On Robustness Properties of Support Vector Machines for Heavy-Tailed Distributions

Arnout Van Messem Andreas Christmann

Vrije Universiteit Brussel

UNIVERSITÄT BAYREUTH

International Conference on Robust Statistics,
June 14-19, 2009, Parma, Italy

Introduction

Known

Support Vector Machines (SVMs) are **consistent** and **robust**, if based on Lipschitz continuous loss and bounded kernel.

Christmann & Van Messem '08
Steinwart & Christmann '08
Christmann & Steinwart '07

Question

Can the assumptions $f \in L_1(\mathrm{P}_X)$ and $\int |Y| \, d\mathrm{P} < \infty$ be weakened?

(both for regression and classification problems)

Notation

Assumptions:

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

Aim:

- $f(x) =$ quantity of interest
  e.g., conditional median for robust regression

Assumption:

- Loss function: $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$, $L(x, y, f(x))$
- Properties mentioned are w.r.t. 3^{rd} argument of L

Kernel methods

- **Kernel:** $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, if $\exists$ H-space $\mathcal{H}$, $\Phi : \mathcal{X} \to \mathcal{H}$:
$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}, \quad \forall x, x' \in \mathcal{X}$$
- **Canonical feature map:** $\Phi(x) = k(x, \cdot), \ x \in \mathcal{X}$

Reproducing Kernel Hilbert Space (RKHS)

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

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

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

Assumption: k measurable, e.g., continuous

Risk

Definition

Risk

$$\mathcal{R}_{L,\mathrm{P}}(f) = \mathbb{E}_{\mathrm{P}} L(X, Y, f(X))$$

Regularized risk

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

where $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, $\mathcal{H}$ a RKHS and $\lambda > 0$.

Support Vector Machines

Definition

SVM

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

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

Empirical version

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

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

Trick

Loss function $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ measurable

Definition

$L^\star : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ with

$$L^\star(x, y, t) := L(x, y, t) - L(x, y, 0).$$

Huber, 1967

$L^\star$ can be negative!

Properties

- L (strictly) convex, then $L^\star$ (strictly) convex.
- L Lipschitz continuous, then $L^\star$ Lipschitz continuous.

Reason

Conditions for finite risk

For L Lipschitz continuous

- $\mathbb{E}_{\mathrm{P}} L(X, Y, f(X)) < \infty$ if $f \in L_1(\mathrm{P}_X)$ and $Y \in L_1(\mathrm{P}_{Y|x})$.

$$\mathcal{R}_{L,\mathrm{P}}(f) \leq |L|_1 \left(\int_{\mathcal{X}} |f(x)| d\mathrm{P}_X(x) + \int_{\mathcal{X}} \int_{\mathcal{Y}} |y| d\mathrm{P}(y|x)\, d\mathrm{P}_X(x) \right)$$

- $\mathbb{E}_{\mathrm{P}} L^\star(X, Y, f(X)) < \infty$ if $f \in L_1(\mathrm{P}_X)$.

$$\mathcal{R}_{L,\mathrm{P}}(f) \leq |L|_1 \int_{\mathcal{X}} |f(x)|\, d\mathrm{P}_X(x)$$

Properties

- If L Lipschitz continuous, then

$$|\mathcal{R}_{L^\star,\mathrm{P}}(f)| \leq |L|_1 \mathbb{E}_{\mathrm{P}_X}|f(X)|.$$

$$|\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f)| \leq |L|_1 \mathbb{E}_{\mathrm{P}_X}|f(X)| + \lambda \|f\|^2_{\mathcal{H}}.$$

- If L Lipschitz continuous and $f_{L^\star,\mathrm{P},\lambda}$ exists, then

$$\|f_{L^\star,\mathrm{P},\lambda}\|^2_{\mathcal{H}} \leq \lambda^{-1} \min\Big\{|L|_1 \mathbb{E}_{\mathrm{P}_X}|f_{L^\star,\mathrm{P},\lambda}(X)|, \mathcal{R}_{L,\mathrm{P}}(0)\Big\}.$$

If additionally k is bounded, then $\|f_{L^\star,\mathrm{P},\lambda}\|_{\mathcal{H}} < \infty$.

Existence and Uniqueness of SVM solution

Existence and Uniqueness

The SVM $f_{L^\star,\mathrm{P},\lambda}$ exists and is unique, if

- L **Lipschitz continuous** and **convex**,
- $\mathcal{H}$ RKHS of a **bounded**, measurable kernel k,
- $\mathcal{R}_{L^\star,\mathrm{P}}(f) < \infty$ for some $f \in \mathcal{H}$,
- $\mathcal{R}_{L^\star,\mathrm{P}}(f) > -\infty$ for all $f \in \mathcal{H}$.

For more details, see talk by A.C. "Some recent results on support vector machines" on Monday

When is trick not needed?

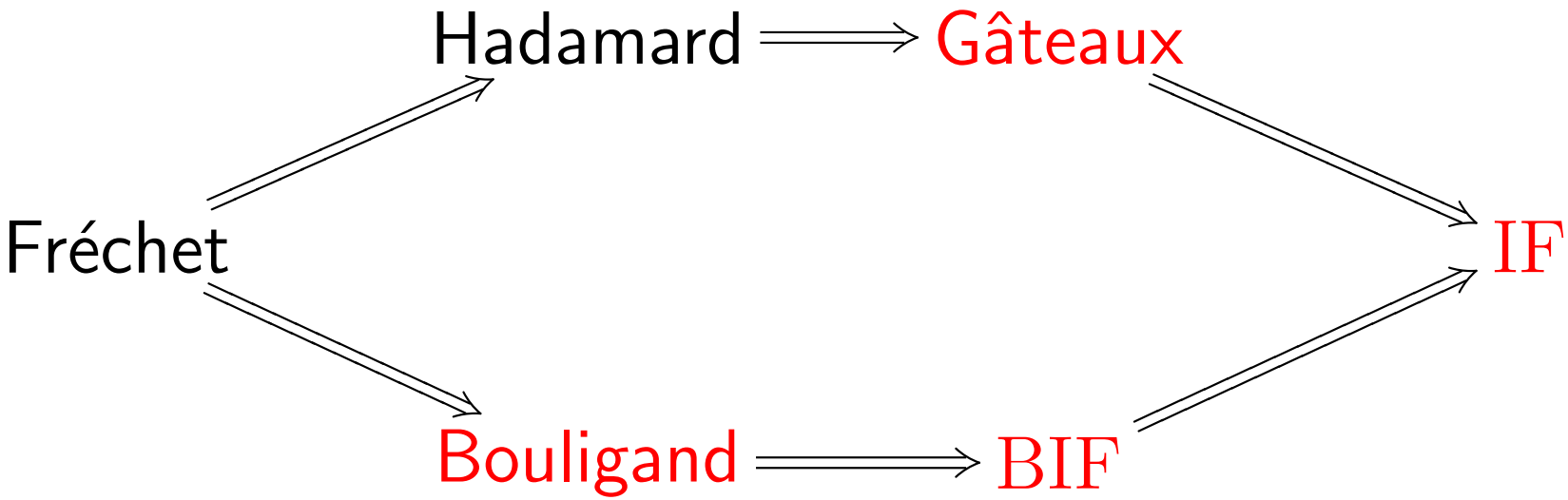
If $\mathcal{R}_{L,\mathrm{P}}(0) = \mathbb{E}_{\mathrm{P}} L(X, Y, 0) < \infty$, then

$$\begin{aligned}\mathcal{R}^{reg}_{L^\star,\mathrm{P},\lambda}(f_{L^\star,\mathrm{P},\lambda}) &:= \inf_{f\in\mathcal{H}} \mathbb{E}_{\mathrm{P}}\Big(L(X,Y,f(X)) - L(X,Y,0)\Big) + \lambda \left\|f\right\|^2_{\mathcal{H}} \\ &= \inf_{f\in\mathcal{H}} \Big[\mathbb{E}_{\mathrm{P}} L(X,Y,f(X)) + \lambda \left\|f\right\|^2_{\mathcal{H}}\Big] - \mathbb{E}_{\mathrm{P}} L(X,Y,0). \\ &= \mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f_{L,\mathrm{P},\lambda}) - \mathbb{E}_{\mathrm{P}} L(X,Y,0).\end{aligned}$$

Therefore

- $f_{L^\star,\mathrm{P},\lambda} = f_{L,\mathrm{P},\lambda}$, both exist and are unique

Derivatives and Influence Functions



Christmann & Van Messem (2008)

Notation: ∇^F, ∇^G, ∇^B, ∇_3^B, etc.

Property: $\nabla_3^F L^\star = \nabla_3^F L$, $\nabla_3^B L^\star = \nabla_3^B L$

Influence Function

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

The **influence function** (IF) of a function $S : \mathcal{M}_1 \to \mathcal{H}$ for a distribution 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 $z := (x, y) \in \mathcal{X} \times \mathcal{Y}$ where this limit exists.

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

Goal: Bounded IF

Problem: Loss function L often not Fréchet-differentiable

Bouligand Influence Function

Definition (C&VM '08)

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 Bouligand-derivative

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

(if it exists).

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

Goal: Bounded BIF

Result for IF

Assumptions

- $\mathcal{H}$ is RKHS with **bounded**, continuous kernel k
- L **convex** and **Lipschitz continuous**
- $\nabla_3^F L(x, y, \cdot)$ and $\nabla_{3,3}^F L(x, y, \cdot)$ continuous with
  $\kappa_1 := \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\|\nabla_3^F L(x, y, \cdot)\right\|_\infty \in (0, \infty)$,
  $\kappa_2 := \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\|\nabla_{3,3}^F L(x, y, \cdot)\right\|_\infty < \infty$

Theorem IF

Then $\mathrm{IF}(z; S, \mathrm{P})$ with $S(\mathrm{P}) := f_{L^\star,\mathrm{P},\lambda}$ and $z := (x, y)$

1. exists,
2. equals

$$\mathbb{E}_\mathrm{P} \nabla_3^F L^\star(X, Y, f_{L^\star,\mathrm{P},\lambda}(X)) T^{-1} \Phi(X)$$
$$-\nabla_3^F L^\star(x, y, f_{L^\star,\mathrm{P},\lambda}(x)) T^{-1} \Phi(x)\,,$$

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

Bounds for bias

Maxbias and IF

- $\mathcal{H}$ is separable RKHS with **bounded**, measurable kernel k
- L **convex** and **Lipschitz continuous**

Then, for all $\lambda > 0$, all $\varepsilon \in [0, 1]$ and all $\mathrm{P}, \mathrm{Q} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$

$$\left\| f_{L^\star,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}} - f_{L^\star,\mathrm{P},\lambda} \right\|_{\mathcal{H}} \leq c_{\mathrm{P},\mathrm{Q}}\varepsilon,$$

where $c_{\mathrm{P},\mathrm{Q}} = \lambda^{-1}\|k\|_\infty |L|_1 \|\mathrm{P} - \mathrm{Q}\|_{\mathcal{M}}$.

- $\mathrm{Q} = \delta_z$ with $z := (x, y)$
- $\mathrm{IF}(z; S, \mathrm{P})$ with $S(\mathrm{P}) := f_{L^\star,\mathrm{P},\lambda}$ exists

Then $$\|\mathrm{IF}(z; S, \mathrm{P})\|_{\mathcal{H}} \leq c_{\mathrm{P},\delta_z}.$$

Result for BIF

Assumptions

- $\mathcal{H}$ is RKHS with **bounded**, continuous kernel k
- L **convex** and **Lipschitz continuous** with $|L|_1 \in (0, \infty)$
- $\nabla_3^B L(x, y, \cdot)$ and $\nabla_{3,3}^B L(x, y, \cdot)$ measurable with
  $\kappa_1 := \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\|\nabla_3^B L(x, y, \cdot)\right\|_\infty \in (0, \infty)$,
  $\kappa_2 := \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\|\nabla_{3,3}^B L(x, y, \cdot)\right\|_\infty < \infty$

Assumptions

- $\delta_1 > 0$, $\delta_2 > 0$
- $\mathcal{N}_{\delta_1}(f_{L^\star,\mathrm{P},\lambda}) := \{f \in \mathcal{H} : \|f - f_{L^\star,\mathrm{P},\lambda}\|_{\mathcal{H}} < \delta_1\}$
- $\lambda > \frac{1}{2}\kappa_2 \|k\|_\infty^3$ ($\kappa_2 = 0$ for eps-insensitive and pinball)
- $\mathrm{P} \neq \mathrm{Q}$, probability measures on $\mathcal{X} \times \mathcal{Y}$
- Define $G : (-\delta_2, \delta_2) \times \mathcal{N}_{\delta_1}(f_{L^\star,\mathrm{P},\lambda}) \to \mathcal{H}$,

$$G(\varepsilon, f) := 2\lambda f + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}} \nabla_3^B L^\star(X, Y, f(X))\Phi(X)$$

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

Theorem BIF

Then $\mathrm{BIF}(\mathrm{Q}; S, \mathrm{P})$ with $S(\mathrm{P}) := f_{L^\star,\mathrm{P},\lambda}$ and $\mathrm{Q} \neq \mathrm{P} \in \mathcal{M}_1$

1. exists,
2. equals

$$T^{-1}\Big(\mathbb{E}_\mathrm{P}\nabla_3^B L^\star(X,Y,f_{L^\star,\mathrm{P},\lambda}(X))\Phi(X) - \mathbb{E}_\mathrm{Q}\nabla_3^B L^\star(X,Y,f_{L^\star,\mathrm{P},\lambda}(X))\Phi(X)\Big),$$

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

Conclusions

SVMs based on $L^\star(x, y, t) := L(x, y, t) - L(x, y, 0)$

1. Weaker assumption on P: only $f \in L_1(\mathrm{P}_X)$ is needed
2. Existence and uniqueness of $f_{L^\star,\mathrm{P},\lambda}$
3. If $\mathbb{E}_\mathrm{P} L(X, Y, 0) < \infty$ then $f_{L^\star,\mathrm{P},\lambda} = f_{L,\mathrm{P},\lambda}$
4. Robustness
   - Existence of IF and BIF
   - $\mathrm{IF}(\mathrm{Q}; S, \mathrm{P})$ bounded if $\nabla_3^F L$, $\nabla_{3,3}^F L$ and k continuous and bounded
   - $\mathrm{BIF}(\mathrm{Q}; S, \mathrm{P})$ bounded if $\nabla_3^B L$, $\nabla_{3,3}^B L$ measurable and bounded as well as k continuous and bounded
   - Bounds for bias

References

- Christmann, Van Messem & Steinwart (2009). Tentatively accepted.
- Christmann & Van Messem (2008). *Journal of Machine Learning Research*, **9**, 915-936.
- Steinwart & Christmann (2008). *Support Vector Machines.* Springer, New York.
- Christmann & Steinwart (2007). *Bernoulli*, **13**, 799-819.
- Hampel (1974). *J. Amer. Statist. Assoc.*, **69**, 383-393.
- Huber (1967). *Proceedings of the* 5^{th} *Berkeley Symposium*.
- Koenker (2005). *Quantile regression.* Cambridge University Press.
- Schölkopf & Smola (2002). *Learning with kernels.* MIT Press.
- Vapnik (1998). *Statistical learning theory.* Wiley.

Sketch of proof for IF

- $G(\varepsilon, f) := 2\lambda f + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\Delta_z} \nabla_2^F L^*(Y, f(X))\Phi(X)$
- $G(\varepsilon, f) = \nabla_2^F \mathcal{R}^{reg}_{L^*,(1-\varepsilon)\mathrm{P}+\varepsilon\Delta_z,\lambda}(f), \qquad \varepsilon \in [0,1]$
- $G(\varepsilon, f)$ fulfills conditions of a standard implicit function theorem on Banach spaces

Sketch: Proof for IF

For the proof of the theorem about the IF we showed:

i. $G(0, f) = 0 \Leftrightarrow f = f_{L^*,\mathrm{P},\lambda}$.

ii. G continuously F-differentiable.

iii. $\frac{\partial G}{\partial \mathcal{H}}(0, f_{L^*,\mathrm{P},\lambda})$ invertible.

iv. Then there exist $\delta > 0$, a neighborhood $\mathcal{N}_\delta(f_{L^*,\mathrm{P},\lambda}) := \{f \in \mathcal{H}; \|f - f_{L^*,\mathrm{P},\lambda}\|_{\mathcal{H}} < \delta\}$, and a function $f^* : (-\delta, \delta) \to \mathcal{N}_\delta(f_{L^*,\mathrm{P},\lambda})$ satisfying

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

iv.2) It holds

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

Sketch of proof for BIF

- $\nabla_2^B L^\star(Y, f(X)) = \nabla_2^B L(Y, f(X))$ hence

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

- $G(\varepsilon, f) = \nabla_2^B \mathcal{R}^{reg}_{L^\star,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda}(f), \qquad \varepsilon \in [0, 1]$
- $G(\varepsilon, f)$ fulfills the conditions of Robinson's (1991) implicit function theorem on Bouligand-derivatives for non-smooth functions in Banach or normed linear spaces

$\Rightarrow$ Rest of proof uses same arguments as Christmann & Van Messem (2008).

Sketch: Proof for BIF

For the proof of the theorem about the BIF we showed:

i. For some χ and each $f \in \mathcal{N}_{\delta_1}(f_{L^*,\mathrm{P},\lambda})$, $G(\cdot\,, f)$ 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 $(0, f_{L^*,\mathrm{P},\lambda})$.

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

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

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

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