Submitted 5/07; Revised 1/08; Published 5/08

Bouligand Derivatives and Robustness of Support Vector Machines for Regression

Andreas Christmann ANDREAS.CHRISTMANN@UNI-BAYREUTH.DE
Department of Mathematics
University of Bayreuth
D-95440 Bayreuth, GERMANY

Arnout Van Messem AVMESSEM@VUB.AC.BE
Department of Mathematics
Vrije Universiteit Brussel
B-1050 Brussels, BELGIUM

Editor: Peter Bartlett

Abstract

We investigate robustness properties for a broad class of support vector machines with non-smooth loss functions. These kernel methods are inspired by convex risk minimization in infinite dimensional Hilbert spaces. Leading examples are the support vector machine based on the $\underline{\varepsilon}$-insensitive loss function, and kernel based quantile regression based on the pinball loss function. Firstly, we propose with the Bouligand influence function (BIF) a modification of F.R. Hampel's influence function. The BIF has the advantage of being positive homogeneous which is in general not true for Hampel's influence function. Secondly, we show that many support vector machines based on a Lipschitz continuous loss function and a bounded kernel have a bounded BIF and are thus robust in the sense of robust statistics based on influence functions.

Keywords: Bouligand derivatives, empirical risk minimization, influence function, robustness, support vector machines

1. Introduction

The goal in non-parametric regression is to estimate a functional relationship between an $\mathbb{R}^d$-valued input random variable X and an $\mathbb{R}$-valued output random variable Y, under the assumption that the joint distribution P of (X,Y) is (almost) completely unknown. In order to model this relationship one typically assumes that one has a training data set $D_{train} = \big((x_1,y_1),\dots,(x_n,y_n)\big)$ from independent and identically distributed (i.i.d.) random variables (X_i,Y_i), $i = 1,\dots,n$, which all have the distribution P. Informally, the aim is to build a predictor $f : \mathbb{R}^d \to \mathbb{R}$ based on these observations such that $f(X)$ is a good approximation of Y. To formalize this aim one uses a continuous *loss function* $L : Y \times \mathbb{R} \to [0,\infty)$ that assesses the quality of a prediction $f(x)$ for an observed output y by $L(y, f(x))$. We follow the convention that the smaller $L(y, f(x))$ is, the better the prediction is. The quality of a predictor f is measured by the *L-risk* $\mathcal{R}_{L,\mathrm{P}}(f) := \mathbb{E}_{\mathrm{P}}L(Y, f(X))$ which of course is unknown, because P is unknown. One tries to find a predictor whose risk is close to the minimal risk, that is to the Bayes risk $\mathcal{R}^*_{L,\mathrm{P}} := \inf\{\mathcal{R}_{L,\mathrm{P}}(f);\ f : \mathbb{R}^d \to \mathbb{R} \text{ measurable}\}$. One way to build a non-parametric predictor f is to use a support vector machine (SVM) which finds a minimizer $f_{\mathrm{P},\lambda}$

of the regularized risk

$$\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f) := \mathbb{E}_{\mathrm{P}} L(Y, f(X)) + \lambda \|f\|^2_{\mathcal{H}} , \tag{1}$$

where $\lambda > 0$ is a regularization parameter to reduce the danger of overfitting, $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) of a measurable kernel $k : X \times X \to \mathbb{R}$, and L is a measurable, *convex* loss function in the sense that $L(y,\cdot) : \mathbb{R} \to [0,\infty)$ is convex for all $y \in Y$, see Vapnik (1998) and Schölkopf and Smola (2002). Since (1) is strictly convex in f, the minimizer $f_{\mathrm{P},\lambda}$ is unique if it exists. We denote the canonical feature map by $\Phi : \mathcal{H} \to \mathcal{H}$, $\Phi(x) := k(\cdot,x)$. The reproducing property gives $f(x) = \langle f, \Phi(x)\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$ and $x \in X$. A kernel k is bounded, if $\|k\|_\infty := \sup\{\sqrt{k(x,x)} : x \in X\} < \infty$. Using the reproducing property and $\|\Phi(x)\|_{\mathcal{H}} = \sqrt{k(x,x)}$, one obtains the well-known inequalities

$$\|f\|_\infty \leq \|k\|_\infty \|f\|_{\mathcal{H}} \quad \text{and} \quad \|\Phi(x)\|_\infty \leq \|k\|_\infty \|\Phi(x)\|_{\mathcal{H}} \leq \|k\|^2_\infty \tag{2}$$

for $f \in \mathcal{H}$ and $x \in X$. The Gaussian radial basis function kernel defined by $k_{RBF}(x,x') = \exp(-\|x-x'\|^2/\gamma^2)$, $\gamma > 0$, is bounded and universal on every compact subset of $\mathbb{R}^d$ (Steinwart, 2001) which partially explains its popularity. The corresponding RKHS of this kernel has infinite dimension. Of course, $\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f)$ is not computable, because P is unknown. However, the empirical distribution $\mathrm{D} = \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_i)}$ corresponding to the training data set D_{train} can be used as an estimator of P. Here $\delta_{(x_i,y_i)}$ denotes the Dirac distribution in (x_i,y_i). If we replace P by D in (1), we obtain the regularized empirical risk

$$\mathcal{R}^{reg}_{L,\mathrm{D},\lambda}(f) := \mathbb{E}_{\mathrm{D}} L\big(Y, f(X)\big) + \lambda \|f\|^2_{\mathcal{H}} .$$

An empirical SVM f_{D,λ_n} with $\lambda_n > 0$ and $\lambda_n \to 0$ if $n \to \infty$, is called L-risk consistent if $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n}) \to \mathcal{R}^*_{L,\mathrm{P}}$ in probability for $n \to \infty$.

Traditionally, research in nonparametric regression is often based on the least squares loss $L_{LS}(y,t) := (y-t)^2$. The least squares loss function is convex in t, is useful to estimate the conditional mean function, and is advantageous from a numerical point of view, but L_{LS} is not Lipschitz continuous. From a practical point of view there are situations in which a different loss function is more appropriate. (i) In some situations one is actually not interested in modeling the conditional mean, but in fitting a conditional quantile function instead. For this purpose the convex pinball loss function $L_{\tau-pin}(y,t) := (\tau-1)(y-t)$, if $y-t<0$, and $L_{\tau-pin}(y,t) := \tau(y-t)$, if $y-t \geq 0$, is used, where $\tau \in (0,1)$ specifies the desired conditional quantile, see Koenker and Bassett (1978) and Koenker (2005) for parametric quantile regression and Takeuchi et al. (2006) for nonparametric quantile regression. (ii) If the goal is to estimate the conditional median function, then the $\underline{\varepsilon}$-insensitive loss given by $L_{\underline{\varepsilon}}(y,t) := \max\{|y-t|-\underline{\varepsilon}, 0\}$, $\underline{\varepsilon} \in (0,\infty)$, promises algorithmic advantages in terms of sparseness compared to the L1-loss function $L_{L1}(y,t) = |y-t|$, see Vapnik (1998) and Schölkopf and Smola (2002). (iii) If the regular conditional distribution of Y given $X = x$ is known to be symmetric, basically all invariant loss functions of the form $L(y,t) = \psi(r)$ with $r = y-t$, where $\psi : \mathbb{R} \to [0,\infty)$ is convex, symmetric and has its only minimum at 0, can be used to estimate the conditional mean, see Steinwart (2007). In this case a less steep loss function such as the Lipschitz continuous Huber loss function given by $L_{c-Huber}(y,t) := \psi(r) = r^2/2$, if $|r| \leq c$, and $\psi(r) = c|r| - c^2/2$, if $|r| > c$ for some $c \in (0,\infty)$, may be more suitable if one fears outliers in y-direction, see Huber (1964) and Christmann and Steinwart (2007).

The deeper reason to consider Lipschitz continuous loss functions is the following. One strong argument in favor of SVMs is that they are L-risk consistent under weak assumptions, that is SVMs are able to “learn”, but it is also important to investigate the robustness properties for such statistical learning methods. In almost all cases statistical models are only approximations to the true random process which generated a given data set. Hence the natural question arises what impact such deviations may have on the results. J.W. Tukey, one of the pioneers of robust statistics, mentioned already in 1960 (Hampel et al., 1986, p. 21): *“A tacit hope in ignoring deviations from ideal models was that they would not matter; that statistical procedures which were optimal under the strict model would still be approximately optimal under the approximate model. Unfortunately, it turned out that this hope was often drastically wrong; even mild deviations often have much larger effects than were anticipated by most statisticians.”*

Let us consider $T(\mathrm{P}) := \mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f)$, with P a probability measure, as a mapping $T : \mathrm{P} \mapsto \mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f)$. In robust statistics we are interested in smooth and bounded functions T, because this will give stable regularized risks within small neighborhoods of P. If an appropriate derivative $\nabla T(\mathrm{P})$ of $T(\mathrm{P})$ is bounded, then the function $T(\mathrm{P})$ cannot increase or decrease unlimited in small neighborhoods of P. Several notions of differentiability have been used for this purpose.

Let us therefore take a look at the following results from Averbukh and Smolyanov (1967, 1968), Fernholz (1983) and Rieder (1994) on various notions of differentiation to clarify the connections between these notions. For every pair of normed real vector spaces (X,Y) let a subset $\mathcal{S}(X,Y)$ of the functions from X to Y be given. The following conditions are imposed on this system $\mathcal{S}$, which will provide the (Landau) o remainder of the first-order Taylor approximation of an $\mathcal{S}$-differentiation: (i) $\rho(0) = 0$, $\rho \in \mathcal{S}(X,Y)$, (ii) $\mathcal{S}(X,Y)$ is a real vector subspace of all functions from X to Y, (iii) $\mathcal{S}(X,Y) \cap \mathcal{L}(X,Y) = \{0\}$ where $\mathcal{L}(X,Y)$ is the space of continuous linear mappings from X to Y, and 0 stands for the zero operator and (iv) moreover, in case $X = \mathbb{R}$, it is required that $\mathcal{S}(\mathbb{R},Y) = \{\rho : \mathbb{R} \to Y | \lim_{t\to 0} \rho(t)/t = 0\}$. If $\mathcal{S}$ fulfills (i) to (iv), then some mapping $T : X \to Y$ is called *$\mathcal{S}$-differentiable* at x if there exists some $A \in \mathcal{L}(X,Y)$ and $\rho \in \mathcal{S}(X,Y)$ such that for all $h \in X$, $T(x+h) = T(x) + Ah + \rho(h)$. The continuous linear mapping $\nabla^{\mathcal{S}} T(x) = A$ is called *$\mathcal{S}$-derivative* of T at x. The set of all functions $T : X \to Y$ which are $\mathcal{S}$-differentiable at x is denoted by $\mathcal{D}_{\mathcal{S}}(X,Y;x)$. From conditions (ii) and (iii) it is seen that the $\mathcal{S}$-derivative $\nabla^{\mathcal{S}} T(x)$ is uniquely defined. Condition (iv) ensures that $\mathcal{S}$-differentiability in case $X = \mathbb{R}$ coincides with the usual notion of differentiability. The function $T \mapsto \nabla^{\mathcal{S}} T(x)$ is a linear mapping from $\mathcal{D}_{\mathcal{S}}(X,Y;x)$ to $\mathcal{L}(X,Y)$.

$\mathcal{S}$-differentiations may be constructed in a special way by means of coverings $\mathcal{C}$, whose elements are naturally assumed to be bounded sets C (so that $th \to 0$ uniformly for $h \in C$ as $t \to 0$). For every normed real vector space X let a covering $\mathcal{C}_X$ of X be given which consists of bounded subsets of X. If Y is another normed real vector space, define $\mathcal{S}_{\mathcal{C}}(X,Y) = \{\rho : X \to Y \,|\, \lim_{t\to 0} \sup_{h\in C} \frac{\|\rho(th)\|}{t} = \rho(0) = 0\ \forall C \in \mathcal{C}_X\}$. Then the class $\mathcal{S}_{\mathcal{C}}$ satisfies the conditions (i) to (iv). With X ranging through all normed real vector spaces, we can then define the following concepts of differentiation by varying the covering $\mathcal{C}_X$. *Gâteaux*-differentiation is defined by the choices $\mathcal{C}_{GX} = \{C \subset X | C \text{ finite}\}$. For *Hadamard*-differentiation, $\mathcal{C}_{HX} = \{C \subset X | C \text{ compact}\}$ and *Fréchet*-differentiation uses the covering $\mathcal{C}_{FX} = \{C \subset X | C \text{ bounded}\}$. The three differentiations will be indicated by the corresponding authors’ initials. From these definitions it is clear that ∇^F implies ∇^H which implies ∇^G. It can be shown that ∇^H is actually the weakest $\mathcal{S}$-derivative which fulfills the chain rule.

One general approach to robustness (Hampel, 1968, 1974) is the one based on influence functions which are related to Gâteaux-derivatives. Let $\mathcal{M}_1$ be the set of probability distributions on some measurable space $(Z, \mathcal{B}(Z))$ and let $\mathcal{H}$ be a reproducing kernel Hilbert space. The influence function (IF) of $T : \mathcal{M}_1 \to \mathcal{H}$ at a point $z \in Z$ for a distribution P is defined as

$$\mathrm{IF}(z;T,\mathrm{P}) = \lim_{\varepsilon \downarrow 0} \frac{T((1-\varepsilon)\mathrm{P} + \varepsilon\delta_z) - T(\mathrm{P})}{\varepsilon}, \tag{3}$$

if the limit exists. Within this approach robust estimators are those which have a bounded influence function.[1] The influence function is neither supposed to be linear nor continuous. If the influence functions exists for all points $z \in Z$ and if it is continuous and linear, then the IF is a special Gâteaux-derivative.

Christmann and Steinwart (2004, 2007) and Steinwart and Christmann (2008b) showed that SVMs have a bounded influence function in binary classification and in regression problems provided that the kernel is bounded and continuous, L is twice Fréchet-differentiable w.r.t. the second argument, and the first and second F-derivative of L is bounded. Hence Lipschitz continuous loss functions are of special interest from a robustness point of view. An example of a loss function with these properties is the logistic loss given by $L_{log}(y,t) := -\log\bigl(4\Lambda(y-t)(1-\Lambda(y-t))\bigr)$, $y,t \in \mathbb{R}$, where $\Lambda(y-t) = 1/\bigl(1+e^{-(y-t)}\bigr)$. However the important special cases L_ε, $L_{\tau-pin}$, and $L_{c-Huber}$ are excluded in these results, because these loss functions are not everywhere Fréchet-differentiable.

The present paper tries to fill this gap: we will propose in Definition 1 an alternative to the influence function. This alternative is based on Bouligand-derivatives whereas Hampel's influence function was defined having Gâteaux-derivatives in mind. The second goal of this paper is to use this new notion of robustness to show that SVMs for regression are robust in this sense even if the loss function has no Fréchet-derivative.

Let us now recall some facts on Bouligand-derivatives and strong approximation of functions. For the rest of the introduction let X, Y, W, and Z be normed linear spaces, and we consider neighborhoods $\mathcal{N}(x_0)$ of x_0 in X, $\mathcal{N}(y_0)$ of y_0 in Y, and $\mathcal{N}(w_0)$ of w_0 in W. Let F and G be functions from $\mathcal{N}(x_0) \times \mathcal{N}(y_0)$ to Z, h_1 and h_2 functions from $\mathcal{N}(w_0)$ to Z, f a function from $\mathcal{N}(x_0)$ to Z and g a function from $\mathcal{N}(y_0)$ to Z. A function f *approximates* F in x at (x_0,y_0), written as $f \sim_x F$ at (x_0,y_0), if $F(x,y_0) - f(x) = o(x-x_0)$. Similarly, $g \sim_y F$ at (x_0,y_0) if $F(x_0,y) - g(y) = o(y-y_0)$. A function h_1 *strongly approximates* h_2 at w_0, written as $h_1 \approx h_2$ at w_0, if for each $\varepsilon > 0$ there exists a neighborhood $\mathcal{N}(w_0)$ of w_0 such that whenever w and w' belong to $\mathcal{N}(w_0)$, $\bigl\|\bigl(h_1(w) - h_2(w)\bigr) - \bigl(h_1(w') - h_2(w')\bigr)\bigr\| \le \varepsilon\|w - w'\|$. A function f *strongly approximates* F in x at (x_0,y_0), written as $f \approx_x F$ at (x_0,y_0), if for each $\varepsilon > 0$ there exist neighborhoods $\mathcal{N}(x_0)$ of x_0 and $\mathcal{N}(y_0)$ of y_0 such that whenever x and x' belong to $\mathcal{N}(x_0)$ and y belongs to $\mathcal{N}(y_0)$ we have $\bigl\|\bigl(F(x,y) - f(x)\bigr) - \bigl(F(x',y) - f(x')\bigr)\bigr\| \le \varepsilon\|x - x'\|$. Strong approximation amounts to requiring $h_1 - h_2$ to have a strong Fréchet-derivative of 0 at w_0, though neither h_1 nor h_2 is assumed to be differentiable in any sense. A similar definition is made for strong approximation in y. We define strong approximation for functions of several groups of variables, for example $G \approx_{(x,y)} F$ at (x_0,y_0), by replacing W by $X \times Y$ and making the obvious substitutions. Note that one has both $f \approx_x F$ and $g \approx_y F$ at (x_0,y_0) exactly if $f(x) + g(y) \approx_{(x,y)} F$ at (x_0,y_0).

Recall that a function $f : X \to Z$ is called *positive homogeneous* if

$$f(\alpha x) = \alpha f(x) \quad \forall \alpha \ge 0, \forall x \in X.$$

Following Robinson (1987) we can now define the *Bouligand-derivative*. Given a function f from an open subset $\mathcal{X}$ of a normed linear space X into another normed linear space Z, we say that

[1] In the following we use the term "robust" in this sense, unless otherwise stated.

f is *Bouligand-differentiable* at a point $x_0 \in \mathcal{X}$, if there exists a positive homogeneous function $\nabla^B f(x_0) : \mathcal{X} \to Z$ such that

$$f(x_0 + h) = f(x_0) + \nabla^B f(x_0)(h) + o(h). \tag{4}$$

We can write (4) also as

$$\lim_{h \to 0} \left\| f(x_0 + h) - f(x_0) - \nabla^B f(x_0)(h) \right\|_Z / \|h\|_X = 0. \tag{5}$$

Let $F : X \times Y \to Z$, and suppose that F has a partial B-derivative[2] $\nabla_1^B F(x_0, y_0)$ with respect to x at (x_0, y_0). We say $\nabla_1^B F(x_0, y_0)$ is *strong* if $F(x_0, y_0) + \nabla_1^B F(x_0, y_0)(x - x_0) \approx_x F$ at (x_0, y_0). Robinson (1987) showed that the chain rule holds for Bouligand-derivatives. Let f be a Lipschitzian function from an open set $\Omega \subset \mathbb{R}^m$ to $\mathbb{R}^k$, $x_0 \in \Omega$, and f B-differentiable at x_0. Let g be a Lipschitzian function from an open set $\Gamma \subset \mathbb{R}^k$, with $f(x_0) \in \Gamma$, to $\mathbb{R}^l$ be B-differentiable at $f(x_0)$. Then $g \circ f$ is B-differentiable at x_0 and $\nabla^B (g \circ f)(x_0) = \nabla^B g\big(f(x_0)\big) \circ \nabla^B f(x_0)$. The fact that B-derivatives, just as F- and H-derivatives, fulfill the chain rule is no contradiction to the before mentioned fact that H-differentiability is the *weakest $\mathcal{S}$-differentiation* which fulfills the chain rule (Rieder, 1994, p. 4) because the B-derivative is not necessarily a continuous linear function.

In general Gâteaux- and Bouligand-differentiability are not directly comparable, because B-derivatives are by definition positive homogeneous, but not necessarily linear. We will show that the existence of the BIF implies the existence of the IF and that in that case BIF=IF. Please note that this in general does not imply that the IF is a Gâteaux-derivative.

In this paper, we will prove that many SVMs based on Lipschitz continuous loss functions have a bounded Bouligand influence function. To formulate our results we will use Bouligand-derivatives in the sense of Robinson (1991) as defined above. These directional derivatives were to our best knowledge not used in robust statistics so far, but are successfully applied in approximation theory for non-smooth functions. Section 2 covers our definition of the Bouligand influence function (BIF) and contains the main result which gives the BIF for support vector machines based on a bounded kernel and a B-differentiable Lipschitz continuous convex loss function. In Section 3 it is shown that this result covers the loss functions $L_{\underline{\varepsilon}}$, $L_{\tau-pin}$, $L_{c-Huber}$, and L_{log} as special cases. Section 4 contains the conclusions. All proofs are given in the Appendix.

2. Main Result

This section contains our two main results: the definition of the Bouligand influence function and a theorem which shows that a broad class of support vector machines based on a Lipschitz continuous, but not necessarily Fréchet-differentiable loss function have a bounded Bouligand influence function. We denote the set of all probability distributions on some measurable space $(Z, \mathcal{B}(Z))$ by $\mathcal{M}_1$ and let $\mathcal{H}$ be a Hilbert space.

Definition 1 *The* ***Bouligand influence function (BIF)*** *of the function* $T : \mathcal{M}_1 \to \mathcal{H}$ *for a distribution* P *in the direction of a distribution* $\mathrm{Q} \neq \mathrm{P}$ *is the special Bouligand-derivative (if it exists)*

$$\lim_{\varepsilon \downarrow 0} \frac{\left\| T\big((1-\varepsilon)\mathrm{P} + \varepsilon \mathrm{Q}\big) - T(\mathrm{P}) - \mathrm{BIF}(\mathrm{Q}; T, \mathrm{P}) \right\|_{\mathcal{H}}}{\varepsilon} = 0. \tag{6}$$

[2] Throughout the paper we will denote partial B-derivatives of f by $\nabla_1^B f$, $\nabla_2^B f$, $\nabla_{2,2}^B f := \nabla_2^B(\nabla_2^B f)$ etc.

The BIF has the interpretation that it measures the impact of an infinitesimal small amount of contamination of the original distribution P in the direction of Q on the quantity of interest $T(\mathrm{P})$. It is thus desirable that the function T has a *bounded* BIF.

Note that (6) is indeed a special B-derivative, because we consider the directions $h = \varepsilon(\mathrm{Q}-\mathrm{P})$ and $x_0 = \mathrm{P}$. If Q equals the Dirac distribution δ_z in a point $z \in Z$, that is $\delta_z(\{z\}) = 1$, we write $\mathrm{BIF}(z;T,\mathrm{P})$. The choice of the metric on $\mathcal{M}_1$ is not important for the definition of the BIF, because $\|\varepsilon(\mathrm{Q}-\mathrm{P})\| = \varepsilon\,\|\mathrm{Q}-\mathrm{P}\|$ and $\|\mathrm{Q}-\mathrm{P}\|$ is a positive constant. For the norm of total variation we obtain for example,

$$\lim_{\varepsilon(\mathrm{Q}-\mathrm{P})\downarrow 0} \frac{\left\| T\big(\mathrm{P}+\varepsilon(\mathrm{Q}-\mathrm{P})\big) - T(\mathrm{P}) - \mathrm{BIF}(\mathrm{Q};T,\mathrm{P})\right\|_{\mathcal{H}}}{\|\varepsilon(\mathrm{Q}-\mathrm{P})\|_{\mathrm{tv}}} = 0,$$

(cf., Equation 5). Since $\varepsilon(\mathrm{Q}-\mathrm{P}) \to 0$ iff $\varepsilon \to 0$ and by assumption $\mathrm{Q} \neq \mathrm{P}$ we obtain (6).

The Bouligand influence function is a modification of the influence function given by (3). Recall that the Gâteaux-derivative of some mapping f at a point x_0 equals $\nabla^G f(x_0)(h) = \lim_{\varepsilon\downarrow 0}\big(f(x_0 + \varepsilon h) - f(x_0)\big)/\varepsilon$ if it exists for every $h \in X$. Hence the influence function is the special Gâteaux-derivative with $\mathrm{Q} = \delta_z$ and $h = \delta_z - \mathrm{P}$, if the IF is continuous and linear. However, the BIF is always positive homogeneous because it is a Bouligand-derivative, which is in general not true for the influence function. As will be shown in (13), this property leads to the result that for $\alpha \geq 0$ and $h := \varepsilon(\mathrm{Q}-\mathrm{P})$ the asymptotic bias $T((1-\alpha\varepsilon)\mathrm{P} + \alpha\varepsilon\mathrm{Q}) - T(\mathrm{P})$ equals $\alpha\,\mathrm{BIF}(\mathrm{Q};T,\mathrm{P}) + o(h)$.

The following simple calculations clarify the connection between the BIF and the IF. In general we have for B-derivatives with $h = \varepsilon\tilde{h}$, where $\varepsilon \in (0,\infty)$ and $\tilde{h} \in X$ with $0 < \|\tilde{h}\| \leq 2$,

$$\begin{aligned} 0 &= \lim_{h\to 0} \frac{\| f(x_0+h) - f(x_0) - \nabla^B f(x_0)(h)\|}{\|h\|} \\ &= \lim_{\varepsilon\downarrow 0} \frac{\| f(x_0+\varepsilon\tilde{h}) - f(x_0) - \varepsilon\nabla^B f(x_0)(\tilde{h})\|}{\varepsilon\|\tilde{h}\|} \\ &= \lim_{\varepsilon\downarrow 0} \left\| \frac{f(x_0+\varepsilon\tilde{h}) - f(x_0)}{\varepsilon} - \nabla^B f(x_0)(\tilde{h}) \right\|. \end{aligned}$$

Hence $\lim_{\varepsilon\downarrow 0}\big(f(x_0+\varepsilon\tilde{h}) - f(x_0)\big)/\varepsilon = \nabla^B f(x_0)(\tilde{h})$. In particular we obtain for $\mathrm{Q} \neq \mathrm{P}$ and taking $0 < \|\mathrm{Q}-\mathrm{P}\| \leq 2$ into account that, if $\mathrm{BIF}(\mathrm{Q};T,\mathrm{P})$ exists, then $\mathrm{BIF}(\mathrm{Q};T,\mathrm{P}) = \lim_{\varepsilon\downarrow 0}\big(T((1-\varepsilon)\mathrm{P} + \varepsilon\mathrm{Q}) - T(P)\big)/\varepsilon$, which is the definition of the IF, if we choose $\mathrm{Q} = \delta_z$.

We can now give a general result on the BIF of the support vector machine $T(\mathrm{P}) := f_{\mathrm{P},\lambda}$. We restrict attention to Lipschitz continuous loss functions, because the growth behavior of L plays an important role to obtain consistency and robustness results as was shown by Christmann and Steinwart (2007). For notational convenience we shall often write $\nabla_2^B L(Y, f(X))$ instead of $\nabla_2^B L(Y,\cdot)(f(X))$, because $f(X) \in \mathbb{R}$. We will sometimes explicitly write "$\cdot$" for multiplication to avoid misunderstandings.

Theorem 2 *Let $X \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}$ be closed sets, $\mathcal{H}$ be a RKHS with a bounded, continuous kernel k, $f_{\mathrm{P},\lambda} \in \mathcal{H}$, and $L : Y \times \mathbb{R} \to [0,\infty)$ be a convex loss function which is Lipschitz continuous w.r.t. the second argument with uniform Lipschitz constant $|L|_1 := \sup_{y\in Y} |L(y,\cdot)|_1 \in (0,\infty)$. Further, assume that L has measurable partial B-derivatives w.r.t. to the second argument with*

$$\kappa_1 := \sup_{y\in Y} \left\| \nabla_2^B L(y,\cdot) \right\|_\infty \in (0,\infty)\ ,\ \kappa_2 := \sup_{y\in Y} \left\| \nabla_{2,2}^B L(y,\cdot) \right\|_\infty < \infty\,. \qquad (7)$$

Let $\delta_1 > 0$, $\delta_2 > 0$, $\mathcal{N}_{\delta_1}(f_{\mathrm{P},\lambda}) := \{f \in \mathcal{H};\, \|f - f_{\mathrm{P},\lambda}\|_{\mathcal{H}} < \delta_1\}$, $\lambda > \frac{1}{2}\kappa_2 \|k\|_\infty^3$, *and* P,Q *be probability measures*[3] *on* $\big(X \times Y, \mathcal{B}(X \times Y)\big)$ *with* $\mathbb{E}_{\mathrm{P}}|Y| < \infty$ *and* $\mathbb{E}_{\mathrm{Q}}|Y| < \infty$*. Define* $G : (-\delta_2, \delta_2) \times \mathcal{N}_{\delta_1}(f_{\mathrm{P},\lambda}) \to \mathcal{H}$,

$$G(\varepsilon, f) := 2\lambda f + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}} \nabla_2^B L(Y, f(X)) \cdot \Phi(X)\,, \tag{8}$$

and assume that $\nabla_2^B G(0, f_{\mathrm{P},\lambda})$ *is strong. Then the Bouligand influence function of* $T(\mathrm{P}) := f_{\mathrm{P},\lambda}$ *in the direction of* $\mathrm{Q} \neq \mathrm{P}$ *exists,*

$$\begin{aligned} \mathrm{BIF}(\mathrm{Q}; T, \mathrm{P}) \quad &= \quad S^{-1}\big(\mathbb{E}_{\mathrm{P}} \nabla_2^B L(Y, f_{\mathrm{P},\lambda}(X)) \cdot \Phi(X)\big) && (9)\\ &\quad\ -S^{-1}\big(\mathbb{E}_{\mathrm{Q}} \nabla_2^B L(Y, f_{\mathrm{P},\lambda}(X)) \cdot \Phi(X)\big)\,, && (10) \end{aligned}$$

where $S : \mathcal{H} \to \mathcal{H}$ *with*

$$S(\cdot) := \nabla_2^B G(0, f_{\mathrm{P},\lambda})(\cdot) = 2\lambda\, \mathrm{id}_{\mathcal{H}}(\cdot) + \mathbb{E}_{\mathrm{P}} \nabla_{2,2}^B L(Y, f_{\mathrm{P},\lambda}(X)) \cdot \langle \Phi(X), \cdot \rangle_{\mathcal{H}} \Phi(X),$$

and $\mathrm{BIF}(\mathrm{Q}; T, \mathrm{P})$ *is bounded.*

Remark 3 *We additionally show that under the assumptions of Theorem 2 we have:*

1. *For some* χ *and each* $f \in \mathcal{N}_{\delta_1}(f_{\mathrm{P},\lambda})$, $G(\cdot, f)$ *is Lipschitz continuous on* $(-\delta_2, \delta_2)$ *with Lipschitz constant* χ.

2. *G has partial B-derivatives with respect to* ε *and* f *at* $(0, f_{\mathrm{P},\lambda})$.

3. $\nabla_2^B G(0, f_{\mathrm{P},\lambda})(h - f_{\mathrm{P},\lambda})$ *lies in a neighborhood of* $0 \in \mathcal{H}$, $\forall h \in \mathcal{N}_{\delta_1}(f_{\mathrm{P},\lambda})$.

4. $d_0 := \inf_{h_1,h_2 \in \mathcal{N}_{\delta_1}(f_{\mathrm{P},\lambda}) - f_{\mathrm{P},\lambda};\, h_1 \neq h_2} \frac{\|\nabla_2^B G(0, f_{\mathrm{P},\lambda})(h_1) - \nabla_2^B G(0, f_{\mathrm{P},\lambda})(h_2)\|_{\mathcal{H}}}{\|h_1 - h_2\|_{\mathcal{H}}} > 0\,.$

5. *For each* $\xi > d_0^{-1}\chi$ *there exist constants* $\delta_3, \delta_4 > 0$*, a neighborhood* $\mathcal{N}_{\delta_3}(f_{\mathrm{P},\lambda}) := \{f \in \mathcal{H};\, \|f - f_{\mathrm{P},\lambda}\|_{\mathcal{H}} < \delta_3\}$*, and a function* $f^* : (-\delta_4, \delta_4) \to \mathcal{N}_{\delta_3}(f_{\mathrm{P},\lambda})$ *satisfying*

   v.1) $f^*(0) = f_{\mathrm{P},\lambda}$.

   v.2) f^* *is Lipschitz continuous on* $(-\delta_4, \delta_4)$ *with Lipschitz constant* $|f^*|_1 = \xi$.

   v.3) For each $\varepsilon \in (-\delta_4, \delta_4)$ *is* $f^*(\varepsilon)$ *the unique solution of* $G(\varepsilon, f) = 0$ *in* $\mathcal{N}_{\delta_3}(f_{\mathrm{P},\lambda})$.

   v.4) $\nabla^B f^*(0)(u) = \big(\nabla_2^B G(0, f_{\mathrm{P},\lambda})\big)^{-1}\big(-\nabla_1^B G(0, f_{\mathrm{P},\lambda})(u)\big)$, $u \in (-\delta_4, \delta_4)$.

   The function f^* *is the same as in the implicit function theorem by Robinson (1991), see Theorem 7.*

Remark 4 *It will be shown that* $\kappa_2 = 0$ *for* $L = L_{\underline{\varepsilon}}$ *and* $L = L_{\tau-pin}$ *and thus the regularization condition only states that* $\lambda > \frac{1}{2}\kappa_2 \|k\|_\infty^3 = 0$.

[3] Because X and Y are assumed to be closed, P can be split up into the marginal distribution P_X and the regular conditional probability $\mathrm{P}(\cdot\,|x)$, $x \in X$, on Y. Same for Q.

Note that S can be interpreted as the (Bouligand-)Hessian of the regularized risk, see (14) and (17). Further the formula in (9) and (10) is similar to the one obtained by Christmann and Steinwart (2007) for the IF of $T(\mathrm{P}) = f_{\mathrm{P},\lambda}$. The difference is that we used B-derivatives instead of F-derivatives, because we allow non-smooth L.

Note that the first summand of the BIF given in (9) does *not* depend on the contaminating distribution Q. In contrast to that, the second summand of the BIF given in (10) depends on Q and consists of two factors. The first factor depends on the partial B-derivative of the loss function, and is hence bounded due to (7). For many loss functions this factor depends only on the residual term $y - f_{\mathrm{P},\lambda}(x)$. The second factor is the feature map $\Phi(x)$ which is bounded, because k is bounded. For the Gaussian RBF kernel we expect that the second factor is not only bounded, but that the impact of $\mathrm{Q} \neq \mathrm{P}$ on the BIF is approximately local, because $k(x,x')$ converges exponentially fast to zero if $||x-x'||_2$ is large.

3. Examples

In this section we show that our main theorem covers some SVMs widely used in practice. The following result treats SVMs based on the $\underline{\varepsilon}$-insensitive loss function or Huber's loss function for regression, and SVMs based on the pinball loss function for nonparametric quantile regression. These loss functions have uniformly bounded first and second partial B-derivatives w.r.t. the second argument, see the Appendix.

Corollary 5 *Let $X \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}$ be closed, and P,Q be distributions on $X \times Y$ with $\mathbb{E}_{\mathrm{P}}|Y| < \infty$ and $\mathbb{E}_{\mathrm{Q}}|Y| < \infty$.*

1. *For $L \in \{L_{\tau-pin}, L_{\underline{\varepsilon}}\}$, assume that for all $\delta > 0$ there exist positive constants ξ_{P}, ξ_{Q}, c_{P}, and c_{Q} such that for all $t \in \mathbb{R}$ with $|t - f_{\mathrm{P},\lambda}(x)| \leq \delta \|k\|_\infty$ the following inequalities hold for all $a \in [0, 2\delta \|k\|_\infty]$ and $x \in X$:*

$$\mathrm{P}\big(Y \in [t, t+a] \,\big|\, x\big) \leq c_{\mathrm{P}} a^{1+\xi_{\mathrm{P}}} \text{ and } \mathrm{Q}\big(Y \in [t, t+a] \,\big|\, x\big) \leq c_{\mathrm{Q}} a^{1+\xi_{\mathrm{Q}}}\,. \tag{11}$$

2. *For $L = L_{c-Huber}$, assume for $x \in X$:*

$$\mathrm{P}\big(Y \in \big\{ f_{\mathrm{P},\lambda}(x) - c,\, f_{\mathrm{P},\lambda}(x) + c \big\} \,\big|\, x\big) = \mathrm{Q}\big(Y \in \big\{ f_{\mathrm{P},\lambda}(x) - c,\, f_{\mathrm{P},\lambda}(x) + c \big\} \,\big|\, x\big) = 0\,. \tag{12}$$

Then the assumptions of Theorem 2 are valid: $\mathrm{BIF}(\mathrm{Q}; T, \mathrm{P})$ *of* $T(\mathrm{P}) := f_{\mathrm{P},\lambda}$ *exists, is given by (9) to (10), and is bounded.*

For the somewhat smoother Huber loss function we only need to exclude by (12) that the conditional probabilities of Y given X with respect to P and Q have no point probabilities at the two points $f_{\mathrm{P},\lambda}(x) - c$ and $f_{\mathrm{P},\lambda}(x) + c$. Therefore, for this loss function Q can be a Dirac distribution and in this case we have BIF = IF.

For the pinball loss function some calculations give

$$\begin{aligned}\mathrm{BIF}(\mathrm{Q}; T, \mathrm{P}) \;=\;& \frac{1}{2\lambda} \int_X \big(\mathrm{P}\big(Y \leq f_{\mathrm{P},\lambda}(x) \,\big|\, x\big) - \tau\big) \Phi(x) \, d\mathrm{P}_X(x) \\ & - \frac{1}{2\lambda} \int_X \big(\mathrm{Q}\big(Y \leq f_{\mathrm{P},\lambda}(x) \,\big|\, x\big) - \tau\big) \Phi(x) \, d\mathrm{Q}_X(x),\end{aligned}$$

if the BIF exists. We expect the first integral to be small, because $f_{\mathrm{P},\lambda}(x)$ approximates the τ-quantile of $\mathrm{P}(\cdot\,|x)$ and even rates of convergence are known (Steinwart and Christmann, 2008a,b). As will become clear from the proof, (11) and (12) guarantee that the regular conditional probabilities $\mathrm{P}(\cdot\,|x)$ and $\mathrm{Q}(\cdot\,|x)$ do not have large point masses at those points where the Lipschitz continuous loss function L is *not* F-differentiable or in small neighborhoods around these points. Even for the case of parametric quantile regression, that is for $L = L_{\tau-pin}$, $\lambda = 0$ and the unbounded linear kernel $k(x,x') := \langle x, x' \rangle$, some assumptions on the distribution P seem to be necessary for the existence of the IF, see Koenker (2005, p. 44). He assumes that P has a continuous density which is strictly positive where needed.

Nevertheless, the question arises whether Theorem 2 and Corollary 5 can be shown without any assumption on the distributions P and Q. This is—at least with the techniques we used—not possible for non-smooth loss functions as the following counterexample shows. Let us consider kernel based quantile regression based on the Gaussian RBF kernel, that is $L = L_{\tau-pin}$, $k = k_{RBF}$, and $\lambda > 0$. Hence the set $\mathfrak{D}$ of discontinuity points of $\nabla_2^B L$ is $\mathfrak{D} = \{0\}$. Fix $x \in X$ and $y, y^* \in Y$ with $y \neq y^*$. Define $\mathrm{P} = \delta_{(x,y)}$ and $\mathrm{Q} = \delta_{(x,y^*)}$. Consider $f_1, f_2 \in \mathcal{N}_{\delta_1}(f_{\mathrm{P},\lambda})$ with $f_1(x) \neq f_2(x)$, $y - f_1(x) > 0$, $y - f_2(x) < 0$, $y^* - f_1(x) > 0$, and $y^* - f_2(x) < 0$. Hence, $\nabla_2^B L(y, f_1(x)) = \nabla_2^B L(y^*, f_1(x)) = -\tau$ and $\nabla_2^B L(y, f_2(x)) = \nabla_2^B L(y^*, f_2(x)) = 1 - \tau$. Note that $\nabla_{2,2}^B L(y,t) = 0$ for all $y, t \in \mathbb{R}$. We thus obtain for the $\mathcal{H}$-norm in (19) that $\big\|\mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}}\big(\nabla_2^B L(Y, f_1(X)) - \nabla_2^B L(Y, f_2(X))\big) \cdot \Phi(X)\big\|_{\mathcal{H}} = \|\Phi(x)\|_{\mathcal{H}} > 0$. Hence $\nabla_2^B G(0, f_{\mathrm{P},\lambda})$ is not strong in this special case, because $\|\Phi(x)\|_{\mathcal{H}}$ is in general greater than $\varepsilon^* \, \|f_1 - f_2\|_{\mathcal{H}}$ for arbitrarily small values of ε^*.

Now we shall show for L_{log} that the assumptions (11) or (12) are not needed to obtain a bounded BIF. It is easy to see that L_{log} is strictly convex w.r.t. the second argument and Fréchet-differentiable with $\nabla_2^F L_{log}(y,t) = 1 - 2\Lambda(y-t)$, $\nabla_{2,2}^F L_{log}(y,t) = 2\Lambda(y-t)[1-\Lambda(y-t)]$, and $\nabla_{2,2,2}^F L_{log}(y,t) = -2\Lambda(y-t)[1-\Lambda(y-t)][1-2\Lambda(y-t)]$. Obviously, these partial derivatives are bounded for all $y, t \in \mathbb{R}$. Furthermore, $\kappa_1 = \sup_{y\in\mathbb{R}} |\nabla_2^F L_{log}(y,\cdot)|_1 = 1/2$ and $\kappa_2 = \sup_{y\in\mathbb{R}} |\nabla_{2,2}^F L_{log}(y,\cdot)|_1 \leq 1/2$, because an everywhere F-differentiable function g is Lipschitz continuous with $|g|_1 = ||\nabla^F g||_\infty$ if $\nabla^F g$ is bounded.

Corollary 6 *Let $X \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}$ be closed, $L = L_{log}$, and P, Q be distributions on $X \times Y$ with $\mathbb{E}_{\mathrm{P}}|Y| < \infty$ and $\mathbb{E}_{\mathrm{Q}}|Y| < \infty$. Then the assumptions of Theorem 2 are valid, and $\mathrm{BIF}(\mathrm{Q};T,\mathrm{P})$ of $T(\mathrm{P}) := f_{\mathrm{P},\lambda}$ exists, is given by (9) to (10), and $\mathrm{BIF}(\mathrm{Q};T,\mathrm{P})$ is bounded.*

Corollary 6 is of course also valid for empirical distributions D_n and Q_m consisting of n and m data points, because no specific assumptions on P and Q are made.

The influence function of $T(\mathrm{P}) = f_{\mathrm{P},\lambda}$ based on L_{log} and error bounds of the type

$$\big\|T\big((1-\varepsilon)\mathrm{P} + \varepsilon\delta_{(x,y)} - T(\mathrm{P})\big)\big\|_{\mathcal{H}} \leq c^* \varepsilon$$

where the constant c^* is known and depends only on P, $\mathrm{Q} := \delta_{(x,y)}$, and λ, were recently derived by Christmann and Steinwart (2007). We like to mention that Corollary 6 shows that this influence function is even a Bouligand-derivative, hence *positive homogeneous* in $h = \varepsilon(\mathrm{Q}-\mathrm{P})$. Therefore, we immediately obtain from the existence of the BIF that the asymptotic bias of SVMs has the form

$$\begin{aligned}
f_{(1-\alpha\varepsilon)\mathrm{P}+\alpha\varepsilon\mathrm{Q},\lambda} - f_{\mathrm{P},\lambda} &= T(\mathrm{P}+\alpha h) - T(\mathrm{P}) \\
&= \alpha\,\mathrm{BIF}(\mathrm{Q};T,\mathrm{P}) + o(\alpha h) \qquad (13)\\
&= \alpha\big(T(\mathrm{P}+h) - T(\mathrm{P}) + o(h)\big) + o(\alpha h) \\
&= \alpha\big(f_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda} - f_{\mathrm{P},\lambda}\big) + o(\alpha\varepsilon(\mathrm{Q}-\mathrm{P})), \quad \alpha \geq 0.
\end{aligned}$$

This equation nicely describes the behavior of the asymptotic bias term $f_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda} - f_{\mathrm{P},\lambda}$ if we consider the amount $\alpha\varepsilon$ of contamination instead of ε.

4. Discussion

Bouligand-derivatives and strong Bouligand-derivatives were successfully used in approximation theory, see for example Clarke (1983), Robinson (1987, 1991), Ip and Kyparisis (1992), and the references cited therein. To our best knowledge however, these concepts were not used so far to investigate robustness properties of statistical operators.

Therefore, we defined the Bouligand influence function (BIF) as a modification of the influence function (IF), the latter being related to Gâteaux-derivatives and a cornerstone of robust statistics, see Hampel (1974), Hampel et al. (1986), and Maronna et al. (2006). If the BIF exists, then it is identical to the IF. The BIF is a positive homogeneous function by definition. This is in general not true for the IF. We used the BIF to show that support vector machines for regression, which play an important role in modern statistical learning theory, are robust in the sense of influence functions, if a bounded continuous kernel is used and if the convex loss function is Lipschitz continuous and twice Bouligand-differentiable, but not necessarily twice Fréchet-differentiable. The result covers the important special cases of SVMs based on the $\underline{\varepsilon}$-insensitive, Huber or logistic loss function for regression, and kernel based quantile regression based on the pinball loss function. The IF of SVMs based on the logistic loss was recently derived by Christmann and Steinwart (2007) and Steinwart and Christmann (2008b).

From our point of view, the Bouligand-derivative is a promising concept for robust statistics for the following reason. Many robust estimators proposed in the literature are implicitly defined as solutions of minimization problems where the objective function or loss function is continuous or Lipschitz continuous, but not necessarily twice Fréchet-differentiable. Examples are not only SVMs treated in this paper, but also M-estimators of Huber-type and certain maximum likelihood estimators under non-standard conditions. Bouligand-differentiation nicely fills the gap between Fréchet-differentiation, which is too strong for many robust estimators, and Gâteaux-differentiation which is the basis for the robustness approach based on influence functions. Bouligand-derivatives fulfill a chain rule and a theorem of implicit functions which is in general not true for Gâteaux-derivatives.

Acknowledgments

We thank S.M. Robinson for drawing our attention to his paper from 1987. We also thank the editor and three anonymous referees.

Appendix A. Proofs

This appendix contains all the proofs of the previous sections.

A.1 Proofs for the Results in Section 2

For the proof of Theorem 2 we shall use the following implicit function theorem for B-derivatives, see Robinson (1991, Cor. 3.4). For a function f from a metric space (X, d_X) to another metric space

(Y,d_Y), we define

$$\delta(f,X) = \inf\{d_Y\big(f(x_1),f(x_2)\big)\,/\,d_X(x_1,x_2) \mid x_1 \neq x_2; x_1,x_2 \in X\}\,.$$

Theorem 7 *Let Y be a Banach space and X and Z be normed linear spaces. Let x_0 and y_0 be points of X and Y, respectively, and let $\mathcal{N}(x_0)$ be a neighborhood of x_0 and $\mathcal{N}(y_0)$ be a neighborhood of y_0. Suppose that G is a function from $\mathcal{N}(x_0) \times \mathcal{N}(y_0)$ to Z with $G(x_0,y_0) = 0$. In particular, for some ϕ and each $y \in \mathcal{N}(y_0)$, $G(\cdot,y)$ is assumed to be Lipschitz continuous on $\mathcal{N}(x_0)$ with modulus ϕ. Assume that G has partial B-derivatives with respect to x and y at (x_0,y_0), and that: (i) $\nabla_2^B G(x_0,y_0)(\cdot)$ is strong. (ii) $\nabla_2^B G(x_0,y_0)(y-y_0)$ lies in a neighborhood of $0 \in Z$, $\forall y \in \mathcal{N}(y_0)$. (iii) $\delta(\nabla_2^B G(x_0,y_0),\mathcal{N}(y_0)-y_0) \;=:\; d_0 > 0$. Then for each $\xi > d_0^{-1}\phi$ there are neighborhoods U of x_0 and V of y_0, and a function $f^* : U \to V$ satisfying (a) $f^*(x_0) = y_0$. (b) f^* is Lipschitz continuous on $\mathcal{N}(x_0)$ with modulus ξ. (c) For each $x \in U$, $f^*(x)$ is the unique solution in V of $G(x,y) = 0$. (d) The function f^* is B-differentiable at x_0 with $\nabla^B f^*(x_0)(u) = \big(\nabla_2^B G(x_0,y_0)\big)^{-1}\big(-\nabla_1^B G(x_0,y_0)(u)\big)$.*

We will also need the following consequence of the open mapping theorem, see Lax (2002, p. 170).

Theorem 8 *Let X and Y be Banach spaces, $A : X \to Y$ be a bounded, linear, and bijective function. Then the inverse $A^{-1} : Y \to X$ is a bounded linear function.*

The key ingredient of our proof of Theorem 2 is of course the map $G : \mathbb{R} \times \mathcal{H} \to \mathcal{H}$ defined by (8). If $\varepsilon < 0$ the integration is w.r.t. a signed measure. The $\mathcal{H}$-valued expectation used in the definition of G is well-defined for all $\varepsilon \in (\delta_2,\delta_2)$ and all $f \in \mathcal{N}_{\delta_1}(f_{\mathrm{P},\lambda})$, because $\kappa_1 \in (0,\infty)$ by (7) and $\|\Phi(x)\|_\infty \leq \|k\|_\infty^2 < \infty$ by (2). For F- and B-derivatives holds a chain rule and F-differentiable functions are also B-differentiable. For $\varepsilon \in [0,1]$ we thus obtain

$$G(\varepsilon,f) = \frac{\partial \mathcal{R}^{reg}_{L,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda}}{\partial \mathcal{H}}(f) = \nabla_2^B \mathcal{R}^{reg}_{L,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda}(f)\,. \tag{14}$$

Since $f \mapsto \mathcal{R}^{reg}_{L,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda}(f)$ is convex and continuous for all $\varepsilon \in [0,1]$ equation (14) shows that we have $G(\varepsilon,f) = 0$ if and only if $f = f_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda}$ for such ε. Hence

$$G(0,f_{\mathrm{P},\lambda}) = 0\,. \tag{15}$$

We shall show that Theorem 7 is applicable for G and that there exists a B-differentiable function $\varepsilon \mapsto f_\varepsilon$ defined on a small interval $(-\delta_2,\delta_2)$ for some $\delta_2 > 0$ satisfying $G(\varepsilon,f_\varepsilon) = 0$ for all $\varepsilon \in (-\delta_2,\delta_2)$. From the existence of this function we shall obtain $\mathrm{BIF}(\mathrm{Q};T,\mathrm{P}) = \nabla^B f_\varepsilon(0)$.

Proof of Theorem 2. The existence of $f_{\mathrm{P},\lambda}$ follows from the convexity of L and the penalizing term, see also Christmann and Steinwart (2007, Prop. 8). The assumption that $G(0,f_{\mathrm{P},\lambda}) = 0$ is valid by (15). Let us now prove the results of Remark 3 parts 1 to 5.

Remark 3 part 1. For $f \in \mathcal{H}$ fixed let $\varepsilon_1, \varepsilon_2 \in (-\delta_2, \delta_2)$. Using $\|k\|_\infty < \infty$ and (15) we obtain

$$\begin{aligned}
& \left|\mathbb{E}_{(1-\varepsilon_1)\mathrm{P}+\varepsilon_1\mathrm{Q}} \nabla_2^B L(Y, f(X)) \cdot \Phi(X) - \mathbb{E}_{(1-\varepsilon_2)\mathrm{P}+\varepsilon_2\mathrm{Q}} \nabla_2^B L(Y, f(X)) \cdot \Phi(X)\right| \\
= & \left|(\varepsilon_1 - \varepsilon_2) \mathbb{E}_{\mathrm{Q}-\mathrm{P}} \nabla_2^B L(Y, f(X)) \cdot \Phi(X)\right| \\
\leq & |\varepsilon_1 - \varepsilon_2| \int \left|\nabla_2^B L(y, f(x)) \cdot \Phi(x)\right| \, d|\mathrm{Q}-\mathrm{P}|(x,y) \\
\leq & |\varepsilon_1 - \varepsilon_2| \int \sup_{y \in Y} |\nabla_2^B L(y, f(x))| \sup_{x \in X} |\Phi(x)| \, d|\mathrm{Q}-\mathrm{P}|(x,y) \\
\leq & |\varepsilon_1 - \varepsilon_2| \, \|\Phi(x)\|_\infty \sup_{y \in Y} \left\|\nabla_2^B L(y, \cdot)\right\|_\infty \int d|\mathrm{Q}-\mathrm{P}|(x,y) \\
\leq & 2 \, \|k\|_\infty^2 \sup_{y \in Y} \left\|\nabla_2^B L(y, \cdot)\right\|_\infty |\varepsilon_1 - \varepsilon_2| \\
= & 2 \, \|k\|_\infty^2 \, \kappa_1 \, |\varepsilon_1 - \varepsilon_2| < \infty .
\end{aligned}$$

Remark 3 part 2. We have

$$\begin{aligned}
\nabla_1^B G(\varepsilon, f) &= \nabla_1^B \Big(\mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}} \nabla_2^B L(Y, f(X)) \cdot \Phi(X) \Big) \\
&= \nabla_1^B \Big(\mathbb{E}_{\mathrm{P}} \nabla_2^B L(Y, f(X)) \cdot \Phi(X) + \varepsilon \mathbb{E}_{\mathrm{Q}-\mathrm{P}} \nabla_2^B L(Y, f(X)) \cdot \Phi(X) \Big) \\
&= \mathbb{E}_{\mathrm{Q}-\mathrm{P}} \nabla_2^B L(Y, f(X)) \cdot \Phi(X) \\
&= \mathbb{E}_{\mathrm{Q}} \nabla_2^B L(Y, f(X)) \cdot \Phi(X) - \mathbb{E}_{\mathrm{P}} \nabla_2^B L(Y, f(X)) \cdot \Phi(X) . \qquad (16)
\end{aligned}$$

This expectation exists due to (2) and (7). Furthermore, we obtain

$$\begin{aligned}
& \nabla_2^B G(0, f_{\mathrm{P},\lambda})(h) + o(h) \\
= & G(0, f_{\mathrm{P},\lambda} + h) - G(0, f_{\mathrm{P},\lambda}) \\
= & 2\lambda h + \mathbb{E}_{\mathrm{P}} \nabla_2^B L(Y, (f_{\mathrm{P},\lambda}(X) + h(X))) \cdot \Phi(X) - \mathbb{E}_{\mathrm{P}} \nabla_2^B L(Y, f_{\mathrm{P},\lambda}(X)) \cdot \Phi(X) \\
= & 2\lambda h + \mathbb{E}_{\mathrm{P}} \Big(\nabla_2^B L\big(Y, (f_{\mathrm{P},\lambda}(X) + h(X))\big) - \nabla_2^B L\big(Y, f_{\mathrm{P},\lambda}(X)\big) \Big) \cdot \Phi(X) .
\end{aligned}$$

This expectation exists, as the term $\nabla_2^B L\big(Y, (f_{\mathrm{P},\lambda}(X) + h(X))\big) - \nabla_2^B L\big(Y, f_{\mathrm{P},\lambda}(X)\big)$ is bounded due to (2), (7), and $\|k\|_\infty < \infty$. Using $\langle \Phi(X), \cdot \rangle_{\mathcal{H}} \in \mathcal{H}$, we get

$$\nabla_2^B G(0, f_{\mathrm{P},\lambda})(\cdot) = 2\lambda \mathrm{id}_{\mathcal{H}}(\cdot) + \mathbb{E}_{\mathrm{P}} \nabla_{2,2}^B L(Y, f_{\mathrm{P},\lambda}(X)) \cdot \langle \Phi(X), \cdot \rangle_{\mathcal{H}} \Phi(X) . \qquad (17)$$

Note that $\mathbb{E}_{\mathrm{P}} \nabla_{2,2}^B L\big(Y, f(X)\big) = \nabla_2^B \mathbb{E}_{\mathrm{P}} \nabla_2^B L\big(Y, f(X)\big)$, because

$$\begin{aligned}
& \nabla_2^B \mathbb{E}_{\mathrm{P}} \nabla_2^B L\big(Y, f(X)\big) - \mathbb{E}_{\mathrm{P}} \nabla_{2,2}^B L\big(Y, f(X)\big) \\
= & \mathbb{E}_{\mathrm{P}} \big(\nabla_2^B L(Y, (f(X) + h(X))) - \nabla_2^B L(Y, f(X)) \big) - \mathbb{E}_{\mathrm{P}} \nabla_{2,2}^B L(Y, f(X)) + o(h) \\
= & \mathbb{E}_{\mathrm{P}} \big(\nabla_2^B L(Y, (f(X) + h(X))) - \nabla_2^B L(Y, f(X)) - \nabla_{2,2}^B L(Y, f(X)) \big) + o(h) = o(h)
\end{aligned}$$

by definition of the B-derivative.

Remark 3 part 3. Let $\mathcal{N}_{\delta_1}(f_{\mathrm{P},\lambda})$ be a δ_1-neighborhood of $f_{\mathrm{P},\lambda}$. Because $\mathcal{H}$ is a RKHS and hence a vector space it follows for all $h \in \mathcal{N}_{\delta_1}(f_{\mathrm{P},\lambda})$ that $\left\| f_{\mathrm{P},\lambda} - h - 0 \right\|_{\mathcal{H}} \leq \delta_1$ and hence $h - f_{\mathrm{P},\lambda} \in \mathcal{N}_{\delta_1}(0) \subset$

$\mathcal{H}$. Note that $\nabla_2^B G(0, f_{\mathrm{P},\lambda})(\,\cdot\,)$ computed by (17) is a mapping from $\mathcal{H}$ to $\mathcal{H}$. For $\xi := h - f_{\mathrm{P},\lambda}$ we have $\|\xi\|_{\mathcal{H}} \le \delta_1$ and the reproducing property yields

$$\nabla_2^B G(0, f_{\mathrm{P},\lambda})(\xi) = 2\lambda\xi + \mathbb{E}_{\mathrm{P}} \nabla_{2,2}^B L(Y, f_{\mathrm{P},\lambda}(X)) \cdot \xi \Phi(X).$$

Using (2) and (7) we obtain

$$\begin{aligned}
& \left\| 2\lambda\xi + \mathbb{E}_{\mathrm{P}} \nabla_{2,2}^B L(Y, f_{\mathrm{P},\lambda}(X)) \cdot \xi\Phi(X) - 0 \right\|_{\mathcal{H}} \\
\le\ & 2\lambda \|\xi\|_{\mathcal{H}} + \left\| \mathbb{E}_{\mathrm{P}} \nabla_{2,2}^B L(Y, f_{\mathrm{P},\lambda}(X)) \cdot \xi\Phi(X) \right\|_{\mathcal{H}} \\
\le\ & 2\lambda \|\xi\|_{\mathcal{H}} + \sup_{y \in Y} \left\| \nabla_{2,2}^B L(y, \cdot) \right\|_{\infty} \|\xi\|_{\infty} \|\Phi(x)\|_{\infty} \\
\le\ & 2\lambda \|\xi\|_{\mathcal{H}} + \kappa_2 \, \|\xi\|_{\mathcal{H}} \, \|k\|_{\infty}^3 \\
\le\ & \left(2\lambda + \kappa_2 \, \|k\|_{\infty}^3\right) \delta_1 \,,
\end{aligned}$$

which shows that $\nabla_2^B G(0, f_{\mathrm{P},\lambda})(h - f_{\mathrm{P},\lambda})$ lies in a neighborhood of $0 \in \mathcal{H}$, for all $h \in \mathcal{N}_{\delta_1}(f_{\mathrm{P},\lambda})$.

Remark 3 part 4. Due to (17) we have to prove that

$$d_0 := \inf_{f_1 \neq f_2} \frac{\left\| 2\lambda(f_1 - f_2) + \mathbb{E}_{\mathrm{P}} \nabla_{2,2}^B L\big(Y, f_{\mathrm{P},\lambda}(X)\big) \cdot (f_1 - f_2)\Phi(X) \right\|_{\mathcal{H}}}{\| f_1 - f_2 \|_{\mathcal{H}}} > 0 \,.$$

If $f_1 \neq f_2$, then (2), (7), and $\lambda > \frac{1}{2}\kappa_2 \, \|k\|_{\infty}^3$ yield that

$$\begin{aligned}
& \left\| 2\lambda(f_1 - f_2) + \mathbb{E}_{\mathrm{P}} \nabla_{2,2}^B L\big(Y, f_{\mathrm{P},\lambda}(X)\big) \cdot (f_1 - f_2)\Phi(X) \right\|_{\mathcal{H}} / \| f_1 - f_2 \|_{\mathcal{H}} \\
\ge\ & \big(\| 2\lambda(f_1 - f_2) \|_{\mathcal{H}} - \left\| \mathbb{E}_{\mathrm{P}} \nabla_{2,2}^B L(Y, f_{\mathrm{P},\lambda}(X)) \cdot (f_1 - f_2)\Phi(X) \right\|_{\mathcal{H}} \big) / \| f_1 - f_2 \|_{\mathcal{H}} \\
\ge\ & 2\lambda - \kappa_2 \, \|k\|_{\infty}^3 > 0
\end{aligned}$$

by our assumption, which gives the assertion.

Remark 3 part 5. The assumptions of Robinson's implicit function theorem, see Theorem 7, are valid for G due to the results of Remark 3 parts 1 to 4 and the assumption that $\nabla_2^B G(0, f_{\mathrm{P},\lambda})$ is strong. This gives part 5.

The result of Theorem 2 now follows from inserting (16) and (17) into Remark 3 part 5(v.4). Using (7) we see that S is bounded. The linearity of S follows from its definition and the inverse of S does exist by Theorem 7. If necessary we can restrict the range of S to $S(\mathcal{H})$ to obtain a bijective function $S_* : \mathcal{H} \to S(\mathcal{H})$ with $S_*(f) = S(f)$ for all $f \in \mathcal{H}$. Hence S^{-1} is also bounded and linear by Theorem 8. This gives the existence of a bounded BIF specified by (9) and (10). ■

A.2 Calculations for the Results in Section 3

For the proof of Corollary 5 we need the partial B-derivatives for the three loss functions and also have to check that $\nabla_2^B G(0, f_{\mathrm{P},\lambda})$ is strong. We shall compute the partial B-derivatives for these loss functions in advance.

A.2.1 $\underline{\varepsilon}$-Insensitive Loss

We shall show for the $\underline{\varepsilon}$-insensitive loss $L = L_{\underline{\varepsilon}}$ that

$$\nabla_2^B L(y,t)(h) = \begin{cases} -h & \text{if} \quad \{t < y - \underline{\varepsilon}\} \text{ or } \{y - t = \underline{\varepsilon}, h < 0\} \\ 0 & \text{if} \quad \{y - \underline{\varepsilon} < t < y + \underline{\varepsilon}\} \text{ or } \{y - t = \underline{\varepsilon}, h \geq 0\} \\ & \quad\ \text{or } \{y - t = -\underline{\varepsilon}, h < 0\} \\ h & \text{if} \quad \{t > y + \underline{\varepsilon}\} \text{ or } \{y - t = -\underline{\varepsilon}, h \geq 0\} \end{cases}$$

and $\nabla_{2,2}^B L(y,t)(h) = 0$.

For the derivation of $\nabla_2^B L(y,t)$ we need to consider 5 cases.

1. If $t > y + \underline{\varepsilon}$, we have $t + h > y + \underline{\varepsilon}$ as long as h is small enough. Therefore,

$$\nabla_2^B L(y,t)(h) + o(h) = L(y,t+h) - L(y,t) = t + h - y - \underline{\varepsilon} - (t - y - \underline{\varepsilon}) = h.$$

2. If $t < y - \underline{\varepsilon}$, we have $t + h < y + \underline{\varepsilon}$ if h is sufficiently small. Thus

$$\nabla_2^B L(y,t)(h) + o(h) = y - t - h - \underline{\varepsilon} - (y - t - \underline{\varepsilon}) = -h.$$

3. If $y - t \in (-\underline{\varepsilon}, \underline{\varepsilon})$ we have $y - t - h \in (-\underline{\varepsilon}, \underline{\varepsilon})$ for $h \to 0$. This yields $\nabla_2^B L(y,t)(h) + o(h) = 0 - 0 = 0$.

4. If $y - t = \underline{\varepsilon}$ we have to consider 2 cases. If $h \geq 0$ and small, then $-\underline{\varepsilon} < y - t - h < \underline{\varepsilon}$ and hence $\nabla_2^B L(y,t)(h) + o(h) = 0 - 0 = 0$.
   If $h < 0$, we have $y - t - h > \underline{\varepsilon}$ and thus

$$\nabla_2^B L(y,t)(h) + o(h) = y - t - h - \underline{\varepsilon} - 0 = -h.$$

5. If $y - t = -\underline{\varepsilon}$ we have again to consider 2 cases. If $h \geq 0$, we have $y - t - h < -\underline{\varepsilon}$. Hence

$$\nabla_2^B L(y,t)(h) + o(h) = t + h - y - \underline{\varepsilon} - 0 = h.$$

   If $h < 0$, we get $-\underline{\varepsilon} < y - t - h < \underline{\varepsilon}$ which gives $\nabla_2^B L(y,t)(h) + o(h) = 0 - 0 = 0$.

This gives the assertion for the first partial B-derivative. Using the same reasoning we obtain $\nabla_{2,2}^B L(y,t)(h) = 0$.

A.2.2 Pinball-Loss

It will be shown that for the pinball loss $L = L_{\tau-pin}$ we get

$$\nabla_2^B L(y,t)(h) = \begin{cases} (1-\tau)h & \text{if} \quad \{y - t < 0\} \text{ or } \{y - t = 0, h \geq 0\} \\ -\tau h & \text{if} \quad \{y - t > 0\} \text{ or } \{y - t = 0, h < 0\} \end{cases}$$

and $\nabla_{2,2}^B L(y,t)(h) = 0$.

For the calculation of $\nabla_2^B L(y,t)$ we consider 3 cases.

1. If $y-t<0$ we have $y-t-h<0$ for sufficiently small values of $|h|$. Hence
$$\begin{aligned}\nabla_2^B L(y,t)(h)+o(h) &= L(y,t+h)-L(y,t)\\ &= (\tau-1)(y-t-h)-(\tau-1)(y-t)=(1-\tau)h.\end{aligned}$$

2. If $y-t>0$ we have $y-t-h>0$ for sufficiently small values of $|h|$ which yields
$$\nabla_2^B L(y,t)(h)+o(h)=\tau(y-t-h)-\tau(y-t)=-\tau h.$$

3. Assume $y-t=0$. If $y-t-h<0$ we have
$$\nabla_2^B L(y,t)(h)+o(h)=(1-\tau)h.$$
If $y-t-h>0$ it follows
$$\nabla_2^B L(y,t)(h)+o(h)=\tau(y-t-h)-\tau(y-t)=-\tau h.$$

Together this gives the assertion for $\nabla_2^B L(y,t)(h)$. In the same way we get $\nabla_{2,2}^B L(y,t)(h)=0$.

A.2.3 HUBER LOSS

It will be shown that for the Huber loss $L=L_{c-Huber}$ we have
$$\nabla_2^B L(y,t)(h)=\begin{cases}-c\,\mathrm{sign}(y-t)h & \text{if} \quad |y-t|>c\\ -(y-t)h & \text{if} \quad |y-t|\le c\end{cases}$$

and
$$\nabla_{2,2}^B L(y,t)(h) \quad = \quad \begin{cases} h & \text{if} \quad \{y-t=c, h\ge 0\} \text{ or } \{y-t=-c, h<0\}\\ & \quad\;\; \text{or } \{|y-t|<c\}\\ 0 & \text{if} \quad \text{else}\,.\end{cases}$$

For the derivation of $\nabla_2^B L(y,t)$ we consider the following 5 cases.

1. Let $y-t=c$. If $h\ge 0$ or $y-t-h\le c$ then
$$\begin{aligned}\nabla_2^B L(y,t)(h)+o(h) &= L(y,t+h)-L(y,t)\\ &= \frac{1}{2}(y-t-h)^2-\frac{1}{2}(y-t)^2=-(y-t)h+\frac{h^2}{2}.\end{aligned}$$
If $h<0$ or $y-t-h>c>0$ we have
$$\begin{aligned}\nabla_2^B L(y,t)(h)+o(h) &= c|y-t-h|-\frac{c^2}{2}-\frac{1}{2}(y-t)^2\\ &= c(y-t-h)-\frac{c^2}{2}-\frac{c^2}{2}\\ &= c(c-h)-c^2=-(y-t)h.\end{aligned}$$

2. Now we consider the case $y-t=-c$. If $h \geq 0$ or $y-t-h \leq -c < 0$ we obtain

$$\begin{aligned}\nabla_2^B L(y,t)(h)+o(h) &= c|y-t-h|-\frac{c^2}{2}-\frac{1}{2}(y-t)^2 \\ &= c(c+h)-\frac{c^2}{2}-\frac{c^2}{2}=-(y-t)h.\end{aligned}$$

   If $h<0$ or $y-t-h>-c$ we get

$$\nabla_2^B L(y,t)(h)+o(h)=\frac{1}{2}(y-t-h)^2-\frac{1}{2}(y-t)^2=-(y-t)h+\frac{h^2}{2}.$$

3. If $y-t>c$, we have $y-t-h>c$ and thus

$$\begin{aligned}\nabla_2^B L(y,t)(h)+o(h) &= c|y-t-h|-\frac{c^2}{2}-c|y-t|+\frac{c^2}{2} \\ &= c(y-t-h)-c(y-t)=-ch=-c\,\mathrm{sign}(y-t)h.\end{aligned}$$

4. If $y-t<-c$, we have $y-t-h<-c$ and obtain analogously to (iii) that

$$\begin{aligned}\nabla_2^B L(y,t)(h)+o(h) &= c|y-t-h|-\frac{c^2}{2}-c|y-t|+\frac{c^2}{2} \\ &= c(-y+t+h)-c(-y+t)=ch=-c\,\mathrm{sign}(y-t)h.\end{aligned}$$

5. If $-c<y-t<c$, then $-c<y-t-h<c$ and

$$\nabla_2^B L(y,t)(h)+o(h)=\frac{1}{2}(y-t-h)^2-\frac{1}{2}(y-t)^2=-(y-t)h+\frac{h^2}{2}.$$

This gives the assertion for $\nabla_2^B L(y,t)(h)$. Only the first two cases, where $y-t=\pm c$, were necessary to compute, since in the other 3 parts the function is already F-differentiable, and thus also B-differentiable. For the second partial B-derivative we consider 3 cases.

1. Assume $y-t=c$. If $y-t-h<c$ then

$$\nabla_{2,2}^B L(y,t)(h)+o(h)=\nabla_2^B L(y,t+h)-\nabla_2^B L(y,t)=-(y-t-h)-(-(y-t))=h.$$

   If $y-t-h>c$ then $\nabla_{2,2}^B L(y,t)(h)+o(h)=-c-(-(y-t))=0$.

2. Assume $y-t=-c$. If $y-t-h<-c$ we obtain $\nabla_{2,2}^B L(y,t)(h)+o(h)=c-(-(y-t))=0$. If $y-t-h>-c$ then

$$\nabla_{2,2}^B L(y,t)(h)+o(h)=-(y-t-h)-(-(y-t))=h.$$

3. Assume that $|y-t|\neq c$. Then $\nabla_2^B L(y,t+h)=\nabla_2^B L(y,t)$. The difference, and consequently $\nabla_{2,2}^B L(y,t)(h)=0$.

This gives the assertion for Huber's loss function.

Proof of Corollary 5. Now that we have shown that these loss functions have bounded first and second partial B-derivatives, we are ready to check if $\nabla_2^B G(0, f_{\mathrm{P},\lambda})$ is strong in these cases. Recall that $\nabla_2^B G(0, f_{\mathrm{P},\lambda})$ is strong, if for all $\varepsilon^* > 0$ there exist a neighborhood $\mathcal{N}_{\delta_1}(f_{\mathrm{P},\lambda})$ and an interval $(-\delta_2, \delta_2)$ with $\delta_1, \delta_2 > 0$ such that for all $f_1, f_2 \in \mathcal{N}_{\delta_1}(f_{\mathrm{P},\lambda})$ and for all $\varepsilon \in (-\delta_2, \delta_2)$ we have

$$\big\|\big(G(\varepsilon, f_1) - g(f_1)\big) - \big(G(\varepsilon, f_2) - g(f_2)\big)\big\|_{\mathcal{H}} \leq \varepsilon^* \left\| f_1 - f_2 \right\|_{\mathcal{H}}, \tag{18}$$

where

$$\begin{aligned} g(f) \;&=\; 2\lambda f_{\mathrm{P},\lambda}(X) + \mathbb{E}_{\mathrm{P}} \nabla_2^B L\big(Y, f_{\mathrm{P},\lambda}(X)\big) \cdot \Phi(X) + 2\lambda\, \mathrm{id}_{\mathcal{H}}(f(X) - f_{\mathrm{P},\lambda}(X)) \\ &\quad + \mathbb{E}_{\mathrm{P}} \nabla_{2,2}^B L\big(Y, f_{\mathrm{P},\lambda}(X)\big) \cdot \langle (f(X) - f_{\mathrm{P},\lambda}(X)), \Phi(X)\rangle_{\mathcal{H}} \Phi(X)\,, \;\; f \in \mathcal{H}. \end{aligned}$$

Fix $\varepsilon^* > 0$. Obviously, (18) is valid for $f_1 = f_2$. For the rest of the proof we therefore fix arbitrary functions $f_1, f_2 \in \mathcal{N}_{\delta_1}(f_{\mathrm{P},\lambda})$ with $f_1 \neq f_2$. We obtain for the term on the left hand side of (18) that

$$\begin{aligned}
& \bigg\| \Big(2\lambda f_1(X) + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}} \nabla_2^B L(Y, f_1(X)) \cdot \Phi(X) \\
& \qquad - 2\lambda f_{\mathrm{P},\lambda}(X) - \mathbb{E}_{\mathrm{P}} \nabla_2^B L(Y, f_{\mathrm{P},\lambda}(X)) \cdot \Phi(X) \\
& \qquad - 2\lambda (f_1(X) - f_{\mathrm{P},\lambda}(X)) - \mathbb{E}_{\mathrm{P}} \nabla_{2,2}^B L(Y, f_{\mathrm{P},\lambda}(X)) \cdot (f_1(X) - f_{\mathrm{P},\lambda}(X)) \Phi(X) \Big) \\
& \quad - \Big(2\lambda f_2(X) + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}} \nabla_2^B L(Y, f_2(X)) \cdot \Phi(X) \\
& \qquad - 2\lambda f_{\mathrm{P},\lambda}(X) - \mathbb{E}_{\mathrm{P}} \nabla_2^B L(Y, f_{\mathrm{P},\lambda}(X)) \cdot \Phi(X) \\
& \qquad - 2\lambda (f_2(X) - f_{\mathrm{P},\lambda}(X)) - \mathbb{E}_{\mathrm{P}} \nabla_{2,2}^B L(Y, f_{\mathrm{P},\lambda}(X)) \cdot (f_2(X) - f_{\mathrm{P},\lambda}(X)) \Phi(X) \Big) \bigg\|_{\mathcal{H}} \\
= \;\; & \bigg\| \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}} \Big(\nabla_2^B L(Y, f_1(X)) - \nabla_2^B L(Y, f_2(X)) \Big) \cdot \Phi(X) \qquad\qquad\qquad\qquad\qquad (19) \\
& \qquad - \mathbb{E}_{\mathrm{P}} \nabla_{2,2}^B L(Y, f_{\mathrm{P},\lambda}(X)) \cdot (f_1(X) - f_2(X)) \Phi(X) \bigg\|_{\mathcal{H}} \\
\leq \;\; & |1-\varepsilon| \, \bigg\| \mathbb{E}_{\mathrm{P}} \Big(\nabla_2^B L(Y, f_1(X)) - \nabla_2^B L(Y, f_2(X)) \\
& \qquad\qquad - \nabla_{2,2}^B L(Y, f_{\mathrm{P},\lambda}(X)) \cdot (f_1(X) - f_2(X)) \Big) \Phi(X) \bigg\|_{\mathcal{H}} \\
& + |\varepsilon| \, \bigg\| \mathbb{E}_{\mathrm{Q}} \Big(\nabla_2^B L(Y, f_1(X)) - \nabla_2^B L(Y, f_2(X)) \Big) \cdot \Phi(X) \bigg\|_{\mathcal{H}} \\
& + |\varepsilon| \, \bigg\| \mathbb{E}_{\mathrm{P}} \nabla_{2,2}^B L(Y, f_{\mathrm{P},\lambda}(X)) \cdot (f_1(X) - f_2(X)) \Phi(X) \bigg\|_{\mathcal{H}} \\
=: \;\; & |1-\varepsilon| A + |\varepsilon| B + |\varepsilon| C\,. \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (20)
\end{aligned}$$

We shall show that (20) is bounded from above by $\varepsilon^* \left\| f_1 - f_2 \right\|_{\mathcal{H}}$. When we look at the first partial B-derivatives of our loss functions, we see that we can separate them in 2 cases: for $L_{\underline{\varepsilon}}$ and $L_{\tau-pin}$ there are one or more discontinuities in $\nabla_2^B L$, whereas $\nabla_2^B L$ is continuous for $L_{c-Huber}$. Recall that the set $\mathfrak{D}$ of points where Lipschitz continuous functions are *not* Fréchet-differentiable, has Lebesgue measure zero by Rademacher's theorem (Rademacher, 1919). Define the function $h\big(y, f_1(x), f_2(x)\big) := \nabla_2^B L\big(y, f_1(x)\big) - \nabla_2^B L\big(y, f_2(x)\big)$. For $L \in \{L_{\underline{\varepsilon}}, L_{\tau-pin}\}$, denote the set of discontinuity points of $\nabla_2^B L$ by $\mathfrak{D}$. Take $f_1,\ f_2 \in \mathcal{N}_{\delta_1}(f_{\mathrm{P},\lambda})$. For $\nabla_2^B L(Y, f_{\mathrm{P},\lambda}(x)) \notin \mathfrak{D}$ we obtain $\nabla_2^B L(Y, f_1(x)) = \nabla_2^B L(Y, f_2(x))$ for sufficiently small δ_1 and hence $h(y, f_1(x), f_2(x)) = 0$. If, on the other hand, $\nabla_2^B L(Y, f_{\mathrm{P},\lambda}(x)) \in \mathfrak{D}$ and $f_1(x) < f_{\mathrm{P},\lambda}(x) < f_2(x)$ or $f_2(x) < f_{\mathrm{P},\lambda}(x) < f_1(x)$, then $\nabla_2^B L(Y, f_1(x)) \neq \nabla_2^B L(Y, f_2(x))$ and hence $h(y, f_1(x), f_2(x)) \neq 0$. Define $m = 2|\mathfrak{D}|$.

A.2.4 PINBALL LOSS

Using the first part of this proof we see that for the pinball loss $L = L_{\tau-pin}$ we obtain $|h(y, f_1(x), f_2(x))| \leq c_1$, with $c_1 = 1$, $\mathfrak{D} = \{0\}$, $m = 2$, and $\nabla^B_{2,2}L(y,t) = 0$, for all $t \in \mathbb{R}$. For all $f \in \mathcal{N}_{\delta_1}(f_{\mathrm{P},\lambda})$ we get

$$|f(x) - f_{\mathrm{P},\lambda}(x)| \leq \left\| f - f_{\mathrm{P},\lambda} \right\|_\infty \leq \|k\|_\infty \left\| f - f_{\mathrm{P},\lambda} \right\|_{\mathcal{H}} \leq \|k\|_\infty \delta_1 . \tag{21}$$

Further

$$|f_1(x) - f_2(x)| \leq \|f_1 - f_2\|_\infty \leq \|k\|_\infty \|f_1 - f_2\|_{\mathcal{H}} \leq 2\|k\|_\infty \delta_1 . \tag{22}$$

Using (21), (22), and (11) we obtain

$$\begin{aligned}
A &= \left\| \mathbb{E}_{\mathrm{P}}(\nabla^B_2 L(Y, f_1(X)) - \nabla^B_2 L(Y, f_2(X))) \cdot \Phi(X) \right\|_{\mathcal{H}} \\
&\leq \mathbb{E}_{\mathrm{P}} |h(Y, f_1(X), f_2(X))| \, |\Phi(X)| \\
&\leq \|k\|^2_\infty \mathbb{E}_{\mathrm{P}} |h(Y, f_1(X), f_2(X))| \mathbf{1}_{\{h \neq 0\}} \\
&\leq \|k\|^2_\infty c_1 \mathrm{P}\big(\nabla^B_2 L(Y, f_1(X)) \neq \nabla^B_2 L(Y, f_2(X))\big) \\
&= \|k\|^2_\infty \Big(\mathrm{P}\big(\{Y - f_1(X) < 0\} \wedge \{Y - f_2(X) > 0\}\big) \\
&\qquad\qquad + \mathrm{P}\big(\{Y - f_2(X) < 0\} \wedge \{Y - f_1(X) > 0\}\big) \Big) \\
&= \|k\|^2_\infty \int_X \mathrm{P}\big(Y \in (f_2(x), f_1(x)) \,|\, x\big) + \mathrm{P}\big(Y \in (f_1(x), f_2(x)) \,|\, x\big) d\mathrm{P}_X(x) \\
&= \|k\|^2_\infty \int_X \mathrm{P}\big(Y \in (f_2(x), f_2(x) + [f_1(x) - f_2(x)]) \,|\, x\big) \\
&\qquad\qquad + \mathrm{P}\big(Y \in (f_1(x), f_1(x) + [f_2(x) - f_1(x)]) \,|\, x\big) d\mathrm{P}_X(x) \\
&\leq m \|k\|^2_\infty \int_X c_{\mathrm{P}} |f_1(x) - f_2(x)|^{1+\xi_{\mathrm{P}}} d\mathrm{P}_X(x) \\
&\leq m \|k\|^2_\infty c_{\mathrm{P}} \|f_1 - f_2\|^{1+\xi_{\mathrm{P}}}_\infty \\
&\leq m c_{\mathrm{P}} \|k\|^{3+\xi_{\mathrm{P}}}_\infty \|f_1 - f_2\|^{1+\xi_{\mathrm{P}}}_{\mathcal{H}} ,
\end{aligned}$$

where P_X denotes the marginal distribution of X. Similar calculations give that $B \leq m c_{\mathrm{Q}} \|k\|^{3+\xi_{\mathrm{Q}}}_\infty \|f_1 - f_2\|^{1+\xi_{\mathrm{Q}}}_{\mathcal{H}}$. We obtain $C = 0$, because $\nabla^B_{2,2} L(Y, f_{\mathrm{P},\lambda}(X)) = 0$. Hence, the term in (20) is less than or equal to

$$\begin{aligned}
& |1 - \varepsilon| m c_{\mathrm{P}} \|k\|^{3+\xi_{\mathrm{P}}}_\infty \|f_1 - f_2\|^{1+\xi_{\mathrm{P}}}_{\mathcal{H}} + |\varepsilon| m c_{\mathrm{Q}} \|k\|^{3+\xi_{\mathrm{Q}}}_\infty \|f_1 - f_2\|^{1+\xi_{\mathrm{Q}}}_{\mathcal{H}} \\
= \;& \big(|1 - \varepsilon| c_{\mathrm{P}} \|k\|^{\xi_{\mathrm{P}}}_\infty \|f_1 - f_2\|^{\xi_{\mathrm{P}}}_{\mathcal{H}} + |\varepsilon| c_{\mathrm{Q}} \|k\|^{\xi_{\mathrm{Q}}}_\infty \|f_1 - f_2\|^{\xi_{\mathrm{Q}}}_{\mathcal{H}} \big) m \|k\|^3_\infty \|f_1 - f_2\|_{\mathcal{H}} \\
\leq \;& \varepsilon^* \|f_1 - f_2\|_{\mathcal{H}} ,
\end{aligned}$$

where $\varepsilon^* = (|1 - \varepsilon| c_{\mathrm{P}} \|k\|^{\xi_{\mathrm{P}}}_\infty 2^{\xi_{\mathrm{P}}} \delta_1^{\xi_{\mathrm{P}}} + |\varepsilon| c_{\mathrm{Q}} \|k\|^{\xi_{\mathrm{Q}}}_\infty 2^{\xi_{\mathrm{Q}}} \delta_1^{\xi_{\mathrm{Q}}}) m \|k\|^3_\infty$.

A.2.5 $\underline{\varepsilon}$-INSENSITIVE LOSS

The proof for the $\underline{\varepsilon}$-insensitive loss $L = L_{\underline{\varepsilon}}$ is analogous to the proof for $L_{\tau-pin}$, but with $c_1 = 2$, $\mathfrak{D} = \{-\underline{\varepsilon}, +\underline{\varepsilon}\}$, $m = 4$ and thus we must consider 4 cases instead of 2 where $h(y, f_1(x), f_2(x)) \neq 0$.

A.2.6 HUBER LOSS

For Huber's loss function $L = L_{c-Huber}$ we have $|\nabla^B_{2,2}L(y,t)| \leq 1 := c_2$ and $h(y, f_1(x), f_2(x))$ is bounded by $c_1 = 2c$. Let us define

$$\begin{aligned} h^*(y, f_{\mathrm{P},\lambda}(x), f_1(x), f_2(x)) \quad := \quad & \nabla^B_2 L(y, f_1(x)) - \nabla^B_2 L(y, f_2(x)) \\ & -\nabla^B_{2,2}L(y, f_{\mathrm{P},\lambda}(x)) \cdot (f_1(x) - f_2(x)). \end{aligned}$$

Somewhat tedious calculations show that there are 8 cases where $h^*(y, f_{\mathrm{P},\lambda}(x), f_1(x), f_2(x)) \neq 0$ and 6 cases where $h^*(y, f_{\mathrm{P},\lambda}(x), f_1(x), f_2(x)) = 0$. In each of the 8 cases, $y - f_{\mathrm{P},\lambda}(x) \in \{-c, c\}$ and $|h^*(y, f_{\mathrm{P},\lambda}(x), f_1(x), f_2(x))| \leq |f_1(x) - f_2(x)|$. Due to symmetry of the Huber loss function, the calculations are quite similar, therefore we only consider here some cases.

If $-c < Y - f_{\mathrm{P},\lambda}(x) < c$, then $\nabla^B_{2,2}L(Y, f_{\mathrm{P},\lambda}(x)) \cdot (f_1(x) - f_2(x)) = f_1(x) - f_2(x)$ and for sufficiently small δ_1, $\nabla^B_2 L(Y, f_1(x)) = -(Y - f_1(x))$ and $\nabla^B_2 L(Y, f_2(x)) = -(Y - f_2(x))$. A small calculation shows that $h^*(Y, f_{\mathrm{P},\lambda}(x), f_1(x), f_2(x)) = 0$.

By straightforward calculations we also obtain that $h^*(Y, f_{\mathrm{P},\lambda}(x), f_1(x), f_2(x)) = 0$ for the following 5 cases:

1. $Y - f_{\mathrm{P},\lambda}(x) < -c$ or $Y - f_{\mathrm{P},\lambda}(x) > c$,
2. $Y - f_{\mathrm{P},\lambda}(x) = -c$ and $f_{\mathrm{P},\lambda}(x) > f_2(x) > f_1(x)$,
3. $Y - f_{\mathrm{P},\lambda}(x) = -c$ and $f_1(x) > f_2(x) > f_{\mathrm{P},\lambda}(x)$,
4. $Y - f_{\mathrm{P},\lambda}(x) = c$ and $f_{\mathrm{P},\lambda}(x) > f_2(x) > f_1(x)$,
5. $Y - f_{\mathrm{P},\lambda}(x) = c$ and $f_1(x) > f_2(x) > f_{\mathrm{P},\lambda}(x)$.

If $Y - f_{\mathrm{P},\lambda}(x) = -c$ and $f_1(x) > f_{\mathrm{P},\lambda}(x) > f_2(x)$, we get $\nabla^B_2 L(Y, f_1(X)) = c$, $\nabla^B_2 L(Y, f_2(x)) = -(Y - f_2(x))$ and $\nabla^B_{2,2}L(Y, f_{\mathrm{P},\lambda}(x)) \cdot (f_1(x) - f_2(x)) = 0$. Hence,

$$h^*(Y, f_{\mathrm{P},\lambda}(x), f_1(x), f_2(x)) = c + Y - f_2(x) = f_{\mathrm{P},\lambda}(x) - f_2(x) \neq 0,$$

since $f_2(x) < f_{\mathrm{P},\lambda}(x)$.

Analogously, some calculations show that $h^*(Y, f_{\mathrm{P},\lambda}(x), f_1(x), f_2(x)) \neq 0$ for the following 7 cases:

1. $Y - f_{\mathrm{P},\lambda}(x) = -c$ and $f_2(x) > f_{\mathrm{P},\lambda}(x) > f_1(x)$,
2. $Y - f_{\mathrm{P},\lambda}(x) = -c$ and $f_{\mathrm{P},\lambda}(x) > f_1(x) > f_2(x)$,
3. $Y - f_{\mathrm{P},\lambda}(x) = -c$ and $f_2(x) > f_1(x) > f_{\mathrm{P},\lambda}(x)$,
4. $Y - f_{\mathrm{P},\lambda}(x) = c$ and $f_1(x) > f_{\mathrm{P},\lambda}(x) > f_2(x)$,
5. $Y - f_{\mathrm{P},\lambda}(x) = c$ and $f_2(x) > f_{\mathrm{P},\lambda}(x) > f_1(x)$,
6. $Y - f_{\mathrm{P},\lambda}(x) = c$ and $f_{\mathrm{P},\lambda}(x) > f_1(x) > f_2(x)$,
7. $Y - f_{\mathrm{P},\lambda}(x) = c$ and $f_2(x) > f_1(x) > f_{\mathrm{P},\lambda}(x)$.

Using (12) in (20) we get for the term A in (20) that

$$\begin{aligned} A &= \big\|\mathbb{E}_{\mathrm{P}} h^*(Y, f_{\mathrm{P},\lambda}(X), f_1(X), f_2(X))\Phi(X)\big\|_{\mathcal{H}} \\ &\le \|k\|_\infty^2 \int |h^*(y, f_{\mathrm{P},\lambda}(x), f_1(x), f_2(x))| \mathbf{1}_{\{h^*\neq 0\}} d\mathrm{P}(x,y) \\ &\le \|k\|_\infty^2 \int |f_1(x) - f_2(x)| \mathrm{P}\big(Y \in \{-c + f_{\mathrm{P},\lambda}(x), c + f_{\mathrm{P},\lambda}(x)\}\big|x\big) d\mathrm{P}_X(x) = 0. \end{aligned}$$

Also $C = \Big\|\mathbb{E}_{\mathrm{P}} \nabla^B_{2,2} L(Y, f_{\mathrm{P},\lambda}(X)) \cdot (f_1(X) - f_2(X))\Phi(X)\Big\|_{\mathcal{H}} \le \kappa_2 \|k\|_\infty^3 \|f_1 - f_2\|_{\mathcal{H}}$. One can compute the analogous terms to A and C, say $A(\mathrm{Q})$ and $C(\mathrm{Q})$, respectively, where the integration is with respect to Q instead of P. Combining these expressions we obtain

$$\begin{aligned} B &= \big\|\mathbb{E}_{\mathrm{Q}}(\nabla^B_2 L(Y, f_1(X)) - \nabla^B_2 L(Y, f_2(X))) \cdot \Phi(X)\big\|_{\mathcal{H}} \\ &\le \mathbb{E}_{\mathrm{Q}}\big|\nabla^B_2 L(Y, f_1(X)) - \nabla^B_2 L(Y, f_2(X)) - \\ &\qquad \nabla^B_{2,2} L(Y, f_{\mathrm{P},\lambda}(X)) \cdot (f_1(X) - f_2(X))\big| \, |\Phi(X)| \\ &\quad + \mathbb{E}_{\mathrm{Q}}\big|\nabla^B_{2,2} L(Y, f_{\mathrm{P},\lambda}(X)) \cdot (f_1(X) - f_2(X))\big| \, |\Phi(X)| \\ &= A(\mathrm{Q}) + C(\mathrm{Q}) \le \kappa_2 \|k\|_\infty^3 \|f_1 - f_2\|_{\mathcal{H}}. \end{aligned}$$

Hence, the term in (20) is less than or equal to $\varepsilon^* \|f_1 - f_2\|_{\mathcal{H}}$ where $\varepsilon^* = 2|\varepsilon|\kappa_2 \|k\|_\infty^3$. This gives the assertion, because $|\varepsilon|$ can be chosen arbitrarily small. ■

Proof of Corollary 6. Both partial F-derivatives $\nabla^F_2 L_{log}(y,t) = 1 - 2\Lambda(y-t)$ and $\nabla^F_{2,2} L_{log}(y,t) = 2\Lambda(y-t)[1-\Lambda(y-t)]$ are clearly bounded, because $\Lambda(z) \in (0,1)$, $z \in \mathbb{R}$. We only have to show that $\nabla^B_2 G(0, f_{\mathrm{P},\lambda})$ is strong for $L = L_{log}$, that is that the term in (19) is bounded by $\varepsilon^* \|f_1 - f_2\|_{\mathcal{H}}$ for arbitrary chosen $\varepsilon^* > 0$. A Taylor expansion gives for arbitrary $y, t_1, t_2 \in \mathbb{R}$ that

$$\Lambda(y - t_2) = \Lambda(y - t_1) + (t_1 - t_2)\Lambda(y - t_1)\big(1 - \Lambda(y - t_1)\big) + O((t_1 - t_2)^2). \tag{23}$$

Combining (2), (21), (22), and (23) we obtain

$$\begin{aligned} &\big|\mathbb{E}_{\mathrm{P}}\big(\nabla^B_2 L(Y, f_1(X)) - \nabla^B_2 L(Y, f_2(X)) - \nabla^B_{2,2} L(Y, f_{\mathrm{P},\lambda}) \cdot (f_1(X) - f_2(X))\big)\Phi(X)\big| \\ \le\ & 2\|k\|_\infty^2 \mathbb{E}_{\mathrm{P}}\big|\Lambda(Y - f_2(X)) - \Lambda(Y - f_1(X)) \\ &\qquad - \Lambda(Y - f_{\mathrm{P},\lambda}(X))(1 - \Lambda(Y - f_{\mathrm{P},\lambda}(X))\big(f_1(X) - f_2(X)\big)\big| \\ \le\ & 2\|k\|_\infty^2 \mathbb{E}_{\mathrm{P}}\big|\big(f_1(X) - f_2(X)\big)\big[\Lambda(Y - f_1(X))(1 - \Lambda(Y - f_1(X))) \\ &\qquad - \Lambda(Y - f_{\mathrm{P},\lambda}(X))(1 - \Lambda(Y - f_{\mathrm{P},\lambda}(X)))\big] + O((f_1(X) - f_2(X))^2)\big| \\ \le\ & 2\|k\|_\infty^2 \mathbb{E}_{\mathrm{P}}\big(\|f_1 - f_2\|_\infty \big|\Lambda(Y - f_1(X))(1 - \Lambda(Y - f_1(X))) \\ &\qquad - \Lambda(Y - f_{\mathrm{P},\lambda}(X))(1 - \Lambda(Y - f_{\mathrm{P},\lambda}(X)))\big| + c_3 \|f_1 - f_2\|_\infty^2\big). \end{aligned} \tag{24}$$

A Taylor expansion around $f_{\mathrm{P},\lambda}(x)$ shows that $\Lambda(y - f_1(x))(1 - \Lambda(y - f_1(x)))$ equals

$$\begin{aligned} &\Lambda(y - f_{\mathrm{P},\lambda}(x))(1 - \Lambda(y - f_{\mathrm{P},\lambda}(x))) \\ +\ & \big(f_{\mathrm{P},\lambda}(x) - f_1(x)\big)\Lambda(y - f_{\mathrm{P},\lambda}(x))(1 - \Lambda(y - f_{\mathrm{P},\lambda}(x)))(1 - 2\Lambda(y - f_{\mathrm{P},\lambda}(x))) \\ +\ & O((f_1(x) - f_{\mathrm{P},\lambda}(x))^2). \end{aligned}$$

Using this expansion and (2), (21), and (22) it follows that the term in (24) is bounded by

$$\begin{aligned} & 2\,\|k\|_\infty^2\,\mathbb{E}_{\mathrm{P}}\big(\|f_1-f_2\|_\infty\,(\big\|f_1-f_{\mathrm{P},\lambda}\big\|_\infty/4+c_4\delta_1^2\,\|k\|_\infty^2)+c_3\,\|f_1-f_2\|_\infty^2\big) \\ \leq\quad & \|k\|_\infty^4\,\big(\delta_1/2+2c_4\delta_1^2\,\|k\|_\infty+4c_3\delta_1\big)\,\|f_1-f_2\|_{\mathcal{H}}\,. \qquad (25)\end{aligned}$$

Using the Lipschitz continuity of $\nabla_2^B L(y,\cdot)$, (2), and (23) we obtain

$$\begin{aligned} & |\varepsilon|\,\mathbb{E}_{\mathrm{Q-P}}\big|\big(\nabla_2^B L(Y,f_1(X))-\nabla_2^B L(Y,f_2(X))\big)\cdot\Phi(X)\big| \\ \leq\quad & |\varepsilon|\;\|k\|_\infty^2\,\mathbb{E}_{|\mathrm{Q-P}|}\big|\nabla_2^B L(Y,f_1(X))-\nabla_2^B L(Y,f_2(X))\big| \\ \leq\quad & |\varepsilon|\;\|k\|_\infty^3\,\|f_1-f_2\|_{\mathcal{H}}\,. \qquad (26)\end{aligned}$$

Combining (25) and (26) shows that the term in (19) is bounded by $\varepsilon^*\,\|f_1-f_2\|_{\mathcal{H}}$ with the positive constant $\varepsilon^*=\|k\|_\infty^3\,\big(\delta_1\,\|k\|_\infty/2+2c_4\delta_1^2\,\|k\|_\infty^2+4c_3\delta_1\,\|k\|_\infty+|\varepsilon|\big)$, where $\delta_1>0$ and $\varepsilon>0$ can be chosen as small as necessary. ■

References

V.I. Averbukh and O.G. Smolyanov. The theory of differentiation in linear topological spaces. *Russian Mathematical Surveys*, 22:201–258, 1967.

V.I. Averbukh and O.G. Smolyanov. The various definitions of the derivative in linear topological spaces. *Russian Mathematical Surveys*, 23:67–113, 1968.

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.

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

F.H. Clarke. *Optimization and Nonsmooth Analysis*. Wiley & Sons, New York, 1983.

L.T. Fernholz. *Von Mises Calculus for Statistical Functionals*, volume 19 of *Lecture Notes in Statistics*. Springer, New York, 1983.

F.R. Hampel. Contributions to the theory of robust estimation. Unpublished Ph.D. thesis, Dept. of Statistics, University of California, Berkeley, 1968.

F.R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69:383–393, 1974.

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

P.J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35:73–101, 1964.

C. Ip and J. Kyparisis. Local convergence of quasi-Newton methods for B-differentiable equations. *Mathematical Programming*, 56:71–89, 1992.

R. Koenker. *Quantile Regression*. Cambridge University Press, New York, 2005.

R.W. Koenker and G.W. Bassett. Regression quantiles. *Econometrica*, 46:33–50, 1978.

P.D. Lax. *Functional Analysis*. Wiley & Sons, New York, 2002.

R.A. Maronna, R.D. Martin, and V.J. Yohai. *Robust Statistics. Theory and Methods*. Wiley & Sons, New York, 2006.

H. Rademacher. Über partielle und totale Differenzierbarkeit. *Math. Ann.*, 79:254–269, 1919.

H. Rieder. *Robust Asymptotic Statistics*. Springer, New York, 1994.

S.M. Robinson. Local structure of feasible sets in nonlinear programming, Part III: Stability and sensitivity. *Mathematical Programming Study*, 30:45–66, 1987.

S.M. Robinson. An implicit-function theorem for a class of nonsmooth functions. *Mathematics of Operations Research*, 16:292–309, 1991.

B. Schölkopf and A.J. Smola. *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, Massachusetts, 2002.

I. Steinwart. How to compare different loss functions. *Constrained Approximation*, 26:225–287, 2007.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

I. Steinwart and A. Christmann. How SVMs can estimate quantiles and the median. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*. MIT Press, Cambridge, Massachusetts, 2008a.

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

I. Takeuchi, Q.V. Le, T.D. Sears, and A.J. Smola. Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7:1231–1264, 2006.

V.N. Vapnik. *Statistical Learning Theory*. Wiley & Sons, New York, 1998.