Estimation in Reproducing Kernel Hilbert Spaces with Dependent Data

Alessio Sancetta[*]

November 23, 2020

Abstract

This paper derives consistency results for estimation in the finite direct sum of reproducing kernel Hilbert spaces (RKHS) for dependent data. The link between penalized and constrained estimation is established. We consider the relation between topological equivalent norms for direct sums of RKHS. These norms have different implications for estimation. Estimation in a ball of the RKHS defined by these norms essentially results in estimation with a ridge and Lasso penalty, respectively. A greedy algorithm for the solution of the hesitation problem under these two norms is discussed for general loss functions.

Key Words: Constrained estimation, convergence rates, nonlinear model, reproducing kernel Hilbert space.

Contents

[*] Acknowledgments: I am grateful to the referees and the Associate Editor Radu Balan for suggestions and remarks that stimulated substantial improvements both in content and presentation. E-mail: <asancetta@gmail.com>, URL: <http://sites.google.com/site/wwwsancetta/>. Address for correspondence: Department of Economics, Royal Holloway University of London, Egham TW20 0EX, UK

1 Introduction

This paper studies estimation of additive models in reproducing kernel Hilbert spaces (RKHS) when the data are dependent and there is possibly no true model. Instead of a true model, the target is the minimizer of a population objective function. For the sake of definiteness suppose that we want to estimate the number of event arrivals Y in the next one minute, conditioning on a vector of covariates X known at the start of the interval. We decide to minimize the negative log-likelihood for Poisson arrivals with conditional intensity $\exp\{\mu(X)\}$ for some function μ. For observation i, the negative loglikelihood is proportional to

$$\exp\{\mu(X_i)\} - Y_i\mu(X_i). \tag{1}$$

To avoid the curse of dimensionality, we choose μ to lie in the space $\mathcal{H}^K$ where each element can be written as

$$\mu(X_i) = \sum_{k=1}^{K} f^{(k)}\left(X_i^{(k)}\right) \tag{2}$$

where $X_i^{(k)}$ denotes the k^{th} element in the K-dimensional covariate X_i, and the univariate functions $f^{(k)} \in \mathcal{H}$ where $\mathcal{H}$ is a RKHS, possibly infinite dimensional, $k = 1, 2, ..., K$. The notation implies that $\mathcal{H}^K = \bigoplus_{k=1}^{K} \mathcal{H}$ is the direct sum of RKHS. The target parameter is the minimizer of the expectation of (1) with respect to μ in $\mathcal{H}^K(B)$, the ball of radius B in $\mathcal{H}^K$ centered at zero. This is the constrained population parameter. The sample constrained estimator is the minimizer in $\mathcal{H}^K(B)$ of the sample mean of (1). We also consider the penalized sample estimator, which is an alternative to the constrained estimator. In this case, the target parameter is the minimizer of the expectation of (1) with respect to μ in $\mathcal{H}^K$, the unconstrained population parameter. This latter target may not lie in $\mathcal{H}^K(B)$ for any $B < \infty$.

A RKHS can be generated by the measure of a Gaussian process (Li and Linde, 1999, for precise definitions). Results on the small probability of Gaussian processes (Li and Linde, 1999) provide an estimate of the metric entropy of the RKHS. This estimate can then be used in a well known maximal inequality for beta mixing random variables (Doukhan et al., 1995). This provides the control of the estimator and allows us to derive convergence rates under mixing assumptions. While the literature provides the tools, the existing proof for consistency of estimators in RKHS needs to be modified in order to derive convergence rates. We also provide consistency under the uniform norm using complementary weak assumptions including only stationarity and ergodicity of the data. We study both the penalized and the constrained estimation problem. The penalized estimation problem is usually referred to as estimation using support vector machines (Christmann and Steinwart, 2007). On the other hand estimation in a ball of fixed radius under the RKHS norm is traditionally referred as estimation in RKHS (Mendelson, 2002). By duality, there is a link between the penalized and constrained estimation, and we provide details of this relation. We consider the linear functional defining the first order condition in the estimation problem, and show its convergence to a Gaussian process. We then consider two different norms in order to define either the constraint or the penalty and discuss the variable selection and shrinkage properties of both for the estimation of additive models. One norm is the standard norm for the direct sum of RHKS. Estimation in a ball defined under such norm is equivalent to estimation with a ridge penalty. The second is the ℓ_1 norm of the individual RKHS norms and estimation under such norm mimics Lasso. The estimation results are obtained allowing for estimators that are asymptotic minimizers of the objective function rather than exact ones. Hence, we provide an algorithm that can solve the estimation problem for either norms in $O\left(n^2\epsilon^{-1}\right)$ time, where ϵ is the resulting error of the algorithm and n is the sample size. Hence, relatively to other algorithms it is less resource efficient. Nevertheless, the algorithm is simple to implement, does not rely on a randomly selected subset of inputs (Rasmussen and Williams, 2006, Ch.8), and provides a solution to the constrained estimator without assuming sparseness. For such algorithm we derive convergence rates.

1.1 Relation to the Literature

Estimation in RKHS has been addressed in many places in the literature (see the monographs of Wahba, 1990, and Steinwart and Christmann, 2008). Inference is usually confined to consistency (Mendelson, 2002, Christmann and Steinwart, 2007), though there are exceptions (Hable, 2012, in the frequentist framework). Estimation of additive models has been extensively studied by various authors using different techniques (Buja et al., 1989, Mammen et al., 1999, Meier et al., 2009, Christmann and Hable, 2012). The last reference considers estimation in RKHS. These references, including the present paper, focus on the case where the number of additive components is fixed. Recently, Suzuki and Sugiyama (2013), Lv et al. (2018) and Suzuki (2018) have considered penalized estimation in RKHS when the number of additive components can diverge to infinity, also allowing for different penalties.

Estimation in RKHS under dependence has been the subject of study of some researchers (inter alia, Steinwart et al., 2009a, 2009b, Hang and Steinwart, 2014, 2017). These references bound the difference between the expected risk evaluated at the sample and population parameters. Optimal rates have also been derived under rather technical conditions. However, this is not sufficient to derive sharp convergence rates under say the L_2 norm. In this case, on top of a concentration inequality, the argument requires a chaining argument to explicitly bound the local behaviour of the centered empirical risk when the parameter space is uncountable (van der Vaart and Wellner, 2000, Ch.3.2).

The assumptions and estimation results presented here are not overall directly comparable to the reviewed results. This paper adds to this existing literature as the focus is on consistency of the estimator under different norms, and on convergence rates for the L_2 norm. In particular, Theorem 1 shows consistency of the estimator under the uniform norm under the sole condition of stationarity and ergodicity of the data. Theorem 2 uses mixing conditions to show consistency under the RKHS norm. Theorem 3 derives convergence rates under the L_2 norm using mixing conditions. In Theorem 4 we also show a weak convergence result that complements the one of Hable (2012).

For relatively large sample sizes (e.g. greater than 10,000) estimation in RKHS can be challenging. For example, for the regression problem under the square error loss, estimation would require inversion of a high dimensional matrix whose size grows with the sample size. Computational aspects in RKHS have received a lot of attention in the literature, though mostly for the regression problem under the square error loss (Lázaro-Gredilla et al., 2010, Banerjee et al., 2013, and references therein). Here we discuss a greedy algorithm, which is simple to implement and is not restricted to the regression problem (Jaggi, 2013, Sancetta, 2016). Greedy algorithms have been applied by various authors (Smola and Bartlett, 2001, Nair et al., 2002, and references therein). The algorithm discussed herein allows us to solve the constrained estimation problem for general loss functions, and for this algorithm we derive convergence rates in Theorem 5.

1.2 Outline

The plan for the paper is as follows. Section 2.2 reviews some basics of RKHS, and Section 2.3 defines the estimation problem. Section 2.4 contains the consistency and weak convergence results. Section 3 discusses the conditions used to derive the asymptotic results and introduces an alternative constraint. The algorithms for computational implementation under either of these constraints can be found in Section 4. The proofs are in Section 5.

2 The Inference Problem

2.1 Problem Setup

The explanatory variable $X^{(k)}$ takes values in $\mathcal{X}$, a compact subset of a separable Banach space ($k = 1, 2, ..., K$). The most basic example of $\mathcal{X}$ is $[0, 1]$. The vector covariate $X = \left(X^{(1)}, ..., X^{(K)}\right)$ takes values in the Cartesian product $\mathcal{X}^K$, e.g., $[0, 1]^K$. The dependent variable takes values in $\mathcal{Y}$ usually $\mathbb{R}$. Let $Z = (Y, X)$ and this takes values in $\mathcal{Z} = \mathcal{Y} \times \mathcal{X}^K$. If no dependent variable Y can be defined, as for unsupervised learning, or certain likelihood estimators, we set $Z = X$. Let P be the law of Z, and use linear functional notation: for any $f : \mathcal{Z} \to \mathbb{R}$, $Pf = \int_{\mathcal{Z}} f(z)\, dP(z)$. Let $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$, where δ_{Z_i} is the point mass at Z_i, implying that $P_n f = \frac{1}{n}\sum_{i=1}^n f(Z_i)$ is the sample mean of $f(Z)$. For $p \in [1, \infty]$, let $|\cdot|_p$ be the L_p norm (w.r.t. the measure P): for $f : \mathcal{Z} \to \mathbb{R}$, $|f|_p = (P|f|^p)^{1/p}$, with the obvious modification to sup norm when $p = \infty$.

We let $\mathcal{H}^K$ be a vector space of real valued functions on $\mathcal{X}^K$, equipped with a norm $|\cdot|_{\mathcal{H}^K}$. The loss function is defined as $L : \mathcal{Z} \times \mathbb{R} \rightarrow \mathbb{R}$. We are interested in the case where the second argument is $\mu(x)$, i.e. $L(z, \mu(x))$ with $\mu \in \mathcal{H}^K$. Therefore, to keep notation compact, let $\ell_\mu(Z) = L(Z, \mu(X))$. For the special case of the square error loss we would have $\ell_\mu(z) = L(z, \mu(x)) = |y - \mu(x)|^2$ $(z = (y, x))$. The use of ℓ_μ makes it more natural to use linear functional notation so that $P_n \ell_\mu$ is the empirical risk at μ.

2.2 Basic Facts about Reproducing Kernel Hilbert Spaces

Recall that a real RKHS $\mathcal{H}$ on some set $\mathcal{X}$ is a Hilbert space where the evaluation functional e_x which associates f with $f(x)$ is a bounded linear functional: $f(x) = e_x f$, $f \in \mathcal{H}$ (Wahba, 1990, p.2). A RKHS of bounded functions is uniquely generated by a centered Gaussian measure with covariance C (Li and Linde, 1999) and C is usually called the (reproducing) kernel of $\mathcal{H}$. We consider covariance functions with representation

$$C(s, t) = \sum_{v=1}^{\infty} \lambda_v^2 \varphi_v(s) \varphi_v(t), \tag{3}$$

for linearly independent functions $\varphi_v : \mathcal{X} \rightarrow \mathbb{R}$ and coefficients λ_v such that $\sum_{v=1}^{\infty} \lambda_v^2 \varphi_v^2(s) < \infty$. Here, linear independent means that if there is a sequence of real numbers $(f_v)_{v \geq 1}$ such that $\sum_{v=1}^{\infty} f_v^2 / \lambda_v^2 < \infty$ and $\sum_{v=1}^{\infty} f_v \varphi_v(s) = 0$ for all $s \in \mathcal{X}$, then $f_v = 0$ for all $v \geq 1$. The coefficients λ_v^2 would be the eigenvalues of (3) if the functions φ_v were orthonormal, but this is not implied by the above definition of linear independence. The RKHS $\mathcal{H}$ is the completion of the set of functions representable as $f(x) = \sum_{v=1}^{\infty} f_v \varphi_v(x)$ for real valued coefficients f_v as above. Equivalently, $f(x) = \sum_{j=1}^{\infty} \alpha_j C(s_j, x)$, for coefficients s_j in $\mathcal{X}$ and real valued coefficients α_j satisfying $\sum_{j=1}^{\infty} \alpha_i \alpha_j C(s_i, s_j) < \infty$. Moreover, for C in (3),

$$\sum_{j=1}^{\infty} \alpha_j C(s_j, x) = \sum_{v=1}^{\infty} \left(\sum_{j=1}^{\infty} \alpha_j \lambda_v^2 \varphi_v(s_j) \right) \varphi_v(x) = \sum_{v=1}^{\infty} f_v \varphi_v(x) \tag{4}$$

by obvious definition of the coefficients f_v. The change of summation is possible by the aforementioned restrictions on the coefficients λ_v and functions φ_v. The inner product in $\mathcal{H}$ is denoted by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and satisfies $f(x) = \langle f, C(x, \cdot) \rangle_{\mathcal{H}}$. This implies the reproducing kernel property $C(s, t) = \langle C(s, \cdot), C(t, \cdot) \rangle_{\mathcal{H}}$. Therefore, the square of the RKHS norm is defined in the two following equivalent ways

$$|f|_{\mathcal{H}}^2 = \sum_{v=1}^{\infty} \frac{f_v^2}{\lambda_v^2} = \sum_{i,j=1}^{\infty} \alpha_i \alpha_j C(s_i, s_j). \tag{5}$$

Throughout, the unit ball of $\mathcal{H}$ will be denoted by $\mathcal{H}(1) := \{f \in \mathcal{H} : |f|_{\mathcal{H}} \leq 1\}$.

The additive RKHS is generated by the Gaussian measure with covariance function $C_{\mathcal{H}^K}(s, t) = \sum_{k=1}^{K} C\left(s^{(k)}, t^{(k)}\right)$, where C is as in (3), and $s^{(k)}$ is the k^{th} element in $s \in \mathcal{X}^K$. The RKHS of additive functions is denoted by $\mathcal{H}^K$, which is the set of functions as in (2) such that $f^{(k)} \in \mathcal{H}$ and $\sum_{k=1}^{K} \left| f^{(k)} \right|_{\mathcal{H}_k}^2 < \infty$. For such functions, the inner product is $\langle f, g \rangle_{\mathcal{H}^K} = \sum_{k=1}^{K} \left\langle f^{(k)}, g^{(k)} \right\rangle_{\mathcal{H}_k}$. The norm $|\cdot|_{\mathcal{H}^K}$ on $\mathcal{H}^K$ is the one induced by the inner product. For notational simplicity we have set the individual RKHS to be the same.

Within this scenario, the space $\mathcal{H}^K$ restricts functions to be additive, where these additive functions in $\mathcal{H}$ can be multivariate functions.

Example 1 *Suppose that $K = 1$ and $\mathcal{X} = [0,1]^d$ $(d > 1)$ (only one additive function, which is multivariate). Let $C(s,t) = \exp\left\{-a\sum_j |s_j - t_j|^2\right\}$ where s_j is the j^{th} element in $s \in [0,1]^d$, and $a > 0$. Then, the RKHS $\mathcal{H}$ is dense in the space of continuous bounded functions on $[0,1]^d$ (e.g., Christmann and Steinwart, 2007). A (kernel) C with such property is called universal.*

The framework also covers the case of functional data because $\mathcal{X}$ is a compact subset of a Banach space. Most problems of interest where the unknown parameter μ is a smooth function are covered by the current scenario.

2.3 Estimation

Estimation will be considered for models in $\mathcal{H}^K(B) := \left\{f \in \mathcal{H}^K : |f|_{\mathcal{H}^K} \leq B\right\}$, where $B < \infty$ is a fixed constant. The goal is to find

$$\mu_n = \arg\inf_{\mu\in\mathcal{H}^K(B)} P_n\ell_\mu, \tag{6}$$

i.e. the minimizer with respect to $\mu \in \mathcal{H}^K(B)$ of the loss function $P_n\ell_\mu$.

Example 2 *Let $\ell_\mu(z) = |y - \mu(x)|^2$ so that*

$$P_n\ell_\mu = \frac{1}{n}\sum_{i=1}^n \ell_\mu(Z_i) = \frac{1}{n}\sum_{i=1}^n |Y_i - \mu(X_i)|^2.$$

By duality, we can also use $P_n\ell_\mu + \rho_{B,n}|\mu|^2_{\mathcal{H}^K}$ with sample dependent Lagrange multiplier $\rho_{B,n}$ such that the solution is in $\mathcal{H}^K(B)$.

For the square error loss the solution is just a ridge regression estimator with (random) ridge parameter $\rho_{B,n} \geq 0$. Interest is not restricted to least square problems.

The Representer Theorem (Steinwart and Christmann, 2008, Theorem 5.8) says that the solution to the penalized problem takes the form $\mu_n(x) = \sum_{i=1}^n \alpha_i C(X_i, x)$ for real valued coefficients α_i. Even if $\mathcal{H}^K$ is infinite dimensional, μ_n lies in a finite dimensional space.

The target constrained population estimator is

$$\mu_B = \arg\inf_{\mu\in\mathcal{H}^K(B)} P\ell_\mu. \tag{7}$$

We shall show that this minimizer always exists and is unique, under regularity conditions on the loss, because $\mathcal{H}^K(B)$ is closed. The unconstrained population estimator is the minimizer of

$$\mu_0 = \arg\inf_{\mu\in\mathcal{H}^K} P\ell_\mu. \tag{8}$$

This quantity is not necessarily well defined in the sense that we can have that $|\mu_0|_{\mathcal{H}^K} = \infty$.

In what follows we assume K to be a bounded integer. In consequence, the asymptotic results for the estimator μ_n hold irrespective of the value of K and we could take $K = 1$. However, there are practical differences regarding estimation between the case $K > 1$ and $K = 1$. These are discussed in Sections 3.2 and 4.

2.4 Asymptotic Analysis

Throughout the paper, $\lesssim$ means that the l.h.s. is bounded by an absolute constant times the r.h.s.. We use Big-O, little-o notation and add a subscript P when the relation holds in probability. We recall the definition of beta mixing. Suppose that $(Z_i)_{i\in\mathbb{Z}}$ is a strictly stationary sequence of random variables and let $\sigma(Z_i : i \leq 0)$, $\sigma(Z_i : i \geq k)$ be the sigma algebra generated by $(Z_i)_{i\leq 0}$ and $(Z_i)_{i\geq k}$, respectively, for integer k. For any $k \geq 1$, the beta mixing coefficient $\beta(k)$ for $(Z_i)_{i\in\mathbb{Z}}$ is

$$\beta(k) := \mathbb{E} \sup_{A\in\sigma(Z_i:i\geq k)} |\Pr(A|\sigma(Z_i : i \leq 0)) - \Pr(A)|$$

(Doukhan, 1995, for an equivalent definition). We now introduce the following technical conditions.

Condition 1 *The set $\mathcal{H}$ is a RKHS on a compact subset of a separable Banach space $\mathcal{X}$, with continuous uniformly bounded kernel C admitting an expansion (3), where $\lambda_v^2 \lesssim v^{-2\eta}$ with exponent $\eta > 1/2$ and with linearly independent continuous functions $\varphi_v : \mathcal{X} \to \mathbb{R}$, uniformly bounded, uniformly in $v \geq 1$.*

Recall the definition of the loss $L(z, t)$ in Section 2. Let $\bar{B} := c_K B$ where $c_K := \max_{s\in\mathcal{X}^K} \sqrt{C_{\mathcal{H}^K}(s, s)}$. Define $\Delta_k(z) := \max_{|t|\leq\bar{B}} \left|\partial^k L(z, t) / \partial t^k\right|$ for $k = 0, 1, 2, \ldots$ if the derivative exists. Attention is restricted to loss functions satisfying the following.

Condition 2 *The loss $L(z, t)$ is non-negative, twice continuously differentiable for real t in an open set containing $\left[-\bar{B}, \bar{B}\right]$, and $\inf_{z,t} d^2L(z, t)/dt^2 > 0$ for $z \in \mathcal{Z}$ and $t \in \left[-\bar{B}, \bar{B}\right]$. Moreover, $P(\Delta_0 + \Delta_1^p + \Delta_2^p) < \infty$ for some $p > 2$.*

The following dependence condition will be used.

Condition 3 *The sequence $(Z_i)_{i\in\mathbb{Z}}$ ($Z_i = (Y_i, X_i)$) is strictly stationary with beta mixing coefficients $\beta(i)$ satisfying $\beta(i) \lesssim i^{-\beta_0}$ with $\beta_0 > r/(r-2)$ for some $r > 2$, for all $i \geq 1$.*

Remarks on the conditions, including examples regarding the beta mixing condition can be found in Section 3.1. Throughout, we may omit the qualifier strictly when mentioning stationarity.

Consistency. This section shows the consistency of the constrained estimator. We also provide details regarding the relation between constrained, and penalized estimators and convergence rates. The usual penalized estimator is defined as

$$\mu_{n,\rho} = \arg\inf_{\mu\in\mathcal{H}^K} P_n\ell_\mu + \rho\,|\mu|^2_{\mathcal{H}^K} \tag{9}$$

for $\rho \geq 0$. As mentioned in Example 2, suitable choice of ρ leads to the constrained estimator. In particular, we can choose the largest ρ such that $\mu_{n,p}$ still lies in $\mathcal{H}^K(B)$ so that $\mu_{n,p} = \mu_n$ with the r.h.s. as in (6). The results we shall show will remain true if the empirical minimizers are replaced with approximate minimizer. In particular, for some $\epsilon_n \to 0$ in some mode of convergence to be specified by the application, we can consider any μ_n satisfying

$$P_n \ell_{\mu_n} = \inf_{\mu \in \mathcal{H}^K(B)} P_n \ell_\mu + \epsilon_n \tag{10}$$

and any $\mu_{n,\rho}$ satisfying

$$P_n \ell_{\mu_{n,\rho}} + \rho \left|\mu_{n,\rho}\right|_{\mathcal{H}^K} = \inf_{\mu \in \mathcal{H}^K} \left\{P_n \ell_\mu + \rho \left|\mu\right|_{\mathcal{H}^K}\right\} + \rho \epsilon_n. \tag{11}$$

At first we show that the constrained estimator is consistent under the minimal condition of stationarity and ergodicity: strictly stationary processes whose invariant sets are trivial (Kallenberg, 1997, Ch.9).

Theorem 1 *Consider the problem in (6) with fixed $B < \infty$. Suppose Condition 1, Condition 2 with $p = 1$, and that the random variables $(Z_i)_{i \in \mathbb{Z}}$ are stationary and ergodic. Then, $\left|\mu_n - \mu_B\right|_\infty = o(1)$ almost surely, where the population minimizer μ_B in (7) is unique up to an L_2 equivalence. The result continues to hold for any μ_n satisfying (10) with $\epsilon_n = o(1)$ almost surely.*

We now derive stronger results for the penalized estimator (9) under relatively stronger conditions. Throughout, $\mathrm{int}\left(\mathcal{H}^K(B)\right)$ will denote the interior of $\mathcal{H}^K(B)$.

Theorem 2 *Suppose Condition 1 with $\eta > 1$, and Conditions 2, and 3 with $p \geq r$.*

1. *If $\mu_0 \in \mathcal{H}^K(B)$, there is a random $\rho = \rho_{B,n}$ such that $\rho n^{1/2} = O_P(1)$, and $\mu_{n,\rho} = \mu_n$ (μ_n and $\mu_{n,\rho}$ as in (6) and (9)).*

2. *Consider possibly random $\rho = \rho_n$ such that $\rho \to 0$ and $\rho n^{1/2} \to \infty$ in probability. Suppose that there is a finite B such that $\mu_0 \in \mathrm{int}\left(\mathcal{H}^K(B)\right)$. Then, $\left|\mu_{n,\rho} - \mu_0\right|_{\mathcal{H}^K} \to 0$ in probability, and in consequence $\left|\mu_{n,\rho}\right|_{\mathcal{H}^K} < B$ with probability going to one.*

3. *If $\mathcal{H}^K$ is infinite dimensional and $\mu_0 \in \mathcal{H}^K$, then there is a $\rho = \rho_n$ such that $\rho \to 0$, $\rho n^{1/2} \to c < \infty$, and $\left|\mu_{n,\rho} - \mu_0\right|_\infty \to 0$ in probability, but $\left|\mu_{n,\rho} - \mu_0\right|_{\mathcal{H}^K}$ does not converge to zero in probability.*

4. *All the above statements also hold if μ_n and $\mu_{n,\rho}$ in (6) and (9) are approximate minimizers as in (10) and (11), respectively with $\epsilon_n = o_P(1)$.*

Point 1 establishes the connection between the constrained estimator μ_n in (6) and the penalized estimator $\mu_{n,\rho}$ in (9). To establish the connection, we need $\mu_0 = \mu_B$. In this case, it is worth noting that whether $\mathcal{H}^K$ is finite or infinite dimensional, the estimator μ_n is equivalent to a penalized estimator with penalty parameter ρ going to zero relatively fast, as n goes to infinity. In particular, we rule out $\rho n^{1/2} \to \infty$. The condition $\rho n^{1/2} = O_P(1)$ only ensures consistency under the uniform norm, but not consistency under the RKHS norm $\left|\cdot\right|_{\mathcal{H}^K}$ (Points 2-3). Consistency under the RKHS norm requires ρ going to infinity slowly enough. In this case, the constrained and penalized estimator are not the same. In particular, the constrained estimator is not necessarily consistent under the RKHS norm (Point 3). When $\mathcal{H}^K$ is infinite dimensional, this happens because μ_n lies at the boundary of $\mathcal{H}^K(B)$.

We now focus on rates of convergence under the L_2 norm.

Theorem 3 *Suppose Condition 1 with $\eta > 1$, and Conditions 2, 3 with $p > r$. Consider μ_n in (6). We have that $|\mu_n - \mu_B|_2 = O_P\left(s_n^{-1}\right)$ where $s_n = n^{\frac{\gamma}{2}\left(\frac{2\eta-1}{2\eta+(\gamma-1)}\right)}$ and $\gamma := \frac{r}{2(r-1)}\left(\frac{p-1}{p}\right)$. The result also holds true for any μ_n satisfying (10) with $\epsilon_n = O_P\left(s_n^{-2}\right)$. Moreover, if $\mu_0 \in \operatorname{int}\left(\mathcal{H}^K(B)\right)$, we also have that $|\mu_n - \mu_0|_2 = O_P\left(\min\left\{s_n^{-1}, n^{-1/4}\right\}\right)$.*

As usual for RKHS estimators, the convergence rate does not depend on the dimension of $\mathcal{X}$. This is because, the restriction to $\mathcal{H}^K(B)$ implicitly imposes regularity conditions. To see what we mean, take $K = 1$ and $\mathcal{X} = [0,1]^d$ as in Example 1. As d increases the functions will have to be more regular in order to be in $\mathcal{H}(B)$. Because of additivity, the bounds only depend linearly on K, but this is not made explicit, as K is bounded.

The term $\gamma \in (0,1)$ is a penalty for dependence and the fact that the first derivative of the loss function is not bounded. Such derivative allow us to link the loss function to μ. When the dependence is arbitrarily weak and Δ_1 has a finite p moment for any p, we can essentially take $\gamma = 1$ ($r \to 2$ and $p \to \infty$). Then, the rate of convergence is $n^{-\frac{2\eta-1}{4\eta}}$. To put this rate of convergence into perspective, recall that the best rate of convergence of nonparametric estimators for V differentiable functions is $n^{-V/(2V+d)}$ where d is the dimension of $\mathcal{X}$ as in Example 1 (Stone, 1982). Let us consider a univariate case for ease of comparison. In Example 4 in Section 3 we shall recall that the Sobolev space of functions on $\mathcal{X} = [0,1]$ with V square integrable weak derivatives is a RKHS with a covariance kernel admitting the expansion (3) with $\lambda_v \lesssim v^{-V}$. For example, when $V = 2$, the optimal rate would be $n^{-2/5}$. On the other hand we can see that $s_n^{-1} = n^{-3/8}$ when $\gamma \to 1$, in Theorem 3.

Weak Convergence. We shall only consider the constrained estimator μ_n. To ease notation, for any arbitrary, but fixed real valued functions g and g' on $\mathcal{Z}$ define $P_{1,j}(g, g') = \mathbb{E}g(Z_1)\, g'(Z_{1+j})$. For suitable g and g', the quantity $\sum_{j\in\mathbb{Z}} P_{1,j}(g, g')$ will be used as short notation for sums of population covariances.

Theorem 4 *Suppose Condition 1 with $\eta > 1$, and Conditions 2, and 3 with $p \geq r$. If $\mu_0 \in \operatorname{int}\left(\mathcal{H}^K(B)\right)$, then*

$$\sqrt{n} P_n \partial\ell_{\mu_0} h \to G(h),\; h \in \mathcal{H}^K(1)$$

weakly, where $\left\{G(h) : h \in \mathcal{H}^K(1)\right\}$ is a mean zero Gaussian process with covariance function

$$\mathbb{E}G(h)\, G(h') = \sum_{j\in\mathbb{Z}} P_{1,j}\left(\partial\ell_{\mu_0} h, \partial\ell_{\mu_0} h'\right)$$

for any $h, h' \in \mathcal{H}^K(1)$.

In addition to the above, also suppose that $|\Delta_3|_p < \infty$ (in Condition 2) and that $|\mu_n - \mu_0|_2 = o_P\left(n^{-\frac{1}{4}\left(\frac{p}{p-1}\right)}\right)$. If $\mu_n \in \mathcal{H}^K(B)$ satisfies (10) with $\epsilon_n = o_P\left(n^{-1}\right)$, and $\sup_{h\in\mathcal{H}^K(1)} P_n \partial\ell_{\mu_n} h = o_P\left(n^{-1/2}\right)$, then,

$$\sqrt{n} P \partial^2\ell_{\mu_0}(\mu_n - \mu_0)\, h = \sqrt{n} P_n \partial\ell_{\mu_0} h + o_P(1),\; h \in \mathcal{H}^K(1).$$

We can use Theorem 3 to verify the condition on $|\mu_n - \mu_0|_2$. The second statement in Theorem 4 cannot be established for the penalized estimator with penalty satisfying $\rho \to 0$ such that $\rho n^{1/2} \to \infty$.

This is because in the first order conditions, the contribution from the penalty is non-negligible. The restriction $\sup_{h\in\mathcal{H}^K(1)} P_n\partial\ell_{\mu_n}h = o_p\left(n^{-1/2}\right)$ holds for finite dimensional models as long as $\mu_0 \in \operatorname{int}\left(\mathcal{H}^K(B)\right)$. For infinite dimensional models this is no longer true as the constraint is eventually binding even if $\mu_0 \in \operatorname{int}\left(\mathcal{H}^K(B)\right)$ (μ_n lies at the boundary of $\mathcal{H}^K(B)$ when the sample size is large enough). Then, it can be shown that the $o_P\left(n^{-1/2}\right)$ term has to be replaced with $O_P\left(n^{-1/2}\right)$ (Lemma 8, in the Appendix). The asymptotic distribution of the estimator is immediately derived if $\mathcal{H}^K(B)$ is finite dimensional.

Example 3 *Consider the rescaled square error loss so that $\partial^2\ell_{\mu_0} = 1$. Defining $\nu = \lim_n \sqrt{n}(\mu_n - \mu_0)$, Theorem 4 gives*

$$G(h) = P\nu h,$$

in distribution, where G is as in Theorem 4 as long as $\mu_0 \in \operatorname{int}\left(\mathcal{H}^K(B)\right)$. The distribution of ν is then given by the solution to the above display when $\mathcal{H}^K(B)$ is finite dimensional.

In the infinite dimensional case, Hable (2012) has shown that $\sqrt{n}(\mu_{n,\rho}(x) - \mu_{0,\rho}(x))$ converges to a Gaussian process whose covariance function would require the solution of some Fredholm equation of the second type. Here, $\mu_{n,\rho}$ is as in (9), while we use $\mu_{0,\rho}$ to denote its population version. The penalty $\rho = \rho_n$ needs to satisfy $\sqrt{n}(\rho_n - \rho_0) = o_P(1)$ for some fixed constant $\rho_0 > 0$. When $\mu_0 \in \operatorname{int}\left(\mathcal{H}^K(B)\right)$, we have $\mu_0 = \arg\min_{\mu\in\mathcal{H}} P\ell_\mu$. Hence, there is no $\rho_0 > 0$ such that $\mu_0 = \mu_{0,\rho_0}$. When the penalty does not go to zero, the approximation error is non-negligible, e.g. for the square loss the estimator is biased.

Theorem 4 requires $\mu_0 \in \operatorname{int}\left(\mathcal{H}^K(B)\right)$. The distribution of the estimator when μ_0 lies on the boundary of $\mathcal{H}^K(B)$ is not standard (Geyer, 1994, for the finite dimensional case) and is implicitly defined as the solution of a stochastic quadratic programming problem, similar to Example 3.

3 Discussion

3.1 Remarks on Conditions

The minimal decay condition for the coefficients λ_v is $\lambda_v \lesssim v^{-\eta}$ with $\eta > 1/2$ as this is essentially required for $\sum_{v=1}^{\infty}\lambda_v^2\varphi_v^2(s) < \infty$ for any $s \in \mathcal{X}$. Mendelson (2002) derives consistency of the empirical risk under this minimal condition in the i.i.d. case, but does not give convergence rates. Theorem 1 gives consistency under the uniform norm under this same minimal condition allowing for dependence. Theorems 2 and 3 show convergence under the RKHS norm, and derive L_2 convergence rates for $\eta > 1$. This stronger condition on η is not necessarily restrictive in practice. The covariance in Example 1 satisfies Condition 1 with exponentially decaying coefficients λ_v (Rasmussen and Williams, 2006, Ch. 4.3.1). Other covariance kernels satisfy this condition.

Example 4 *Suppose that $\mathcal{H}^K$ is an additive space of univariate functions, where each univariate function is an element in the Sobolev Hilbert space of index V on $[0,1]$, i.e. functions with V square integrable weak derivatives. Then, $C_{\mathcal{H}^K}(s,t) = \sum_{k=1}^{K} C\left(s^{(k)}, t^{(k)}\right)$ where $C\left(s^{(k)}, t^{(k)}\right) = \sum_{v=1}^{V-1}\left(s^{(k)}t^{(k)}\right)^v/(v!)^2 + H_V\left(s^{(k)}, t^{(k)}\right)$ and where H_V is the covariance function of the $(V-1)$-fold* *integrated Brownian motion. In particular,*

$$H_V(\cdot,\cdot) = \int_0^1 G_V(\cdot,u)\, G_V(\cdot,u)\, du \text{ with } G_V(r,u) := \max\left\{\frac{(r-u)^{V-1}}{(V-1)!}, 0\right\},$$

where $r,u \in [0,1]$ (Wahba, 1990, p.7-8). Then, the covariance C admits an expansion as in (3) with $\lambda_v \lesssim v^{-\eta}$ where $\eta = V$ (Ritter et al., 1995, Corollary 2, and 523-524).

When the coefficients λ_v^2 are the eigenvalues of C, the restriction on their decay rate is usually referred to as spectral assumption and $2\eta > 1$ has been assumed by other authors to bound the covering numbers of subsets of $\mathcal{H}$ (Steinwart et al., 2009, Suzuki and Sugiyama, 2013, Lv et al., 2018, Suzuki, 2018).

It is not difficult to see that many loss functions (or negative log-likelihoods) of interest satisfy Condition 2, using the fact that $|\mu|_\infty$ is bounded (square error loss, logistic, negative log-likelihood of Poisson, etc.). Nevertheless, interesting loss functions such as absolute deviation for conditional median estimation do not satisfy Condition 2. The extension to such loss functions requires arguments that are specific to the problem together with additional restrictions to compensate for the lack of smoothness. In the interest of space, this shall not be discussed here.

Condition 3 is a common dependence condition. Essentially, this condition is satisfied by any model that can be written as a Markov chain with smooth conditional distribution (Doukhan, 1995, for a review; Basrak et al., 2002, for GARCH). Models with innovations that do not have a smooth density function may not be covered (Bradley, 1986, Example 6.2).

Example 5 *(Regression) Suppose that $Y_i = \sum_{k=1}^K f^{(k)}\left(X_i^{(k)}\right) + \varepsilon_i$, where the sequence of random variable $(\varepsilon_i)_{i\in\mathbb{Z}}$ and $(X_i)_{i\in\mathbb{Z}}$ are independent of each other. By independence, the mixing coefficients of $(Y_i, X_i)_{i\in\mathbb{Z}}$ are bounded by the sum of the mixing coefficients of $(\varepsilon_i)_{i\in\mathbb{Z}}$ and $(X_i)_{i\in\mathbb{Z}}$ (Bradley, 1986, Theorem 3.2). Suppose that the variables ε_i and X_i are positive recurrent Markov chains with innovations with continuous conditional density function. Under additional mild regularity conditions, Condition 3 is satisfied with geometric mixing rates (Doukhan, 1995, section 2.4.0.1). Examples include GARCH and ARMA processes, as discussed in the aforementioned references.*

Example 6 *(Classification) Suppose that $Y_i \in \{-1, 1\}$. A classification model based on the regressors X_i can be generated via the random utility model*

$$Y_i^* = \mu(X_i) + \varepsilon_i$$

where $Y_i = sign(Y_i^*)$. The sigma algebra generated by $\{Y_i : i \in \mathcal{A}\}$ for any subset $\mathcal{A}$ of the integers is contained in the sigma algebra generated by $\{Y_i^* : i \in \mathcal{A}\}$. Hence, for errors ε_i and covariates X_i as in Example 5, the data are beta mixing with geometric mixing rate.

Example 7 *(Functional Data) In Example 5 let X_i and ε_i be continuous random functions from $[0,1]$ to $\mathbb{R}$. Suppose that there is a finite positive integer R such that $X_i = \sum_{r=1}^R \theta_r U_i^{(r)} g_r$, where for $r = 1, 2, ..., R$, $g_r : [0,1] \to \mathbb{R}$ is continuous, θ_r is a scalar, and $U^{(r)} := \left(U_i^{(r)}\right)_{i\in\mathbb{Z}}$ is independent across r. Also suppose that $U^{(r)}$ is strictly stationary and beta mixing $r = 1, 2, ..., R$. Then, the beta mixing* *coefficients of $(X_i)_{i\in\mathbb{Z}}$ are bounded by the sum of the mixing coefficients of $\left(U_i^{(r)}\right)_{i\in\mathbb{Z}}$, $r = 1, 2, ..., R$. With the same conditions imposed on $(\varepsilon_i)_{i\in\mathbb{Z}}$ independent of $(X_i)_{i\in\mathbb{Z}}$, we can cover the problem of regression using (finite dimensional) functional data.*

We conclude summarising the implications of the above remarks. When estimation is in a Sobolev space of functions on $[0, 1]$, Condition 1 does not impose any restriction as we can take $V = \eta$ in Example 4 and Theorem 1 applies. However, in Theorems 2, 3 and 4, we further require that $\eta > 1$. Given that V is an integer, we need to restrict the scope to functions with two square integrable derivatives (Sobolev spaces of index 2) if we want to use our results on consistency under the RKHS norm or derive convergence rates under the L_2 norm. These remarks are independent of Conditions 2 and 3. The two latter conditions have an effect on the convergence rates. As discussed, common models of interest satisfy Condition 3 with geometric rates. In such cases, we can take r in Condition 3 arbitrarily close to 2. Using this in Theorem 3, it becomes clear that higher convergence rates are achieved if the loss function is smooth and has moments of high order, i.e. $p \to \infty$ in Condition 2. In the context of the square error loss of Example 2, we would require the target variable Y_i to have a moment generating function in order to satisfy Condition 2 with p arbitrarily large.

3.2 Alternative Constraints

As an alternative to the norm $|\cdot|_{\mathcal{H}^K}$, define the norm $|f|_{\mathcal{L}^K} := \sum_{k=1}^K \left|f^{(k)}\right|_{\mathcal{H}}$. Estimation in $\mathcal{L}^K(B) := \left\{f \in \mathcal{H}^K : |f|_{\mathcal{L}^K} \le B\right\}$ is also of interest for variable screening. The following lists some details about the two different constraints.

Lemma 1 *Suppose an additive kernel $C_{\mathcal{H}^K}$ as in Section 2.2 and $K > 1$. The following hold.*
1. $|\cdot|_{\mathcal{H}^K}$ and $|\cdot|_{\mathcal{L}^K}$ are norms on $\mathcal{H}^K$.
2. We have the inclusion

$$K^{-1/2}\mathcal{H}^K(1) \subset \mathcal{L}^K(1) \subset \mathcal{H}^K(1).$$

3. For any $B > 0$, $\mathcal{H}^K(B)$ and $\mathcal{L}^K(B)$ are convex sets.
4. Let $c := \max_{s\in\mathcal{X}} \sqrt{C(s,s)}$. If $\mu \in \mathcal{H}^K(B)$, then, $\sup_{\mu\in\mathcal{H}^K(B)} |\mu|_p \le c\sqrt{K}B$ for any $p \in [1, \infty]$, while $\sup_{\mu\in\mathcal{L}^K(B)} |\mu|_p \le cB$.

By the inclusion in Lemma 1, all the results derived for $\mathcal{H}^K(B)$ also apply to $\mathcal{L}^K\left(K^{1/2}B\right)$. In this case, we still need to suppose that $\mu_0 \in \mathrm{int}\left(\mathcal{H}^K(B)\right)$. Both norms are of interest. When interest lies in variable screening, estimation in $\mathcal{L}^K(B)$ inherits the properties of the l_1 norm, as for Lasso. The estimation algorithms discussed in Section 4 cover estimation in both subsets of $\mathcal{H}^K$.

4 Computation Algorithm

As mentioned in Section 1.1, estimation in an RKHS poses computational difficulties when the sample size n is large. Simplifications are possible when the covariance $C_{\mathcal{H}^K}$ admits a series expansion as in (3).

Estimation for functions in $\mathcal{L}^K(B)$ rather than in $\mathcal{H}^K(B)$ is even more challenging because the norm $|\cdot|_{\mathcal{L}^K}$ is not everywhere differentiable. In the case of the square error loss, estimation in $\mathcal{L}^K(B)$ resembles Lasso, while estimation in $\mathcal{H}^K(B)$ resembles ridge regression.

A greedy algorithm can be used to solve both problems. In virtue of Lemma 1 and the fact that estimation in $\mathcal{H}^K(B)$ has been considered extensively, only estimation in $\mathcal{L}^K(B)$ will be address in details. The minor changes required for estimation in $\mathcal{H}^K(B)$ will be discussed in Section 4.2.

4.1 Estimation in $\mathcal{L}^K(B)$

Estimation of μ_n in $\mathcal{L}^K(B)$ is carried out according to the following Frank-Wolfe algorithm. Let $f_m^{(s(m))}$ be the solution to

$$\min_{k\leq K}\min_{f^{(k)}\in\mathcal{H}(1)} P_n\partial\ell_{F_{m-1}}f^{(k)}. \tag{12}$$

Here, $F_0=0$, $F_m=(1-\tau_m)F_{m-1}+c_mf_m^{(s(m))}$, and $c_m=B\tau_m$, where τ_m is the solution to the line search

$$\min_{\tau\in[0,1]} P_n\ell\left((1-\tau)F_{m-1}+\tau Bf_m^{(s(m))}\right), \tag{13}$$

writing $\ell(\mu)$ instead of ℓ_μ for typographical reasons. Details on how to solve (12) will be given in Section 4.1.1; the line search in (12) is elementary. The algorithm produces a set of functions $\left\{f_j^{(s(j))}:j=1,2,...,m\right\}$ and coefficients $\{c_j:j=1,2,...,m\}$. Note that $s(j)\in\{1,2,...,K\}$ identifies which of the K additive functions will be selected at the j^{th} iteration.

To map the results of the algorithm into functions with representation in $\mathcal{H}^K$, one uses simple algebraic manipulations. A simpler variant of the algorithm sets $\tau_m=1/m$. In this case, the solution at the m^{th} iteration, takes the particularly simple form $F_m=\sum_{j=1}^m\frac{B}{m}f_j^{(s(j))}$ (Sancetta, 2016) and the k^{th} additive function can be written as $\tilde{f}^{(k)}=\frac{B}{m}\sum_{j\leq m:s(j)=k}f_j^{(s(j))}$.

To avoid cumbersome notation, the dependence on the sample size n has been suppressed in the quantities defined in the algorithm. The algorithm can find a solution with arbitrary precision as the number of iterations m increases.

Theorem 5 *For F_m derived from the above algorithm,*

$$P_n\ell_{F_m}\leq\inf_{\mu\in\mathcal{L}^K(B)}P_n\ell_\mu+\epsilon_m$$

where,

$$\epsilon_m\lesssim\begin{cases}\frac{B^2\sup_{|t|\leq B}P_nd^2L(\cdot,t)/dt^2}{m} & \textit{if } \tau_m=\frac{2}{m+2} \textit{ or line search in (13)}\\ \frac{B^2\sup_{|t|\leq B}\left[P_nd^2L(\cdot,t)/dt^2\right]\ln(1+m)}{m} & \textit{if } \tau_m=\frac{1}{m}\end{cases}.$$

For the sake of clarity, recall that $P_nd^2L(\cdot,t)/dt^2=\frac{1}{n}\sum_{i=1}^nd^2L(Z_i,t)/dt^2$.

4.1.1 Solving for the Additive Functions

The solution to (12) is found by minimizing the Lagrangian

$$P_n\partial\ell_{F_{m-1}}f^{(k)}+\rho\left|f^{(k)}\right|_{\mathcal{H}}^2. \tag{14}$$

Let $\Phi^{(k)}\left(x^{(k)}\right)=C\left(\cdot,x^{(k)}\right)$ be the canonical feature map (Lemma 4.19 in Steinwart and Christmann, 2008); $\Phi^{(k)}$ has image in $\mathcal{H}$ and the superscript k is only used to stress that it corresponds to the k^{th} additive component. The first derivative w.r.t. $f^{(k)}$ is $P_n\partial\ell_{F_{m-1}}\Phi^{(k)}+2\rho f^{(k)}$, using the fact that $f^{(k)}\left(x^{(k)}\right)=\left\langle f^{(k)},\Phi^{(k)}\left(x^{(k)}\right)\right\rangle_{\mathcal{H}}$, by the reproducing kernel property. Then, the solution is

$$f^{(k)}=-\frac{1}{2\rho}P_n\partial\ell_{F_{m-1}}\Phi^{(k)},$$

where ρ is such that $\left|f^{(k)}\right|_{\mathcal{H}}^2=1$. If $P_n\partial\ell_{F_{m-1}}\Phi^{(k)}=0$, set $\rho=1$. Explicitly, using the properties of RKHS (see (5))

$$\left|f^{(k)}\right|_{\mathcal{H}}^2=\frac{1}{(2\rho)^2}\sum_{i,j=1}^{n}\frac{\partial\ell_{F_{m-1}}(Z_i)}{n}\frac{\partial\ell_{F_{m-1}}(Z_j)}{n}C\left(X_i^{(k)},X_j^{(k)}\right)$$

which is trivially solved for ρ. With this choice of ρ, the constraint $\left|f^{(k)}\right|_{\mathcal{H}}\leq 1$ is satisfied for all integer $k\leq K$, and the algorithm, simply selects k such that $P_n\partial\ell_{F_{m-1}}f^{(k)}$ is minimized.

The above calculations together with Theorem 5 imply the following, which for simplicity, is stated using the update $\tau_m=m^{-1}$ instead of the line search.

Theorem 6 *Let ρ_j be the Lagrange multiplier estimated at the j^{th} iteration of the algorithm in (12) with $\tau_m=m^{-1}$ instead of the line search (13). Then,*

$$\mu_n=\lim_{m\rightarrow\infty}\sum_{j=1}^{m}\left(-\frac{B}{2m\rho_j}\right)P_n\partial\ell_{F_{j-1}}\Phi^{(s(j))},$$

is the solution of the constrained estimation problem in $\mathcal{L}^K(B)$.

4.2 The Algorithm for Estimation in $\mathcal{H}^K(B)$

When estimation is constrained in $\mathcal{H}^K(B)$, the algorithm has to be modified. Let $\Phi(x)=C_{\mathcal{H}^K}(\cdot,x)$ be the canonical feature map of $\mathcal{H}^K$ (do not confuse Φ with $\Phi^{(k)}$ in the previous section). Then, (12) is replaced by

$$\min_{f\in\mathcal{H}^K(1)}P_n\partial\ell_{F_{m-1}}f,$$

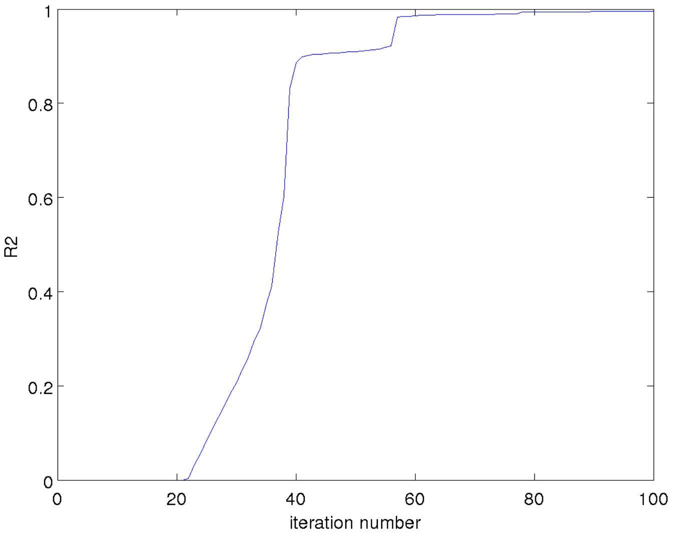
and we denote by $f_m\in\mathcal{H}^K(B)$ the solution at the m^{th} iteration. This solution can be found replacing the minimization of (14) with minimization of $P_n\partial\ell_{F_{m-1}}f+\rho|f|_{\mathcal{H}^K}^2$. The solution is then $f_m=-\frac{1}{2\rho}P_n\partial\ell_{F_{m-1}}\Phi$ where ρ is chosen to satisfy the constraint $|f|_{\mathcal{H}^K}^2\leq 1$ (Steinwart and Christmann, 2008, Corollary 5.11). No other change in the algorithm is necessary and the details are left to the reader.

4.3 Numerical Illustration

To gauge the rate at which the algorithm converges to a solution, we consider a numerical illustration using the SARCOS data set (`http://www.gaussianprocess.org/gpml/data/`), which comprises a sample of 44484 observations with 21 input variables and a continuous response variable. We standardize the variables by their Euclidean norm, use the square error loss and the Gaussian covariance kernel of Example 1 with $d = 21$ and $a^{-1} = 0.75$. Hence for this example, the kernel is not additive. Given that the kernel is universal, we shall be able to interpolate the data if B is chosen large enough: we choose $B = 1000$. The aim is not to find a good statistical estimator, but to evaluate the computational algorithm. Figure 1, plots the R^2 as a function of the number of iterations m. After approximately 20 iterations, the algorithm starts to fit the data better than a constant, and after about 80-90 iterations the R^2 is very close to one: $R^2 = 99.55\%$. However, the number of operations per iteration is $O\left(n^2\right)$ (Rasmussen and Williams, 2006, Ch.8, for a comparison of methods).

We also use the test sample from the same dataset to evaluate the out of sample R^2 as we vary B. The sample size of the test sample is 4489 observations. To distinguish it from this, we call estimation sample the one with 44484 observations used to estimate the function. The estimation algorithm is used with $m = 100$. Given the estimated function, we compute the standardized mean square error (SMSE) on the test sample. This is just the mean square error divided by the variance of the target variable in the test sample. The out of sample R^2 is one minus the SMSE. We also use cross-validation (CV) to find an estimate of the generalization error. This is useful to choose B. In particular CV is computed randomly sampling without replacement 67% of the estimation sample and using the remaining 33% to evaluate the error. To reduce dependence on the random split, this procedure is repeated 5 times to find an estimate of the out of sample mean square error. We can then compute the cross-validated SMSE and R^2 from this quantity. With such large sample size, we find that we can choose B relatively large and CV would have led to relatively good performance (Table 1).

Figure 1: Estimation Algorithm R^2 as Function of Number of Iterations. The R^2 is computed for each iteration m of the estimation algorithm. Negative R^2 have been set to zero.



4.4 Selection of B and Variable Screening

The parameter B uniquely identifies the Lagrange multiplier $\rho_{B,n}$ in the penalized version of the optimization problem (6) (see Example 2). If the loss is non-negative, we have that $|\mu_n|^2_{\mathcal{H}^K} \leq \rho^{-1}_{B,n} P_n \ell_{\mu_B}$

Table 1: Test Sample Results. Out of sample and cross-validated R^2 (R^2_{oos} and R^2_{CV}) are reported as a function of B. The values of B are chosen to be multiples of the variance of the target variable on the estimation sample.

B	108.04	216.07	432.15	864.29	1728.58	3457.17	6914.33
R^2_{oos}	-0.0811	0.6805	0.9884	0.9866	0.9964	0.8043	0.7670
R^2_{CV}	-0.2343	0.4323	0.7430	0.7735	0.7635	0.5869	0.5299

(Steinwart and Christmann, 2008, Section 5.1). The exact same argument holds for $\mathcal{L}^K(B)$ in place of $\mathcal{H}^K(B)$. When the constraint $\mu \in \mathcal{L}^K(B)$ is considered, the solution via the greedy algorithm allows us to keep track of the iterations at which selected variables are included. Variables included at the early stage of the algorithm will be clearly included even when B is increased. Hence, exploration for the purpose of feature selection (using the constraint $\mu \in \mathcal{L}^K(B)$) can be carried out using a large B to reduce the computational burden.

Selection of B is usually based on cross-validation or penalized estimation, where the penalty estimates the "degrees of freedom".

5 Proofs

Recall that $\ell_\mu(Z) = L(Z, \mu(X))$ and $\partial^k \ell_\mu(Z) = \partial^k L(Z,t)/\partial t^k\big|_{t=\mu(X)}$, $k \geq 1$. Condition 2 implies Fréchet differentiability of $P\ell_\mu$ and $P\partial\ell_\mu$ (as functions of μ, from L_∞ to $\mathbb{R}$) at $\mu \in \mathcal{H}^K$ in the direction of $h \in \mathcal{H}^K$. The derivative can be weakened to μ and h elements in L_∞, the space of uniformly bounded function. It can be shown that these two derivatives are $P\partial\ell_\mu h$ and $P\partial^2\ell_\mu hh$, respectively. For this purpose, we view $P\ell_\mu$ as a map from the set of uniformly bounded functions on $\mathcal{X}^K$ - to be denoted by $\ell^\infty(\mathcal{X}^K)$ - to $\mathbb{R}$. The details can be derived following the steps in the proof of Lemma 2.21 in Steinwart and Christmann (2008) or the proof of Lemma A.4 in Hable (2012). The application of those proofs to the current scenario, essentially requires that the loss function $L(Z,t)$ is differentiable w.r.t. real t, and that μ is uniformly bounded, together with integrability of Δ_k, $k = 0, 1, 2$, as in Condition 2. It will also be necessary to take the Fréchet derivative of $P_n\ell_\mu$ and $P_n\partial\ell_\mu h$ conditioning on the sample data. By Condition 2 this will also hold because Δ_0, and Δ_1 are finite. This will also allow us to apply Taylor's Theorem in Banach spaces. These derivatives will be used in the proofs. Moreover, for notational simplicity, we shall tacitly suppose that $\sup_{x\in\mathcal{X}^K}\sqrt{C_{\mathcal{H}^K}(x,x)} = 1$ so that $h \in \mathcal{H}^K(B)$ implies that $|h|_\infty \leq B$ for any $B > 0$.

5.1 Complexity and Gaussian Approximation

The reader can skip this section and refer to it when needed. Recall that the ϵ-covering number of a set $\mathcal{F}$ under the L_p norm, denoted by $N\left(\epsilon, \mathcal{F}, |\cdot|_p\right)$, is the minimum number of balls of L_p radius ϵ needed to cover $\mathcal{F}$. The entropy is the logarithm of the covering number. The ϵ-bracketing number of the set $\mathcal{F}$ under the L_p norm is the minimum number of ϵ-brackets under the L_p norm needed to cover $\mathcal{F}$. Given two functions $f_L \leq f_U$ such that $|f_L - f_U|_p \leq \epsilon$, an L_p ϵ-bracket $[f_L, f_U]$ is the set of all functions $f \in \mathcal{F}$ such that $f_L \leq f \leq f_U$. Denote the L_p ϵ-bracketing number of $\mathcal{F}$ by $N_{[]}\left(\epsilon, \mathcal{F}, |\cdot|_p\right)$. Under the uniform norm, $N(\epsilon, \mathcal{F}, |\cdot|_\infty) = N_{[]}(\epsilon, \mathcal{F}, |\cdot|_\infty)$ and this will be tacitly used in what follows.

In this section, let $(G(x))_{x\in\mathcal{X}}$ be a centered Gaussian process on $\mathcal{X}$ with covariance C as in (3). In what follows we refer to Li and Linde (1999) for details. The space $\mathcal{H}$ is generated by the measure of the Gaussian process $(G(x))_{x\in\mathcal{X}}$ with covariance function C. In particular, $G(x)=\sum_{v=1}^{\infty}\lambda_v\xi_v\varphi_v(x)$, where the $(\xi_v)_{v\geq1}$ is a sequence of i.i.d. standard normal random variables, and the equality holds in distribution. For any positive integer V, the l-approximation number $l_V(G,|\cdot|_\infty)$ (Li and Linde, 1999, p. 1560) is bounded above by $\left(\mathbb{E}\left|\sum_{v>V}\lambda_v\xi_v\varphi_v\right|_\infty^2\right)^{1/2}$. Under Condition 1, we deduce that

$$l_V(G,|\cdot|_\infty)\lesssim\sum_{v>V}\lambda_v\lesssim V^{-(\eta-1)}. \tag{15}$$

There is a link between the $l_V(G,|\cdot|_\infty)$ approximation number of the centered Gaussian process G with covariance C and the L_∞ ϵ-entropy of the class of functions $\mathcal{H}(1)$. These quantities are related by $-\ln\Pr(|G|_\infty<\epsilon)$, which is determined by the small ball probability of G under the uniform norm.

Entropy bounds. We have the following bound on the ϵ-entropy of $\mathcal{H}(1)$ under the uniform norm $|\cdot|_\infty$.

Lemma 2 *Under Condition 1,* $\ln N(\epsilon,\mathcal{H}(1),|\cdot|_\infty)\lesssim\epsilon^{-2/(2\eta-1)}$.

Proof. As previously remarked, the space $\mathcal{H}(1)$ is generated by the law of the Gaussian process G with covariance function C. For any integer $V<\infty$, the l-approximation number of G, $l_V(G,|\cdot|_\infty)$ is bounded as in (15). Proposition 4.1 in Li and Linde (1999) says that in this case $-\ln\Pr(|G|_\infty<\epsilon)\lesssim\epsilon^{-1/(\eta-1)}$. Moreover, Theorem 1.2 in Li and Linde (1999) links the small ball probability to the entropy of the RKHS, and in this case, it gives the estimate $\ln N(\epsilon,\mathcal{H}(1),|\cdot|_\infty)\lesssim\epsilon^{-2/(2\eta-1)}$. ■

Lemma 3 *Under Condition 1,*

$$\ln N\left(\epsilon,\mathcal{H}^K(B),|\cdot|_\infty\right)\lesssim K(B/\epsilon)^{2/(2\eta-1)}.$$

Proof. If $\mu\in\mathcal{H}^K(B)$, then $\mu(x)=\sum_{k=1}^K f^{(k)}\left(x^{(k)}\right)$ for some $f^{(k)}\in\mathcal{H}(B)$. Hence, the covering numbers of $\left\{\mu\in\mathcal{H}^K(B)\right\}$ are bounded by the product of the covering numbers of the sets $\mathcal{F}_k:=\left\{f^{(k)}\in\mathcal{H}(B)\right\}$, $k=1,2,...,K$. By Lemma 2, the ϵ-covering number of each $\mathcal{F}_k$ is given by $\exp\left\{(B/\epsilon)^{2/(2\eta-1)}\right\}$. The statement of the lemma follows by taking logs and summing over $k=1,2,...,K$. ■

We link the entropy of $\mathcal{H}(1)$ under the uniform norm to the entropy with bracketing of $\ell_\mu h$.

Lemma 4 *Suppose Condition 1 holds. For the set* $\mathcal{F}:=\left\{\partial\ell_\mu h:\mu\in\mathcal{H}^K(B),h\in\mathcal{H}^K(1)\right\}$, *for any* $p\in[1,\infty]$ *satisfying Condition 2, the* L_p ϵ*-entropy with bracketing is*

$$\ln N_{[]}\left(\epsilon,\mathcal{F},|\cdot|_p\right)\lesssim K(B/\epsilon)^{2/(2\eta-1)}.$$

The same exact result holds for $\mathcal{F}:=\left\{\ell_\mu:\mu\in\mathcal{H}^K(B)\right\}$ *under Condition 2.*

Proof. In the interest of conciseness, we only prove the result for

$$\mathcal{F}:=\left\{\partial\ell_\mu h:\mu\in\mathcal{H}^K(B),h\in\mathcal{H}^K(1)\right\}.$$

To this end, note that by Condition 2 and the triangle inequality,

$$\left|\partial\ell_{\mu}h-\partial\ell_{\mu'}h'\right|\leq\left|\partial\ell_{\mu}-\partial\ell_{\mu'}\right|\sup_{h\in\mathcal{H}^{K}(1)}|h|+\sup_{\mu\in\mathcal{H}^{K}(B)}\left|\partial\ell_{\mu}\right|\left|h-h'\right|.$$

By Condition 2, $\left|\partial\ell_{\mu}(z)\right|\leq\Delta_{1}(z)$, and $\left|\partial\ell_{\mu}(z)-\partial\ell_{\mu'}(z)\right|\leq\Delta_{2}(z)\left|\mu(x)-\mu'(x)\right|$, and $P\left(\Delta_{1}^{p}+\Delta_{2}^{p}\right)<\infty$. By Lemma 1, $|h|_{\infty}\leq1$. By these remarks, the previous display is bounded by by a constant multiple of

$$\Delta_{2}\left|\mu-\mu'\right|_{\infty}+\Delta_{1}\left|h-h'\right|_{\infty}.$$

Theorem 2.7.11 in van der Vaart and Wellner (2000) says that the L_{p} ϵ-bracketing number of class of functions satisfying the above Lipschitz kind of condition is bounded by the L_{∞} ϵ'-covering number of $\mathcal{H}^{K}(B)\times\mathcal{H}^{K}(1)$ with $\epsilon'=\epsilon/\left[2\left(P\left|\Delta_{1}+\Delta_{2}\right|^{p}\right)^{1/p}\right]$. Using Lemma 3, the statement of the lemma is deduced because the product of the covering numbers is the sum of the entropies. ■

Convergence to a Gaussian process and maximal inequality. The following will be used in the proof of Theorem 4.

Lemma 5 *Under Condition 1 with $\eta>1$, and Conditions 2, and 3, with $p\geq r$,*

$$\sqrt{n}\left(P_{n}-P\right)\partial\ell_{\mu}h\to G\left(\partial\ell_{\mu},h\right)$$

weakly, where $G\left(\partial\ell_{\mu},h\right)$ is a mean zero Gaussian process indexed by $\left(\partial\ell_{\mu},h\right)\in\left\{\partial\ell_{\mu}:\mu\in\mathcal{H}^{K}(B)\right\}\times\mathcal{H}^{K}(1)$, with a.s. continuous sample paths and covariance function

$$\mathbb{E}G\left(\partial\ell_{\mu},h\right)G\left(\partial\ell_{\mu'},h'\right)=\sum_{j\in\mathbb{Z}}P_{1,j}\left(\partial\ell_{\mu}h,\partial\ell_{\mu}h'\right).$$

Proof. The proof shall use the main result in Doukhan et al. (1995). Let $\mathcal{F}:=\left\{\partial\ell_{\mu}h:\mu\in\mathcal{H}^{K}(B),h\in\mathcal{H}^{K}(1)\right\}$. The elements in $\mathcal{F}$ have finite L_{r} norm because $P\left|\partial\ell_{\mu}\right|^{p}\leq P\Delta_{1}^{p}$ by Condition 2, and $|h|_{\infty}\leq1$ by Lemma 1. The entropy integral in Doukhan et al. (1995, Theorem 1, eq. 2.10) is implied by

$$\int_{0}^{1}\sqrt{\ln N_{[]}\left(\epsilon,\mathcal{F},|\cdot|_{r}\right)}d\epsilon<\infty,\tag{16}$$

and $\beta(i)\lesssim i^{-\beta_{0}}$ with $\beta_{0}>r/(r-2)$ and $r>2$ as in Condition 3 (see their discussion on page 405 to relate the L_{r} norm to their norm). When (16) holds, Theorem 1 in Doukhan et al. (1995) shows that the empirical process indexed in $\mathcal{F}$ converges weakly to the Gaussian one given in the statement of the present lemma. By Lemma 4, (16) holds because $\eta>1$. ■

The following is a corollary to Theorem 2 in Doukhan et al. (1995) and is used in the proof of Theorems 2 and 3.

Lemma 6 *Suppose Condition 3 and let r be as defined there. Let $\mathcal{F}$ is a class of real valued measurable functions such that $|f|_{r}\leq\delta$ for any $f\in\mathcal{F}$ and, $\int_{0}^{\delta}\sqrt{\ln N_{[]}\left(\epsilon,\mathcal{F},|\cdot|_{r}\right)}d\epsilon\lesssim\delta^{\alpha}$, for some $\alpha\in(0,1]$. Let*

F be a function such that $|f| \leq F$, for any $f \in \mathcal{F}$, and $PF^p < \infty$ for some $p \geq r$. Then, we have that

$$\sqrt{n}\mathbb{E}\sup_{f\in\mathcal{F}}|(P_n - P)f| \lesssim \delta^{\alpha} + \sqrt{n}\left(\delta^{2\alpha-2}n^{-1}\right)^{\frac{r}{2(r-1)}\frac{p-1}{p}}$$

for any δ such that $\delta^{-1} \lesssim n^{1/(2(1-\alpha))}$.

Proof. We shall follow similar arguments to the ones in the proof of Theorem 3 in Doukhan et al. (1995). Let $\mathcal{F}_M := \left\{f1_{\{F\leq M\}} : f \in \mathcal{F}\right\}$ for a constant $M > 0$, where $1_{\{\cdot\}}$ is the indicator function: one if the argument is true and zero otherwise. Clearly,

$$\sqrt{n}\mathbb{E}\sup_{f\in\mathcal{F}}|(P_n - P)f| \leq \sqrt{n}\mathbb{E}\sup_{f\in\mathcal{F}_M}|(P_n - P)f| + 2\sqrt{n}PF1_{\{F>M\}} =: \mathrm{I} + \mathrm{II}. \tag{17}$$

Under the conditions of the lemma, Theorem 2 in Doukhan et al. (1995) says that

$$\mathrm{I} \lesssim \delta^{\alpha} + \frac{Mq\delta^{2\alpha-2}}{n^{1/2}} + \sqrt{n}M\beta(q) \tag{18}$$

for arbitrary $q \geq 1$. Set an $\epsilon \in (0,1)$ to be chosen in due course, and set $q = \beta^{-1}(\epsilon)$ where $\beta^{-1}(\cdot)$ is the inverse of $\beta(\lfloor\cdot\rfloor)$ where $\lfloor\cdot\rfloor$ is the integer part of its argument. Then, $\beta^{-1}(\epsilon)$ is the smallest integer q such that $\beta(q) \leq \epsilon$. To balance the last two terms in (18) we can set $\sqrt{n}\epsilon \asymp q\delta^{2\alpha-2}n^{-1/2}$, where $\asymp$ means that the l.h.s. is bounded above and below by constants times the r.h.s.. Let $Q : [0,1] \to \mathbb{R}$ be the quantile function of F, i.e. $Q(u) := \inf\{x > 0 : \Pr(F > x) \leq u\}$. Then, set $M = Q(\epsilon)$. Therefore, $\mathrm{I} \lesssim \delta^{\alpha} + \sqrt{n}Q(\epsilon)\epsilon$. Now, from the proof of Lemma 4 in Doukhan et al. (1995) we have that $PF1_{\{F>M\}} \leq \int_0^{\epsilon} Q(\epsilon)\,d\epsilon$ and given that F has a finite p moment we have that $Q(\epsilon) \lesssim \epsilon^{-1/p}$. By these remarks, we can deduce that $\mathrm{II} \lesssim 2\sqrt{n}\epsilon^{\frac{p-1}{p}}$. Using the upper bound for $Q(\epsilon)$ in I, we also deduce that $\mathrm{I} + \mathrm{II} \lesssim \delta^{\alpha} + \sqrt{n}\epsilon^{\frac{p-1}{p}}$. By Condition 3, $q \lesssim \epsilon^{-\frac{r-2}{r}}$, so that $\epsilon \asymp \left(\delta^{2\alpha-2}n^{-1}\right)^{\frac{r}{2(r-1)}}$ and in consequence that (17) is bounded above by $\delta^{\alpha} + \sqrt{n}\left(\delta^{2\alpha-2}n^{-1}\right)^{\frac{r}{2(r-1)}\frac{p-1}{p}}$. Note that the specific choice of ϵ also guarantees that $\epsilon \in (0,1)$ as long as $\delta^{-1} \lesssim n^{1/(2(1-\alpha))}$ as stated in the lemma, and this concludes its proof. ∎

When the data is i.i.d. (take $r \downarrow 2$ in Condition 3) and F bounded (take $p \to \infty$) the above inequality becomes the usual inequality obtained for i.i.d. data (van der Vaart and Wellner, 2000, Lemma 3.4.4).

5.2 Proof of Theorem 1

We shall apply Corollary 3.2.3 in van der Vaart and Wellner (2000) replacing their in probability result with almost sure convergence (a.s.). The result requires an identification condition and uniform convergence of the empirical loss function. By Taylor's Theorem in Banach spaces,

$$P\ell_{\mu} \quad - P\ell_{\mu_B} = \quad P\partial\ell_{\mu_B}(\mu - \mu_B) + \frac{1}{2}P\partial^2\ell_{\mu_t}(\mu - \mu_B)^2 \tag{19}$$

for $\mu_t = \mu + t(\mu_B - \mu)$ with some $t \in [0,1]$ and arbitrary $\mu \in \mathcal{H}^K(B)$. The variational inequality $P\partial\ell_{\mu_B}(\mu - \mu_B) \geq 0$ holds by definition of μ_B and the fact that $\mu \in \mathcal{H}^K(B)$. Therefore, the previous display implies that $P\ell_{\mu} - P\ell_{\mu_B} \gtrsim P(\mu - \mu_B)^2$ because $P\partial^2\ell_{\mu_t}(\mu - \nu)^2 \gtrsim P(\mu - \nu)^2 \geq 0$ by Condition

2. The right hand most inequality holds with equality if and only if $\mu = \mu_B$ in L_2. This shows identifiability of the estimator.

We show that $\sup_{\mu \in \mathcal{H}^K(B)} |(P_n - P)\,\ell_\mu| \rightarrow 0$ a.s., which then implies $|\mu_n - \mu_0|_2 \rightarrow 0$ a.s.. For any fixed μ, $|(P_n - P)\,\ell_\mu| \rightarrow 0$ a.s., by the ergodic theorem, because $P\,|\ell_\mu| < \infty$ by Condition 2. Hence, it is sufficient to show that $\left\{\ell_\mu : \mu \in \mathcal{H}^K(B)\right\}$ has finite ϵ-bracketing number under the L_1 norm (see the proof of Theorem 2.4.1 in van der Vaart and Wellner, 2000). This is the case by Lemma 4. We have shown that $|\mu_n - \mu_B|_2 \rightarrow 0$ a.s.. To turn the L_2 convergence into uniform, note that $\mathcal{H}^K(B)$ is compact under the uniform norm and functions in $\mathcal{H}^K(B)$ are continuous and defined on a compact domain $\mathcal{X}^K$. In consequence, any convergent sequence in $\mathcal{H}^K(B)$ converges uniformly.

Above, we have shown that the population loss function is convex and coercive. Moreover, $\mathcal{H}^K(B)$ is a closed convex set, as $\mathcal{H}^K$ is a Hilbert space. Hence, the population minimizer μ_B exists and is unique up to an L_2 equivalence class.

5.3 Proof of Theorem 2

We prove Points 1 to 3. The validity of the results when using asymptotic minimizers is in Section 5.7.

The following lemma puts together crucial results for estimation in RKHS (Steinwart and Christmann, 2008, Theorems 5.9 and 5.17 for a proof). The cited results make use of the definition of integrable Nemitski loss of finite order p (Steinwart and Christmann, 2008, Def. 2.16). However, under Condition 2, the proofs of those results still hold.

Lemma 7 *Under Condition 2,*

$$|\mu_{0,\rho} - \mu_{n,\rho}|_{\mathcal{H}^K} \leq \frac{1}{\rho}\left|P\partial\ell_{\mu_{0,\rho}}\Phi - P_n\partial\ell_{\mu_{0,\rho}}\Phi\right|_{\mathcal{H}^K}, \tag{20}$$

where $\Phi(x) = C_{\mathcal{H}^K}(\cdot, x)$ is the canonical feature map. Moreover, if $\mu_{0,\rho}$ is bounded for $\rho \rightarrow 0$, then $|\mu_{0,\rho} - \mu_0|_{\mathcal{H}^K} \rightarrow 0$.

We apply Lemma 7 and the results in Section 5.1 to derive the following.

Lemma 8 *Suppose Condition 1 with $\eta > 1$, and Conditions 2 and 3 with $p \geq r$, and $\mu_0 \in \mathcal{H}^K$. The following statements hold.*

1. *There is a finite B such that $|\mu_{0\rho}|_{\mathcal{H}^K} \leq |\mu_0|_{\mathcal{H}^K} < B$ for any $\rho \geq 0$.*
2. *If $\mu_0 \in \operatorname{int}\left(\mathcal{H}^K(B)\right)$, we have that $|\mu_{n,\rho} - \mu_{0,\rho}|^2_{\mathcal{H}^K} = O_P\left(\rho^{-2}n^{-1}\right)$, and $|\mu_{n,\rho}|_{\mathcal{H}^K} < B$ eventually in probability for any $\rho \rightarrow 0$ such that $\rho n^{1/2} \rightarrow \infty$.*
3. *There is a $\rho = O_P\left(n^{-1/2}\right)$ such that $|\mu_{n,\rho}|_{\mathcal{H}^K} \leq B$ and*

$$\sup_{h \in \mathcal{H}^K(1)} P_n\partial\ell_{\mu_{n,\rho}}h = O_P\left(n^{-1/2}B\right).$$

Proof. Given that K is finite and the kernel is additive, there is no loss in restricting attention to $K = 1$ in order to reduce the notational burden. Given that $\mu_0 \in \mathcal{H}^K$, there is a finite B such that $\mu_0 \in \operatorname{int}\left(\mathcal{H}^K(B)\right)$ (this proves Point 1 in the lemma). By this remark, it follows that, uniformly in

$\rho \geq 0$, there is an $\epsilon > 0$ such that $|\mu_{0,\rho}|_{\mathcal{H}^K} \leq B - \epsilon$. We shall need a bound for the r.h.s. of (20). By (3), the canonical feature map can be written as $\Phi(x) = \sum_{v=1}^{\infty} \lambda_v^2 \varphi_v(\cdot)\varphi_v(x)$. This implies that,

$$(P_n - P)\,\partial\ell_{\mu_{0,\rho}}\Phi(x) = \sum_{v=1}^{\infty}\left[\lambda_v^2 (P_n - P)\,\partial\ell_{\mu_{0,\rho}}\varphi_v\right]\varphi_v(x).$$

By Lemma 7, (5), and the above,

$$\left|(P_n - P)\,\partial\ell_{\mu_{0,\rho}}\Phi\right|_{\mathcal{H}^K}^2 = \sum_{v=1}^{\infty}\frac{\left[\lambda_v^2 (P_n - P)\,\partial\ell_{\mu_{0,\rho}}\varphi_v\right]^2}{\lambda_v^2} = \sum_{v=1}^{\infty}\lambda_v^2\left[(P_n - P)\,\partial\ell_{\mu_{0,\rho}}\varphi_v\right]^2.$$

In consequence of the above display, by the triangle inequality,

$$\begin{aligned}|\mu_{0,\rho} - \mu_{n,\rho}|_{\mathcal{H}^K} &\leq \frac{1}{\rho}\left[\sum_{v=1}^{\infty}\lambda_v^2\left|(P_n - P)\,\partial\ell_{\mu_{0,\rho}}\varphi_v\right|^2\right]^{1/2}\\ &\leq \frac{1}{\rho}\sum_{v=1}^{\infty}\lambda_v\left|(P_n - P)\,\partial\ell_{\mu_{0,\rho}}\varphi_v\right|.\end{aligned}$$

Using the maximal inequality in the first display on page 410 of Doukhan et al. (1995) we deduce that

$$\mathbb{E}\sup_{\mu\in\mathcal{H}^K(B)}\left|\sqrt{n}\,(P_n - P)\,\partial\ell_{\mu}\varphi_v\right| \leq c_1 \tag{21}$$

for some finite constant c_1, for any $v \geq 1$, because the entropy integral (16) is finite in virtue of Lemma 4 and φ_v is uniformly bounded, uniformly in $v \geq 1$. Define

$$L_n := \sum_{v=1}^{\infty}\lambda_v \sup_{\mu\in\mathcal{H}^K(B)}\left|\sqrt{n}\,(P_n - P)\,\partial\ell_{\mu_{0,\rho}}\varphi_v\right|.$$

Given that the coefficients λ_v are summable by Condition 1 when $\eta > 1$, deduce from (21) that $(L_n)_{n\geq 1}$ is a tight random sequence. Using the above display, we have shown that (20) is bounded by $L_n/\left(\rho n^{1/2}\right)$. This proves Point 2 in the lemma. For any fixed $\epsilon > 0$, we can choose $\rho = \rho_n := L_n/\left(\epsilon n^{1/2}\right)$ so that $|\mu_{0,\rho} - \mu_{n,\rho}|_{\mathcal{H}^K} \leq \epsilon$. By the triangle inequality and the above calculations, deduce that,

$$|\mu_{n,\rho}|_{\mathcal{H}^K} \leq |\mu_{0,\rho}|_{\mathcal{H}^K} + |\mu_{0,\rho} - \mu_{n,\rho}|_{\mathcal{H}^K} \leq B$$

by the aforementioned choice of ρ. By tightness of L_n, deduce that $\rho_n = O_p\left(n^{-1/2}\right)$. Also, the first order condition for the sample estimator $\mu_{n,\rho}$ reads

$$P_n\partial\ell_{\mu_{n,\rho}}h = -2\rho\left\langle\mu_{n,\rho}, h\right\rangle_{\mathcal{H}^K} \leq 2\rho\,|\mu_{n,\rho}|_{\mathcal{H}^K}\,|h|_{\mathcal{H}^K} \tag{22}$$

for any $h \in \mathcal{H}^K(1)$. In consequence, $\sup_{h\in\mathcal{H}^K(1)} P_n\partial\ell_{\mu_{n,\rho}}h \leq 2\rho\,|\mu_{n,\rho}|_{\mathcal{H}^K}$. These calculations prove Point 3 in the lemma for some $\rho = O_P\left(n^{-1/2}\right)$. ■

We start the proof of the theorem.

Proof of Point 1. The penalized objective function is increasing with ρ. In the Lagrangian formulation of the constrained minimization, interest lies in finding the smallest value of ρ such that the constraint is still satisfied. When ρ equals such smallest value $\rho_{B,n}$, we have $\mu_n = \mu_{n,\rho}$. From Point 3 in Lemma 8 deduce that $\rho_{B,n} = O_P\left(n^{-1/2}\right)$.

Proof of Point 2. Point 1 in Lemma 8 together with the last statement of Lemma 7 gives that $\left|\mu_{n,\rho} - \mu_0\right|^2_{\mathcal{H}^K} = o(1)$. Then, Point 2 in Lemma 8 together with the triangle inequality complete the proof.

Proof of Point 3. If $\mathcal{H}^K$ is infinite dimensional, the constraint is eventually binding for n large enough, so that $|\mu_n|_{\mathcal{H}^K} = B$. Hence, if $\mu_0 \in \operatorname{int}\left(\mathcal{H}^K(B)\right)$ there is an $\epsilon > 0$ such that $|\mu_0|_{\mathcal{H}^K} = B - \epsilon$. By the triangle inequality, we deduce that $|\mu_n - \mu_0|^2_{\mathcal{H}^K} \geq \epsilon$. This means that μ_n cannot converge under the norm $|\cdot|_{\mathcal{H}^K}$.

The statement concerning approximate minimizers will be proved in Section 5.7.

5.4 Proof of Theorem 3

For reasons that will become clear, we show convergence rates for $|\mu - \mu_B|_{rp/(p-r)}$ where p and r are as in the statement of the theorem. To this end, we verify the conditions of Theorem 3.2.5 van der Vaart and Wellner (2000). Define $\mathcal{F}_\delta := \left\{(\ell_\mu - \ell_{\mu_B}) : |\mu - \mu_B|_{rp/(p-r)} \leq \delta, \mu \in \mathcal{H}^K(B)\right\}$. It is sufficient to show that (i) $P\ell_\mu - P\ell_{\mu_B} \gtrsim |\mu - \mu_B|_{rp/(p-r)}$, (ii) $\sqrt{n}\mathbb{E}\sup_{f\in\mathcal{F}_\delta}|(P_n - P)f| \leq \phi_n(\delta)$, for any $\delta \in (0,1)$, where $\phi_n(\delta)$ is a function that grows slower than δ^2, and (iii) to find an increasing sequence s_n such that $s_n^2\phi_n\left(s_n^{-1}\right) \lesssim \sqrt{n}$. Then, $|\mu - \mu_B|_{rp/(p-r)} = O_P\left(s_n^{-1}\right)$. Note that δ can be taken less than one because, by Theorem 1, the estimator is consistent in L_∞ as soon as $\eta > 1/2$. The uniform bound for elements in $\mathcal{H}^K(B)$ implies that $B^{(2/v)-1}|\mu - \mu_B|_v \leq |\mu - \mu_B|_2^{2/v}$ for any $v \in (2,\infty)$. Therefore, the arguments just below (19) verify (i). We now focus on (ii). By Holder inequality, $P\Delta_1^r|\mu - \mu_B|^r \leq \left(P\Delta_1^p\right)^{r/p}\left(P|\mu - \mu_B|^{rp/(p-r)}\right)^{(p-r)/p}$. Using the fact that $|\ell_\mu - \ell_{\mu_B}| \leq \Delta_1|\mu - \mu_B|$, we can conclude that $|f|_r \leq |\Delta_1|_p\,\delta$ because $f \in \mathcal{F}_\delta$. We also deduce that $\Delta_1 B$ has finite p moment, and is an envelope function for $\mathcal{F}_\delta$. By Lemma 4, we have that $\int_0^\delta \sqrt{\ln N_{[]}\left(\epsilon, \mathcal{F}_\delta, |\cdot|_r\right)}d\epsilon \lesssim \delta^\alpha$ with $\alpha = 2(\eta - 1)/(2\eta - 1)$. Hence, an application of Lemma 6 shows that $\phi_n(\delta) \lesssim \delta^\alpha + \sqrt{n}\left(\delta^{2\alpha-2}n^{-1}\right)^{\frac{r}{2(r-1)}\frac{p-1}{p}}$ as long as $\delta^{-1} \lesssim n^{1/(2(1-\alpha))}$. In verifying (iii), we see that this imposes the constraints $s_n \lesssim n^{1/(2(1-\alpha))}$. Now, $s_n^{2-\alpha} \lesssim n^{1/2}$ implies that $s_n \lesssim n^{(2\eta-1)/4\eta}$, while $s_n^2\left(s_n^{2-2\alpha}n^{-1}\right)^{\frac{r}{2(r-1)}\frac{p-1}{p}} \lesssim 1$ implies that $s_n \lesssim n^{\frac{\gamma(2\eta-1)}{4\eta+2(\gamma-1)}}$ where $\gamma = \frac{r}{2(r-1)}\frac{p-1}{p} < 1$. Hence, we have that $s_n \lesssim n^{\frac{\gamma(2\eta-1)}{4\eta+(\gamma-1)}}$and by definition of α, we also see that $s_n \lesssim n^{1/(2(1-\alpha))}$ as required. This proves the first statement of the theorem. Lemma 9 in Section 5.7 shows that the result also holds for approximate minimizers. Finally, the proof in the last statement of the theorem is deferred to Section 5.6.

5.5 Proof of Theorem 4

We introduce additional notation. Let $l^\infty\left(\mathcal{H}^K\right)$ be the space of uniformly bounded functions on $\mathcal{H}^K$. Let $\Psi(\mu)$ be the operator in $l^\infty\left(\mathcal{H}^K\right)$ such that $\Psi(\mu)h = P\partial\ell_\mu h$, $h \in \mathcal{H}^K$. When $\mu_0 \in \operatorname{int}\left(\mathcal{H}^K(B)\right)$, it holds that $\Psi(\mu_0)h = 0$, for any $h \in \mathcal{H}^K(1)$. The empirical counterpart of $\Psi(\mu)$ is the operator

$\Psi_n(\mu)$ such that $\Psi_n(\mu)h = P_n\partial\ell_\mu h$. Finally, write $\dot{\Psi}_{\mu_0}(\mu-\mu_0)$ for the Fréchet derivative of $\Psi(\mu)$ at μ_0 tangentially to $(\mu-\mu_0)$, where $\mu,\mu_0 \in \mathcal{H}^K(B)$. Then, $\dot{\Psi}_{\mu_0}$ is an operator from $\mathcal{H}^K$ to $l^\infty(\mathcal{H}^K)$. This same notation is used in van der Vaart and Wellner (2000, ch.3.3).

By the conditions of the theorem, $\mu_0 \in \operatorname{int}(\mathcal{H}^K(B))$, hence by the first order conditions, $\Psi(\mu_0)h = 0$ for any $h \in \mathcal{H}^K(1)$. This remark and Lemma 5 prove the first part in the theorem. By this remark again, and basic algebra,

$$\begin{aligned}\sqrt{n}\Psi_n(\mu_n) &= \sqrt{n}\Psi_n(\mu_0) + \sqrt{n}\left[\Psi(\mu_n) - \Psi(\mu_0)\right] \\ &\quad + \sqrt{n}\left[\Psi_n(\mu_n) - \Psi(\mu_n)\right] - \sqrt{n}\left[\Psi_n(\mu_0) - \Psi(\mu_0)\right]. \qquad (23)\end{aligned}$$

To bound the last two terms, we verify that

$$\sup_{h\in\mathcal{H}^K(1)} \sqrt{n}\left[\left(\Psi_n(\mu_n) - \Psi(\mu_n)\right) - \left(\Psi_n(\mu_0) - \Psi(\mu_0)\right)\right]h = o_P(1). \qquad (24)$$

This follows if (i) $\sqrt{n}(\Psi_n(\mu) - \Psi(\mu))h$, $\mu \in \mathcal{H}^K(B)$, $h \in \mathcal{H}^K(1)$, converges weakly to a Gaussian process with continuous sample paths, (ii) $\mathcal{H}^K(B)$ is compact under the uniform norm, and (iii) μ_n is consistent for μ_0 in $|\cdot|_\infty$. Point (i) is satisfied by Lemma 5, which also controls the first term on the r.h.s. of (23). Point (ii) is satisfied by Lemma 3. Point (iii) is satisfied by Theorem 1. Hence, by continuity of the sample paths of the Gaussian process, as $|\mu_n - \mu_0|_\infty \to 0$ in probability (using Point iii), the above display holds true. We now control the second term on the r.h.s. of (23). For any $h \in \mathcal{H}^K(1)$,

$$\left|\left[\Psi(\mu_n) - \Psi(\mu_0)\right]h - \dot{\Psi}_{\mu_0}(\mu_n - \mu_0)h\right| \leq \sup_{t\in(0,1)}\left|P\partial^3\ell_{\mu_0+t(\mu_n-\mu_0)}(\mu_n-\mu_0)^2 h\right| \qquad (25)$$

using differentiability of the loss function and Taylor's theorem in Banach spaces. By the condition that $|\Delta_3|_p < \infty$, and the fact that h is uniformly bounded, using Holder inequality, the r.h.s. is a constant multiple of $\left|(\mu-\mu_0)^2\right|_{p/(p-1)}$. By Lemma 1, $|\mu-\mu_0|_\infty \leq 2B$, so that $\left|(\mu-\mu_0)^2\right|_{p/(p-1)} \lesssim |\mu-\mu_0|_2^{2(p-1)/p}$. Hence, the r.h.s of (25) is $o_P\left(n^{-1/2}\right)$ if $|\mu-\mu_0|_2 = o_P\left(n^{\frac{1}{4}\left(\frac{p}{p-1}\right)}\right)$, which is the case by assumption. These calculations show that

$$\sqrt{n}\left[\Psi(\mu_n) - \Psi(\mu_0)\right] = \sqrt{n}\dot{\Psi}_{\mu_0}(\mu_n - \mu_0) + o_P(1).$$

Inserting the above in (23), and using (24), we deduce that

$$\begin{aligned}\sqrt{n}\Psi_n(\mu_n) - \sqrt{n}\Psi_n(\mu_0) &= \sqrt{n}\left(\Psi(\mu_n) - \Psi(\mu_0)\right) + o_P(1) \\ &= \sqrt{n}\dot{\Psi}_{\mu_0}(\mu_n - \mu_0) + o_P(1). \qquad (26)\end{aligned}$$

By Lemma 5, $\sqrt{n}\Psi_n(\mu_0) = O_P(1)$. For the moment, suppose that μ_n is the exact solution to the minimization problem in (6). By Lemma 8, $\sup_{h\in\mathcal{H}^K(1)}\sqrt{n}\Psi_n(\mu_n)h = O_P(1)$, so that $\sup_{h\in\mathcal{H}^K(1)}\sqrt{n}\dot{\Psi}_{\mu_0}(\mu_n - \mu_0)h = O_P(1)$. Finally, if $\sup_{h\in\mathcal{H}^K(1)}\sqrt{n}\Psi_n(\mu_n)h = o_P(1)$, (26) together with the previous displays imply that $-\lim_n\sqrt{n}(\Psi_n(\mu_0) - \Psi(\mu_0)) = \lim_n \dot{\Psi}_{\mu_0}\sqrt{n}(\mu_n - \mu_0)$ in probability, where the l.h.s. has same distribution as the Gaussian process G given in the statement of the theorem. It remains to show that if we use an approximate minimizer say ν_n to distinguish it here from μ_n in (6), the result still holds. Lemma 9, in Section 5.7, shows that this is true, hence completing the proof of Theorem 4.

5.6 Lower Bound on L_2 Convergence Rates

We use the same notation as in (23) and rely on the arguments that followed that display. When $\mu_0 \in \operatorname{int}\left(\mathcal{H}^K(B)\right)$, we deduce that $\sqrt{n}\left[\Psi(\mu_n) - \Psi(\mu_0)\right]h = O_P(1)$ for any $h \in \mathcal{H}^K(1)$. We choose $h = (2B)^{-1}(\mu_n - \mu_0)$. By definition of $\Psi(\mu)h$, using Taylor's theorem as in (19) and the notation defined there, $\sqrt{n}\left[\Psi(\mu_n) - \Psi(\mu_0)\right] = \sqrt{n}P\partial^2\ell_{\mu_t}(\mu_u - \mu_0)h$. By Condition 2 and the specific choice of h, the r.h.s. is lower bounded by a constant multiple of $\sqrt{n}P(\mu_u - \mu)^2$. This implies that $|\mu_n - \mu_0|_2 = O_P\left(n^{-1/4}\right)$. This bound verifies the last statement of Theorem 3.

5.7 Asymptotic Minimizers

The following lemma collects results on asymptotic minimizers.

Lemma 9 *Let $(\epsilon_n)_{n\geq 1}$ be an $o_p(1)$ sequence. Suppose that ν_n satisfies $P_n\ell_{\nu_n} = P_n\ell_{\mu_n} + \epsilon_n$, where μ_n is as in (6). Also suppose that $\nu_{n,\rho}$ satisfies $P_n\ell_{\nu_{n,\rho}} + \rho\left|\nu_{n,\rho}\right|^2_{\mathcal{H}^K} = P_n\ell_{\mu_{n,\rho}} + \rho\left|\mu_{n,\rho}\right|^2_{\mathcal{H}^K} + \rho\epsilon_n$, where $\mu_{n,\rho}$ is as in (9) and $\rho n^{1/2} \to \infty$.*

1. *Under the conditions of Theorem 1, $|\mu_n - \nu_n|_\infty = o(1)$ almost surely if $\epsilon_n \to 0$ almost surely or in probability if $\epsilon_n = o_P(1)$.*

2. *Under Condition 2, $|\mu_{n,\rho} - \nu_{n\rho}|_{\mathcal{H}^K} = \epsilon_n$ in probability, and there is a finite B such that $|\nu_{n,\rho}|_{\mathcal{H}^K} \leq B$ eventually in probability.*

3. *Under the conditions of Theorem 3, $|\mu_n - \nu_n|_2 = O_P\left(s_n^{-1}\right)$ if $\epsilon_n = O_P\left(s_n^{-2}\right)$.*

4. *If $\epsilon_n = o_P\left(n^{-1}\right)$, under the Conditions of Theorem 4, $\sup_{h\in\mathcal{H}^K(1)}\left|\Psi_n(\mu_n)h - \Psi_n(\nu_n)h\right| = o_P\left(n^{-1/2}\right)$.*

Proof. We prove each statement separately.

Proof of Point 1. Consider the constrained estimator. For the uniform convergence, by assumption we replace $P_n\ell_{\nu_n}$ with $P_n\ell_{\mu_n}$ with an error $o(1)$ almost surely. Hence, the proof of Theorem 1 is not altered and this implies Point 1 in the lemma.

Proof of Point 2. Consider the penalized estimator. To this end, follow the same steps in the proof of 5.14 in Theorem 5.9 of Steinwart and Christmann (2008). Mutatis mutandis, the argument in their second paragraph on page 174 gives

$$\begin{aligned}&\left\langle \nu_{n,\rho} - \mu_{n,\rho}, P_n\partial\ell_{\mu_{n,\rho}}\Phi + 2\rho\mu_{n,\rho}\right\rangle_{\mathcal{H}^K} + \rho\left|\mu_{n,\rho} - \nu_{n,\rho}\right|_{\mathcal{H}^K}\\ \leq &P_n\ell_{\nu_{n,\rho}} + \rho\left|\nu_{n,\rho}\right|^2_{\mathcal{H}^K} - \left(P_n\ell_{\mu_{n,\rho}} + \rho\left|\mu_{n,\rho}\right|^2_{\mathcal{H}^K}\right).\end{aligned}$$

Derivation of this display requires convexity of $L(z,t)$ w.r.t. t, which is the case by Condition 2. By assumption, the r.h.s. is $\rho\epsilon_n$. Note that $\mu_{n,\rho}$ is the exact minimizer of the penalized empirical risk. Hence, eq. (5.12) in Theorem 5.9 of Steinwart and Christmann (2008) says that $\mu_{n,\rho} = -(2\rho)^{-1} P_n \partial \ell_{\mu_{n,\rho}} \Phi$ for any $\rho > 0$, implying that the inner product in the display is zero. By these remarks, we deduce that the above display simplifies to $\rho \left|\mu_{n,\rho} - \nu_{n,\rho}\right|_{\mathcal{H}^K} = \rho\epsilon_n$. In consequence, $\left|\mu_{n,\rho} - \nu_{n,\rho}\right|_{\mathcal{H}^K} = o_P(1)$ so that by the triangle inequality, and Lemma 8, $\left|\nu_{n,\rho}\right|_{\mathcal{H}^K} \leq B$ eventually, in probability for some $B < \infty$.

Proof of Point 3. By Point 1 with $\epsilon_n = o_P(1)$, we obtain consistency for ν_n. Once consistency is ensured, the rates of convergence are not altered according to Theorem 3.2.5 of van der Vaart and Wellner (2000) as long as $\epsilon_n = O_P\left(s_n^{-2}\right)$ where s_n is as in Theorem 3. By the triangle inequality we obtain the result.

Proof of Point 4. Conditioning on the data, by definition of μ_n, the variational inequality $P_n \partial \ell_{\mu_n}(\nu_n - \mu_n) \geq 0$ holds because $\nu_n - \mu_n$ is an element of the tangent cone of $\mathcal{H}^K(B)$ at μ_n. Conditioning on the data, by Taylor's theorem in Banach spaces, and the fact that $\inf_{z\in\mathcal{Z},|t|\leq B} \partial^2 L(z,t) > 0$ by Condition 2, we have that$\left|P_n \ell_{\nu_n} - P_n \ell_{\mu_n}\right| \gtrsim P_n(\mu_n - \nu_n)^2$. By the conditions of the lemma, and the previous inequality deduce that $P_n(\mu_n - \nu_n)^2 = O_P(\epsilon_n)$. Now, conditioning on the data, by Fréchet differentiability,

$$\begin{aligned} \left|\Psi_n(\mu_n) h - \Psi_n(\nu_n) h\right| &= \left|P_n \partial \ell_{\nu_n} - P_n \partial \ell_{\mu_n}\right| \\ &\leq P_n \left| \sup_{\mu \in \mathcal{H}^K(B)} \partial^2 \ell_\mu (\nu_n - \mu_n) h \right|. \end{aligned}$$

By Holder's inequality, and the fact that $h \in \mathcal{H}^K(1)$ is bounded, the r.h.s. is bounded by a constant multiple of

$$\left[P_n \left| \sup_{\mu \in \mathcal{H}^K(B)} \partial^2 \ell_\mu \right|^2\right]^{1/2} \left[P_n(\nu_n - \mu_n)^2\right]^{1/2} \lesssim \left[P_n \Delta_2^2\right]^{1/2} \left[P_n(\nu_n - \mu_n)^2\right]^{1/2}.$$

By Condition 2, deduce that $P_n \Delta_2^2 = O_P(1)$ so that, by the previous calculations, the r.h.s. is bounded by a quantity $O_P\left(\epsilon_n^{1/2}\right) = o_P\left(n^{-1/2}\right)$ under the conditions of the lemma. ■

The first two points in the lemma prove Point 4 in Theorem 2 and the last part of Theorem 1. The third point proves the last statement in Theorem 3. The fourth point is used in the proof of Theorem 4 to show that for an asymptotic minimizer the first order condition remains $o_P\left(n^{-1/2}\right)$.

5.8 Proof of Theorem 5

Only here, for typographical reasons, write $\ell(\mu)$ instead of ℓ_μ and similarly for $\partial \ell(\mu)$. Let

$$h_m := \arg \min_{h \in \mathcal{L}^K(B)} P_n \partial \ell(F_{m-1}) h.$$

Note that by linearity, and the l_1 constraint imposed by $\mathcal{L}^K(B)$, the minimum is obtained by an additive function with $K-1$ additive components equal to zero and a non-zero one in $\mathcal{H}$ with norm

$|\cdot|_{\mathcal{H}}$ equal to B, i.e. $Bf^{s(m)}$, where $f^{s(m)} \in \mathcal{H}(1)$. Define,

$$D(F_{m-1}) := \min_{h \in \mathcal{L}^K(B)} P_n \partial \ell (F_{m-1})(h - F_{m-1}),$$

so that for any $\mu \in \mathcal{L}^K(B)$,

$$P_n \ell(\mu) - P_n \ell(F_{m-1}) \geq D(F_{m-1}) \tag{27}$$

by convexity. For $m \geq 1$, define $\tilde{\tau}_m = 2/(m+2)$ if τ_m is chosen by line search, or $\tilde{\tau}_m = \tau_m$ if $\tau_m = m^{-1}$. By convexity, again,

$$P_n \ell(F_m) = \inf_{\tau \in [0,1]} P_n \ell(F_{m-1} + \tau(h_m - F_{m-1})) \leq P_n \ell(F_{m-1}) + P_n \partial \ell(F_{m-1})(h_m - F_{m-1}) \tilde{\tau}_m + \frac{Q}{2} \tilde{\tau}_m^2$$

where

$$Q := \sup_{h, F \in \mathcal{L}^K(B), \tau \in [0,1]} \frac{2}{\tau^2} \left[P_n \ell(F + \tau(h - F)) - P_n \ell(F) - \tau P_n \partial \ell(F)(h - F) \right].$$

The above two displays together with the definition of $D(F_{m-1}) = P_n \partial \ell(F_{m-1})(h_m - F_{m-1})$ imply that for any $\mu \in \mathcal{L}^K(B)$,

$$\begin{aligned} P_n \ell(F_m) &\leq P_n \ell(F_{m-1}) + \tilde{\tau}_m D(F_{m-1}) + \frac{Q}{2} \tilde{\tau}_m^2 \\ &\leq P_n \ell(F_{m-1}) + \tilde{\tau}_m (P_n \ell(\mu) - P_n \ell(F_{m-1})) + \frac{Q}{2} \tau_m^2, \end{aligned}$$

where the second inequality follows from (27). Subtracting $P_n \ell(\mu)$ on both sides and rearranging, we have the following recursion

$$P_n \ell(F_m) - P_n \ell(\mu) \leq (1 - \tilde{\tau}_m)(P_n \ell(F_{m-1}) - P_n \ell(\mu)) + \frac{Q}{2} \tau_m^2.$$

The result is proved by bounding the above recursion for the two different choices of $\tilde{\tau}_m$. When, $\tilde{\tau}_m = 2/(m+1)$, the proof of Theorem 1 in Jaggi (2013) bounds the recursion by $2Q/(m+2)$. If $\rho_m = m^{-1}$, then, Lemma 2 in Sancetta (2016) bounds the recursion by $4Q \ln(1+m)/m$ for any $m \geq 1$. It remains to bound Q. By Taylor expansion of $\ell(F + \tau(h - F))$ at $\tau = 0$,

$$\ell(F + \tau(h - F)) = \ell(F) + \partial \ell(F)(h - F)\tau + \frac{\partial^2 \ell(F + t(h - F))(h - F)^2 \tau^2}{2}$$

for some $t \in [0,1]$. It follows that

$$\begin{aligned} Q &\leq \max_{t \in [0,1]} \sup_{h, F \in \mathcal{L}^K(B), \tau \in [0,1]} P_n \partial^2 \ell(F + t(h - F))(h - F)^2 \\ &\leq 4B^2 \sup_{|t| < B} P_n d^2 L(\cdot, t)/dt^2. \end{aligned}$$

5.9 Proof of Lemma 1

Point 1 is obvious. By the relation between the l_1 and l_2 norms (derived using Minkowski and the Cauchy-Schwarz inequality), $|\mu|_{\mathcal{H}^K} \leq |\mu|_{\mathcal{L}^K} \leq \sqrt{K} |\mu|_{\mathcal{H}^K}$ and this shows the inclusion in Point 2.

Every Hilbert space is uniformly convex, hence the ball of radius B is a convex set, and this proves Point 3. By the RKHS property $f^{(k)}\left(x^{(k)}\right)=\left\langle f^{(k)},C\left(\cdot,x^{(k)}\right)\right\rangle_{\mathcal{H}}$, for $\mu\left(x\right)=\sum_{k=1}^{K}f^{(k)}\left(x^{(k)}\right)$,

$$\left|\mu\left(x\right)\right|=\left|\sum_{k=1}^{K}\left\langle f^{(k)},C\left(\cdot,x^{(k)}\right)\right\rangle_{\mathcal{H}}\right|.$$

When $\mu\in\mathcal{L}^{K}\left(B\right)$, by the Cauchy-Schwarz inequality and the RKHS property again, the display is bounded by

$$\sum_{k=1}^{K}\left|f^{(k)}\right|_{\mathcal{H}}\left|C\left(\cdot,x^{(k)}\right)\right|_{\mathcal{H}}=\sum_{k=1}^{K}\left|f^{(k)}\right|_{\mathcal{H}}\sqrt{C\left(x^{(k)},x^{(k)}\right)}\leq cB,$$

using the definition of $\mathcal{L}^{K}\left(B\right)$ and the assumed bound on the kernel. The above two displays imply that $\left|\mu\right|_{\infty}\leq cB$. This shows the result for $p=\infty$. For any $p\in[1,\infty)$, use the trivial inequality $P\left|\mu\right|^{p}\leq\left|\mu\right|_{\infty}^{p}P\left(\mathcal{X}^{K}\right)=\left|\mu\right|_{\infty}^{p}$. When $\mu\in\mathcal{H}^{K}\left(B\right)$, by Cauchy-Schwarz inequality it is simple to deduce from the above two displays that $\left|\mu\right|_{\infty}\leq c\sqrt{K}B$. These remarks prove Point 4.

References

[1] Banerjee, A., D. Dunson and S. Todkar (2013) Efficient Gaussian Process Regression for Large Data Sets. Biometrika 94, 1-16.

[2] Basrak, B., R.A. Davis and T. Mikosch (2002) Regular Variation of GARCH Processes, Stochastic Processes and their Applications 99, 95-115.

[3] Bradley, R.C. (1986) Basic Properties of Strong Mixing Conditions. In E. Eberlein and M.S. Taqqu (eds.), Dependence in Probability and Statistics, 165-192. Boston: Birkhauser.

[4] Buja, A., T. Hastie and R. Tibshirani (1989) Linear Smoothers and Additive Models (with discussion). Annals of Statistics 17, 453-555.

[5] Christmann, A. and I. Steinwart (2007) Consistency and Robustness of Kernel-Based Regression in Convex Risk Minimization. Bernoulli 13, 799-719.

[6] Christmann, A. and R. Hable (2012) Consistency of Support Vector Machines Using Additive Kernels for Additive Models. Computational Statistics and Data Analysis 56, 854-873.

[7] Doukhan, P. (1994) Mixing: Properties and Examples. New York: Springer.

[8] Doukhan, P., P. Massart and E. Rio (1995) Invariance Principles for Absolutely Regular Empirical Processes. Annales de l'Institut Henri Poincar'e: Probabilit'es et Statistiques 31, 393-427

[9] Geyer, C.J. (1994) On the Asymptotics of Constrained M-Estimation. Annals of Statistics 22, 1993-2010.

[10] Hable, R. (2012) Asymptotic Normality of Support Vector Machine Variants and Other Regularized Kernel Methods. Journal of Multivariate Analysis 106, 92-117.

[11]] Hang, H. and I. Steinwart (2014) Fast Learning from α-Mixing Observations. Journal of Multivariate Analysis 127, 184-199.

[12] Hang, H. and I. Steinwart (2017) A Bernstein-type Inequality for Some Mixing Processes and Dynamical Systems with an Application to Learning. Annals of Statistics 45, 708-743.

[13] Kallneberg, O. (1997) Foundations of Modern Probability. New York: Springer.

[14] Lázaro-Gredilla, M., J. Quiñonero-Candela, C.E. Rasmussen and A.R. Figueiras-Vidal (2010) Sparse Spectrum Gaussian Process Regression. Journal of Machine Learning Research 11, 1865-1881.

[15] Li, W.V. and W. Linde (1999) Approximation, Metric Entropy and Small Ball Estimates for Gaussian Measures. Annals of Probability 27, 1556-1578.

[16] Lv, S., H. Lin, H. Lian, and J. Huang (2018) Oracle Inequalities for Sparse Additive Quantile Regression in Reproducing Kernel Hilbert Space. Annals of Statistics 46, 781-813.

[17] Jaggi, M. (2013) Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization. Journal of Machine Learning Research (Proceedings ICML 2013). URL: <http://jmlr.org/proceedings/papers/v28/jaggi13-supp.pdf>.

[18] Mammen, E., O.B. Linton, and J.P. Nielsen (1999) The Existence and Asymptotic Properties of a Backfitting Projection Algorithm under Weak Conditions. Annals of Statistics 27, 1443-1490.

[19] McDonald, D.J. and C.R. Shalizi (2017) Rademacher Complexity of Stationary Sequences. https://arxiv.org/abs/1106.0730.

[20] Meier, L., P. Bühlmann and S. van de Geer (2009). High-Dimensional Additive Modeling. Annals of Statistics 37, 3779-3821.

[21] Mendelson, S. (2002) Geometric Parameters of Kernel Machines. In: Kivinen J., Sloan R.H. (eds) Computational Learning Theory (COLT). Lecture Notes in Computer Science 2375. Berlin: Springer.

[22] Nair, P.B., A. Choudhury and A.J. Keane (2002) Some Greedy Learning Algorithms for Sparse Regression and Classification with Mercer Kernels. Journal of Machine Learning Research 3, 781-801..

[23] Rakhlin, A., K. Sridharan and A. Tewari (2011) Online Learning: Stochastic and Constrained Adversaries. In J. Shawe-Taylor, R. S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger (eds.) Advances in Neural Information Processing Systems 24 [NIPS 2011], 1764–1772. https://arxiv.org/abs/1104.5070

[24] Rasmussen, C. and C.K.I. Williams (2006) Gaussian Processes of Machine Learning. Cambridge, MA: MIT Press.

[25] Ritter, K., G.W. Wasilkowski and H. Wozniakowski (1995) Multivariate Integration and Approximation for Random Fields Satisfying Sacks-Ylvisaker Conditions. Annals of Applied Probability 5, 518-540.

[26] Sancetta A. (2016) Greedy Algorithms for Prediction. Bernoulli 22, 1227-1277.

[27] Smola, A.J., and P.L. Bartlett (2001) Sparse Greedy Gaussian Process Regression. Advances in Neural Information Processing Systems 13, 619-625.

[28] Steinwart, I., and A. Christmann (2008) Support Vector Machines. Berlin: Springer.

[29] Steinwart, I., D. Hush and C. Scovel (2009a) Learning From Dependent Observations, Journal of Multivariate Analysis 100, 175-194.

[30] Steinwart, I., D. Hush, and C. Scovel (2009b) Optimal Rates for Regularized Least Squares Regression. Proceedings of the Annual Conference on Learning Theory, 79-93.

[31] Stone, C. (1982) Optimal Global Rates of Convergence for Nonparametric Regression. Annals of Statistics 10, 1040-1053

[32] Suzuki, T. (2018) Fast Learning Rate of Non-Sparse Multiple Kernel Learning and Optimal Regularization Strategies. Electronic Journal of Statistics 12, 2141-2192.

[33] Suzuki, T. and M. Sugiyama (2013) Fast Learning Rate of Multiple Kernel Learning: Tradeoff Between Sparsity and Smoothness. Annals of Statistics 41, 1381-1405.

[34] Wahba, G. (1990) Spline Models for Observational Data. Philadelphia: SIAM.

[35] van der Vaart, A.W. and J.A. Wellner (2000) Weak Convergence and Empirical Processes. New York: Springer.