-1} (A^\mathsf{T} Y)_i^2,$$

then the a above and $s = 0$ minimise L for $\mu = \mu(r) > 0$ and satisfy the norm constraint, where $\mu(r)$ satisfies

$$\sum_{i=1}^{m} \frac{D_{i,i}}{(D_{i,i} + n\mu(r))^2} (A^\mathsf{T} Y)_i^2 = r^2.$$

Otherwise, the a above and

$$s = r^2 - \sum_{i=1}^{m} D_{i,i}^{-1} (A^\mathsf{T} Y)_i^2 \geq 0$$

minimise L for $\mu = \mu(r) = 0$ and satisfy the norm constraint. Hence, the Lagrangian sufficiency theorem shows

$$\hat{h}_r = \sum_{i=1}^{n} a_i k_{X_i}$$

for the a above with $\mu = \mu(r)$ for $r > 0$. We also have $\hat{h}_0 = 0$.

Since $\mu(r) > 0$ is strictly decreasing for

$$r^2 < \sum_{i=1}^{m} D_{i,i}^{-1} (A^\mathsf{T} Y)_i^2$$

and $\mu(r) = 0$ otherwise, we find

$$\{\mu(r) \leq \mu\} = \left\{ \sum_{i=1}^{m} \frac{D_{i,i}}{(D_{i,i} + n\mu)^2} (A^\mathsf{T} Y)_i^2 \leq r^2 \right\}$$

for $\mu \in [0, \infty)$. Therefore, $\mu(r)$ is measurable on $(\Omega \times [0, \infty), \mathcal{F} \otimes \mathcal{B}((0, \infty)))$, where r varies in $(0, \infty)$. Hence, the a above with $\mu = \mu(r)$ for $r > 0$ is measurable on $(\Omega \times [0, \infty), \mathcal{F} \otimes \mathcal{B}((0, \infty)))$, where r varies in $(0, \infty)$. By Lemma 4.25 of Steinwart and Christmann (2008), the function $\Phi : S \to H$ by $\Phi(x) = k_x$ is a $(H, \mathcal{B}(H))$-valued measurable function on $(S, \mathcal{S})$. Hence, k_{X_i} for $1 \leq i \leq n$ are $(H, \mathcal{B}(H))$-valued random variables on $(\Omega, \mathcal{F})$. Together, these show 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 $[0, \infty)$, recalling that $\hat{h}_0 = 0$. ∎

We prove a continuity result about our estimator.

Lemma 32 *Let $r, s \in [0, \infty)$. We have $\|\hat{h}_r - \hat{h}_s\|_H^2 \leq |r^2 - s^2|$.*

Proof Recall the diagonalisation of $K = ADA^\mathsf{T}$ from Lemma 3. If $u, v \in \mathbb{R}^n$ and

$$h_1 = \sum_{i=1}^n u_i k_{X_i} \text{ and } h_2 = \sum_{i=1}^n v_i k_{X_i},$$

then $\langle h_1, h_2 \rangle_H = u^\mathsf{T} K v = (A^\mathsf{T} u)^\mathsf{T} D (A^\mathsf{T} v)$. Let $s > r$. If $r > 0$ then, by Lemma 3, we have

$$\begin{aligned}
\langle \hat{h}_r, \hat{h}_s \rangle_H &= \sum_{i=1}^m \frac{D_{i,i}}{(D_{i,i} + n\mu(r))(D_{i,i} + n\mu(s))} (A^\mathsf{T} Y)_i^2 \\
&\geq \sum_{i=1}^m \frac{D_{i,i}}{(D_{i,i} + n\mu(r))^2} (A^\mathsf{T} Y)_i^2 \\
&= \|\hat{h}_r\|_H^2.
\end{aligned}$$

Furthermore, again by Lemma 3, if $\mu(r) > 0$ then $\|\hat{h}_r\|_H^2 = r^2$ and

$$\begin{aligned}
\|\hat{h}_r - \hat{h}_s\|_H^2 &= \|\hat{h}_r\|_H^2 + \|\hat{h}_s\|_H^2 - 2\langle \hat{h}_r, \hat{h}_s \rangle_H \\
&\leq \|\hat{h}_s\|_H^2 - \|\hat{h}_r\|_H^2 \\
&= \|\hat{h}_s\|_H^2 - r^2 \\
&\leq s^2 - r^2.
\end{aligned}$$

Otherwise, $\mu(r) = 0$ and so $\mu(s) = 0$ by Lemma 3, which means $\hat{h}_r = \hat{h}_s$. If $r = 0$ then $\hat{h}_r = 0$ and $\|\hat{h}_r - \hat{h}_s\|_H^2 = \|\hat{h}_s\|_H^2 \leq s^2$. Hence, whenever $r < s$, we have $\|\hat{h}_r - \hat{h}_s\|_H^2 \leq s^2 - r^2$. The result follows. ∎

We also have the estimator $\hat{r}$ when performing validation.

Lemma 33 *We have that $\hat{r}$ is a random variable on $(\Omega, \mathcal{F})$.*

Proof Let

$$W(s) = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} (V\hat{h}_s(\tilde{X}_i) - \tilde{Y}_i)^2$$

for $s \in R$. Note that $W(s)$ is a random variable on $(\Omega, \mathcal{F})$ and continuous in s by Lemma 32. Since $R \subseteq \mathbb{R}$, it is separable. Let R_0 be a countable dense subset of R. Then $\inf_{s \in R} W(s) = \inf_{s \in R_0} W(s)$ is a random variable on $(\Omega, \mathcal{F})$ as the right-hand side is the infimum of countably many random variables on $(\Omega, \mathcal{F})$. Let $r \in [0, \rho]$. By the definition of $\hat{r}$, we have

$$\{\hat{r} \leq r\} = \bigcup_{s \in R \cap [0, r]} \{W(s) \leq \inf_{t \in R} W(t)\}.$$

Since $R \cap [0, r] \subseteq \mathbb{R}$, it is separable. Let A_r be a countable dense subset of $R \cap [0, r]$. By the sequential compactness of $R \cap [0, r]$ and continuity of $W(s)$, we have

$$\{\hat{r} \leq r\} = \bigcap_{a=1}^{\infty} \bigcup_{s \in A_r} \{W(s) \leq \inf_{t \in R} W(t) + a^{-1}\}.$$

This set is an element of $\mathcal{F}$. ■

Appendix G. Subgaussian Random Variables

We need the definition of a sub-σ-algebra for the next result. The σ-algebra $\mathcal{G}$ is a sub-σ-algebra of the σ-algebra $\mathcal{F}$ if $\mathcal{G} \subseteq \mathcal{F}$. The following lemma relates a quadratic form of subgaussians to that of centred normal random variables.

Lemma 34 *Let ε_i for $1 \leq i \leq n$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ which are independent conditional on some sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$ and let*

$$\mathbb{E}(\exp(t\varepsilon_i)|\mathcal{G}) \leq \exp(\sigma^2 t^2/2)$$

almost surely for all $t \in \mathbb{R}$. Also, let δ_i for $1 \leq i \leq n$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ which are independent of each other and $\mathcal{G}$ with $\delta_i \sim \mathrm{N}(0, \sigma^2)$. Let M be an $n \times n$ non-negative-definite matrix which is an $(\mathbb{R}^{n \times n}, \mathcal{B}(\mathbb{R}^{n \times n}))$-valued measurable matrix on $(\Omega, \mathcal{G})$. We have

$$\mathbb{E}(\exp(z\varepsilon^\mathsf{T} M \varepsilon)|\mathcal{G}) \leq \mathbb{E}(\exp(z\delta^\mathsf{T} M \delta)|\mathcal{G})$$

almost surely for all $z \geq 0$.

Proof This proof method uses techniques from the proof of Lemma 9 of Abbasi-Yadkori, Pál, and Szepesvári (2011). We have

$$\mathbb{E}(\exp(t_i \varepsilon_i/\sigma)|\mathcal{G}) \leq \exp(t_i^2/2)$$

almost surely for all $1 \leq i \leq n$ and $t_i \in \mathbb{R}$. Furthermore, the ε_i are independent conditional on $\mathcal{G}$, so

$$\mathbb{E}(\exp(t^\mathsf{T} \varepsilon/\sigma)|\mathcal{G}) \leq \exp(\|t\|_2^2/2)$$

almost surely for all $t \in \mathbb{R}^n$. By Lemma 31 with $\mathcal{F}$ replaced by $\mathcal{G}$, 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 $(\Omega, \mathcal{G})$ such that $M = ADA^\mathsf{T}$. Hence, M has a square root $M^{1/2} = AD^{1/2}A^\mathsf{T}$ which is an $(\mathbb{R}^{n \times n}, \mathcal{B}(\mathbb{R}^{n \times n}))$-valued measurable matrix on $(\Omega, \mathcal{G})$, where $D^{1/2}$ is the diagonal matrix with entries equal to the square root of those of D. Note that these entries are non-negative because M is non-negative definite. We can then replace t with $sM^{1/2}u$ for $s \in \mathbb{R}$ and $u \in \mathbb{R}^n$ to get

$$\mathbb{E}(\exp(su^\mathsf{T} M^{1/2} \varepsilon/\sigma)|\mathcal{G}) \leq \exp(s^2 \|M^{1/2}u\|_2^2/2)$$

almost surely. Integrating over u with respect to the distribution of δ gives

$$\mathbb{E}(\exp(s\delta^\mathsf{T} M^{1/2}\varepsilon/\sigma)|\mathcal{G}) \leq \mathbb{E}(\exp(s^2\delta^\mathsf{T} M\delta/2)|\mathcal{G})$$

almost surely. The moment generating function of δ gives the left-hand side as being

$$\mathbb{E}(\exp(s^2\varepsilon^\mathsf{T} M\varepsilon/2)|\mathcal{G})$$

almost surely. The result follows. ■

Having established this relationship, we can now obtain a probability bound on a quadratic form of subgaussians by using Chernoff bounding. The following result is a conditional subgaussian version of the Hanson–Wright inequality.

Lemma 35 *Let ε_i for $1 \leq i \leq n$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ which are independent conditional on some sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$ and let*

$$\mathbb{E}(\exp(t\varepsilon_i)|\mathcal{G}) \leq \exp(\sigma^2 t^2/2)$$

almost surely for all $t \in \mathbb{R}$. Let M be an $n \times n$ non-negative-definite matrix which is an $(\mathbb{R}^{n\times n}, \mathcal{B}(\mathbb{R}^{n\times n}))$-valued measurable matrix on $(\Omega, \mathcal{G})$ and $z \geq 0$. We have

$$\varepsilon^\mathsf{T} M\varepsilon \leq \sigma^2 \operatorname{tr}(M) + 2\sigma^2\|M\|z + 2\sigma^2(\|M\|^2 z^2 + \operatorname{tr}(M^2)z)^{1/2}$$

with probability at least $1 - e^{-z}$ almost surely conditional on $\mathcal{G}$. Here, $\|M\|$ is the operator norm of M, which is a random variable on $(\Omega, \mathcal{G})$.

Proof This proof method follows that of Theorem 3.1.9 of Giné and Nickl (2016). By Lemma 31 with $\mathcal{F}$ replaced by $\mathcal{G}$, 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 $(\Omega, \mathcal{G})$ such that $M = ADA^\mathsf{T}$. Let δ_i for $1 \leq i \leq n$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ which are independent of each other and $\mathcal{G}$, with $\delta_i \sim \mathrm{N}(0, \sigma^2)$. By Lemma 34 and the fact that $A^\mathsf{T}\delta$ has the same distribution as δ, we have

$$\mathbb{E}(\exp(t\varepsilon^\mathsf{T} M\varepsilon)|\mathcal{G}) \leq \mathbb{E}(\exp(t\delta^\mathsf{T} M\delta)|\mathcal{G}) = \mathbb{E}(\exp(t\delta^\mathsf{T} D\delta)|\mathcal{G})$$

almost surely for all $t \geq 0$. Furthermore,

$$\mathbb{E}(\exp(t\delta_i^2/\sigma^2)) = \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{1/2}} \exp(tx^2 - x^2/2)dx = \frac{1}{(1-2t)^{1/2}}$$

for $0 \leq t < 1/2$ and $1 \leq i \leq n$, so

$$\mathbb{E}(\exp(t(\delta_i^2/\sigma^2 - 1))) = \exp(-(\log(1-2t) + 2t)/2).$$

We have

$$-2(\log(1-2t) + 2t) \leq \sum_{i=2}^{\infty} (2t)^i(2/i) \leq 4t^2/(1-2t)$$

for $0 \leq t < 1/2$. Therefore, since the δ_i are independent of $\mathcal{G}$, we have

$$\mathbb{E}(\exp(tD_{i,i}(\delta_i^2 - \sigma^2))|\mathcal{G}) \leq \exp\left(\frac{\sigma^4 D_{i,i}^2 t^2}{1 - 2\sigma^2 D_{i,i} t}\right)$$

almost surely for $0 \leq t < 1/(2\sigma^2 D_{i,i})$ and $1 \leq i \leq n$. Since the $D_{i,i}$ are random variables on $(\Omega, \mathcal{G})$ and the $D_{i,i}\delta_i$ for $1 \leq i \leq n$ are independent conditional on $\mathcal{G}$, we have

$$\mathbb{E}(\exp(t(\delta^\mathsf{T} D\delta - \sigma^2 \operatorname{tr}(D)))|\mathcal{G}) \leq \exp\left(\frac{\sigma^4 \operatorname{tr}(D^2)t^2}{1 - 2\sigma^2(\max_i D_{i,i})t}\right)$$

almost surely for $0 \leq t < 1/(2\sigma^2(\max_i D_{i,i}))$. Combining this with $\mathbb{E}(\exp(t\varepsilon^\mathsf{T} M\varepsilon)|\mathcal{G}) \leq \mathbb{E}(\exp(t\delta^\mathsf{T} D\delta)|\mathcal{G})$, we find

$$\mathbb{E}(\exp(t(\varepsilon^\mathsf{T} M\varepsilon - \sigma^2 \operatorname{tr}(M)))|\mathcal{G}) \leq \exp\left(\frac{\sigma^4 \operatorname{tr}(M^2)t^2}{1 - 2\sigma^2\|M\|t}\right)$$

almost surely for $0 \leq t < 1/(2\sigma^2\|M\|)$. By Chernoff bounding, we have

$$\varepsilon^\mathsf{T} M\varepsilon - \sigma^2 \operatorname{tr}(M) > s$$

for $s \geq 0$ with probability at most

$$\exp\left(\frac{\sigma^4 \operatorname{tr}(M^2)t^2}{1 - 2\sigma^2\|M\|t} - ts\right)$$

almost surely conditional on $\mathcal{G}$ for $0 \leq t < 1/(2\sigma^2\|M\|)$. Letting

$$t = \frac{s}{2\sigma^4 \operatorname{tr}(M^2) + 2\sigma^2\|M\|s}$$

gives the bound

$$\exp\left(-\frac{s^2}{4\sigma^4 \operatorname{tr}(M^2) + 4\sigma^2\|M\|s}\right).$$

Rearranging gives the result. ■

Appendix H. Covering Numbers

The following lemma gives a bound on the covering numbers of F.

Lemma 36 *Let $\varepsilon > 0$. We have*

$$N(F, \|\cdot\|_\infty, \varepsilon) \leq 1 + \frac{\|k\|_\infty^2 \rho^2}{2\varepsilon^2}.$$

Proof Let $a \geq 1$ and $r_i \in R$ and $f_i = V\hat{h}_{r_i} \in F$ for $1 \leq i \leq a$. Also, let $f = V\hat{h}_r \in F$ for $r \in R$. Since V is a contraction, we have $\|f - f_i\|_\infty \leq \varepsilon$ whenever $\|\hat{h}_r - \hat{h}_{r_i}\|_\infty \leq \varepsilon$. By Lemma 32, we have $\|\hat{h}_r - \hat{h}_{r_i}\|_\infty \leq \varepsilon$ whenever $|r^2 - r_i^2| \leq \varepsilon^2/\|k\|_\infty^2$. Hence, if we let $r_i^2 = \varepsilon^2(2i-1)/\|k\|_\infty^2$ and let ρ be such that

$$\rho^2 - \varepsilon^2(2a-1)/\|k\|_\infty^2 \leq \varepsilon^2/\|k\|_\infty^2,$$

then we find $N(F, \|\cdot\|_\infty, \varepsilon) \leq a$. Rearranging the above shows that we can choose

$$a = \left\lceil \frac{\|k\|_\infty^2 \rho^2}{2\varepsilon^2} \right\rceil$$

and the result follows. ■

We also calculate integrals of these covering numbers.

Lemma 37 *Let $a \geq 1$. We have*

$$\int_0^L (\log(aN(F, \|\cdot\|_\infty, \varepsilon)))^{1/2} d\varepsilon \leq \left(\log\left(\left(1 + \frac{\|k\|_\infty^2 \rho^2}{2L^2}\right) a\right)\right)^{1/2} L + \left(\frac{\pi}{2}\right)^{1/2} L$$

for $L \in (0, \infty)$. When $a = 1$, we have

$$\int_0^L (\log(N(F, \|\cdot\|_\infty, \varepsilon)))^{1/2} d\varepsilon \leq 2\left(\log\left(1 + \frac{\|k\|_\infty^2 \rho^2}{8C^2}\right)\right)^{1/2} C + (2\pi)^{1/2} C$$

for $L \in (0, \infty]$.

Proof Let $L \in (0, \infty)$. Then

$$\int_0^L (\log(aN(F, \|\cdot\|_\infty, \varepsilon)))^{1/2} d\varepsilon \leq \int_0^L \left(\log\left(a\left(1 + \frac{\|k\|_\infty^2 \rho^2}{2\varepsilon^2}\right)\right)\right)^{1/2} d\varepsilon$$

by Lemma 36. Changing variables to $u = \varepsilon/L$ gives

$$\begin{aligned} &L \int_0^1 \left(\log\left(a\left(1 + \frac{\|k\|_\infty^2 \rho^2}{2L^2 u^2}\right)\right)\right)^{1/2} du \\ \leq\ &L \int_0^1 \left(\log\left(a\left(1 + \frac{\|k\|_\infty^2 \rho^2}{2L^2}\right)\frac{1}{u^2}\right)\right)^{1/2} du \\ =\ &L \int_0^1 \left(\log\left(a\left(1 + \frac{\|k\|_\infty^2 \rho^2}{2L^2}\right)\right) + \log\left(\frac{1}{u^2}\right)\right)^{1/2} du. \end{aligned}$$

For $b, c \geq 0$ we have $(b + c)^{1/2} \leq b^{1/2} + c^{1/2}$, so the above is at most

$$\begin{aligned} &L \int_0^1 \left(\log\left(a\left(1 + \frac{\|k\|_\infty^2 \rho^2}{2L^2}\right)\right)\right)^{1/2} du + L \int_0^1 \left(\log\left(\frac{1}{u^2}\right)\right)^{1/2} du \\ =\ &L\left(\log\left(a\left(1 + \frac{\|k\|_\infty^2 \rho^2}{2L^2}\right)\right)\right)^{1/2} + L \int_0^1 \left(\log\left(\frac{1}{u^2}\right)\right)^{1/2} du. \end{aligned}$$

Changing variables to

$$s = \left(\log\left(\frac{1}{u^2}\right)\right)^{1/2}$$

shows

$$\int_0^1 \left(\log\left(\frac{1}{u^2}\right)\right)^{1/2} du = \int_0^\infty s^2 \exp(-s^2/2) ds$$
$$= \frac{1}{2}\int_{-\infty}^\infty s^2 \exp(-s^2/2) ds$$
$$= \left(\frac{\pi}{2}\right)^{1/2},$$

since the last integral is a multiple of the variance of an $\mathrm{N}(0,1)$ random variable. The first result follows. Note that $N(F, \|\cdot\|_\infty, \varepsilon) = 1$ whenever $\varepsilon \geq 2C$, as the ball of radius $2C$ about any point in F is the whole of F. Hence, when $a = 1$, we have

$$\int_0^L (\log(N(F, \|\cdot\|_\infty, \varepsilon)))^{1/2} d\varepsilon \leq \int_0^\infty (\log(N(F, \|\cdot\|_\infty, \varepsilon)))^{1/2} d\varepsilon$$
$$= \int_0^{2C} (\log(N(F, \|\cdot\|_\infty, \varepsilon)))^{1/2} d\varepsilon$$
$$\leq 2\left(\log\left(1 + \frac{\|k\|_\infty^2 \rho^2}{8C^2}\right)\right)^{1/2} C + (2\pi)^{1/2} C$$

for $L \in (0, \infty]$. ■

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, pages 2312–2320, 2011.

Jöran Bergh and Jörgen Löfström. *Interpolation Spaces. An Introduction.* Springer–Verlag, Berlin–New York, 1976.

Andrea Caponnetto and Ernesto de Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 7(3):331–368, 2007.

Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithm. *arXiv preprint arXiv:1702.07254*, 2017.

Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models.* Cambridge University Press, New York, 2016.

Shahar Mendelson and Joseph Neeman. Regularization in kernel learning. *Ann. Statist.*, 38 (1):526–565, 2010.

Luca Oneto, Sandro Ridella, and Davide Anguita. Tikhonov, Ivanov and Morozov regularization for support vector machine learning. *Mach. Learn.*, 103(1):103–136, 2016.

Yamilet Quintana and José M. Rodríguez. Measurable diagonalization of positive definite matrices. *J. Approx. Theory*, 185:91–97, 2014.

Steve Smale and Ding-Xuan Zhou. Estimating the approximation error in learning theory. *Anal. Appl. (Singap.)*, 1(1):17–41, 2003.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constr. Approx.*, 26(2):153–172, 2007.

Bharath Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839–1893, 2016.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer–Verlag, New York, 2008.

Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constr. Approx.*, 35(3):363–417, 2012.

Ingo Steinwart, Don R. Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *Proceedings of the 22nd Conference on Learning Theory*, 2009.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer–Verlag, New York, 1996.

David Williams. *Probability with Martingales*. Cambridge University Press, Cambridge, 1991.