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

Stephen Page
STOR-i
Lancaster University
Lancaster, LA1 4YF, United Kingdom

Steffen Grünewälder
Department of Mathematics and Statistics
Lancaster University
Lancaster, LA1 4YF, United Kingdom

Abstract

We study kernel least-squares estimation under a norm constraint. This form of regularisation is known as Ivanov regularisation and it provides far better control of the regression function than the well-established Tikhonov regularisation. In particular, the smoothness and the scale of the regression function are directly controlled by the constraint. This choice of estimator also allows us to dispose of the standard assumption that the reproducing kernel Hilbert space (RKHS) has a Mercer kernel. The Mercer assumption is restrictive as it usually requires compactness of the covariate set. Instead, we make only the minimal assumption 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 complement this result with a high probability bound under the stronger assumption that the response variables have subgaussian errors and that the regression function is bounded. Finally, we derive an adaptive version of the high probability bound under the assumption that the response variables are bounded. The rates we achieve are close to the optimal rates attained under the stronger Mercer kernel assumption.

Keywords: Ivanov Regularisation, RKHSs, Mercer Kernels

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. However, 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 space then the bound forces the estimator to lie in a ball of predefined radius inside the function space. In our case the function space is a reproducing kernel Hilbert space (RKHS). An RKHS norm measures the smoothness of a function and the norm constraint bounds the smoothness of the estimator. Beside enforcing these properties, the norm constraint also allows us to control empirical processes evaluated at the estimator by controlling the supremum of the process over a ball in the RKHS. This is vital to the proofs of the bounds in this paper. For example, see Lemmas 4 and 10 on pages 8 and 11.

Tikhonov regularisation has been extensively applied to RKHS regression in the context of support vector machines (SVMs) and other least-squares estimators (Steinwart and Christmann, 2008; Caponnetto and de Vito, 2007; Mendelson and Neeman, 2010; Steinwart, Hush, and Scovel, 2009). In these cases the regularisation function is some function of the norm of the RKHS. In this paper we instead use an Ivanov-regularised least-squares estimator with regularisation function equal to the norm of the RKHS. This choice of regularisation automatically gives us a bound on the norm of our estimator in the RKHS. Having a tight bound on the norm of an estimator in the RKHS is important for its analysis, so it is surprising that Ivanov-regularised least-squares estimators have not attracted more attention in the kernel setting. Although there is a relatively trivial bound on the norm of a Tikhonov-regularised least-squares estimator in an RKHS, it is often difficult to make this bound tight. For example, see the start of Chapter 7 of Steinwart and Christmann (2008). Furthermore, such estimators are often clipped so as to impose a uniform bound despite already being regularised (Steinwart and Christmann, 2008; Steinwart et al., 2009).

Using Ivanov regularisation we can avoid clipping of the estimator in Sections 3 and 4. Furthermore, this form of regularisation allows us also to dispose of the restrictive Mercer kernel assumption. This is because we control empirical processes over balls in the RKHS instead. On the other hand, the current theory of RKHS regression only applies when the RKHS $H$ has a Mercer kernel $k$ with respect to the covariate distribution $P$. This allows for a simple decomposition of $k$ and succinct representation of $H$ as a subspace of $L^2(P)$. These descriptions are in terms of the eigenfunctions and eigenvalues of the kernel operator $T$ on $L^2(P)$. Many results have assumed a fixed rate of decay of these eigenvalues in order to produce estimators whose squared $L^2(P)$ error decreases quickly with the number of data points (Mendelson and Neeman, 2010; Steinwart et al., 2009). However, the assumptions necessary for $H$ to have a Mercer kernel are in general quite restrictive. The usual set of assumptions is that the covariate set $S$ is compact, the kernel $k$ of $H$ is continuous on $S \times S$ and the covariate distribution satisfies $\text{supp } P = S$ (see Section 4.5 of Steinwart and Christmann, 2008). In particular, the assumption that the covariate set $S$ is compact is inconvenient and there has been some research into how to relax this condition by Steinwart and Scovel (2012). By contrast, we provide results that hold under the significantly weaker assumption that the RKHS is separable with a bounded and measurable kernel.
In this paper we 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 $\beta$ parameterises the interpolation space containing the regression function. 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 $P$. We then move away from the average behaviour of the error towards something which tells us more about what happens 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 that the regression function is bounded. Finally, we assume that the response variables are bounded to analyse an adaptive version of our estimator which does not require us to know which interpolation space contains the regression function. We use the high probability bound from our previous result to obtain a high probability bound in this setting which is of order $\log(n)^{1/2}n^{-\beta/2}$. For small $\beta > 0$ the order $n^{-\beta/2}$ is around a power of $1/2$ larger than the order of $n^{-\beta/((\beta+1))}$ of Steinwart et al. (2009) when $p = 1$. However, we do not make the restrictive assumption that $k$ is a Mercer kernel of $H$ as discussed above.

1.1 RKHSs and Their Interpolation Spaces

A Hilbert space $H$ of real-valued functions on $S$ is an RKHS if the evaluation functional $L_x : H \to \mathbb{R}$ by $L_x h = h(x)$ 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 $k(x_1, x_2) = \langle k_{x_1}, k_{x_2} \rangle$ for $x_1, x_2 \in S$ and is symmetric and positive-definite.

We can define a range of spaces between $L^2(P)$ and $H$. Let $(Z, \| \cdot \|_Z)$ be a Banach space and $(V, \| \cdot \|_V)$ be a subspace of $Z$. Then 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 can define

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

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 $\| z \|_{\beta, q} < \infty$ for $\beta \in (0, 1)$ and $1 \leq q \leq \infty$. Smaller values of $\beta$ give larger spaces. When $\beta$ is close to 1 the space is not much larger than $V$, but as $\beta$ decreases we obtain spaces which get closer to $Z$. Hence, we can define the interpolation spaces $[L^2(P), H]_{\beta, q}$. We will work with $q = \infty$ which gives the largest space of functions for a fixed $\beta \in (0, 1)$. Note that although $H$ is not a subspace of $L^2(P)$, the above definitions are still valid as there is a natural embedding of $H$ into $L^2(P)$.
1.2 Literature Review

The current literature assumes that the RKHS $H$ has a Mercer kernel $k$ as discussed above. Earlier research in this area, 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 the regression function be of the form $g = T^{(\beta-1)/2}h$ for $\beta \in [1, 2]$ and $h \in H$. Here $T : L^2(P) \rightarrow L^2(P)$ is the kernel operator and $P$ is the covariate distribution. By assuming that the $i$th 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 using SMVs. 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 using Tikhonov-regularised least-squares estimators. The authors also show that if the eigenfunctions of the kernel operator $T$ are uniformly bounded in $L^\infty(P)$ then the order can be improved to $n^{-\beta/(\beta+p)}$. Steinwart et al. (2009) relax the condition on the eigenfunctions to the condition $\|h\|_\infty \leq C_p \|h\|_H \|h\|_{L^2(P)}^{1-p}$ for all $h \in H$ and some constant $C_p > 0$. The same rate is attained using clipped Tikhonov-regularised least-squares estimators, including clipped SMVs, 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 $L^\infty(P)$ 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 using SVMs.

1.3 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 $n^{-\beta/2}$. Under the stronger assumption that the response variables have subgaussian errors we show that the squared $L^2(P)$ error is of order $n^{-\beta/2}$ with high
probability. Finally, we assume that the response variables are bounded. In this case 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 $\log(n)^{1/2} n^{-\beta/2}$ with high probability.

2. Problem Definition and Assumptions

We now formally define our regression problem and the assumptions that we make in this paper. Assume that $(X_i, Y_i) \in S \times \mathbb{R}$ are i.i.d. for $1 \leq i \leq n$ where $X_i \sim P$ and

$$E(Y_i | X_i) = g(X_i),$$

$$\text{var}(Y_i | X_i) \leq \sigma^2.$$  \hspace{1cm} (1) (2)

The conditional expectation used is that of Kolmogorov, defined using the Radon-Nikodym derivative. Here we assume that $g \in [L^2(P), H]_{\beta, \infty}$ with norm at most $B$ where $H$ is a separable RKHS and $\beta \in (0, 1)$. Finally we assume that $H$ has kernel $k$ which is bounded and measurable with

$$\|k\|_\infty = \sup_{x \in S} k(x, x)^{1/2} < \infty.$$  \hspace{1cm} (3)

Let $B_H$ be the unit ball of $H$ and $r > 0$. Our assumption that $g \in [L^2(P), H]_{\beta, \infty}$ with norm at most $B$ and $\beta \in (0, 1)$ provides us with the following result. Theorem 3.1 of Smale and Zhou (2003) shows that

$$\inf \{ \|h - g\|_{L^2(P)}^2 : h \in rB_H \} \leq \frac{B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}}$$

when $H$ is dense in $L^2(P)$. This additional condition is present because these are the only interpolation spaces considered by the authors and in fact the result holds by the same proof even when $H$ is not dense. Hence, for all $\alpha > 0$ there is some $h_{r, \alpha} \in rB_H$ such that

$$\|h_{r, \alpha} - g\|_{L^2(P)}^2 \leq \frac{B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}} + \alpha.$$  \hspace{1cm} (3)

We will estimate $h_{r, \alpha}$ for small $\alpha > 0$ by

$$\hat{h}_r = \arg \min_{f \in rB_H} \frac{1}{n} \sum_{i=1}^n (f(X_i) - Y_i)^2.$$  \hspace{1cm} (4)

Note that the estimator $\hat{h}_r$ is not uniquely defined. It is a least-squares estimator with norm at most $r$ in $H$ if one exists. However, there is a unique least-squares estimator with minimal norm in $H$ which is measurable and lies in the span of the $k_{X_i}$. It is the limit of SVMs as the regularisation parameter tends to 0. We will use this least-squares estimator in the definition of $\hat{h}_r$. If this estimator has norm strictly bigger than $r$ then $\hat{h}_r$ is the SVM with regularisation parameter selected so that its norm is $r$ in $H$.  \hspace{1cm} (5)
2.1 Measurability

Since $H$ has a measurable kernel $k$ it follows that all functions in $H$ are measurable (Lemma 4.24 of Steinwart and Christmann, 2008). Furthermore, the estimator $\hat{h}_r$ is measurable. The least-squares estimator used in the definition of $\hat{h}_r$ is measurable, as is its norm. An SVM is also measurable as long as its regularisation parameter is measurable. The norm of an SVM in $H$ is measurable in the data and strictly decreasing in its regularisation parameter. Hence, the regularisation parameter for which an SVM has norm $r$ is measurable and so is the corresponding SVM.

We will also require the measurability of certain suprema over subsets of $H$. By definition of $H$ being separable, it has a countable dense subset $H_0$. Furthermore, $H$ is continuously embedded in $L^2(P_n)$ and $L^2(P)$ because its kernel $k$ is bounded. Hence, 

$$\sup_{f \in rB_H} \left| \|f\|_{L^2(P_n)}^2 - \|f\|_{L^2(P)}^2 \right| = \sup_{f \in rB_H \cap H_0} \left| \|f\|_{L^2(P_n)}^2 - \|f\|_{L^2(P)}^2 \right|.$$ 

This is a random variable since the right-hand side is a supremum of countably many random variables. Therefore this quantity has a well-defined expectation and we can also apply Talagrand’s inequality to it.

3. Expectation Bound

In our first results section we prove an upper bound on the expectation of the squared $L^2(P)$ error of our estimator $\hat{h}_r$ under the weak assumption that the response variables have bounded variance. We recall that $g$ is the regression function. The $L^2(P)$ norm is appropriate here as if $X_{n+1}$ is a new independent covariate with the same distribution $P$ then

$$E \left( \|\hat{h}_r - g\|_{L^2(P)}^2 \right) = E \left( (\hat{h}_r(X_{n+1}) - g(X_{n+1}))^2 \right),$$

the expected squared error of $\hat{h}_r$ at $X_{n+1}$. Recall that $r > 0$ is the level of the hard constraint on the norm of our estimator. By selecting $r$ of order $n^{(1-\beta)/4}$ we achieve a bound of order $n^{-\beta/2}$. Note that in the first bound of the following theorem, the second and third terms are the largest asymptotically as we require $r \to \infty$ for our bound to tend to 0.

**Theorem 1** We have

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

Based on this bound the asymptotically optimal choice of $r$ is

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

6
which gives a bound of
\[
\left(10 \left(\frac{32(1 - \beta)}{5\beta}\right)^\beta + 64 \left(\frac{5\beta}{32(1 - \beta)}\right)^{1 - \beta}\right) \|k\|_\infty^{\beta} B^2 n^{-\beta/2} + 8 \left(\frac{5\beta}{32(1 - \beta)}\right)^{(1 - \beta)/2} \|k\|_\infty^{\beta} \sigma B n^{-(1 + \beta)/2}.
\]

We prove this result as follows. We already have an upper bound on the squared \(L^2(P)\) norm of \(h_r - g\) from (3) and so our main aim is to establish an upper bound on the expectation of the squared \(L^2(P_n)\) norm of \(\hat{h}_r - h_{r,\alpha}\). To do this we first bound the squared \(L^2(P_n)\) norm of \(h_r - h_{r,\alpha}\) where \(P_n\) is the empirical distribution of the \(X_i\). The definition of \(\hat{h}_r\) and some rearranging provide us with the following inequality.

**Lemma 2** The definition of \(\hat{h}_r\) shows
\[
\|\hat{h}_r - h_{r,\alpha}\|^2_{L^2(P_n)} \leq \frac{4}{n} \sum_{i=1}^{n} (Y_i - g(X_i))(\hat{h}_r(X_i) - h_{r,\alpha}(X_i)) + 4\|h_{r,\alpha} - g\|^2_{L^2(P_n)}.
\]

We continue by bounding the expectation of the right-hand side of (4). The second term has expectation
\[
E \left(\|h_{r,\alpha} - g\|^2_{L^2(P_n)}\right) = \|h_{r,\alpha} - g\|^2_{L^2(P)} \leq \frac{B^2/(1 - \beta)}{\sigma^{2\beta/(1 - \beta)}} + \alpha
\]
by (3). The first term is more difficult to deal with but we can use the method of Remark 6.1 of Sriperumbudur (2016). It can be bounded using an expression of the form
\[
\sup_{f \in B_H} \frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i)) f(X_i)
\]
which is equal to
\[
\left\|\frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i)) k_{X_i}\right\|_H
\]
by the reproducing kernel property and the Cauchy–Schwarz inequality. By writing the norm as the square root of the corresponding inner product and using the reproducing kernel property again we find that this is equal to
\[
\left(\frac{1}{n^2} \sum_{i,j=1}^{n} (Y_i - g(X_i))(Y_j - g(X_j)) k(X_i, X_j)\right)^{1/2}.
\]

We can take the expectation of this quantity inside the square root using Jensen’s inequality and use the independence and variance assumptions on the data to obtain the following bound.
Lemma 3 By bounding the expectation of the right-hand side of (4) we have
\[
E\left(\|\hat{h}_r - h_{r,\alpha}\|_{L^2(P_n)}^2\right) \leq \frac{4\|k\|_\infty \sigma_r}{n^{1/2}} + \frac{4B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}} + 4\alpha.
\]
The next step is to move our bound on the expectation of the squared \(L^2(P_n)\) norm of \(\hat{h}_r - h_{r,\alpha}\) to the expectation of the squared \(L^2(P)\) norm of \(\hat{h}_r - h_{r,\alpha}\). We can achieve this by bounding
\[
E\left(\sup_{f \in B_H} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \right).
\]
By symmetrisation and the contraction principle for Rademacher processes it is enough to bound the expectation of
\[
\sup_{f \in B_H} \left| \frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i)) f(X_i) \right|
\]
which can be done in the same way as bounding the expectation of
\[
\sup_{f \in B_H} \left| \frac{1}{n} \sum_{i=1}^{n} Y_i f(X_i) - \frac{1}{n} \sum_{i=1}^{n} g(X_i) f(X_i) \right|
\]
as discussed earlier.

Lemma 4 Using Rademacher processes we can show
\[
E\left(\sup_{f \in B_H} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \right) \leq \frac{8\|k\|_\infty^2}{n^{1/2}}.
\]

Since \(h_{r,\alpha}, \hat{h}_r \in rB_H\) we have
\[
\|\hat{h}_r - h_{r,\alpha}\|_{L^2(P)}^2 \leq \|\hat{h}_r - h_{r,\alpha}\|_{L^2(P_n)}^2 + 4r^2 \sup_{f \in B_H} \left| \frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i)) f(X_i) \right|
\]
and the next result follows as a simple consequence of Lemmas 3 and 4.

Corollary 5 We have
\[
E\left(\|\hat{h}_r - h_{r,\alpha}\|_{L^2(P)}^2\right) \leq \frac{4\|k\|_\infty \sigma_r}{n^{1/2}} + \frac{4B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}} + 32\|k\|_\infty^2 r^2 + 4\alpha.
\]

All that remains to prove Theorem 1 is to combine Corollary 5 with (3) by using
\[
\|\hat{h}_r - g\|_{L^2(P)}^2 \leq \left(\|\hat{h}_r - h_{r,\alpha}\|_{L^2(P)} + \|h_{r,\alpha} - g\|_{L^2(P)}\right)^2
\]
\[
\leq 2\|\hat{h}_r - h_{r,\alpha}\|_{L^2(P)}^2 + 2\|h_{r,\alpha} - g\|_{L^2(P)}^2
\]
and let \(\alpha \to 0\).
4. High Probability Bound

In this section we move away from the average behaviour of the squared $L^2(P)$ error of our estimator towards understanding what happens in the worst case. We do this by proving a high probability bound on the squared $L^2(P)$ error of our estimator. Our proof method follows that of the expectation bound. However, we must additionally prove concentration inequalities for all of the quantities involved. In order to achieve these concentration results we make the following further assumptions. We assume that the moment generating function of $Y_i - g(X_i)$ conditional on $X_i$ satisfies

$$E(\exp(t(Y_i - g(X_i)))|X_i) \leq \exp(\sigma^2 t^2/2)$$

for all $t \in \mathbb{R}$ and $1 \leq i \leq n$. Note that the right-hand side of the inequality is the moment generating function of an $N(0, \sigma^2)$ random variable. The random variable $Y_i - g(X_i)$ is said to be $\sigma^2$-subgaussian conditional on $X_i$. This condition implies our weaker assumptions (1) and (2), so our results from Section 3 still apply. We also assume that $\|g\|_\infty \leq C$. Again, we can select $r$ of order $n^{(1-\beta)/4}$ to achieve a bound of order $n^{-\beta/2}$.

**Theorem 6** Let $t > 0$. With probability at least $1 - 4e^{-t}$ we have

$$\|\hat{h}_r - g\|_{L^2(P)}^2 \leq \frac{8\|k\|_\infty \sigma r ((1 + t + 2(t^2 + t)^{1/2})^{1/2} + (2t)^{1/2})}{n^{1/2}} + \frac{10B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}} + 8(C + \|k\|_\infty r)^2(2t)^{1/2} + \frac{64\|k\|_\infty^2 r^2}{n^{1/2}}$

$$+ 8 \left( \frac{2\|k\|_4^4 r^4}{n} + \frac{32\|k\|_\infty^4 r^4}{n^{3/2}} \right)^{1/2} t^{1/2} + \frac{16\|k\|_\infty^2 r^2 t}{3n}.$$

For $t \geq 1$ this can be simplified to

$$\|\hat{h}_r - g\|_{L^2(P)}^2 \leq \frac{8(4C^2 + 5\|k\|_\infty \sigma r + 18\|k\|_\infty^2 r^2)t^{1/2}}{n^{1/2}} + \frac{10B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}} + \frac{16\|k\|_\infty^2 r^2 t}{3n}.$$

Based on this bound the asymptotically optimal choice of $r$ is

$$\left( \frac{5\beta}{72(1-\beta)} \right)^{(1-\beta)/2} \|k\|_\infty^{-(1-\beta)} B t^{-(1-\beta)/4} n^{(1-\beta)/4}$$

which ignoring factors involving only $\beta$ gives a bound of asymptotic order

$$\|k\|_\infty^{2\beta} B^2 t^{\beta/2} n^{-(\beta/2)}.$$

The concentration results we need are reasonably simple to show for normal random variables but are more complicated to derive in the general subgaussian case. We need the following result which relates a quadratic form of subgaussians to that of centred normal random variables.
Lemma 7 For $1 \leq i \leq n$ let $\varepsilon_i$ be independent conditional on some sigma algebra $\mathcal{G}$ and let
\[ \mathbb{E}(\exp(t\varepsilon_i)|\mathcal{G}) \leq \exp(\sigma^2t^2/2) \]
for all $t \in \mathbb{R}$. Also let $\delta_i$ be independent of each other and $\mathcal{G}$ with $\delta_i \sim \mathcal{N}(0, \sigma^2)$. If $A$ is an $n \times n$ and $\mathcal{G}$-measurable non-negative-definite matrix then
\[ \mathbb{E}(\exp(t\varepsilon^T A \varepsilon)|\mathcal{G}) \leq \mathbb{E}(\exp(t\delta^T A \delta)|\mathcal{G}) \]
for all $t \geq 0$.

The above lemma can be proved by integrating the argument of the conditional moment generating function of the vector of $\varepsilon_i$ over a Gaussian kernel. Having established this relationship we can now obtain a probability bound on a quadratic form of subgaussians by using Chernoff bounding. The following is a conditional subgaussian version of the Hanson–Wright inequality.

Lemma 8 For $1 \leq i \leq n$ let $\varepsilon_i$ be independent conditional on some sigma algebra $\mathcal{G}$ and let
\[ \mathbb{E}(\exp(t\varepsilon_i)|\mathcal{G}) \leq \exp(\sigma^2t^2/2) \]
for all $t \in \mathbb{R}$. If $A$ is an $n \times n$ and $\mathcal{G}$-measurable non-negative-definite matrix and $t \geq 0$ then
\[ \varepsilon^T A \varepsilon \leq \sigma^2 \text{tr}(A) + 2\sigma^2 \left( \max_{1 \leq i \leq n} \lambda_i \right) t + 2 \left( \sigma^4 \left( \max_{1 \leq i \leq n} \lambda_i \right)^2 t^2 + \sigma^4 \text{tr}(A^2) t \right)^{1/2} \]
with probability at least $1 - e^{-t}$ conditional on $\mathcal{G}$. Here $\lambda_i$ are the eigenvalues of $A$ for $1 \leq i \leq n$ and they are $\mathcal{G}$-measurable.

We now proceed with the main part of the proof. As in Section 3 we first bound the squared $L^2(P_n)$ norm and then the squared $L^2(P)$ norm of $\hat{h}_r - h_{r,\alpha}$. Finally, we combine this with the bound on the squared $L^2(P)$ norm of $h_{r,\alpha} - g$ from (3) to achieve our goal. In order to prove the probability bound on the squared $L^2(P_n)$ norm of $\hat{h}_r - h_{r,\alpha}$ we use a quadratic form of subgaussians from Lemma 8. The second term can be bounded using Hoeffding’s inequality. This is because $h_{r,\alpha} \in rB_H$ implies that $\|h_{r,\alpha}\|_{\infty} \leq \|k\|_{\infty} r$ and we assume that $\|g\|_{\infty} \leq C$.

Lemma 9 Let $t > 0$. With probability at least $1 - 3e^{-t}$ we have
\[ \|\hat{h}_r - h_{r,\alpha}\|_{L^2(P_n)}^2 \leq \frac{4\|k\|_{\infty} r \sigma r \left( (1 + t + 2(t^2 + t^{1/2})^{1/2} + (2t)^{1/2} \right)^{n/2}}{ \left( \frac{4B^2(1-\beta)}{r^{2\beta/(1-\beta)}} + \frac{4(C + \|k\|_{\infty} r)^2(2t)^{1/2}}{n^{1/2}} \right)^{n/2} + 4\alpha. \]
We now move our probability bound on the squared $L^2(P)$ norm of $\hat{h}_r - h_{r,\alpha}$ to the squared $L^2(P)$ norm of $h_r - h_{r,\alpha}$. We can achieve this by bounding
\[
\sup_{f \in rB_H} \|f\|_{L^2(P)}^2 - \|f\|_{L^2(P_n)}^2 \leq \frac{8\|k\|_\infty^2 r^2}{n^{1/2}} + \left(\frac{2\|k\|_\infty^4 r^4}{n} + \frac{32\|k\|_\infty^4 r^4}{n^{3/2}}\right)^{1/2} r^{1/2} + \frac{2\|k\|_\infty^4 r^4 t}{3n}.
\]

The proof of Theorem 6 is then as follows. Since $h_{r,\alpha}, \hat{h}_r \in rB_H$ we have
\[
\|\hat{h}_r - h_{r,\alpha}\|_{L^2(P)}^2 \leq \|\hat{h}_r - h_{r,\alpha}\|_{L^2(P_n)}^2 + \sup_{f \in rB_H} \|f\|_{L^2(P)}^2 - \|f\|_{L^2(P_n)}^2,
\]
which can be bounded using Lemmas 9 and 10. Finally, we combine this bound with (3) in the same way as for Theorem 1 and let $\alpha \to 0$.

5. Adaptive Probability Bound

We now use training and validation on the data in order to select $r > 0$ without knowing which interpolation space contains the regression function. In order to do this we have to make the additional assumption that the response variables $Y_i$ are bounded in $[-M, M]$. Because of this our regression function satisfies $\|g\|_\infty \leq M$, which in turn implies that $|Y_i - g(X_i)| \leq 2M$. Hoeffding’s Lemma then implies that $Y_i - g(X_i)$ is $4M^2$-subgaussian conditional on $X_i$, as defined by (5). This means the results from Section 4 hold with $\sigma^2 = 4M^2$ and $C = M$. Similarly to Chapter 7 of Steinwart and Christmann (2008) and Steinwart et al. (2009), we will define the projection $V : \mathbb{R} \to [-M, M]$. Since $|Y_1| \leq M$ and
\[
\|h - g\|_{L^2(P)}^2 = \mathbb{E}((h(X_1) - Y_1)^2) - \mathbb{E}((g(X_1) - Y_1)^2)
\]
for $h \in H$ we find
\[
\|Vh - g\|_{L^2(P)}^2 \leq \|h - g\|_{L^2(P)}^2.
\]
In particular we find that our results for $\hat{h}_r$ in Section 4 also hold for $V\hat{h}_r$. We then have the following result based on Theorem 7.24 of Steinwart and Christmann (2008).

Theorem 11 Let $a \geq b > 0$, $I = [(a/b)^2(n^{1/2} - 1)]$ and
\[
R = \{r_i = (a^2 + b^2i)^{1/2} : 0 \leq i \leq I - 1\} \cup \{r_I = an^{1/4}\}.
\]
Also let $p \in (0, 1)$. For $n > p^{-1}$ the following result holds. Suppose that we only use the first $m = \lfloor pn \rfloor$ data points to calculate the $\hat{h}_r$ for $r \in R$. Let
\[
\hat{r} = \arg\min_{r \in R} \frac{1}{n - m} \sum_{i=m+1}^n (V\hat{h}_r(X_i) - Y_i)^2
\]
with ties broken towards the smallest value of \( r \in \mathbb{R} \). Note that \( \hat{r}, \hat{h}_r \) and \( \hat{V}_r \) are measurable. Then \( \| \hat{V}_r - g \|^2_{L^2(P)} \) is of order at most

\[
\log(n)^{1/2} n^{-\beta/2} + t^{1/2} n^{-\beta/2}
\]

in \( n \) and \( t \geq 1 \) with probability at least \( 1 - e^{-t} \).

6. 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 consider the case in which the regression function lies in an interpolation space between \( L^2(P) \) and an RKHS. We achieve these bounds without the standard assumption that the RKHS has a Mercer kernel, which means that our results apply even when the covariate set is not compact. In fact, our only assumption on the RKHS is that it 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 \( n^{-\beta/2} \). Under the stronger assumption that the response variables have subgaussian errors we achieve a probability bound of the same order. Finally, we assume that the response variables are bounded and use training and validation on the data set. This allows us to select the size of the norm constraint for our Ivanov regularisation without knowing which interpolation space contains the regression function.

For small \( \beta > 0 \) the order \( n^{-\beta/2} \) is around a power of \( 1/2 \) larger than the order of \( n^{-\beta/(\beta+1)} \) of Steinwart et al. (2009) when \( p = 1 \). However, we do not make the restrictive assumption that \( k \) is a Mercer kernel of \( H \). This is because we use Ivanov regularisation instead of Tikhonov regularisation and control empirical processes over balls in the RKHS. By contrast, the current theory uses the embedding of \( H \) in \( L^2(P) \) from Mercer’s theorem to achieve bounds on Tikhonov-regularised estimators (Mendelson and Neeman, 2010; Steinwart et al., 2009). For this reason, it seems unlikely that Tikhonov-regularised estimators such as SVMs would perform well in the absence of a Mercer kernel, although it would be interesting to investigate whether or not this is the case.

Lower bounds on estimation rates are useful for complementing upper bounds. Together they give a range in which the optimal estimation rate lies and this can be used to assess the performance of estimators with good upper bounds. The lower bound of order \( n^{-\beta/(\beta+1)} \) of Steinwart et al. (2009) when \( p = 1 \) certainly applies to our setting. Improving this lower bound would require constructing an element of an interpolation space between \( L^2(P) \) and an RKHS which is difficult to estimate.

It is restrictive to assume that the response variables are bounded in order to achieve our adaptive estimation result and it would be useful to be able to relax this assumption. We need the boundedness assumption to make use of the training and validation approach based on Theorem 7.24 of Steinwart and Christmann (2008). It is not obvious that this training and validation approach can be made to work without the boundedness assumption and a different tool might be needed to remove this restriction.
Appendix A. Proof of Expectation Bound

Proof of Lemma 2 Since \( h_{r,\alpha} \in rB_H \), the definition of \( \hat{h}_r \) gives

\[
\frac{1}{n} \sum_{i=1}^{n} (\hat{h}_r(X_i) - Y_i)^2 \leq \frac{1}{n} \sum_{i=1}^{n} (h_{r,\alpha}(X_i) - Y_i)^2.
\]

Expanding

\[
(\hat{h}_r(X_i) - Y_i)^2 = \left( (\hat{h}_r(X_i) - h_{r,\alpha}(X_i)) + (h_{r,\alpha}(X_i) - Y_i) \right)^2,
\]

substituting into the above and rearranging gives

\[
\frac{1}{n} \sum_{i=1}^{n} (\hat{h}_r(X_i) - h_{r,\alpha}(X_i))^2 \leq \frac{2}{n} \sum_{i=1}^{n} (Y_i - h_{r,\alpha}(X_i))(\hat{h}_r(X_i) - h_{r,\alpha}(X_i)).
\]

Substituting

\[ Y_i - h_{r,\alpha}(X_i) = (Y_i - g(X_i)) + (g(X_i) - h_{r,\alpha}(X_i)) \]

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

\[
\|\hat{h}_r - h_{r,\alpha}\|_{L^2(P_n)}^2 \leq \frac{2}{n} \sum_{i=1}^{n} (Y_i - g(X_i))(\hat{h}_r(X_i) - h_{r,\alpha}(X_i)) + 2\|g - h_{r,\alpha}\|_{L^2(P_n)} \|\hat{h}_r - h_{r,\alpha}\|_{L^2(P_n)}.
\]

For constants \( a, b \in \mathbb{R} \) we have

\[ x^2 \leq a + 2bx \implies x^2 \leq 2a + 4b^2 \]

for \( x \in \mathbb{R} \) by completing the square and rearranging. Applying this result to the above inequality proves the lemma.

Proof of Lemma 3 We have

\[
\mathbb{E} \left( \|h_{r,\alpha} - g\|_{L^2(P_n)}^2 \right) = \|h_{r,\alpha} - g\|_{L^2(P)}^2 \leq \frac{B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}} + \alpha
\]

by (3) and

\[
\mathbb{E} \left( \frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i))h_{r,\alpha}(X_i) \right) = \mathbb{E} \left( \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(Y_i - g(X_i)|X_i)h_{r,\alpha}(X_i) \right) = 0.
\]
The remainder of this proof method is due to Remark 6.1 of Sriperumbudur (2016). Since \( \hat{h}_r \in rB_H \) we have
\[
\frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i)) \hat{h}_r(X_i) \leq \sup_{f \in rB_H} \frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i)) f(X_i)
\]
\[
= \sup_{f \in rB_H} \left( \frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i)) k_{X_i, f} \right)_H
\]
\[
= r \left( \frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i)) k_{X_i} \right)_H
\]
\[
= r \left( \frac{1}{n^2} \sum_{i,j=1}^{n} (Y_i - g(X_i))(Y_j - g(X_j)) k(X_i, X_j) \right)^{1/2}
\]
by the Cauchy–Schwarz inequality. Denote the vector of the \( X_i \in S \) by \( X \). Then by Jensen’s inequality and the independence of the \((X_i, Y_i)\) we have
\[
\mathbb{E} \left( \frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i)) \hat{h}_r(X_i) \bigg| X \right) \leq r \left( \frac{1}{n^2} \sum_{i,j=1}^{n} \text{cov}(Y_i, Y_j|X) k(X_i, X_j) \right)^{1/2}
\]
\[
\leq r \left( \frac{\sigma^2}{n^2} \sum_{i=1}^{n} k(X_i, X_i) \right)^{1/2}
\]
and again by Jensen’s inequality we have
\[
\mathbb{E} \left( \frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i)) \hat{h}_r(X_i) \right) \leq r \left( \frac{\sigma^2 \|k\|_\infty^3}{n} \right)^{1/2}.
\]
The result follows from Lemma 2. ■

**Proof of Lemma 4** Let the \( \varepsilon_i \) be i.i.d. Rademacher random variables independent of the \( X_i \). Lemma 2.3.1 of van der Vaart and Wellner (2013) shows
\[
\mathbb{E} \sup_{f \in B_H} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i)^2 - \int f^2 dP \right| \leq 2 \mathbb{E} \sup_{f \in B_H} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i)^2 \right|
\]
by symmetrisation. Since
\[
f(X_i) = \langle k_{X_i}, f \rangle_H \leq \|k_{X_i}\|_H = k(X_i, X_i)^{1/2} \leq \|k\|_\infty
\]
for all \( f \in B_H \) we find
\[
f(X_i)^2 \leq \frac{2\|k\|_\infty}{\|k\|_\infty}
\]
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
\[
\mathbb{E} \left( \sup_{f \in B_H} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i)^2 \right| \bigg| X \right) \leq 2 \mathbb{E} \left( \sup_{f \in B_H} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \bigg| X \right) .
\]
Therefore
\[
\mathbb{E} \sup_{f \in B_H} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i)^2 - \int f^2 dP \right| \leq 8\|k\|_{\infty} \mathbb{E} \left( \sup_{f \in B_H} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \right).
\]

We now follow a similar argument to the proof of Lemma 3. We have
\[
\sup_{f \in B_H} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| = \sup_{f \in B_H} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i k(X_i, f) \right|
= \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i k(X_i) \right\|_H
= \left( \frac{1}{n^2} \sum_{i,j=1}^{n} \varepsilon_i \varepsilon_j k(X_i, X_j) \right)^{1/2}.
\]

By Jensen’s inequality we have
\[
\mathbb{E} \left( \sup_{f \in B_H} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \right) \leq \left( \frac{1}{n^2} \sum_{i,j=1}^{n} \text{cov} \left( \varepsilon_i, \varepsilon_j \right) k(X_i, X_j) \right)^{1/2}
= \left( \frac{1}{n^2} \sum_{i=1}^{n} k(X_i, X_i) \right)^{1/2}
\]
and again by Jensen’s inequality we have
\[
\mathbb{E} \left( \sup_{f \in B_H} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \right) \leq \left( \frac{\|k\|_{\infty}^2}{n} \right)^{1/2}.
\]

The result follows.

Appendix B. Proof of High Probability Bound

**Proof of Lemma 7** This proof method uses techniques from the proof of Lemma 9 of Abbasi-Yadkori, Pál, and Szepesvári (2011). We have
\[
\mathbb{E}(\exp(t_i \varepsilon_i / \sigma) | \mathcal{G}) \leq \exp(t_i^2 / 2)
\]
for all $1 \leq i \leq n$ and $t_i \in \mathbb{R}$. Furthermore, the $\varepsilon_i$ are independent conditional on $\mathcal{G}$ so
\[
\mathbb{E}(\exp(t^T \varepsilon / \sigma) | \mathcal{G}) \leq \exp(||t||_2^2 / 2).
\]

Quintana and Rodríguez (2014) show that a measurable strictly-positive-definite matrix can be diagonalised by a measurable orthogonal matrix and a measurable diagonal matrix. The result holds for non-negative-definite matrices by adding the identity matrix before
diagonalisation, so let \( A \) have the \( \mathcal{G} \)-measurable square root \( A^{1/2} \). Therefore we can replace \( t \) with \( sA^{1/2}u \) for \( s \in \mathbb{R} \) and \( u \in \mathbb{R}^n \) to get
\[
\mathbb{E}(\exp(su^T A^{1/2} \varepsilon / \sigma) | \mathcal{G}) \leq \exp(s^2 \| A^{1/2}u \|_2^2 / 2).
\]
Integrating \( u \) with respect to the distribution of \( \delta \) gives
\[
\mathbb{E}(\exp(s^2 \varepsilon^T A \varepsilon / 2) | \mathcal{G}) \leq \mathbb{E}(\exp(s^2 \delta^T A \delta / 2) | \mathcal{G}).
\]
The result follows.

**Proof of Lemma 8** This proof method follows that of Theorem 3.1.9 of Giné and Nickl (2016). Quintana and Rodríguez (2014) show that a measurable strictly-positive-definite matrix can be diagonalised by a measurable orthogonal matrix and a measurable diagonal matrix. The result holds for non-negative-definite matrices by adding the identity matrix before diagonalisation, so let
\[
A = QDQ^T
\]
where \( Q \) is an \( n \times n \) and \( \mathcal{G} \)-measurable orthogonal matrix and \( D \) is an \( n \times n \) and \( \mathcal{G} \)-measurable diagonal matrix with non-negative entries. Also let \( \delta_i \) be independent of each other and \( \mathcal{G} \) with \( \delta_i \sim \mathcal{N}(0, \sigma^2) \). Then by Lemma 7 we have
\[
\mathbb{E}(\exp(t \varepsilon^T A \varepsilon) | \mathcal{G}) \leq \mathbb{E}(\exp(t \delta^T A \delta) | \mathcal{G}) = \mathbb{E}(\exp(t \delta^T D \delta) | \mathcal{G})
\]
for all \( t \geq 0 \). Furthermore,
\[
\mathbb{E}(\exp(t \delta^2_i / \sigma^2)) = \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{1/2}} \exp(tx^2 - x^2/2) dx = \frac{1}{(1 - 2t)^{1/2}}
\]
so
\[
\mathbb{E}(\exp(t(\delta^2_i / \sigma^2 - 1))) = \exp(-(\log(1 - 2t) + 2t)/2).
\]
Hence,
\[
-2(\log(1 - 2t) + 2t) \leq \sum_{i=2}^{\infty} (2t)^i (2/i) \leq 4t^2 / (1 - 2t)
\]
for \( 0 \leq t \leq 1/2 \). Therefore
\[
\mathbb{E}(\exp(tD_{i,i}(\delta^2_i - \sigma^2)) | \mathcal{G}) \leq \exp\left( \frac{\sigma^4 D^2_{i,i} t^2}{1 - 2\sigma^2 D_{i,i} t} \right)
\]
for \( 0 \leq t \leq 1/(2\sigma^2 D_{i,i}) \) for all \( 1 \leq i \leq n \) and
\[
\mathbb{E}(\exp(t(\delta^T D \delta - \sigma^2 \text{tr}(D))) | \mathcal{G}) \leq \exp\left( \frac{\sigma^4 \text{tr}(D^2) t^2}{1 - 2\sigma^2 (\max_i D_{i,i}) t} \right)
\]
for \( 0 \leq t \leq 1/(2\sigma^2 (\max_i D_{i,i})) \). By Chernoff bounding we have
\[
\varepsilon^T A \varepsilon - \sigma^2 \text{tr}(A) > s
\]
for $s \geq 0$ with probability at most
\[
\exp \left( -\frac{\sigma^4 \text{tr}(A^2)t^2}{1 - 2\sigma^2(\max_i \lambda_i)t} - ts \right)
\]
conditional on $\mathcal{G}$ for $0 \leq t \leq 1/(2\sigma^2(\max_i \lambda_i))$. Letting
\[
t = \frac{s}{2\sigma^4 \text{tr}(A^2) + 2\sigma^2(\max_i \lambda_i)s}
\]
gives the bound
\[
\exp \left( -\frac{s^2}{4\sigma^4 \text{tr}(A^2) + 4\sigma^2(\max_i \lambda_i)s} \right).
\]
Rearranging gives the result.

We prove Lemma 9 by proving the bound on each of the two terms on the right-hand side of (4) individually in the following two lemmas.

**Lemma 12** Let $t > 0$. With probability at least $1 - 2e^{-t}$ we have
\[
\frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i))(\hat{h}_r(X_i) - h_{r,\alpha}(X_i)) \leq \|k\|_{\infty} \sigma r \left( \left(1 + t + 2(t^2 + t)^{1/2}\right)^{1/2} + (2t)^{1/2} \right) \left(\frac{1}{n^{1/2}}\right).
\]

**Proof** From the proof of Lemma 3 we have
\[
\frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i))\hat{h}_r(X_i) \leq r \left( \frac{1}{n^2} \sum_{i,j=1}^{n} (Y_i - g(X_i))(Y_j - g(X_j))k(X_i, X_j) \right)^{1/2}.
\]

Let $K$ be the $n \times n$ matrix with $K_{i,j} = k(X_i, X_j)$ for $1 \leq i, j \leq n$ and let $\varepsilon$ be the vector of the $Y_i - g(X_i)$. Then
\[
\frac{1}{n^2} \sum_{i,j=1}^{n} (Y_i - g(X_i))(Y_j - g(X_j))k(X_i, X_j) = \varepsilon^T (n^{-2}K)\varepsilon.
\]

Furthermore, since $k$ is a measurable kernel we have that $n^{-2}K$ is measurable and non-negative-definite. Let $\lambda_i$ for $1 \leq i \leq n$ be the eigenvalues of $n^{-2}K$. Then
\[
\max_i \lambda_i \leq \text{tr}(n^{-2}K) \leq n^{-1}\|k\|_{\infty}^2
\]
and
\[
\text{tr}((n^{-2}K)^2) = \|\lambda\|_2^2 \leq \|\lambda\|_1^2 \leq n^{-2}\|k\|_{\infty}^4.
\]

Therefore by Lemma 8 we have
\[
\varepsilon^T (n^{-2}K)\varepsilon \leq \|k\|_{\infty}^2 \sigma^2 n^{-1}(1 + t + 2(t^2 + t)^{1/2}) \left(\frac{1}{n^{1/2}}\right)
\]
and
\[
\frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i))\hat{h}_r(X_i) \leq \|k\|_{\infty} \sigma r \left(1 + t + 2(t^2 + t)^{1/2}\right)^{1/2} \left(\frac{1}{n^{1/2}}\right).
\]
with probability at least $1 - e^{-t}$. Finally,

$$-\frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i))h_{r,\alpha}(X_i)$$

is subgaussian given $X$ with parameter

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

So for $t > 0$ we have

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

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

Lemma 13 Let $t > 0$. With probability at least $1 - e^{-t}$ we have

$$\|h_{r,\alpha} - g\|_{L^2(P_n)}^2 \leq \frac{B^2(1 - \beta)}{t^{2\beta/(1 - \beta)}} + \frac{(C + \|k\|_\infty^2 r)^2 (2t)^{1/2}}{n^{1/2}} + \alpha$$

Proof We have

$$\frac{1}{n} \sum_{i=1}^{n} (h_{r,\alpha}(X_i) - g(X_i))^2 = \frac{1}{n} \sum_{i=1}^{n} \left( (h_{r,\alpha}(X_i) - g(X_i))^2 - \|h_{r,\alpha} - g\|_{L^2(P)}^2 \right) + \|h_{r,\alpha} - g\|_{L^2(P)}^2.$$

Furthermore,

$$\left| (h_{r,\alpha}(X_i) - g(X_i))^2 - \|h_{r,\alpha} - g\|_{L^2(P)}^2 \right| \leq \max \left\{ (h_{r,\alpha}(X_i) - g(X_i))^2, \|h_{r,\alpha} - g\|_{L^2(P)}^2 \right\} \leq (C + \|k\|_\infty^2 r)^2$$

since $h_{r,\alpha} \in rB_H$ implies that $\|h_{r,\alpha}\|_\infty \leq \|k\|_\infty r$ and we assume that $\|g\|_\infty \leq C$. Hence, by Hoeffding’s inequality we have

$$\frac{1}{n} \sum_{i=1}^{n} \left( (h_{r,\alpha}(X_i) - g(X_i))^2 - \|h_{r,\alpha} - g\|_{L^2(P)}^2 \right) > t$$

for $t > 0$ with probability at most

$$\exp \left( -\frac{n t^2}{2(C + \|k\|_\infty^2 r)^4} \right).$$

Therefore we have

$$\frac{1}{n} \sum_{i=1}^{n} \left( (h_{r,\alpha}(X_i) - g(X_i))^2 - \|h_{r,\alpha} - g\|_{L^2(P)}^2 \right) \leq \frac{(C + \|k\|_\infty^2 r)^2 (2t)^{1/2}}{n^{1/2}}.$$
for $t > 0$ with probability at least $1 - e^{-t}$. The result follows from (3).

**Proof of Lemma 10** Let

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

We have

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

$$n^{-1} \left| f(X_i)^2 - \|f\|_{L^2(P)}^2 \right| \leq \frac{1}{n} ||k||_\infty^4 r^2,$$

$$\mathbb{E}\left( n^{-2} \left( f(X_i)^2 - \|f\|_{L^2(P)}^2 \right)^2 \right) \leq \frac{1}{n^2} ||k||_\infty^4 r^4$$

for all $1 \leq i \leq n$. Furthermore $rB_H$ is separable so we can use Talagrand’s inequality (Theorem A.9.1 of Steinwart and Christmann, 2008) to show

$$Z > \mathbb{E}(Z) + 2t \left( \frac{||k||_\infty^4 r^4}{n} + \frac{2||k||_\infty^2 r^2 \mathbb{E}(Z)}{n} \right)^{1/2} + \frac{2t||k||_\infty^2 r^2}{3n}$$

for $t > 0$ with probability at most $e^{-t}$. From Lemma 4 we have

$$\mathbb{E}(Z) \leq \frac{8||k||_\infty^2 r^2}{n^{1/2}}.$$

The result follows.

**Appendix C. Proof of Adaptive Bound**

**Proof of Theorem 11** We follow the proof of Theorem 7.24 of Steinwart and Christmann (2008). Fix $r \in R$. From Lemma 6 and (6) we have

$$\| \hat{V}_r - g \|_{L^2(P)}^2 \leq \frac{8(4M^2 + 10||k||_\infty Mr + 18||k||_\infty^2 r^2)t^{1/2}}{(pn - 1)} + \frac{10B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}} + \frac{16||k||_\infty^2 r^2 t}{3(pm - 1)}$$

with probability at least $1 - 4e^{-t}$ for $t \geq 1$. Hence, the same inequality holds for all $r \in R$ with probability at least $1 - 4(1+I)e^{-t}$ for $t \geq 1$. Theorem 7.2 and Example 7.3 of Steinwart and Christmann (2008) show

$$\| \hat{V}_r - g \|_{L^2(P)}^2 \leq \min_{r \in R} \| \hat{V}_r - g \|_{L^2(P)}^2 + \frac{512M^2(t + \log(2 + I))}{(1 - p)n}$$

with probability at least $1 - e^{-t}$ for $t > 0$ which gives the upper bound

$$\min_{r \in R} \left( \frac{8(4M^2 + 10||k||_\infty Mr + 18||k||_\infty^2 r^2)t^{1/2}}{(pn - 1)} + \frac{10B^{2/(1-\beta)}}{r^{2\beta/(1-\beta)}} + \frac{16||k||_\infty^2 r^2 t}{3(pm - 1)} \right) + \frac{512M^2(t + \log(2 + I))}{(1 - p)n}$$

19
with probability at least $1-(5+4I)e^{-t}$ for $t \geq 1$. Since $r_0 = a$, $r_I = an^{1/4}$ and $r_i^2 - r_{i-1}^2 \leq b^2$ for all $1 \leq i \leq I$, there is some $0 \leq i \leq I$ such that
\[
a^2n^{(1-\beta)/2} \leq r_i^2 \leq a^2n^{(1-\beta)/2} + b^2.
\]
Bounding the minimum over $R$ using the value at $r_i$ shows
\[
\|V_{\hat{\psi}} - g\|_{L^2(P)}^2 \leq \frac{48(4M^2 + 10a\|k\|_{\infty}Mn^{1-\beta/4} + 18a^2\|k\|_{\infty}^2n^{(1-\beta)/2})t^{1/2}}{(pn-1)^{1/2}} + \frac{60B^{2/(1-\beta)}}{a^{2\beta/(1-\beta)n^{\beta/2}}}
\]
\[
+ \frac{96a^2\|k\|_{\infty}^2n^{(1-\beta)/2}t}{3(pn-1)} + \frac{48(10b\|k\|_{\infty}M + 18b^2\|k\|_{\infty}^2)t^{1/2}}{(pn-1)^{1/2}} + \frac{96b^2\|k\|_{\infty}^2t}{3(pn-1)}
\]
\[
+ \frac{512M^2(t + \log(2 + I))}{(1-p)n}
\]
for $t \geq 1$ with probability at least
\[
1 - (5+4I)e^{-t} \geq 1 - (4(a/b)^2n^{1/2} + 9)e^{-t}
\]
since
\[
I \leq (a/b)^2n^{1/2} + 1.
\]
Replacing $t$ with
\[
\log(4(a/b)^2n^{1/2} + 9) + t
\]
gives a bound on $\|V_{\hat{\psi}} - g\|_{L^2(P)}^2$ of order
\[
\log(n)^{1/2}n^{-\beta/2} + t^{1/2}n^{-\beta/2}
\]
in $n$ and $t \geq 1$ with probability at least $1 - e^{-t}$.

References


