The Bouligand Influence Function: Checking Robustness of Support Vector Machines

Arnout Van Messem Andreas Christmann

7^{th} World Congress in Probability and Statistics, Singapore, July 14-19, 2008

Notation

Assumptions:

- $X \subseteq \mathbb{R}^d$, $Y \subseteq \mathbb{R}$, $X \neq \varnothing$, $Y \neq \varnothing$
- $D = ((x_1, y_1), \ldots, (x_n, y_n))$, $1 \leq i \leq n$
- (X_i, Y_i) i.i.d. $\sim \mathrm{P} \in \mathcal{M}_1$, P (totally) unknown

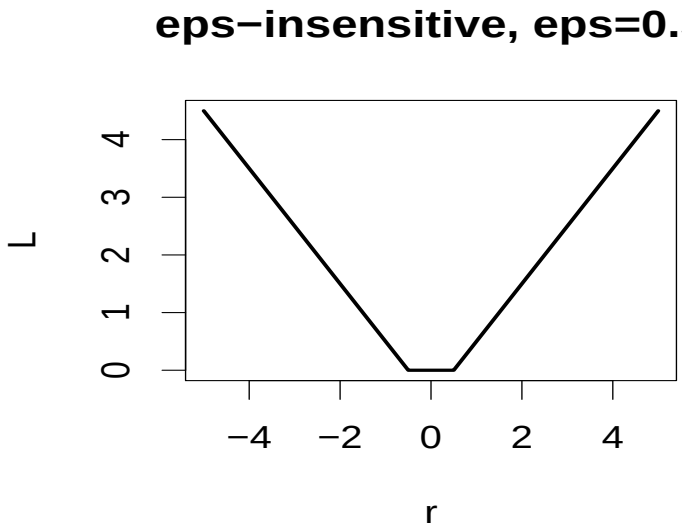
Aim:

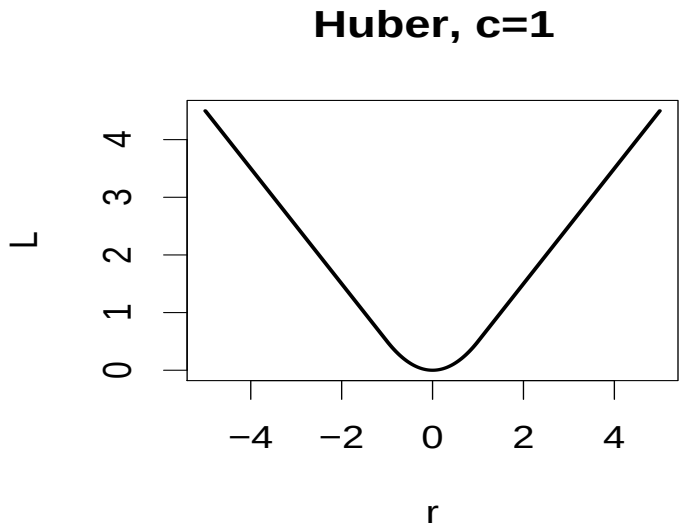
- $f(x_i) =$ quantity of interest of $\mathrm{P}_{Y_i|X_i=x_i}$
  e.g. conditional median for robust regression

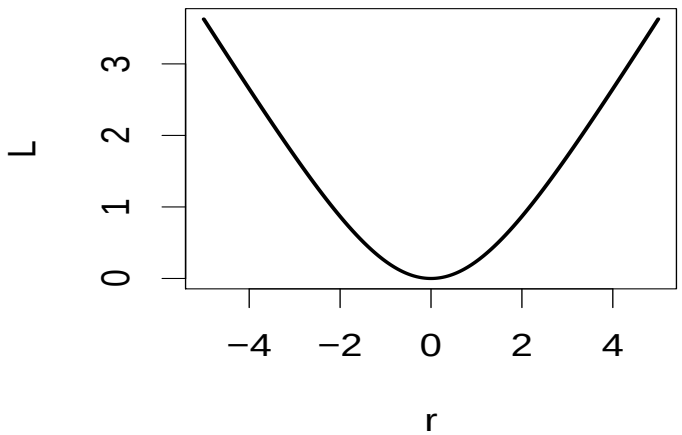
Assumption:

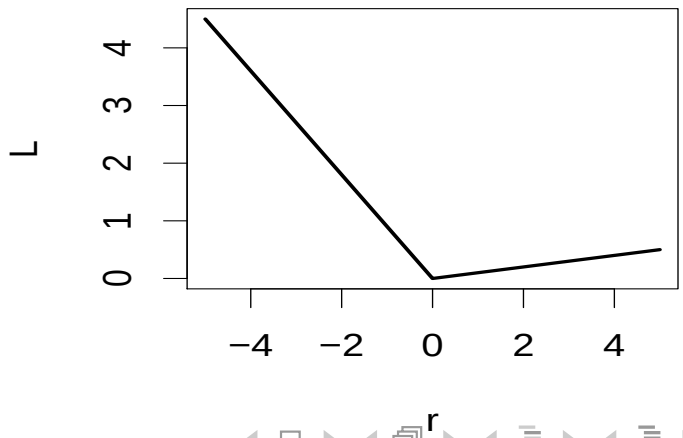
- Loss function: $L : Y \times \mathbb{R} \to [0, \infty)$, $L(y_i, f(x_i))$, convex

Lipschitz loss functions for regression









Kernel methods

- **Kernel**: $k : X \times X \to \mathbb{R}$, if $\exists$ Hilbert space $\mathcal{H}$ and $\Phi : X \to \mathcal{H}$ such that

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle, \quad \forall x, x' \in X$$

Reproducing Kernel Hilbert Space (RKHS)

$\mathcal{H}$ a Hilbert space of functions $f : X \to \mathbb{R}$. A reproducing kernel for $\mathcal{H}$ is a kernel k with

$$f(x) = \langle f, k(x, \cdot) \rangle \quad \forall\, f \in \mathcal{H}, \forall\, x \in X.$$

- **Canonical feature map**: $\Phi(x) = k(x, \cdot), \; x \in X$
- $k \rightleftarrows$ RKHS unique
- **Bounded**: $||k||_\infty := \sqrt{\sup_{x \in \mathcal{X}} k(x, x)} < \infty$
- **Gaussian RBF**: $k(x, x') = e^{-\gamma ||x - x'||_2^2}, \gamma > 0$

Support Vector Machines

Definition

SVM operator

$$S(\mathrm{P}) = f_{\mathrm{P},\lambda} = \arg\min_{f \in \mathcal{H}} \mathbb{E}_{\mathrm{P}} L\big(Y_i, f(X_i)\big) + \lambda \left\| f \right\|_{\mathcal{H}}^2 \ ,$$

where $\mathrm{P} \in \mathcal{M}_1$, $\mathcal{H}$ is a RKHS and $\lambda > 0$.

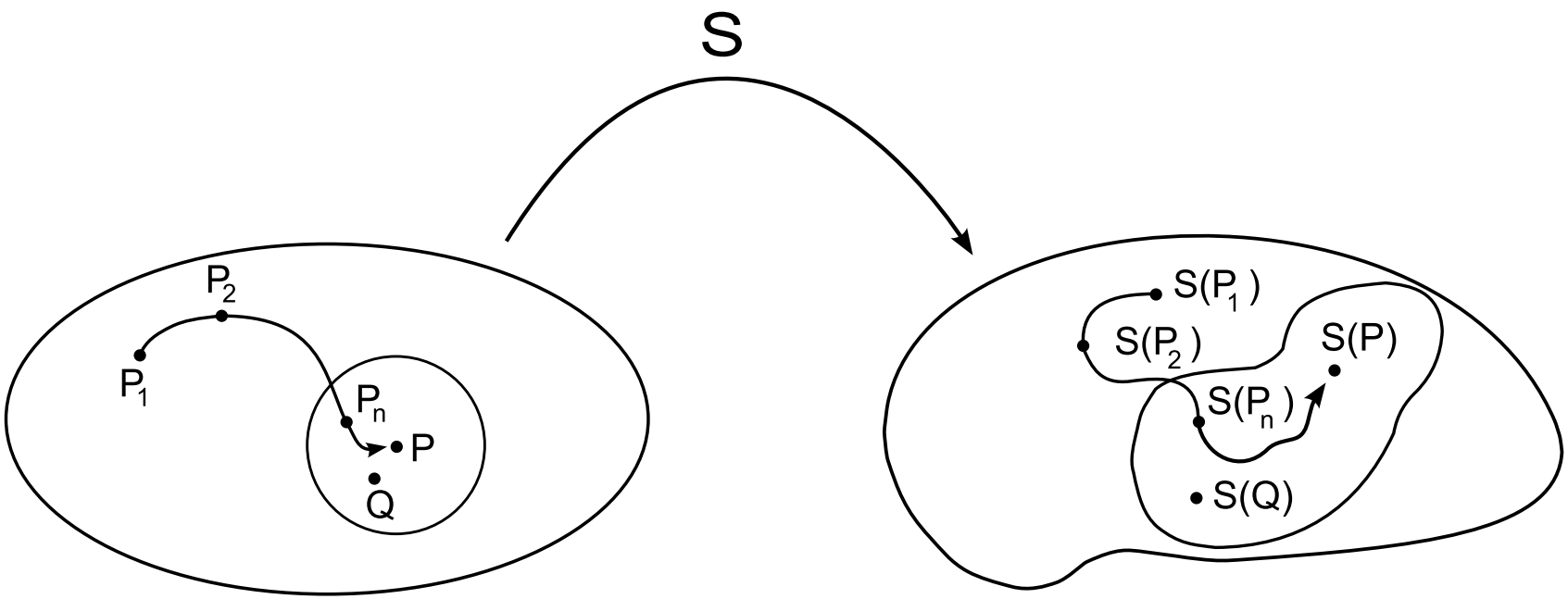
SVM estimator

$$S(\mathrm{P}_n) = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} L\big(Y_i, f(X_i)\big) + \lambda \left\| f \right\|_{\mathcal{H}}^2 \ ,$$

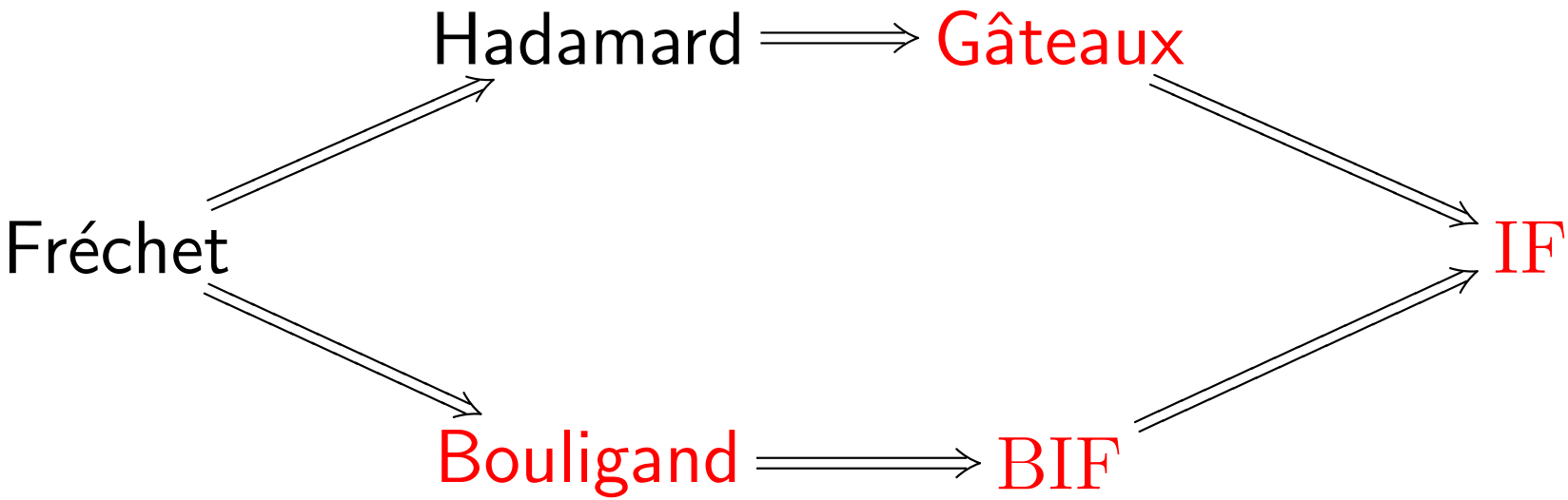
where $\mathrm{P}_n := \frac{1}{n} \sum_{i=1}^{n} \Delta_{(x_i, y_i)}$.

Robustness

1. What if (X_i, Y_i) i.i.d. $\sim \mathrm{P}$, $\mathrm{P} \in \mathcal{M}_1$ unknown is invalid?
2. What is the impact on $S(\mathrm{P}) = f_{\mathrm{P},\lambda}$?



Roadmap



Influence Function

Definition (Hampel, '68, Hampel et al. '86)

The influence function of S at P is given by

$$\mathrm{IF}(z; S, \mathrm{P}) := \lim_{\varepsilon \downarrow 0} \frac{S\big((1-\varepsilon)\mathrm{P} + \varepsilon \Delta_z\big) - S(\mathrm{P})}{\varepsilon},$$

in those z where this limit exists.

If Gâteaux derivative $\nabla^G(z; S, \mathrm{P})$ exists:
$\nabla^G = \mathrm{IF}$ and IF is linear and continuous

Goal: Bounded IF

Bouligand differentiability

Bouligand-derivative

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

$$f(x_0 + h) = f(x_0) + \nabla^B f(x_0)(h) + o(h)\,,$$

i.e.

$$\lim_{h \downarrow 0} \frac{\left\| f(x_0 + h) - f(x_0) - \nabla^B f(x_0)(h) \right\|_Z}{\|h\|_X} = 0.$$

Strong approximation

$f : X \to Z$ **strongly approximates** $F : X \times Y \to Z$ in x at (x_0, y_0) (notation: $f \approx_x F$) if $\forall \varepsilon > 0$ $\exists$ neighborhoods $\mathcal{N}(x_0)$ of x_0 and $\mathcal{N}(y_0)$ of y_0 such that $\forall x, x' \in \mathcal{N}(x_0), \forall y \in \mathcal{N}(y_0)$

$$\left\| \left(F(x,y) - f(x)\right) - \left(F(x',y) - f(x')\right) \right\|_Z \leq \varepsilon \left\| x - x' \right\|_X .$$

Strong Bouligand-derivative

$F : X \times Y \to Z$ has partial B-derivative $\nabla_1^B F(x_0, y_0)$ w.r.t. x at (x_0, y_0). Then $\nabla_1^B F(x_0, y_0)$ 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 (1991)

Bouligand Influence Function

BIF (C&VM '07)

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

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

If BIF exists and $\mathrm{Q} = \Delta_z$: IF exists and $\mathrm{BIF} = \mathrm{IF}$

Goal: Bounded BIF

Bouligand Influence Function

BIF (C&VM '07)

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

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

If BIF exists and $\mathrm{Q} = \Delta_z$: IF exists and $\mathrm{BIF} = \mathrm{IF}$

Goal: Bounded BIF

Main result

Assumptions

- $X \subset \mathbb{R}^d$, $Y \subset \mathbb{R}$ closed sets,
- $\mathcal{H}$ is RKHS with **bounded**, measurable kernel k,
- $f_{\mathrm{P},\lambda} \in \mathcal{H}$,
- $L : Y \times \mathbb{R} \to [0, \infty)$ **convex** and **Lipschitz continuous** w.r.t. the 2^{nd} argument with uniform Lipschitz constant $|L|_1 := \sup_{y \in Y} |L(y, \cdot)|_1 \in (0, \infty)$,
- L has measurable partial B-derivatives w.r.t. the 2^{nd} argument with $\kappa_1 := \sup_{y \in Y} \left\| \nabla_2^B L(y, \cdot) \right\|_\infty \in (0, \infty)$, $\kappa_2 := \sup_{y \in Y} \left\| \nabla_{2,2}^B L(y, \cdot) \right\|_\infty < \infty$,

Assumptions

- $\delta_1 > 0$, $\delta_2 > 0$,
- $\mathcal{N}_{\delta_1}(f_{\mathrm{P},\lambda}) := \{f \in \mathcal{H}; \|f - f_{\mathrm{P},\lambda}\|_{\mathcal{H}} < \delta_1\}$,
- $\lambda > \frac{1}{2}\kappa_2 \|\Phi\|_{\mathcal{H}}^3$, (Note: $\kappa_2 = 0$ for L_ϵ, L_τ)
- P, Q probability measures on $\left(X \times Y, \mathcal{B}(X \times Y)\right)$ with $\mathbb{E}_{\mathrm{P}}|Y| < \infty$ and $\mathbb{E}_{\mathrm{Q}}|Y| < \infty$.
- Define $G : (-\delta_2, \delta_2) \times \mathcal{N}_{\delta_1}(f(\mathrm{P}, \lambda)) \to \mathcal{H}$,

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

- $G(0, f_{\mathrm{P},\lambda}) = 0$ and $\nabla_2^B G(0, f_{\mathrm{P},\lambda})$ is **strong**.

Theorem (C&VM '07)

Then $\mathrm{BIF}(\mathrm{Q}; S, \mathrm{P})$ with $S(\mathrm{P}) := f_{\mathrm{P},\lambda}$

1. exists,
2. equals

$$T^{-1}\Big(\mathbb{E}_{\mathrm{P}}\nabla_2^B L(Y, f_{\mathrm{P},\lambda}(X))\Phi(X) - \mathbb{E}_{\mathrm{Q}}\nabla_2^B L(Y, f_{\mathrm{P},\lambda}(X))\Phi(X)\Big),$$

   where $T : \mathcal{H} \to \mathcal{H}$ with
   $T = 2\lambda\,\mathrm{id}_{\mathcal{H}} + \mathbb{E}_{\mathrm{P}}\nabla_{2,2}^B L(Y, f_{\mathrm{P},\lambda}(X))\langle\Phi(X), \cdot\rangle_{\mathcal{H}}\Phi(X)$, and
3. is bounded.

Sketch of proof

- Existence $f_{\mathrm{P},\lambda}$: convexity of L and penalizing term (Christmann & Steinwart, 2007)
- Define $G(\varepsilon, f)$
- $G(\varepsilon, f)$ fulfills the conditions for an implicit function theorem on B-derivatives (Robinson, 1991)

$$G(\varepsilon, f) = \frac{\partial \mathcal{R}^{reg}_{L,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda}}{\partial \mathcal{H}}(f) = \nabla_2^B \mathcal{R}^{reg}_{L,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda}(f)\,,\ \varepsilon \in [0,1]$$

Examples

The assumptions of the theorem are valid and thus $\mathrm{BIF}(\mathrm{Q}; S, \mathrm{P})$ exists and is bounded, if

ϵ-insensitive loss L_ϵ, pinball loss L_τ

$\forall \delta > 0 \, \exists$ positive constants ξ_P, ξ_Q, c_P, and c_Q such that $\forall t \in \mathbb{R}$ with $|t - f_{\mathrm{P},\lambda}(x)| \leq \delta \|k\|_\infty$ the following inequalities hold $\forall a \in [0, 2\delta\|k\|_\infty]$ and $\forall x \in X$:

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

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

The assumptions of the theorem are valid and thus $\mathrm{BIF}(\mathrm{Q}; S, \mathrm{P})$ exists and is bounded, if

Huber loss L_{Huber}

$$\begin{aligned}\forall x \in X\text{:}\\ &\mathrm{P}\big(Y \in \big\{f_{\mathrm{P},\lambda}(x) - c, f_{\mathrm{P},\lambda}(x) + c\big\} \,\big|\, x\big)\\ =\;&\mathrm{Q}\big(Y \in \big\{f_{\mathrm{P},\lambda}(x) - c, f_{\mathrm{P},\lambda}(x) + c\big\} \,\big|\, x\big)\\ =\;&0\,.\end{aligned}$$

$$\exists\,\mathrm{BIF},\ \mathrm{Q} = \Delta_z,\ z \notin \{f_{\mathrm{P},\lambda}(x) - c, f_{\mathrm{P},\lambda}(x) + c\} : \mathrm{BIF} = \mathrm{IF}$$

Logistic loss L_{log}

No special assumptions on the probabilities needed.

$$\exists\,\mathrm{BIF},\ \mathrm{Q} = \Delta_z : \mathrm{BIF} = \mathrm{IF}$$

Conclusions

Bouligand Influence Function

- If BIF exists and $\mathrm{Q} = \Delta_z$: $\mathrm{BIF} = \mathrm{IF}$
- B-derivative: pos. homog., chain rule, implicit funct. thm

Support Vector Machines

- Non-parametric and flexible
- Able to learn
- Robust: $\mathrm{BIF}(\mathrm{Q}; T, \mathrm{P})$ bounded for regression if $\nabla_2^B L$ and k bounded
- Applications: insurance tariffs, credit scoring in banks, fraud detection, data mining, genomics, ...

References

- Christmann & Van Messem (2008). Bouligand derivatives and robustness of support vector machines for regression. *Journal of Machine Learning Research*, **9**, 915-936
- Christmann & Steinwart (2007). Consistency and robustness of kernel based regression. *Bernoulli*, **13**, 799-819.
- Hampel (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, **69**, 383-393.
- Hampel, Ronchetti, Rousseeuw, Stahel (1986). Robust Statistics. Wiley.
- Robinson (1991). An implicit-function theorem for a class of non-smooth functions. *Mathematics of Operations Research*, **16**, 292-309.
- Schölkopf & Smola (2002). Learning with kernels. MIT Press.
- Vapnik (1998). Statistical learning theory. Wiley.

More on the theorem

For the proof of the theorem we showed:

i. 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.

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

iii. $\nabla_2^B G(0, f_{\mathrm{P},\lambda})\big(\mathcal{N}_{\delta_1}(f_{\mathrm{P},\lambda}) - f_{\mathrm{P},\lambda}\big)$ is a neighborhood of $0 \in \mathcal{H}$.

iv. $\delta\big(\nabla_2^B G(0, f_{\mathrm{P},\lambda}),\, \mathcal{N}_{\delta_1}(f_{\mathrm{P},\lambda}) - f_{\mathrm{P},\lambda}\big) \;=:\; d_0 > 0.$

v. For each $\xi > d_0^{-1}\chi$ there exist $\delta_3, \delta_4 > 0$, 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) $f^*(\cdot)$ 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 $(-\delta_4, \delta_4)$.

v.4) It holds

$$\nabla^B f^*(0)(u) = \left(\nabla_2^B G(0, f_{\mathrm{P},\lambda})\right)^{-1} \left(-\nabla_1^B G(0, f_{\mathrm{P},\lambda})(u)\right).$$