Robustness of Reweighted Least Squares Kernel Based Regression

Michiel Debruyne (corresponding author)
Department of Mathematics, Universiteit Antwerpen
Middelheimlaan 1G, 2020 Antwerpen, Belgium
Tel: +32 32653887
Fax: +32 32653777
Email: michiel.debruyne@ua.ac.be

Andreas Christmann
Department of Mathematics, University of Bayreuth
D-95440 Bayreuth, Germany

Mia Hubert
Department of Mathematics - LStat, K.U.Leuven
Celestijnenlaan 200B, B-3001 Leuven, Belgium

Johan A.K. Suykens
ESAT-SCD/SISTA, K.U.Leuven
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

March 31, 2009

Abstract

Kernel Based Regression (KBR) minimizes a convex risk over a possibly infinite dimensional reproducing kernel Hilbert space. Recently it was shown that KBR with a least squares loss function may have some undesirable properties from a robustness point of view: even very small amounts of outliers can dramatically affect the estimates. KBR with other loss functions is more robust, but often gives rise to more complicated computations (e.g. for Huber or logistic losses). In classical statistics robustness is often improved by reweighting the original estimate. In this paper we provide a theoretical framework for reweighted Least Squares KBR (LS-KBR) and analyze its robustness. Some important differences are found with respect to linear regression, indicating that LS-KBR with a bounded kernel is much more suited for reweighting. In two special cases our results can be translated into practical guidelines for a good choice of weights, providing robustness as well as fast convergence. In particular a logistic weight function seems an appropriate choice, not only to downweight outliers, but also to improve performance at heavy tailed distributions. For the latter some heuristic arguments are given comparing concepts from robustness and stability.

1 Introduction

Kernel Based Regression (KBR) is a popular method belonging to modern machine learning and is based on convex risk minimization. An objective function is optimized consisting of the sum of a data term and a complexity term. The data term represents the loss at the given data points. The optimization is done over a reproducing kernel Hilbert space (RKHS) associated to a kernel function. For some kernels this space is tractable (e.g. with a linear kernel KBR corresponds to linear ridge regression), but for many kernels the corresponding RKHS is a very high (possibly infinite) dimensional space, complicating mathematical analysis.

Recently the robustness of these methods was investigated with respect to outlying observations [Christmann and Steinwart, 2007, Steinwart and Christmann, 2008]. It was found that KBR with a loss function with unbounded first derivative can be heavily affected by the smallest amount of outliers. As such a least squares loss is not a good choice from a robustness point of view, contrary to e.g. an L_1 loss or Vapnik's ϵ-insensitive loss function. From a computational point of view on the other hand, a least squares loss leads to faster algorithms solving a linear system of equations [Wahba, 1990, Evgeniou et al., 2000, Suykens et al., 2002b], whereas an L_1 loss involves solving a quadratic programming problem. Section 2 gives a short overview of these results.

In Section 3 we investigate the possibility of stepwise reweighting Least Squares KBR (LS-KBR) in order to improve its robustness. This is already proposed in Suykens et al. [2002a], where data experiments show how reweighting steps reduce the effect of outliers, whereas the algorithm still only requires solving linear systems of equations. More specifically the following main results are obtained.

- We introduce the weighted regularized risk and show a representer theorem for its minimizer (Theorem 1).

- We define a sequence of successive weighted least squares regularized risk minimizers. It is proven that this sequence converges if the weight function is nonincreasing. Moreover we prove that the solution of KBR with any invariant convex loss function can be obtained as a limit of a sequence of weighted LS-KBR estimators.

- To analyze robustness the influence function of reweighted LS-KBR is obtained (Theorem 3). This shows that the influence function after performing a reweighting step depends on a certain operator evaluated at the influence function before the reweighting step. Since our goal is to reduce the influence function (thereby improving robustness), it is important that the norm of this operator is smaller than one. Under certain assumptions we are able to determine conditions on the weight function such that an operator norm smaller than one is guaranteed. This provides some practical guidelines on how to choose the weights in those special cases.

- If the weight function is well chosen, it is shown that reweighted LS-KBR with a bounded kernel converges to an estimator with a bounded influence function, even if the initial estimator is LS-KBR, which is not robust. This is an important difference compared to linear reweighted LS regression, which converges to an estimator with an unbounded influence function.

Throughout the paper the influence function is used as a tool to assess the robustness of the methods under consideration. It reflects how an estimator changes when a tiny amount of contamination is added to the original distribution. As such it can also be seen as a measure of stability at continuous distributions: it shows how the result changes when the distribution changes slightly. This is very similar to some stability measures that were recently defined. Poggio et al. [2004] for example show that it is very important for a method not to change too much when an additional point is added to a sample. However, these stability measures typically add a point which is generated i.i.d. from the same distribution as the other points. In robust statistics the added contamination can be any possible outcome, even a very unlikely one under the generating distribution. Thus in a way robustness is a stronger requirement than stability. A robust method should give stable results when adding *any* possible point, even an extremely unlikely one. In Section 4 we explore these links and differences between robustness and stability a little bit further. We show how the influence function can be used to approximate traditional stability measures by evaluating it at sample points. A smaller influence function leads to methods that are more stable. Therefore, since reweighting steps reduce the influence function, they also improve the stability of the initial LS-KBR estimator. When the error distribution is Gaussian, this effect is

rather small. At heavy tailed distributions on the other hand the stability can improve quite drastically.

In Section 5 we discuss some practical consequences of our theoretical results. Some weight functions traditionally used in linear regression are examined. It is shown that some weight functions, e.g. Hampel weights, do not satisfy the necessary conditions. Although these conditions are proven to be relevant only in two special cases, examples show that weight functions not satisfying these conditions can also fail in practice, in contrast to e.g. a logistic weight function satisfying all conditions. In the same section we provide some results on the convergence speed. As explained the norm of a certain operator represents an upper bound for the reduction of the influence function in consecutive steps. This norm is calculated in the two special cases considered. Unfortunately it depends on the error distribution, such that this upper bound is not distribution free. In Table 3 results are shown for several error distributions. Again logistic weights give good results in the most common cases.

Finally we analyze some specific data sets. The robustness of reweighted LS-KBR is demonstrated on a data example from astronomy. A small simulation study demonstrates that reweighting leads to better stability at heavy tailed distributions.

2 Kernel based regression

2.1 Kernels

Kernel Based Regression (KBR) methods estimate a functional relationship between a covariate random variable X and a response variable Y, using a sample of n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^d \times \mathbb{R}$ with joint distribution P. The following definitions are taken from Steinwart and Christmann [2008].

Definition 1 *Let $\mathcal{X}$ be a non-empty set. Then a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a* kernel *on $\mathcal{X}$ if there exists a $\mathbb{R}$-Hilbert space $\mathcal{H}$ with an inner product $\langle \cdot, \cdot \rangle$ and a map $\Phi : \mathcal{X} \to \mathcal{H}$ such that for all $x, x' \in \mathcal{X}$ we have*

$$K(x, x') \;=\; \langle \Phi(x), \Phi(x') \rangle \,. \tag{1}$$

We call Φ a feature map and $\mathcal{H}$ a feature space of K.

Frequently used kernels when $\mathcal{X} = \mathbb{R}^d$ include:

- the linear kernel $K(x, x') = x^t x'$. From Definition 1 it is clear that $\mathcal{H}$ equals $\mathbb{R}^d$ itself and Φ is simply the identity map.
- the polynomial kernel of degree p with offset $\tau > 0 : K(x, x') = (\tau + x^t x')^p$.
- the Radial Basis Function (RBF) kernel $K(x, x') = \exp(-\|x - x'\|_2^2/\sigma^2)$ with bandwidth $\sigma > 0$. In this case the feature space $\mathcal{H}$ is infinite dimensional. Also

note that the RBF kernel is bounded, since

$$\sup_{x,x' \in \mathbb{R}^d} K(x, x') = 1.$$

Both the linear and the polynomial kernel are of course unbounded.

Definition 2 *Let $\mathcal{X}$ be a non-empty set and $\mathcal{H}$ be a $\mathbb{R}$-Hilbert function space over X, i.e., a $\mathbb{R}$-Hilbert space that consists of functions mapping from $\mathcal{X}$ into $\mathbb{R}$.*

1. *A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$ if we have $K(\cdot, x) \in \mathcal{H}$ for all $x \in \mathcal{X}$ and the reproducing property $f(x) = \langle f, K(\cdot, x) \rangle$ holds for all $f \in \mathcal{H}$ and all $x \in \mathcal{X}$.*

2. *The space $\mathcal{H}$ is called a reproducing kernel Hilbert space (RKHS) over $\mathcal{X}$ if for all $x \in \mathcal{X}$ the Dirac functional $\delta_x : \mathcal{H} \to \mathbb{R}$ defined by*

$$\delta_x(f) := f(x), f \in \mathcal{H}$$

is continuous.

Note that any reproducing kernel is a kernel in the sense of Definition 1. The RKHS is also a feature space of K, with feature map $\Phi : \mathcal{X} \to \mathcal{H}$ given by

$$\Phi(x) = K(\cdot, x). \qquad\qquad x \in \mathcal{X}.$$

We then call Φ the canonical feature map.

2.2 Empirical regularized risk

Let $L : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a function which is convex with respect to its second argument. Then KBR methods minimize the empirical regularized risk

$$\hat{f}_{n,\lambda} := \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2 \tag{2}$$

where $\lambda \geq 0$ is a regularization parameter and $\mathcal{H}$ is the RKHS of a reproducing kernel K as in Definition 2, see for example Wahba [1990] or Evgeniou et al. [2000].

Results about the form of the solution of KBR methods are known as representer theorems. A well known result in the literature of statistical learning shows that

$$\hat{f}_{n,\lambda} = \frac{1}{n} \sum_{i=1}^{n} \alpha_i \Phi(x_i). \tag{3}$$

The form of the coefficients α_i strongly depends on the loss function. For the squared loss $L(y, t) = (y - t)^2$, Tikhonov and Arsenin [1977] already characterized the coefficients α_i as solutions of a system of linear equations. For arbitrary convex differentiable loss

functions, like the logistic loss $L(y,t) = -\log(4) + |y-t| + 2\log(1+e^{-|y-t|})$, the α_i are the solution of a system of algebraic equations (Girosi [1998], Wahba [1999], Schölkopf et al. [2001]). For arbitrary convex, but possibly non differentiable loss functions, extensions were obtained by Steinwart [2003] and DeVito et al. [2004].

In practice the variational problem (2) and its representation (3) are closely related to the methodology of Support Vector Machines. This method formulates a primal optimization problem and solves it via a corresponding dual formulation. Vapnik [1995] extended this approach to the regression setting introducing Support Vector Regression (SVR) using the ϵ-insensitive loss function. A dual problem similar to (3) is solved, where the coefficients α_i are obtained from a quadratic programming problem. A least squares loss function however leads to a linear system of equations, generally easier to solve (see e.g. Suykens et al. [2002b], where primal-dual problems are formulated, including a bias term as well).

2.3 Theoretical regularized risk

For our theoretical results we will look at the minimization of the theoretical regularized risk

$$f_{P,\lambda} := \arg\min_{f\in\mathcal{H}} \mathbb{E}_P L(Y, f(X)) + \lambda \|f\|_{\mathcal{H}}^2. \tag{4}$$

It is clear that the empirical regularized risk (2) is a stochastic approximation of the theoretical regularized risk.

Two somewhat technical definitions are needed. Firstly we describe the growth behavior of the loss function [Christmann and Steinwart, 2007].

Definition 3 *Let $L : \mathcal{Y}\times\mathbb{R} \to [0,\infty)$ be a loss function, $a : \mathcal{Y} \to [0,\infty)$ be a measurable function and $p \in [0,\infty)$. We say that L is a loss function of type (a,p) if there exists a constant $c > 0$ such that*

$$L(y,t) \le c\,(a(y) + |t|^p + 1)$$

for all $y \in \mathcal{Y}$ and all $t \in \mathbb{R}$. Furthermore we say that L is of strong type (a,p) if the first two partial derivatives $L'(y,r) := \frac{\partial}{\partial r}L(y,r)$ and $L''(y,r) := \frac{\partial^2}{\partial^2 r}L(y,r)$ of L with respect to the second argument of L exist and L, L' and L'' are of (a,p)-type.

Secondly we also need the following definition about the distribution P.

Definition 4 *Let P be a distribution on $\mathcal{X}\times\mathcal{Y}$ with total variation $|P|$ and $a : \mathcal{Y} \to [0,\infty)$ be a measurable function. Then we write*

$$|P|_a := \int_{\mathcal{X}\times\mathcal{Y}} a(y)dP(x,y).$$

If $a(y) = |y|^p$ for $p > 0$ we write $|P|_p$.

In DeVito et al. [2004] the following representation of the theoretical regularized risk was proven.

Proposition 1 *Let $p \geq 1$, L be a convex loss function of strong type (a, p), and P be a distribution on $\mathcal{X} \times \mathcal{Y}$ with $|P|_a < \infty$. Let $\mathcal{H}$ be the RKHS of a bounded, continuous kernel K over $\mathcal{X}$, and $\Phi : \mathcal{X} \to \mathcal{H}$ be the canonical feature map of $\mathcal{H}$. Then with $h(x, y) = L'(y, f_{P,\lambda}(x))$ it holds that*

$$f_{P,\lambda} = -\frac{1}{2\lambda}\mathbb{E}_P\left[h\Phi\right]. \tag{5}$$

Consider the map T which assigns to every distribution P on $\mathcal{X} \times \mathcal{Y}$ with $|P|_a < \infty$, the function $T(P) = f_{P,\lambda} \in \mathcal{H}$. Let $P_{\epsilon,z}$ be a contaminated distribution, i.e. $P_{\epsilon,z} = (1-\epsilon)P + \epsilon\Delta_z$ where Δ_z denotes the Dirac distribution at the point z. Then the influence function of the functional T at the distribution P is defined as [Hampel et al., 1986]

$$IF(z; T, P) = \lim_{\epsilon \downarrow 0} \frac{T(P_{\epsilon,z}) - T(P)}{\epsilon} \tag{6}$$

for any $z \in \mathcal{X} \times \mathcal{Y}$ where this pointwise limit exists. The function $IF(z; T, P)$ measures the effect on T under infinitesimally small contamination at the point z. The following expression for the influence function of T was proven in Christmann and Steinwart [2007].

Proposition 2 *Let $\mathcal{H}$ be a RKHS of a bounded continuous kernel K on $\mathcal{X}$ with canonical feature map $\Phi : \mathcal{X} \to \mathcal{H}$, and $L : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a convex loss function of some strong type (a, p). Furthermore, let P be a distribution on $\mathcal{X} \times \mathcal{Y}$ with $|P|_a < \infty$. Then the influence function of T exists for all $z := (z_x, z_y) \in \mathcal{X} \times \mathcal{Y}$ and we have*

$$IF(z; T, P) = S^{-1}\left(\mathbb{E}_P[L'(Y, f_{P,\lambda}(X))\Phi(X)]\right) - L'(z_y, f_{P,\lambda}(z_x))S^{-1}\Phi(z_x)$$

where $S : \mathcal{H} \to \mathcal{H}$ is defined by $S(f) = 2\lambda f + \mathbb{E}_P\left[L''(Y, f_{P,\lambda}(X))\langle\Phi(X), f\ \rangle\Phi(X)\right]$.

Note that the influence function only depends on z through the term

$$-L'(z_y, f_{P,\lambda}(z_x))S^{-1}\Phi(z_x).$$

From a robustness point of view it is important to bound the influence function. The previous proposition shows that this can be achieved using a bounded kernel, e.g. the Gaussian RBF kernel, and a loss function with bounded first derivative, e.g. the logistic loss. The least squares loss function on the other hand leads to an unbounded influence function.

However, reweighting might improve the robustness of LS-KBR. In the next section we will extend the previous results to the case of reweighted LS-KBR.

Remark: For the special case of the least squares loss function, we provide a slight extension of Proposition 2, including an intercept term. For reasons of simplicity we will however not include this intercept term anymore further on and continue working with the functional part only, as in Christmann and Steinwart [2007].

3 Reweighted LS-KBR

3.1 Definition

For $f \in \mathcal{H}$, let $w(y - f(x)) : \mathbb{R} \to \mathbb{R}^+$ be a weight function depending on the residual $y - f(x)$ with respect to f. We will make the following assumptions about w from now on:

- (w_1) $w(r)$ a non-negative bounded Borel measurable function.
- (w_2) w an even function of r.
- (w_3) w continuous and differentiable with $w'(r) \leq 0$ for $r > 0$.

Then a sequence of successive minimizers of a weighted least squares regularized risk is defined as follows.

Definition 5 *Let $f_{P,\lambda}^{(0)} \in \mathcal{H}$ be an initial fit, e.g. obtained by ordinary unweighted LS-KBR. Let w be a weight function satisfying $(w_1)-(w_3)$. Then the $(k+1)$ step reweighted LS-KBR estimator is defined by*

$$f_{P,\lambda}^{(k+1)} := \arg\min_{f \in \mathcal{H}} \mathbb{E}_P \left[w(Y - f_{P,\lambda}^{(k)}(X))(Y - f(X))^2 \right] + \lambda \|f\|_{\mathcal{H}}^2. \tag{7}$$

3.2 Representation and convergence

The following representation theorem can be derived from Proposition 1 (for full proofs we refer to the Appendix).

Theorem 1 *Let P be a distribution on $\mathcal{X} \times \mathcal{Y}$ with $|P|_2 < \infty$. Then with $h^{(k+1)}(x, y) = (y - f_{P,\lambda}^{(k+1)}(x))$ it holds that*

$$f_{P,\lambda}^{(k+1)} = \frac{1}{\lambda} \mathbb{E}_P \left[w(Y - f_{P,\lambda}^{(k)}(X)) h^{(k+1)}(X, Y) \Phi(X) \right]. \tag{8}$$

Using this representation it can be proven that the sequence $\{f^{(k)}\}$ converges.

Theorem 2 *Let $f_{P,\lambda}^{(0)} \in \mathcal{H}$ be any initial fit and P a distribution on $\mathcal{X} \times \mathcal{Y}$ with $|P|_2 < \infty$. Let w be a weight function satisfying $(w_1) - (w_3)$. Then there exists $f_{P,\lambda}^{(\infty)} \in \mathcal{H}$ such that $f_{P,\lambda}^{(k)} \to f_{P,\lambda}^{(\infty)}$ as $k \to \infty$.*

Note that the limit $f_{P,\lambda}^{(\infty)}$ must satisfy (according to (8))

$$f_{P,\lambda}^{(\infty)} = \frac{1}{\lambda}\mathbb{E}_P\left[w(Y - f_{P,\lambda}^{(\infty)}(X))(Y - f_{P,\lambda}^{(\infty)}(X))\Phi(X)\right]. \tag{9}$$

Let L be a symmetric convex loss function. Suppose L is invariant, which means that there exists a function $l : \mathbb{R} \to [0, +\infty[$ such that $L(y, f(x)) = l(y - f(x))$ for all $y \in \mathcal{Y}$, $x \in \mathcal{X}$, $f \in \mathcal{H}$. Consider the specific choice $w(r) = l'(r)/(2r)$. If l is such that w satisfies conditions $(w_1) - (w_3)$ then it follows from (9) that $f_{P,\lambda}^{(\infty)}$ satisfies equation (5). Thus $f_{P,\lambda}^{(\infty)}$ is the unique minimizer of the theoretical risk (4) with loss L. Consequently the KBR solution for the loss function L can be obtained as the limit of a sequence of reweighted LS-KBR estimators with arbitrary initial fit. Note however that $|P|_2 < \infty$ is required to find the solution by reweighted LS-KBR. This can be more restrictive than the condition $|P|_a < \infty$ required in Proposition 1.

In general of course $f^{(\infty)}$ might depend on the initial fit, hence leading to different solutions for different $f^{(0)}$. This will be the case for instance if L is not a convex loss function. Then $f^{(\infty)}$ can be a local minimum of the regularized risk, depending on the initial start.

3.3 Influence functions

Theorem 3 *Let T_0 denote the map $T_0(P) = f_{P,\lambda}^{(0)} \in \mathcal{H}$. Denote by T_{k+1} the map $T_{k+1}(P) = f_{P,\lambda}^{(k+1)}$. Furthermore, let P be a distribution on $\mathcal{X} \times \mathcal{Y}$ with $|P|_2 < \infty$ and $\int_{\mathcal{X}\times\mathcal{Y}} w(y - f_{P,\lambda}^{(k)}(x))dP(x, y) > 0$. Then the influence function of T_{k+1} exists for all $z := (z_x, z_y) \in \mathcal{X} \times \mathcal{Y}$ and we have*

$$\begin{aligned} IF(z; T_{k+1}, P) &= -S_{w,k}^{-1}(\mathbb{E}_P w(Y - f_{P,\lambda}^{(k)}(X))(Y - f_{P,\lambda}^{(k+1)}(X))\Phi(X)) \\ &+ S_{w,k}^{-1}(C_{w,k}(IF(z; T_k, P))) + w(z_y - f_{P,\lambda}^{(k)}(z_x))(z_y - f_{P,\lambda}^{(k+1)}(z_x))S_{w,k}^{-1}(\Phi(z_x)) \end{aligned}$$

with operators $S_{w,k} : \mathcal{H} \to \mathcal{H}$ and $C_{w,k} : \mathcal{H} \to \mathcal{H}$ given by

$$S_{w,k}(f) = \lambda f + \mathbb{E}_P[w(Y - f_{P,\lambda}^{(k)}(X))\langle f, \Phi(X)\rangle\Phi(X)]$$

and

$$C_{w,k}(f) = -\mathbb{E}_P[w'(Y - f_{P,\lambda}^{(k)}(X))(Y - f_{P,\lambda}^{(k+1)}(X))\langle f, \Phi(X)\rangle\Phi(X)].$$

Note that the expression for $IF(z; T_{k+1}, P)$ consists of three terms. The first one is a constant function independent of z, i.e. it does not depend on the position z where we plug in the contamination. The third one depends on z but not on the influence of the previous step. The second term $(S_{w,k}^{-1} \circ C_{w,k})(IF(z; T_k, P))$ reflects the influence of the previous step. Since $S_{w,k}$ and $C_{w,k}$ are operators independent of z, this term can be unbounded if the influence function of the estimator in the previous step is unbounded, which is the case if we start for instance with LS-KBR as the initial estimator. However, it is possible that this influence of the initial fit is reduced because

the operator $S_{w,k}^{-1} \circ C_{w,k}$ is applied on it. In that case, the second term might vanish if we keep reweighting until convergence. To investigate this iterative reweighting, let us write

$$IF(z; T_{k+1}, P) = S_{w,k}^{-1}(C_{w,k}(IF(z; T_k, P))) + g_k$$

where

$$\begin{aligned} g_k = &-S_{w,k}^{-1}(\mathbb{E}_P w(Y - f_{P,\lambda}^{(k)}(X))(Y - f_{P,\lambda}^{(k+1)}(X))\Phi(X)) \\ &+ w(z_y - f_{P,\lambda}^{(k)}(z_x))(z_y - f_{P,\lambda}^{(k+1)}(z_x))S_{w,k}^{-1}(\Phi(z_x)). \end{aligned}$$

Then solving the recursive relation we have that

$$\begin{aligned} IF(z; T_{k+1}, P) = &\sum_{j=0}^{k} \left((S_{w,k}^{-1} \circ C_{w,k}) \circ \ldots \circ (S_{w,k-j+1}^{-1} \circ C_{w,k-j+1}) \right) (g_{k-j}) \\ &+ \left((S_{w,k}^{-1} \circ C_{w,k}) \circ \ldots \circ (S_{w,1}^{-1} \circ C_{w,1}) \right) (IF(z; T_0, P)). \end{aligned} \tag{10}$$

Assume that the operator norm of $S_{w,\infty}^{-1} \circ C_{w,\infty}$ is bounded by one: $||S_{w,\infty}^{-1} \circ C_{w,\infty}|| < 1$. Thus there exists $k_0 \in \mathbb{N}$ and $\epsilon > 0$ such that $||S_{w,k}^{-1} \circ C_{w,k}|| < 1 - \epsilon$ for all $k > k_0$. Then for $k > k_0$,

$$\begin{aligned} &\left\| \left((S_{w,k}^{-1} \circ C_{w,k}) \circ \ldots \circ (S_{w,1}^{-1} \circ C_{w,1}) \right) (IF(z; T_0, P)) \right\|_{\mathcal{H}} \\ &= \left\| \left((S_{w,k}^{-1} \circ C_{w,k}) \circ \ldots \circ (S_{w,k_0+1}^{-1} \circ C_{w,k_0+1}) \right) \right. \\ &\qquad \left. \left(\left((S_{w,k_0}^{-1} \circ C_{w,k_0}) \circ \ldots \circ (S_{w,1}^{-1} \circ C_{w,1}) \right) (IF(z; T_0, P)) \right) \right\|_{\mathcal{H}} \\ &\leq (1 - \epsilon)^{k-k_0} \left\| \left(\left((S_{w,k_0}^{-1} \circ C_{w,k_0}) \circ \ldots \circ (S_{w,1}^{-1} \circ C_{w,1}) \right) (IF(z; T_0, P)) \right) \right\|_{\mathcal{H}} \end{aligned}$$

Thus the second term in (10) vanishes as $k \to \infty$ and the right hand side converges to

$$\sum_{j=0}^{\infty} \left(S_{w,\infty}^{-1} \circ C_{w,\infty} \right)^j (g_\infty) = \left(\mathrm{id}_{\mathcal{H}} - S_{w,\infty}^{-1} \circ C_{w,\infty} \right)^{-1} (g_\infty)$$

with $\mathrm{id}_{\mathcal{H}}$ the identity operator. This yields the following theorem.

Theorem 4 *Denote by T_{k+1} the map $T_{k+1}(P) = f_{P,\lambda}^{(k+1)}$. Furthermore, let P be a distribution on $\mathcal{X} \times \mathcal{Y}$ with $|P|_2 < \infty$ and $\int_{\mathcal{X}\times\mathcal{Y}} w(y - f_{P,\lambda}^{(\infty)}(x)) dP(x, y) > 0$. Denote by T_∞ the map $T_\infty(P) = f_{P,\lambda}^{(\infty)}$. Denote the operators $S_{w,\infty} : \mathcal{H} \to \mathcal{H}$ and $C_{w,\infty} : \mathcal{H} \to \mathcal{H}$ given by*

$$S_{w,\infty}(f) = \lambda f + \mathbb{E}_P[w(Y - f_{P,\lambda}^{(\infty)}(X))\langle f, \Phi(X)\rangle \Phi(X)]$$

and

$$C_{w,\infty}(f) = -\mathbb{E}_P[w'(Y - f_{P,\lambda}^{(\infty)}(X))(Y - f_{P,\lambda}^{(\infty)}(X))\langle f, \Phi(X)\rangle \Phi(X)].$$

Assume that $||S_{w,\infty}^{-1} \circ C_{w,\infty}|| < 1$. *Then the influence function of* T_∞ *exists for all* $z := (z_x, z_y) \in \mathcal{X} \times \mathcal{Y}$ *and we have*

$$\begin{aligned} IF(z; T_\infty, P) =& (S_{w,\infty} - C_{w,\infty})^{-1} \Big(-(\mathbb{E}_P w(Y - f_{P,\lambda}^{(\infty)}(X))(Y - f_{P,\lambda}^{(\infty)}(X))\Phi(X)) \\ & + w(z_y - f_{P,\lambda}^{(\infty)}(z_x))(z_y - f_{P,\lambda}^{(\infty)}(z_x))\Phi(z_x) \Big) . \end{aligned}$$

A first important conclusion concerns the boundedness of this expression. Since the operators $S_{w,\infty}$ and $C_{w,\infty}$ are independent of the contamination z, the influence function $IF(z; T_\infty, P)$ is bounded if (recall that $\|\Phi(x)\|_{\mathcal{H}}^2 = \langle \Phi(x), \Phi(x) \rangle = K(x,x)$)

$$\|w(r) r \Phi(x)\|_{\mathcal{H}} = w(r)|r|\sqrt{K(x,x)} \text{ is bounded } \forall (x,r) \in \mathbb{R}^d \times \mathbb{R}. \tag{11}$$

Note that for any $f \in \mathcal{H} : \|f\|_\infty \leq \|f\|_{\mathcal{H}} \|K\|_\infty$. Therefore $\|IF(z; T_\infty, P)\|_{\mathcal{H}}$ bounded immediately implies $\|IF(z; T_\infty, P)\|_\infty$ bounded for bounded kernels.

If we take Φ the canonical feature map of a linear kernel, (11) corresponds to the conditions obtained by Dollinger and Staudte [1991] for ordinary linear least squares regression. In that case, the weight function should decrease with the residual r as well as with x to obtain a bounded influence. This is also true for other unbounded kernels, e.g. polynomial, but not for non-linear function estimation using a bounded kernel, like the popular RBF kernel for instance. The latter only requires downweighting the residual, as the influence in x-space is controlled by the kernel. This shows that LS-KBR with a bounded kernel is much more suited for iterative reweighting than linear least squares regression (similar conclusions concerning robustness and the boundedness of the kernel were obtained in Theorem 4 in Christmann and Steinwart [2004] for classification and Corollary 19 in Christmann and Steinwart [2007] for regression).

Let us now restrict ourselves to a weight function of the form

$$w(r) = \frac{\psi(r)}{r} \text{ with } \psi : \mathbb{R} \to \mathbb{R} \text{ a bounded, real, odd function.}$$

From Theorem 4 we know that this is sufficient to bound the influence function of iteratively reweighted LS-KBR with a bounded kernel, if convergence takes place, that is if $\|S_{w,\infty}^{-1} \circ C_{w,\infty}\| < 1$.

Note that if $\psi(r) = L'(r)/2$ for some convex loss function L, then

$$S_{w,\infty}(f) = \lambda f + \mathbb{E}_P \left[\frac{L'(Y - f_{P,\lambda}^{(\infty)}(X))}{Y - f_{P,\lambda}^{(\infty)}(X)} \langle f, \Phi(X) \rangle \Phi(X) \right]$$

and

$$C_{w,\infty}(f) = -\mathbb{E}_P \left[\left(L''(Y - f_{P,\lambda}^{(\infty)}(X)) - \frac{L'(Y - f_{P,\lambda}^{(\infty)}(X))}{Y - f_{P,\lambda}^{(\infty)}(X)} \right) \langle f, \Phi(X) \rangle \Phi(X) \right] .$$

Thus in that case $S_{w,\infty} - C_{w,\infty} = \frac{1}{2} S$ with the operator S as in Proposition 2. In that case the influence function of T_∞ equals the expression in Proposition 2. This is of

course no surprise since in Theorem 2 we have proven that T_∞ with weights $L'(r)/(2r)$ corresponds to KBR with loss function L. Consequently their influence functions should coincide as well.

3.4 On the condition for convergence

In the previous section we showed that it is important that $||S_{w,\infty}^{-1} \circ C_{w,\infty}|| < 1$ to ensure that the influence of the initial estimator ultimately disappears. In this section we further examine this condition in two specific settings. We assume that the distribution P follows a classical regression setting. This means that (i) a function $f_P \in \mathcal{H}$ exists such that the conditional mean $\mathbb{E}_P(Y|x)$ of the response Y given $x \in \mathbb{R}^d$ equals $f_P(x)$, (ii) the error $e = Y - f(X)$ is independent of X and (iii) the distribution P_e of these errors is symmetric about 0 with finite second moment.

3.4.1 Case 1: $\lambda = 0$

Note that in practical data analysis of finite samples one often needs $\lambda > 0$, such that the restriction $\lambda = 0$ might seem very restrictive. Nevertheless it is known that the optimal λ then depends on the sample size, and that a larger sample size requires lower λ. Actually $\lambda \to 0$ is required for consistency of the estimator if the sample size goes to ∞. Thus the case $\lambda = 0$ is still very interesting from an asymptotic point of view.

For distributions P satisfying $(i) - (iii)$, it is easy to see that LS-KBR with $\lambda = 0$ is Fisher consistent, meaning that $f_{P,0} = f_P$ with (see equation (4))

$$f_{P,0} = \arg\min_{f \in \mathcal{H}} \mathbb{E}_P (Y - f(X))^2.$$

Moreover reweighted LS-KBR is also Fisher consistent (see appendix for proof):

$$f_{P,0}^{(k+1)} = \arg\min_{f \in \mathcal{H}} \mathbb{E}_P \left[w(X, Y - f_{P,0}^{(k)}(X))(Y - f(X))^2 \right] = f_P \tag{12}$$

for every $k \in \mathbb{N}$. From its definition in Theorem 4, we know that

$$S_{w,k} = \mathbb{E}_P[w(X, Y - f_P(X))\langle ., \Phi(X)\rangle \Phi(X)].$$

for every $k \in \mathbb{N}$. Since this expression is independent of k we denote this operator by S_w in this section (similarly for C_w). Using the assumed regression structure of P in $(i) - (iii)$, we can decompose P in the error distribution P_e of the errors $e = Y - f_p(X)$ and the distribution P_X of X such that $dP = dP_X dP_e$. This yields

$$S_w = \mathbb{E}_P[\frac{\psi(e)}{e}\langle ., \Phi(X)\rangle \Phi(X)].$$

Defining $d := \mathbb{E}_{P_e} \frac{\psi(e)}{e}$ we have that

$$S_w = d\ \mathbb{E}_{P_X}[\langle ., \Phi(X)\rangle \Phi(X)].$$

Note that d always exists since we assumed errors with finite second moment. Some analogous calculations give a similar result for the operator C_w.

$$C_w = c\, \mathbb{E}_{P_X} \langle ., \Phi(X) \rangle \Phi(X) \text{ with } c := d - E_{P_e} \psi'(e).$$

Thus, denoting $\mathrm{id}_{\mathcal{H}}$ the identity operator in $\mathcal{H}$ such that $\mathrm{id}_{\mathcal{H}}(f) = f$ for all $f \in \mathcal{H}$, we obtain

$$S_w^{-1} \circ C_w = \frac{c}{d}\, \mathrm{id}_{\mathcal{H}} \tag{13}$$

showing that the condition $\|S_w^{-1} \circ C_w\| < 1$ is satisfied if $c < d$, meaning that

$$\mathbb{E}_{P_e} \psi'(e) > 0.$$

Since this condition depends on the error distribution P_e, a stronger but distribution free assumption might be useful, for example taking ψ a strictly increasing function.

Summarizing the previous results we can state: in case of distributions P with a regression structure as defined in the beginning of this section, the influence function of iteratively reweighted LS-KBR with bounded kernel, $\lambda = 0$ and weight function $w(r) = \frac{\psi(r)}{r}$ converges to a bounded function if

$$\begin{array}{ll} (c1) & \psi : \mathbb{R} \to \mathbb{R} \text{ is a measurable, real, odd function.} \\ (c2) & \psi \text{ is continuous and differentiable.} \\ (c3) & \psi \text{ is bounded.} \\ (c4) & \mathbb{E}_{P_e} \psi'(e) > 0. \qquad (c4') \;\; \psi \text{ is strictly increasing.} \end{array} \tag{14}$$

When using unbounded kernels such as linear or polynomial, this is not sufficient. As such the behavior of these reweighted estimators can differ according to the RKHS considered.

3.4.2 Case 2: $f_P = 0$

Consider a distribution P satisfying $(i) - (iii)$ and $f_P \equiv 0$. For such distributions, we have that $f_{P,\lambda}^{(k)} = f_P$ for every k. In this case one can prove (see the appendix) that the operator norm of $S_w^{-1} \circ C_w$ equals

$$\|S_w^{-1} \circ C_w\| = \frac{c}{d + \lambda} \tag{15}$$

where c and d are the same constants as in (13). Since λ is positive, we see that this norm is smaller than 1 if $c < d$, which is exactly the condition found in the case $\lambda = 0$. Now we observe that taking $\lambda > 0$ only relaxes this condition, at least to $c \leq d$. We can thus relax condition $(c4)$ from (14) as well.

$$\begin{array}{ll} (c4) & E_{P_e} \psi'(e) > -\lambda \qquad (c4'') \;\; \psi \text{ is increasing.} \end{array} \tag{16}$$

We conclude that a positive generalization parameter λ improves the convergence of iteratively reweighted LS-KBR. This is plausible, since higher values of λ will lead to smoother fits. Then the method will be less attracted towards an outlier in y-direction, indeed leading to better robustness.

4 Stability

Several measures of stability were recently proposed in the literature. The leave-one-out error often plays a vital role, for example in hypothesis stability [Bousquet and Elisseeff, 2001], partial stability [Kutin and Niyogi, 2002] and CV$_{loo}$-stability [Poggio et al., 2004]. The basic idea is that the result of a learning map T on a full sample should not be very different from the result obtained when removing only one observation. More precisely, denote P_n the empirical distribution associated with a sample $S = \{z_j = (x_j, y_j) \in \mathbb{R}^d \times \mathbb{R},\ j = 1, \ldots, n\}$ of size n, then one can consider

$$D_i = |L(y_i, T(P_n)(x_i)) - L(y_i, T(P_n^i)(x_i))|$$

with P_n^i the empirical distribution of the sample S without the ith observation z_i. Poggio et al. [2004] call the map T CV$_{loo}$-stable if

$$\sup_{i=1,\ldots,n} D_i \to 0 \tag{17}$$

for $n \to \infty$. They show under mild conditions that CV$_{loo}$-stability is required to achieve generalization.

The influence function actually measures something very similar. Recall that this function is defined as

$$IF(z; T, P) = \lim_{\epsilon \downarrow 0} \frac{T(P_{\epsilon,z}) - T(P)}{\epsilon}.$$

It measures how the result of a learning map changes as the original distribution P is changed by adding a small amount of contamination at the point z. In robust statistics it is important to bound the influence function over *all possible points* z in the support of P. This is a major difference with stability, where the supremum is taken over n points *sampled i.i.d. from the distribution* P (as in (17)).

This however suggests a possible approach to analyze stability using the influence function: by evaluating it at n sample points only. For an easy heuristic argument, take $z = z_i$, $P = P_n^i$ and $\epsilon = 1/n$ in the definition of the influence function above. Then for large n we have that

$$IF(z_i; T, P) \approx \frac{T(P_n) - T(P_n^i)}{1/n}.$$

Then it is easy to see that

$$|L(y_i, T(P_n)(x_i) - L(y_i, T(P_n^i)(x_i))| \approx |L'(y_i, T(P)(x_i))| \frac{|IF(z_i; T, P)|}{n}. \tag{18}$$

	Influence function	Leave-one-out
Robustness	$\sup_z \lvert IF(z;T,P)\rvert$ bounded	$\sup_i\{\sup_z D_i^z\} \to 0$
	$\Downarrow$	$\Downarrow$
Stability	$\sup_{z_i} \lvert IF(z_i;T,P)\rvert/n \to 0$	$\sup_i D_i \to 0$

Table 1: Overview of some robustness and stability concepts

As such the influence function can be used in a first order approximation to the quantity D_i which is so important in the concept of CV_{loo}-stability. The influence function *evaluated at the sample point* z_i should be small for every i in order to obtain stability. From equation (18) one might define a new stability criterion, in the spirit of (17) but based on the influence function, as follows:

$$\sup_{i\in\{1,\ldots,n\}} \frac{|IF(z_i;T,P)|}{n} \to 0. \tag{19}$$

If a method is robust, then its influence function is bounded over *all possible points* z in the support of P and thus (19) is obviously satisfied. As such robustness is in a sense a strictly stronger requirement than stability. Robustness can be interpreted as adding *any* point, even points that are very unlikely under the sampling distribution P.

Consider for example unweighted KBR. Recall from Proposition 2 that for any $z = (z_x, z_y)$

$$IF(z;T,P) = S^{-1}\left(\mathbb{E}_P[L'(Y, f_{P,\lambda}(X))\Phi(X)]\right) - L'(z_y, f_{P,\lambda}(z_x))S^{-1}\Phi(z_x).$$

If the first derivative of the loss function L is bounded, this influence function is bounded as well and KBR is then automatically stable as well. For KBR with a squared loss, the influence function is unbounded. Despite this lack of robustness, LS-KBR is stable as long as the distribution P is not too heavy tailed. For example in case of a signal plus noise distribution with Gaussian distributed noise, $\sup_{i=1,\ldots,n}(y_i - T(P)(x_i))$ converges to ∞ as n grows larger. For Gaussian distributed noise however, this convergence will only be at logarithmic speed. Thus the convergence of (19) is of the order $O(\frac{\log(n)}{n})$ and (19) obviously still holds. For a more heavy tailed noise distribution on the other hand, the rate of stability might be much slower than $O(\frac{\log(n)}{n})$.

Since reweighted LS-KBR has a bounded influence function, its rate of stability is always $O(\frac{1}{n})$. Reweighting steps are thus not only helpful when outliers are present in the data. They also lead to a more stable method, especially at heavy tailed distributions.

Table 1 links some of these concepts from robustness and stability. The influence function originated in robust statistics as a tool to assess the robustness of statistical methods (upper left cell of Table 1). The leave-one-out error on the other hand is often used in statistical learning to assess the stability of a learning map (lower right cell

of Table 1). In equation (19) we combined both ideas using the influence function to assess stability (lower left cell of the table). In order to complete the table, the question raises whether a leave-one-out criterion can be formulated to assess robustness. Define $P_n^{z,i}$ the sample P_n where the point z_i is replaced by z and

$$D_i^z = |L(y_i, T(P_n^{z,i})(x_i)) - L(y_i, T(P_n^i)(x_i))|.$$

Then of course $D_i^{z_i} = D_i$, since taking $z = z_i$ returns the original sample P_n. Thus CV_{loo} stability (17) can be written as

$$\sup_{i=1,\ldots,n} D_i^{z_i} \to 0.$$

Now since robustness is concerned with the effect of adding any point z, not only sample points, a possible definition of robustness is

$$\sup_{i=1,\ldots,n} \{\sup_z D_i^z\} \to 0.$$

This could be a sample counterpart for the classical approach of 'bounding the influence function' in robust statistics, completing Table 1 with the upper right cell.

Since we showed that reweighting steps bound the influence function of LS-KBR, it is to be expected that the stability is improved as well. Of course feature research is needed to make these heuristics mathematically rigorous. Replacing P by P_n for instance immediately induces concerns on consistency. Also note that further exploration of these links might be useful in other applications, for example in model selection [Debruyne et al., 2008].

5 Examples

5.1 Weight functions

Many weight functions have been described in the literature, especially for linear regression. We show three of them in Table 2, with corresponding functions $w(r)$, $\psi(r)$ and loss function $L(r)$. Note that only the logistic weight function satisfies all conditions $(c1) - (c4)$ in (14). Huber's weight function [Huber, 1981] does not satisfy $(c4)$ as ψ is not strictly increasing. Simpson et al. [1992] show that this can lead to unstable behavior of M-estimators in linear models. It does however satisfy condition $(c4'')$ in (16). The third weight function in Table 2 is Hampel's [Hampel et al., 1986] suggestion for linear least squares. These weights were also used in the context of least squares support vector regression by Suykens et al. [2002a]. In this case ψ satisfies condition $(c4')$ nor $(c4'')$, but condition $(c4)$ is valid for common error distributions, i.e. normally distributed errors. Also note that the resulting loss function is not convex anymore for these Hampel weights. Although this still leads to satisfactory results in

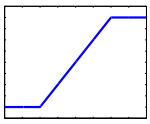
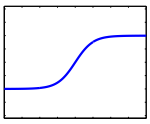
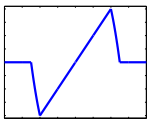
	Huber	logistic	Hampel
$w(r)$	$\begin{array}{ll} 1 & \text{if } \lvert r\rvert < \beta \\ \frac{\beta}{\lvert r\rvert} & \text{if } \lvert r\rvert \geq \beta \end{array}$	$\frac{\tanh(r)}{r}$	$\begin{array}{ll} 1 & \text{if } \lvert r\rvert < b_1 \\ \frac{b_2 - \lvert r\rvert}{b_2 - b_1} & \text{if } b_1 \leq \lvert r\rvert \leq b_2 \\ 0 & \text{if } \lvert r\rvert > b_2 \end{array}$
$\psi(r)$			
$L(r)$	$\begin{array}{ll} r^2 & \text{if } \lvert r\rvert < \beta \\ \beta\lvert r\rvert & \text{if } \lvert r\rvert \geq \beta \end{array}$	$r\tanh(r)$	$\begin{array}{ll} r^2 & \text{if } \lvert r\rvert < b_1 \\ \frac{b_2 r^2 - \lvert r^3\rvert}{b_2 - b_1} & \text{if } b_1 \leq \lvert r\rvert \leq b_2 \\ 0 & \text{if } \lvert r\rvert > b_2 \end{array}$

Table 2: Definitions for Huber, logistic and Hampel weight functions. Only the logistic weight function satisfies all conditions (c1)-(c4).

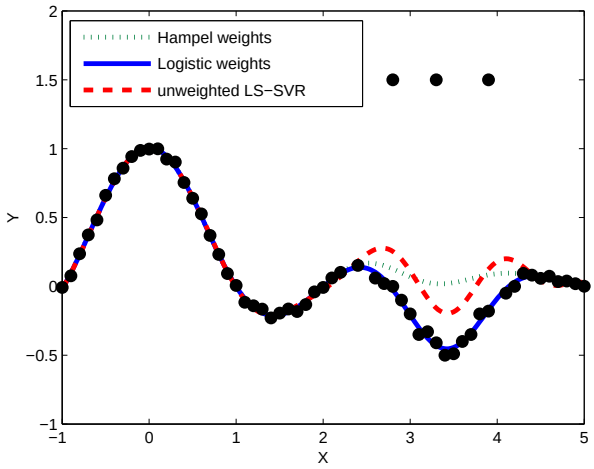


Figure 1: Simulated data example. Dashed curve: original LS-KBR. Dotted curve: reweighted LS-KBR using Hampel weights. Solid curve: reweighted LS-KBR using logistic weights.

many examples, bad fits may occur occasionally. In Figure 1 some data points were simulated including three outliers. Ordinary LS-KBR (dashed curve) is clearly affected by the outlying observations. Reweighting using a logistic weight function (solid curve) improves the fit remarkably well. Using Hampel's weight function (dotted curve) however does not improve the original estimate in this example. In that case all points in the region $x \in [2.5, 4.2]$ receive a weight exactly equal to zero. Thus, locally the outliers do not have a smaller weight than the neighboring "good" data. With logistic weights, all these good data points with $x \in [2.5, 4.2]$ receive a small weight as well, but the outliers get an even smaller weight. Therefore they are also locally recognized as outliers

		$N(0,1)$			t_5			Cauchy		
		c	d	$\frac{c}{d}$	c	d	$\frac{c}{d}$	c	d	$\frac{c}{d}$
Huber	$\beta = 0.5$	0.32	0.71	**0.46**	0.31	0.67	**0.46**	0.26	0.55	**0.47**
	$\beta = 1$	0.22	0.91	**0.25**	0.23	0.87	**0.27**	0.22	0.72	**0.31**
	$\beta = 1.5$	0.11	0.97	**0.11**	0.14	0.94	**0.15**	0.18	0.80	**0.22**
	$\beta = 2$	0.04	0.99	**0.04**	0.08	0.98	**0.08**	0.14	0.85	**0.17**
Logistic		0.22	0.82	**0.26**	0.22	0.79	**0.28**	0.21	0.66	**0.32**

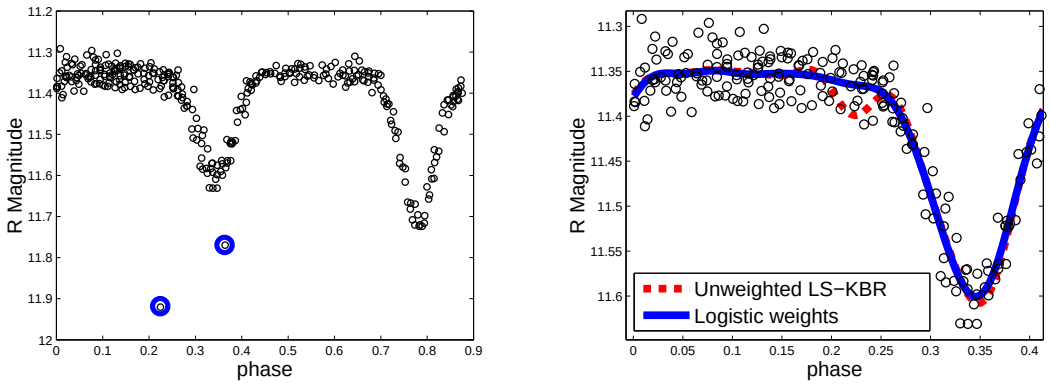
Table 3: Values of the constants c, d and $\frac{c}{d}$ for the Huber weight function with cutoff $\beta = 0.5, 1, 1.5, 2$ and for the logistic weight function, at a standard normal distribution, Student distribution with 5 degrees of freedom, and a Cauchy distribution. The values of c/d (bold) represent an upper bound for the reduction of the influence function at each step.

and thus wLS-KBR with logistic weights performs a lot better in this example. This example clearly shows that it is not trivial to choose a good weight function. Moreover it shows that breaking conditions $(c_1) - (c_4)$ can lead to bad results, also in cases not satisfying the assumptions under which these conditions were derived, since here $\lambda > 0$ and $f_P \neq 0$.

5.2 Convergence

In equations (13) and (15), an upper bound is established on the reduction of the influence function at each step. In Table 3 we calculated this upper bound at a normal distribution, a Student distribution with five degrees of freedom and at a Cauchy distribution. We compare Huber's weight function with several cutoff values β, as well as logistic weights. Note that the convergence of the influence functions is pretty fast, even at heavy tailed distributions such as the Cauchy. For Huber weights, the convergence rate decreases rapidly as β increases. This is quite expected, since the larger β is, the less points are downweighted. Also note that the upper bound on the convergence rate approaches 1 as β goes to 0. The Huber loss function converges to an L_1 loss as β convergence to 0. Thus when reweighting LS-KBR to obtain L_1-KBR no fast convergence is guaranteed by our results, since the upper bound on the reduction factor approaches 1. When β is exactly 0, no results can be given at all, because then the ψ function is discontinuous.

Logistic weights are doing quite well. Even at heavy tailed noise distributions such as a Cauchy, the influence function is reduced to 0.32 of the value at the previous step. This means for example that after k steps, at most 0.32^k is left of the influence of the initial estimator, so fast convergence can be expected.



(a) (b)

Figure 2: Star data. (a) Brightness (expressed in stellar magnitude R) of the binary star versus the phase (with a period of 0.8764 days). The two outliers in the data are circled. (b) Plot of the fit in the region: phase $\in [0, 0.4]$. Initial LS-KBR fit (dashed line), wLS-KBR with logistic weights and one reweighting step (solid line). The fit after four reweighting steps is practically coinciding with the solid line.

5.3 Star data

Variable stars are stars whose brightness periodically changes over time. Such a variable star was analyzed in Oh et al. [2004]. A plot of its brightness versus its phase (with period 0.8764, as found in Oh et al. [2004]) is shown in Figure 2(a). It concerns an eclipsing binary star, with both stars orbiting each other in the plane of the earth. Therefore, if one member of the pair eclipses the other, the combined brightness decreases. This explains the two peaks that are clearly present in the picture. Our goal is now to estimate the light curve, i.e. the functional relationship between brightness and phase, which is useful for classification of stars. In this case for example, the light curve is flat in between two peaks. This feature is associated with the detached type of eclipsing stars.

From Figure 2(a) it is obvious that two outliers are part of the data. When using classical LS-KBR to fit the light curve, these two data points have quite an impact on the result. In Figure 2(b) (dashed line) the LS-KBR fit shows an extra bump at phases in $[0.15, 0.25]$. The solid line represents the one step reweighted LS-KBR with the logistic weight function. The effect of the outliers is severely reduced, leading to quite a nice fit. The two step reweighted LS-KBR is plotted as well (dotted line), but the difference with the one step reweighting is practically invisible. After six steps, all residuals were the same as after five steps up to 0.001, showing the fast convergence properties of weighted LS-KBR.

5.4 Artificial data

This part presents the results of a small simulation study. We consider three well known settings.

- Sinc curve ($d = 1$): $y(x) = \sin(x)/x$.
- Friedman 1 ($d = 10$): $y(x) = 10\sin(\pi x_1 x_2) + 20(x_3 - 1/2)^2 + 10x_4 + 5x_5$
- Friedman 2 ($d = 4$): $y(x) = (x_1^2 + (x_2 x_3 - 1/x_2 x_4)^2))^{1/2}$.

In each replication 100 data points were generated. For the sinc curve, the inputs were taken uniformly on $[-5, 5]$. For the Friedman data sets [Friedman, 1991] inputs were generated uniformly from the unit hypercube. Noise was added to $y(x)$ from two distributions: first, Gaussian with unit variance and second, Student with 2 degrees of freedom.

For each data set, unweighted LS-KBR with RBF kernel was performed. The hyperparameters λ and σ were obtained by 10-fold cross validation using the Mean Squared Error (MSE) as cost criterion. Reweighted LS-KBR with RBF kernel and logistic weights was performed as well, using the same hyperparameters as found in the unweighted case. To compare both methods, the MSE was calculated over 200 noisefree test points. This procedure was repeated in 100 replications. Figure 3 shows boxplots of these 100 MSE's for the six cases.

First consider the left panel of Figure 3 containing the results with Gaussian noise. In that case the difference between reweighting or not is rather small. For Friedman 1, the median MSE is slightly smaller in the case of reweighting, whereas the sinc curve and Friedman 2 give slightly bigger median MSE's.

At the right panel of Figure 3 boxplots are shown for Student distributed noise. In that case reweighting clearly offers an improvement of the results. Not only is the median MSE smaller in all three settings. Also the right skewness of the MSE's clearly diminishes after reweighting, indicating that the method is more stable. This is exactly what we concluded in our theoretical analysis from Section 4, where it was demonstrated that reweighting improves stability at heavy tailed distributions.

Here we see in practice that reweighting leads to improved results at heavy tailed error distributions but retains the quality of unweighted LS-KBR at others such as the Gaussian distribution. Also note that we kept the hyperparameters fixed at their optimal value in the unweighted case, since we also treat the hyperparameters fixed in our theoretical results. Nevertheless, re-optimizing them at each reweighting step might possibly lead to even better results.

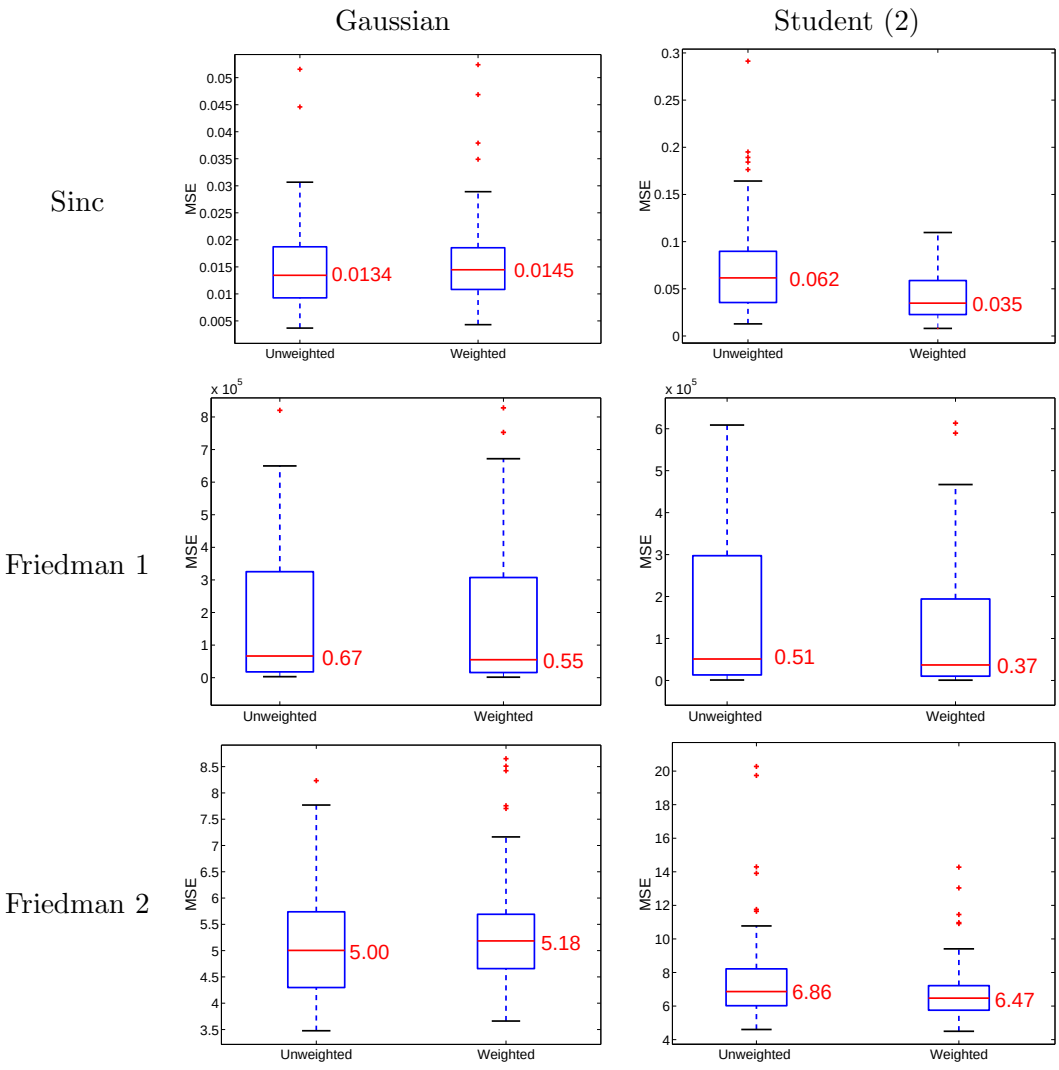


Figure 3: Simulation results for three data sets (sinc, Friedman 1 and Friedman 2). On the left: Gaussian errors. On the right: Student with 2 degrees of freedom. Each time boxplots of 100 MSE's are shown for unweighted LS-KBR and reweighted LS-KBR with logistic weights. For Gaussian errors no clear winner can be seen between unweighted versus reweighted. For Student errors reweighting leads to improvement.

6 Conclusion

We defined a sequence of reweighted LS-KBR estimators. It was shown that this sequence converges if the weight function is non-increasing. As a consequence, any KBR estimator with an invariant convex loss function can be attained as the limit of a sequence of reweighted LS-KBR estimators. We analyzed the series of influence functions of reweighted LS-KBR. A condition on an operator norm was found to guarantee convergence to a bounded influence function if the kernel is bounded, irrespective of the initial estimator. This means for example that reweighted LS-KBR using a RBF-kernel is a robust estimator, even if the initial estimator is obtained by ordinary (non-robust) LS-KBR. In two special cases the mathematical condition on an operator norm was shown to be satisfied if the weight function is chosen as $w(r) = \psi(r)/r$ with ψ satisfying conditions $(c1) - (c4)$. A simple example was given to show that violating these conditions can lead to bad results in practice, even in situations that are not part of the two special cases considered. Therefore we recommend to choose a weight function that satisfies $(c1) - (c4)$ if possible, e.g. logistic weights. Finally we showed that reweighting does not only improve robustness against outliers or gross errors. It also improves the stability of LS-KBR, especially at heavy-tailed distributions.

Acknowledgements: JS acknowledges support from K.U. Leuven, GOA-Ambiorics, CoE EF/05/006, FWO G.0499.04, FWO G.0211.05, FWO G.0302.07, IUAP P5/22. The authors are grateful to the anonymous referees for their useful comments and suggestions leading to an improved version of this paper.

Appendix

Remark on Proposition 1 for least squares

As in the empirical case (3), it is also possible to include an intercept term $b_{P,\lambda} \in \mathbb{R}$ in the theoretical expressions, next to the functional part $f_{P,\lambda}$. For any distribution P and $\lambda > 0$ we denote

$$T(P) = (f_{P,\lambda}, b_{P,\lambda}) \in \mathcal{H} \times \mathbb{R}$$

minimizing the regularized risk:

$$(f_{P,\lambda}, b_{P,\lambda}) = \min_{(f,b)\in\mathcal{H}\times\mathbb{R}} \left(\mathbb{E}_P L(Y, f(X) + b) + \lambda ||f||^2_{\mathcal{H}} \right).$$

The solution of this minimization problem is characterized in DeVito et al. [2004] (main theorem pp. 1369). If the loss function L is the least squares loss function, then this

theorem provides us the following equations:

$$f_{P,\lambda} = \frac{1}{\lambda}\mathbb{E}_P[(Y - f_{P,\lambda}(X) - b_{P,\lambda})\Phi(X)] \tag{20}$$

$$b_{P,\lambda} = \mathbb{E}_P(Y - f_{P,\lambda}(X)). \tag{21}$$

Now we consider the contaminated distribution $P_{\epsilon,z} = (1-\epsilon)P + \epsilon\Delta_z$ with Δ_z a Dirac distribution with all probability mass located at the point z. Then by definition the influence function of the intercept term at $z \in \mathcal{X} \times \mathcal{Y}$ equals

$$IF(z; b, P) = \lim_{\epsilon \downarrow 0} \frac{b_{P_{\epsilon,z},\lambda} - b_{P,\lambda}}{\epsilon}.$$

Using equation (21) for both $b_{P_{\epsilon,z},\lambda}$ and $b_{P,\lambda}$ yields

$$\begin{aligned} IF(z; b, P) &= \lim_{\epsilon \downarrow 0} \frac{\mathbb{E}_{P_{\epsilon,z}}(Y - f_{P_{\epsilon,z},\lambda}(X)) - \mathbb{E}_P(Y - f_{P,\lambda}(X))}{\epsilon} \\ &= \lim_{\epsilon \downarrow 0} \frac{(1-\epsilon)\mathbb{E}_P(Y - f_{P_{\epsilon,z},\lambda}(X)) + \epsilon(z_y - f_{P_{\epsilon,z},\lambda}(z_x)) - \mathbb{E}_P(Y - f_{P,\lambda}(X))}{\epsilon}. \end{aligned}$$

Rearranging terms in the nominator we have

$$\begin{aligned} IF(z; b, P) &= \lim_{\epsilon \downarrow 0} \frac{\mathbb{E}_P(Y - f_{P_{\epsilon,z},\lambda}(X)) - \mathbb{E}_P(Y - f_{P,\lambda}(X))}{\epsilon} \\ &\quad - \lim_{\epsilon \downarrow 0} \frac{\epsilon\mathbb{E}_P(Y - f_{P_{\epsilon,z},\lambda}(X)) + \epsilon(z_y - f_{P_{\epsilon,z},\lambda}(z_x))}{\epsilon} \\ &= \lim_{\epsilon \downarrow 0} \frac{\mathbb{E}_P(f_{P,\lambda}(X) - f_{P_{\epsilon,z},\lambda}(X))}{\epsilon} - \mathbb{E}_P(Y - f_{P,\lambda}(X)) + (z_y - f_{P,\lambda}(z_x)). \end{aligned}$$

Thus for the intercept term we obtain the following expression.

$$IF(z; b, P) = -\mathbb{E}_P IF(z; f, P) - \mathbb{E}_P(Y - f_{P,\lambda}(X)) + (z_y - f_{P,\lambda}(z_x)). \tag{22}$$

For $f_{P,\lambda}$ we have

$$IF(z; b, P) = \lim_{\epsilon \downarrow 0} \frac{f_{P_{\epsilon,z},\lambda} - f_{P,\lambda}}{\epsilon}.$$

Plugging in equation (20) for both $f_{P_{\epsilon,z},\lambda}$ and $f_{P,\lambda}$, it is clear that

$$\begin{aligned} &\lambda IF(z; f, P) + \mathbb{E}_P[IF(z; f, P)(X)\Phi(X)] + \mathbb{E}_P[IF(z; b, P)(X)\Phi(X)] \\ &\quad = -\mathbb{E}_P(Y - f_{P,\lambda}(X) - b_{P,\lambda})\Phi(X) + (z_y - f_{P,\lambda}(z_x) - b_{P,\lambda})\Phi(z_x). \end{aligned} \tag{23}$$

Thus, combining (22) and (23) in matrix notation, we have

$$\begin{aligned} &\begin{pmatrix} \lambda \mathrm{id}_{\mathcal{H}} + \mathbb{E}_P[\langle ., \Phi(X)\rangle\Phi(X)] & \mathbb{E}_P\Phi(X) \\ \mathbb{E}_P\langle ., \Phi(X)\rangle & 1 \end{pmatrix} \begin{pmatrix} IF(z; f, P) \\ IF(z; b, P) \end{pmatrix} \\ &= \begin{pmatrix} -\mathbb{E}_P[(Y - f_{P,\lambda}(X) - b_{P,\lambda})\Phi(X)] + (z_y - f_{P,\lambda}(z_x) - b_{P,\lambda})\Phi(z_x) \\ -\mathbb{E}_P(Y - f_{P,\lambda}(X) - b_{P,\lambda}) + (z_y - f_{P,\lambda}(z_x) - b_{P,\lambda}) \end{pmatrix}. \end{aligned} \tag{24}$$

When not considering the intercept term, the previous expression indeed corresponds to the one already obtained by Christmann and Steinwart [2007]. Also note the similarities to the results obtained in classification [Christmann and Steinwart, 2004]. However, since this intercept term is not essential in explaining the robustness principles of kernel based regression, we will not include it anymore furtheron.

Proof of Theorem 1

Let P be a distribution on $\mathcal{X} \times \mathcal{Y}$ with $|P|_2 < \infty$ and define $\xi = \int_{\mathcal{X}\times\mathcal{Y}} w(y - f_{P,\lambda}^{(k)}(x))dP(x,y)$. Assume $\xi > 0$. Then we can define a distribution P_w by $dP_w(x,y) = \xi^{-1}w(y - f_{P,\lambda}^{(k)}(x))dP(x,y)$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$. Since $\xi > 0$ and w is continuous, P_w is well defined and one can easily see that $f_{P,\lambda}^{(k+1)} = f_{P_w,\lambda/\xi}$. Moreover, since w is nonincreasing, $|P_w|_2 < \infty$ if $|P|_2 < \infty$ and Proposition 1 yields

$$f_{P,\lambda}^{(k+1)} = f_{P_w,\lambda/\xi} = -\frac{\xi}{2\lambda}\mathbb{E}_{P_w}h\Phi = -\frac{1}{2\lambda}\mathbb{E}_P w(Y - f_{P,\lambda}^{(k)}(X))h\Phi.$$

For the least squares loss function we obtain

$$f_{P,\lambda}^{(k+1)} = \frac{1}{\lambda}\mathbb{E}_P w(Y - f_{P,\lambda}^{(k)}(X))(Y - f_{P,\lambda}^{(k+1)}(X))\Phi(X).$$

If $\xi = 0$ then $\mathbb{E}_P[w(Y - f_{P,\lambda}^{(k)}(X))(Y - f(X))^2] = 0$. Thus $f_{P,\lambda}^{(k+1)} = \arg\min_{f\in\mathcal{H}} \lambda||f||_{\mathcal{H}} = 0$. But because $\xi = 0$ we also have that $\frac{1}{2\lambda}\mathbb{E}_P w(Y - f_{P,\lambda}^{(k)}(X))h\Phi = 0$ and therefore Theorem 1 still holds.

Proof of Theorem 2

Define V a real function such that $V'(r) = rw(r)$. Because of condition (w_3) it holds that $V'(r) \geq 0$ if $r > 0$ and $V'(r) \leq 0$ if $r < 0$. Thus V is bounded from below by $V(0)$. Define $g(r^2) = V(r)$, thus $w(r) = 2g'(r^2)$. According to condition $(w3)$ the weights are nonincreasing which implies that the function g is a concave function. Denote

$$R_{P,\lambda,V}(f) = \mathbb{E}_P V(Y - f(x)) + \lambda||f||_{\mathcal{H}}^2.$$

Because of the concavity of g we have that $g(u) - g(v) \leq (u - v)g'(v)$ for all $u, v \in \mathbb{R}$. Thus

$$\begin{aligned}
&R_{P,\lambda,V}(f_{P,\lambda}^{(k+1)}) - R_{P,\lambda,V}(f_{P,\lambda}^{(k)}) \\
&\leq \mathbb{E}_P g'((Y - f_{P,\lambda}^{(k)}(X))^2)\left((Y - f_{P,\lambda}^{(k+1)}(X))^2 - (Y - f_{P,\lambda}^{(k)}(X))^2\right) + \lambda||f_{P,\lambda}^{(k+1)}||_{\mathcal{H}}^2 - \lambda||f_{P,\lambda}^{(k)}||_{\mathcal{H}}^2 \\
&= \frac{1}{2}\mathbb{E}_P w(Y - f_{P,\lambda}^{(k)}(X))\left(f_{P,\lambda}^{(k)}(X) - f_{P,\lambda}^{(k+1)}(X)\right)\left(2Y - f_{P,\lambda}^{(k+1)}(X) - f_{P,\lambda}^{(k)}(X)\right) \\
&\qquad + \lambda||f_{P,\lambda}^{(k+1)}||_{\mathcal{H}}^2 - \lambda||f_{P,\lambda}^{(k)}||_{\mathcal{H}}^2.
\end{aligned}$$

Using the notation in Theorem 1 we can replace Y by $h^{(k+1)}(X,Y) + f_{P,\lambda}^{(k+1)}$.

$$
\begin{aligned}
&R_{P,\lambda,V}(f_{P,\lambda}^{(k+1)}) - R_{P,\lambda,V}(f_{P,\lambda}^{(k)}) \\
&\leq \frac{1}{2}\mathbb{E}_P w(Y - f_{P,\lambda}^{(k)}(X)) \left(f_{P,\lambda}^{(k)}(X) - f_{P,\lambda}^{(k+1)}(X)\right) \left(f_{P,\lambda}^{(k+1)}(X) - f_{P,\lambda}^{(k)}(X)\right) \\
&\qquad + \mathbb{E}_P w(Y - f_{P,\lambda}^{(k)}(X))(f_{P,\lambda}^{(k)}(X) - f_{P,\lambda}^{(k+1)}(X))2h^{k+1}(X,Y) + \lambda||f_{P,\lambda}^{(k+1)}||_{\mathcal{H}}^2 - \lambda||f_{P,\lambda}^{(k)}||_{\mathcal{H}}^2
\end{aligned}
$$

Due to the reproducing property

$$
f_{P,\lambda}^{(k)}(X) - f_{P,\lambda}^{(k+1)}(X) = \langle f_{P,\lambda}^{(k)} - f_{P,\lambda}^{(k+1)}, \Phi(X)\rangle.
$$

Thus

$$
\begin{aligned}
&\mathbb{E}_P w(Y - f_{P,\lambda}^{(k)}(X))(f_{P,\lambda}^{(k)}(X) - f_{P,\lambda}^{(k+1)}(X))2h^{k+1}(X,Y) \\
&\qquad = \langle f_{P,\lambda}^{(k)} - f_{P,\lambda}^{(k+1)}, 2\mathbb{E}_P w(Y - f_{P,\lambda}^{(k)}(X))h^{k+1}(X,Y)\Phi(X)\rangle \\
&\qquad = \langle f_{P,\lambda}^{(k)} - f_{P,\lambda}^{(k+1)}, 2\lambda f_{P,\lambda}^{(k+1)}\rangle \\
&\qquad = 2\lambda\langle f_{P,\lambda}^{(k)}, f_{P,\lambda}^{(k+1)}\rangle - 2\lambda||f_{P,\lambda}^{(k+1)}||_{\mathcal{H}}^2.
\end{aligned}
$$

Consequently

$$
\begin{aligned}
&R_{P,\lambda,V}(f_{P,\lambda}^{(k+1)}) - R_{P,\lambda,V}(f_{P,\lambda}^{(k)}) \\
&\leq -\frac{1}{2}\mathbb{E}_P w(Y - f_{P,\lambda}^{(k)}(X)) \left(f_{P,\lambda}^{(k)}(X) - f_{P,\lambda}^{(k+1)}(X)\right)^2 - \lambda||f_{P,\lambda}^{(k)} - f_{P,\lambda}^{(k+1)}||_{\mathcal{H}}^2.
\end{aligned}
$$

The function $R_{P,\lambda,V}$ decreases in every step with at least $\lambda||f_{P,\lambda}^{(k)} - f_{P,\lambda}^{(k+1)}||_{\mathcal{H}}^2$. Since $R_{P,\lambda V}$ is bounded from below by $V(0)$, this implies that the sequence $\{f_{P,\lambda}^{(k)}\}$ must converge.

Proof of Theorem 3

We can use the representation from Theorem 1 to calculate the influence function in a point $z \in \mathcal{X} \times \mathcal{Y}$

$$
\begin{aligned}
IF(z; T_{k+1}, P) &= \frac{\partial}{\partial\epsilon} T_{k+1}(P_{\epsilon,z})|_{\epsilon=0} \\
&= \frac{1}{\lambda}\frac{\partial}{\partial\epsilon}\mathbb{E}_{P_{\epsilon,z}} w(Y - f_{P_{\epsilon,z},\lambda}^{(k)}(X))(Y - f_{P_{\epsilon,z},\lambda}^{(k+1)}(X))\Phi(X)|_{\epsilon=0} \\
&= -\frac{1}{\lambda}\mathbb{E}_P w(Y - f_{P,\lambda}^{(k)}(X))(Y - f_{P,\lambda}^{(k+1)}(X))\Phi(X) \\
&\quad + \frac{1}{\lambda} w(z_y - f_{P,\lambda}^{(k)}(z_x))(z_y - f_{P,\lambda}^{(k+1)}(z_x))\Phi(z_x) \\
&\quad + \frac{1}{\lambda}\frac{\partial}{\partial\epsilon}\mathbb{E}_P[w(Y - f_{P_{\epsilon,z},\lambda}^{(k)}(X))(Y - f_{P_{\epsilon,z},\lambda}^{(k+1)}(X))\Phi(X)]|_{\epsilon=0}.
\end{aligned}
$$

The last term equals

$$
\begin{aligned}
&-\frac{1}{\lambda}\mathbb{E}_P[IF(z; T_k, P)w'(Y - f_{P,\lambda}^{(k)}(X))(Y - f_{P,\lambda}^{(k+1)}(X))] \\
&\qquad\qquad + \frac{1}{\lambda}\mathbb{E}_P[w(Y - f_{P,\lambda}^{(k)}(X))IF(z; T_{k+1}, P)].
\end{aligned}
$$

Thus defining $S_{w,k}$ and $C_{w,k}$ as in Theorem 3, we have

$$S_{w,k}(IF(z;T_{k+1},P)) = \mathbb{E}_P w(Y - f_{P,\lambda}^{(k)}(X))(Y - f_{P,\lambda}^{(k+1)}(X))\Phi(X) \\ + C_{w,k}(IF(z;T_k,P)) - w(z_y - f_{P,\lambda}^{(k)}(z_x))(z_y - f_{P,\lambda}^{(k+1)}(z_x))\Phi(z_x).$$

Now it suffices to show that $S_{w,k}$ is invertible. In Christmann and Steinwart [2007] this was already proven for the operator S as defined in Proposition 2. However, we can again consider the distribution P_w such that $dP_w(x,y) = \xi^{-1} w(y - f_{P,\lambda}^{(k)}(x))dP(x,y)$ for all $(x,y) \in \mathbb{R}^d \times \mathbb{R}$, with $\xi = \int_{\mathcal{X}\times\mathcal{Y}} w(y - f_{P,\lambda}^{(k)}(x))dP(x,y) > 0$. Then the operator $S_{w,k}$ using distribution P and regularization parameter λ is equivalent to the operator S using the distribution P_w and regularization parameter λ/ξ. Thus, using Christmann and Steinwart [2007] (more specific their proof of Theorem 18), we see that $S_{w,k}$ is invertible.

Proof of equation (12)

Suppose that $k-$step reweighting is Fisher consistent, thus $f_{P,0}^{(k)} = f_P$. Denote P_w the distribution such that $dP_w(x,y) = w(y - f_P(X))dP(x,y)$ for any $(x,y) \in \mathcal{X} \times \mathcal{Y}$. Then

$$f_{P,0}^{(k+1)} = \arg\min_{f\in\mathcal{H}} \mathbb{E}_P \left[w(Y - f_P(x))(Y - f(X))^2 \right] = \arg\min_{f\in\mathcal{H}} \mathbb{E}_{P_w}(Y - f(X))^2 = f_P$$

since unweighted least squares KBR is Fisher consistent if $\lambda = 0$.

Proof of equation (15)

We need two propositions from operator theory in Hilbert spaces.

Proposition 3 *(Spectral Theorem)*
Let T be a compact and self-adjoint operator on a Hilbert space $\mathcal{H}$. Then $\mathcal{H}$ has an orthonormal basis (e_n) consisting of eigenvectors for T. If $\mathcal{H}$ is infinite dimensional, the corresponding eigenvalues (different from 0) (γ_n) can be arranged in a decreasing sequence $|\gamma_1| \geq |\gamma_2| \geq \ldots$ where $\gamma_n \to 0$ for $n \to \infty$, and for $x \in \mathcal{H}$

$$T(x) = \sum_n \gamma_n \langle x, e_n \rangle e_n.$$

Proposition 4 *(Fredholm Alternative)*
Let T be a compact and self-adjoint operator on a Hilbert space $\mathcal{H}$, and consider the equation

$$(T - \gamma\ id_{\mathcal{H}})x = y.$$

If γ is not an eigenvalue of T, then the equation has a unique solution $x = (T - \gamma id_{\mathcal{H}})^{-1} y$.

Recall that we assumed that the distribution P could be decomposed in an error distribution P_e and a distribution in x-space P_x such that $dP = dP_e dP_x$. Using this regression structure of P we can write

$$S_w = \lambda \, \mathrm{id}_{\mathcal{H}} + E_{P_e} \frac{\psi(e)}{e} E_{P_X} \langle ., \Phi(X) \rangle \Phi(X).$$

Denote $T = E_{P_X} \langle ., \Phi(X) \rangle \Phi(X)$, then

$$S_w = \lambda \, \mathrm{id}_{\mathcal{H}} + d \, T$$

with the constant $d = E_{P_e} \frac{\psi(e)}{e}$. In the same way we find

$$C_w = cT$$

with $c = d - E_{P_e} \psi'(e)$.
Now we know T is compact (proven in Christmann and Steinwart [2007]) and self-adjoint. Moreover, T is positive and thus its eigenvalues are positive. As such, $-\frac{\lambda}{d}$ cannot be an eigenvalue, and by the Fredholm alternative, $T - (-\frac{\lambda}{d})\mathrm{id}_{\mathcal{H}}$ is invertible. Thus for any $g \in \mathcal{H}$ the equation

$$\left(T - (-\frac{\lambda}{d})\mathrm{id}_{\mathcal{H}} \right)(f) = \frac{c}{d} T(g)$$

has a unique solution in terms of $f \in \mathcal{H}$. Moreover, from the spectral theorem we know that T has an orthonormal basis (f_i) with corresponding eigenvalues λ_i and we can write our equation as

$$\left(T - (-\frac{\lambda}{d})\mathrm{id}_{\mathcal{H}} \right)(f) = \sum_{i=1}^{\infty} (\lambda_i + \frac{\lambda}{d}) \langle f, f_i \rangle f_i = \sum_{i=1}^{\infty} \frac{c}{d} \lambda_i \langle g, f_i \rangle f_i.$$

Thus we see that

$$\langle f, f_i \rangle = (\lambda_i + \frac{\lambda}{d})^{-1} \lambda_i \frac{c}{d} \langle g, f_i \rangle$$

and so we find

$$(S_w^{-1} \circ C_w)(g) = f = \sum_{i=1}^{\infty} (\lambda_i + \frac{\lambda}{d})^{-1} \lambda_i \frac{c}{d} \langle g, f_i \rangle f_i.$$

Since the operator norm of a compact operator equals the supremum of its eigenvalues, we have that

$$\| S_w^{-1} \circ C_w \| = \sup_i \frac{\lambda_i \frac{c}{d}}{\lambda_i + \frac{\lambda}{d}} = \frac{c}{d} \frac{1}{1 + \frac{\lambda}{d}},$$

proving equation (15). Since $c = d - E_{P_e} \psi'(e)$,

$$\frac{c}{d} \frac{1}{1 + \frac{\lambda}{d}} < 1 \Leftrightarrow 1 - \frac{E_{P_e} \psi'(e)}{d} < 1 + \frac{\lambda}{d}$$

or

$$E_{P_e} \psi'(e) > -\lambda.$$

References

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2001.

A. Christmann and I. Steinwart. Consistency and robustness of kernel based regression. *Bernoulli*, 13:799–819, 2007.

A. Christmann and I. Steinwart. On robust properties of convex risk minimization methods for pattern recognition. *Journal of Machine Learning Research*, 5:1007–1034, 2004.

M. Debruyne, M. Hubert, and J.A.K. Suykens. Model selection in kernel based regression using the influence function. *Journal of Machine Learning Research*, 9: 2377–2400, 2008.

E. DeVito, L. Rosasco, A. Caponnetto, M. Piana, and A. Verri. Some properties of regularized kernel methods. *Journal of Machine Learning Research*, 5:1363–1390, 2004.

M. B. Dollinger and R. G. Staudte. Influence functions of iteratively reweighted least squares estimators. *Journal of the American Statistical Association*, 86:709–716, 1991.

T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13:1–50, 2000.

J.H. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19:1–14, 1991.

F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10:1455–1480, 1998.

F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York, 1986.

P.J. Huber. *Robust Statistics*. Wiley, New York, 1981.

S. Kutin and P. Niyogi. Almost everywhere algorithmic stability and generalization error. In A. Daruich and N. Friedman, editors, *Proceedings of Uncertainty in AI*. Morgan Kaufmann, Edmonton, 2002.

H.S. Oh, D. Nychka, T. Brown, and P. Charbonneau. Period analysis of variable stars by robust smoothing. *Journal of the Royal Statistical Society C (Applied Statistics)*, 53:15–30, 2004.

T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi. General conditions for predictivity in learning theory. *Nature*, 428:419–422, 2004.

B. Schölkopf, R. Herbrich, and A. Smola. A generalized representer theorem. In D. Helmblod and B. Williamson, editors, *Neural Networks and Computational Learning Theory*, pages 416–426, Berlin, 2001. Springer.

D.G. Simpson, D. Ruppert, and R.J. Carroll. On one-step GM-estimates and stability of inferences in linear regression. *Journal of the American Statistical Association*, 87: 439–450, 1992.

I. Steinwart. Sparseness of support vector machines. *Journal of Machine Learning Research*, 4:1071–1105, 2003.

I. Steinwart and A. Christmann. *Support Vector Machines.* Springer, New York, 2008.

J.A.K. Suykens, J. De Brabanter, L. Lukas, and J. Vandewalle. Weighted least squares support vector machines : Robustness and sparse approximation. *Neurocomputing*, 48:85–105, 2002a.

J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002b.

A.N. Tikhonov and V.Y. Arsenin. *Solutions of Ill Posed Problems.* W.H. Winston, Washington D.C., 1977.

V. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag, New-York, 1995.

G. Wahba. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 69–88, Cambridge, MA, 1999. MIT Press.

G. Wahba. Spline models for observational data. Series in applied mathematics, 59, SIAM. Philadelphia, 1990.