On Consistency and Robustness of Support Vector Machines

Andreas Christmann

joint work: Ingo Steinwart (Los Alamos National Lab, USA)

ISI, August 16 - 22, 2009, Durban, South Africa
What are Support Vector Machines?

Support Vector Machines (SVMs) are a kind of penalized non-parametric M-estimators for functions

- applied in many areas, e.g.
  - pattern recognition
  - regression, quantile regression
  - density estimation
  - classification of histograms, images, empirical measures
- even successful for complex, high-dimensional, large data sets
Goals of the talk?

...to convince you that Support Vector Machines

- are consistent and robust solutions of a well-posed problem in Hadamard’s sense
- can “learn” from any totally unknown distribution
- can efficiently solve even non-standard real-life problems

“... a well-posed mathematical problem should have the property that there exists a unique solution that additionally depends continuously on the data”

Hadamard (1902)
Example: SVM for quantile regression

Quantiles $\alpha = 0.95, 0.50, 0.05$

(LIDAR data: $n = 221, d = 1$)
Examples: SVMs for complex and large data sets

SVMs are successfully applied to **general input spaces**, input values are distributions, images, text, websites, ... goal: classification or regression

<table>
<thead>
<tr>
<th></th>
<th>data set</th>
<th>$n$</th>
<th>$d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hein &amp; Bousquet '04</td>
<td>Reuters</td>
<td>8,065</td>
<td>18,635</td>
</tr>
<tr>
<td></td>
<td>WebKB</td>
<td>4,198</td>
<td>24,212</td>
</tr>
<tr>
<td>CHR '05</td>
<td>Corel images</td>
<td>1,400</td>
<td>4,096</td>
</tr>
<tr>
<td>Joachims '06</td>
<td>insurance</td>
<td>2,200,000</td>
<td>108</td>
</tr>
<tr>
<td></td>
<td>Reuters CCAT</td>
<td>804,414</td>
<td>47,236</td>
</tr>
<tr>
<td></td>
<td>Arxiv astro-ph</td>
<td>522,911</td>
<td>99,757</td>
</tr>
</tbody>
</table>
Support Vector Machines

Def. SVM

\[ f_{L,P,\lambda} := \arg \inf_{f \in \mathcal{H}} \mathbb{E}_PL(X,Y,f(X)) + \lambda \|f\|_{\mathcal{H}}^2 \]

- data set \( D = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n \)
- \((X_i, Y_i) \sim P\) are i.i.d. random variables
- \(Y_i|X_i\) depends on an unknown function \( f: \mathcal{X} \to \mathbb{R} \)
- \(\mathcal{H}\) reproducing kernel Hilbert space (RKHS) of \( f: \mathcal{X} \to \mathbb{R} \)
- \(k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}\) kernel used to define \(\mathcal{H}\)
- \(L: \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}\) loss function, usually \(L(x, y, t) \geq 0\)
- \(\lambda > 0\) regularization parameter
- \(f_{L,D,\lambda}\), where \(D\) is empirical distribution of \(D\)
Support Vector Machines

**Definitions**

- **Risk**: $\mathcal{R}_{L,P}(f) = \mathbb{E}_P L(X, Y, f(X))$
- **Bayes Risk**: $\mathcal{R}^*_L,P = \inf_{f: \mathcal{X} \to \mathbb{R}} \text{measurable} \mathcal{R}_{L,P}(f)$
- **Bayes Function**: $f^*_L,P = \arg \inf_{f: \mathcal{X} \to \mathbb{R}} \text{measurable} \mathcal{R}_{L,P}(f)$

Questions

Under which conditions on $X$, $Y$, $L$, $H$, and $k$ we have:

- Existence, uniqueness, representation, computation of $f_{L,D,\lambda}$
- Universal consistency to Bayes risk/function (i.e., $\forall P \exists \mathcal{R}_{L,D,\lambda} \to P \mathcal{R}^*_L,P$)
- $f_{L,D,\lambda} \to P f^*_L,P$ for $|D| = n \to \infty$
- Robustness of $f_{L,D,\lambda}$ and $f_{L,D,\lambda}$
Support Vector Machines

Definitions

risk \( R_{L,P}(f) \) \( \mathbb{E}_P L(X,Y,f(X)) \)
Bayes risk \( R_{L,P}^* \) \( \inf_{f: \mathcal{X} \to \mathbb{R} \text{ measurable}} R_{L,P}(f) \)
Bayes function \( f_{L,P}^* \) \( \arg \inf_{f: \mathcal{X} \to \mathbb{R} \text{ measurable}} R_{L,P}(f) \)

Questions

Under which conditions on \( \mathcal{X}, \mathcal{Y}, L, \mathcal{H}, \) and \( k \) we have:

- \( f_{L,P,\lambda} \): existence, uniqueness, representation, computation
- universal consistency to Bayes risk/function (i.e., \( \forall P \))

\[ R_{L,P}(f_{L,D,\lambda}) \xrightarrow{P} R_{L,P}^* \quad \text{for} \quad |D| = n \to \infty \]
\[ f_{L,D,\lambda} \xrightarrow{P} f_{L,P}^* \quad \text{for} \quad |D| = n \to \infty \]

- robustness of \( f_{L,P,\lambda} \) and \( f_{L,D,\lambda} \) ?
Support Vector Machines

Situation:

- $\mathcal{X}$ input space: complete separable metric space
- $\mathcal{Y} \subset \mathbb{R}$ output space: non-empty and closed
- $\mathbb{P}$ totally unknown distribution on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X} \times \mathcal{Y}))$
  hence: marginal $\mathbb{P}_X$ and cond. prob. $\mathbb{P}(y|x)$
- $(X_i, Y_i) \sim \mathbb{P}$, $i = 1, \ldots, n$
- $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ loss function
- shifted loss function: $L^*(x, y, t) := L(x, y, t) - L(x, y, 0)$

General Assumptions (can often be weakened)

1. $L$ (or $L^*$) convex and Lipschitz continuous w.r.t. 3rd arg.
2. $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ bounded, measurable kernel with
3. separable reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$

Notation: $\Phi : \mathcal{X} \to \mathcal{H}$ canonical feature map, i.e., $\Phi(x) := k(\cdot, x)$
Support Vector Machines

Def. SVMs based on shifted loss

\[ f_{L^\star,P,\lambda} := \arg \inf_{f \in \mathcal{H}} E_P L^\star(X, Y, f(X)) + \lambda \|f\|_H^2 \]

Conditions of finite risk

1. \( \mathcal{R}_{L,P}(f) < \infty \), if \( f \in L_1(P_X) \) and \( E_P|Y| < \infty \).
2. \( \mathcal{R}_{L^\star,P}(f) < \infty \), if \( f \in L_1(P_X) \).

Equality of SVMs

1. If \( f_{L,P,\lambda} \) exists, then \( f_{L^\star,P,\lambda} = f_{L,P,\lambda} \).
2. For any fixed data set: \( f_{L^\star,D,\lambda} = f_{L,D,\lambda} \).
Existence and Uniqueness

Thm: Uniqueness

CHR, Van Messem, Steinwart ’09

- Let $R_{L^*,P}(f) < \infty$ for some $f \in \mathcal{H}$ and $R_{L^*,P}(f) > -\infty$ for all $f \in \mathcal{H}$
- all $f \in L_1(P_X)$.

Then, for all $\lambda > 0$, there exists at most one SVM $f_{L^*,P,\lambda}$. 
Existence and Uniqueness

**Thm: Uniqueness**  
CHR, Van Messem, Steinwart '09

- Let $R_{L^*,P}(f) < \infty$ for some $f \in \mathcal{H}$ and $R_{L^*,P}(f) > -\infty$ for all $f \in \mathcal{H}$
- all $f \in L_1(P_X)$.

Then, for all $\lambda > 0$, there exists at most one SVM $f_{L^*,P,\lambda}$.

**Thm: Existence**  
CHR, Van Messem, Steinwart '09

There exists an SVM $f_{L^*,P,\lambda}$ for all $P \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and for all $\lambda > 0$. 

Andreas Christmann  
http://www.stoch.uni-bayreuth.de  
Some recent results on SVMs
Representation of SVMs

**Thm**

CHR, Van Messem, Steinwart '09

For all $P \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and for all $\lambda > 0$, there exists an $h \in \mathcal{L}_\infty(P)$ with

$$h(x, y) \in \partial L^*(x, y, f_{L^*,P,\lambda}(x)) \quad \forall (x, y)$$

$$\|h\|_\infty \leq |L^*|_1 = |L|_1$$

$$f_{L^*,P,\lambda} = -\frac{2}{\lambda} \mathbb{E}_P h\Phi$$

$$\|f_{L^*,P,\lambda} - f_{L^*,Q,\lambda}\|_\mathcal{H} \leq \frac{1}{\lambda} \|\mathbb{E}_P h\Phi - \mathbb{E}_Q h\Phi\|_\mathcal{H}, \forall Q \in \mathcal{M}_1$$
SVMs are universally consistent

**Thm**
CHR, Van Messem, Steinwart '09

- Let $\mathcal{H}$ dense in $L_1(\mu)$ for all distributions $\mu$ on $\mathcal{X}$,
- $(\lambda_n)$ sequence of strictly positive numbers with $\lambda_n \to 0$.

Then, for all $P \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and all $D$ with $|D| = n$,

1. if $\lambda_n^2 n \to \infty$, then $\mathcal{R}_{L^*,P}(f_{L^*,D},\lambda_n) \xrightarrow{P} \mathcal{R}_{L^*,P}^*$.
2. if $\lambda_n^2 + \delta n \to \infty$ for some $\delta \in (0, \infty)$, then $\mathcal{R}_{L^*,P}(f_{L^*,D},\lambda_n) \xrightarrow{\text{a.s.}} \mathcal{R}_{L^*,P}^*$.
3. for $L = L_\tau$ pinball loss:
   if $\lambda_n^2 n \to \infty$, then $d(f_{L^*,D},\lambda_n, f_{L^*,P}^*) \to 0$.

$d$ is a metric describing convergence in probability.
SVMs: universally consistent ⇒ localizable
SVMs: universally consistent $\Rightarrow$ localizable
SVMs: universally consistent $\Rightarrow$ localizable
SVMs: universally consistent $\Rightarrow$ localizable

\[ x^* = 10, \]
\[ \delta = 0.5 \]
\[ B_\delta(x^*) = (9.5, 10.5) \]
SVMs: universally consistent ⇒ localizable

**Prop**

Assume SVM universally consistent, $L$ Lipschitz continuous and convex, $k$ bounded. Then:

1. Fix $n, m \in \mathbb{N}$. $\forall |D_n| = n, |D_m| = m$:
   $$\|f_{L^*,D_n,\lambda_n} - f_{L^*,D_m,\lambda_m}\|_{\infty} \leq \frac{1}{2}\|k\|_{\infty}^3 \left(\frac{1}{\lambda_n} + \frac{1}{\lambda_m}\right)|L|_1 < \infty.$$

2. $\forall P \in M_1(\mathcal{X} \times \mathcal{Y}) : \mathcal{R}_{L^*,P}(f_{L^*,D,\lambda_n}) \xrightarrow{P} \mathcal{R}_{L^*,P}^*.$

3. Fix any $x^* \in \mathcal{X}$. Let $B_\delta(x^*)$ be $\delta$-ball of $x^*$. Define $P_\delta := \{P \in M_1(\mathcal{X} \times \mathcal{Y}) : \text{supp}(P_X) \subset B_\delta(x^*)\}.$

4. $\forall P \in P_\delta : \mathcal{R}_{L^*,P}(f_{L^*,D,\lambda_n}) \xrightarrow{P} \mathcal{R}_{L^*,P}^*.$

The impact of data points $(x_i, y_i)$ from $D$ with $x_i$ outside of the neighborhood $B_\delta(x^*)$ is asymptotically negligible!

see also: Zakai & Ritov (2009, JMLR)

Andreas Christmann  http://www.stoch.uni-bayreuth.de

Some recent results on SVMs
Universal Kernels for Distributions

**Goal:** Classify histograms, coloured images, distributions, . . .

**Known:** SVMs based on universal kernels are consistent!
(under weak assumptions, see Steinwart ’05, Steinwart & CHR ’08)

**Thm**

CHR & Steinwart ’09

- Let $\Omega \subset \mathbb{R}^d$ compact,
- Prohorov metric $d_{\text{Pro}}$ on $\mathcal{X} := \mathcal{M}_1(\Omega)$ with $\mathcal{B}(\Omega)$,
- $\mathcal{H}$ separable Hilbert space (e.g., $\mathcal{H} = \ell_2$),
- $\rho : \mathcal{X} \rightarrow \mathcal{H}$ continuous and injective.

Then, for all $\gamma \in (0, \infty)$, $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$,

$$k(P_1, P_2) := \exp\left(-\|\rho(P_1) - \rho(P_2)\|_{\mathcal{H}}^2 / \gamma^2\right)$$

is a bounded, continuous, and universal kernel, i.e., the RKHS of $k$ is dense in $C(\mathcal{X}, d_{\text{Pro}})$. 

Andreas Christmann  
http://www.stoch.uni-bayreuth.de  
Some recent results on SVMs
Universal kernels for Distributions

Examples

Let $\Omega \subset \mathbb{R}^d$ be compact and $\gamma > 0$. The following kernels on $\mathcal{X} := \mathcal{M}_1(\Omega, \mathcal{B}(\Omega))$ are universal.

1. Kernels based on characteristic function $\widehat{P}$:

$$k_{F,\gamma}(P_1, P_2) := \exp\left(-\|\widehat{P}_1 - \widehat{P}_2\|_{L_2(\mathcal{X}^d)}/\gamma^2\right)$$
Universal kernels for Distributions

Examples

Let $\Omega \subset \mathbb{R}^d$ be compact and $\gamma > 0$. The following kernels on $X := \mathcal{M}_1(\Omega, \mathcal{B}(\Omega))$ are universal.

1. Kernels based on characteristic function $\widehat{P}$:

$$k_{F, \gamma}(P_1, P_2) := \exp \left( -\| \widehat{P_1} - \widehat{P_2} \|^2_{L_2(\lambda^d)} / \gamma^2 \right)$$

2. Kernels based on continuous, bounded kernel $k_\Omega$ on $\Omega$ with $\rho : \mathcal{M}_1(\Omega) \to \mathcal{H}$, $\rho(P) := \mathbb{E}_P \Phi_\Omega$, injective:

$$k_\gamma(P_1, P_2) := \exp \left( -\| \mathbb{E}_{P_1} \Phi_\Omega - \mathbb{E}_{P_2} \Phi_\Omega \|^2_{H_\Omega} / \gamma^2 \right).$$
SVMs are qualitatively robust

F.R. Hampel: A sequence \((S_n)_{n \in \mathbb{N}}\) of \(\mathcal{H}\)-valued estimates is called **qualitatively robust** for \(P \in \mathcal{M}_1\), if, \(\forall \varepsilon > 0\), \(\exists \delta > 0\)

\[
d_{\text{Pro}}(P, Q) < \delta \iff d_{\text{Pro}}(\mathcal{L}_P(S_n), \mathcal{L}_Q(S_n)) < \varepsilon, \ \forall n \in \mathbb{N}.
\]

**Thm**

Assume \(\partial L^*\) continuous and continuous kernel.

Then, for **all** \(P \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})\) and for all \(\lambda > 0\),

1. \(S : \mathcal{M}_1(\mathcal{X} \times \mathcal{Y}) \to \mathcal{H}, \ S(P) = f_{L^*, P, \lambda}\) is **continuous**.
2. \((S(D_n))_{n \in \mathbb{N}}, \ S(D_n) := f_{L^*, D_n, \lambda},\) is **qualitatively robust**.
Influence Function of SVM is bounded

**Thm**

Assume \( \nabla^F_3 L(x, y, \cdot) \) and \( \nabla^F_{3,3} L(x, y, \cdot) \) are continuous with \( \sup_{(x,y)} \| \nabla^F_3 L(x, y, \cdot) \|_{\infty} \in (0, \infty), \sup_{(x,y)} \| \nabla^F_{3,3} L(x, y, \cdot) \|_{\infty} < \infty \).

Then the influence function of \( S(P) := f_{L^*,P,\lambda} \) at \( z := (x, y) \) exists, is **bounded**, and equals

\[
\text{IF}(z; S, P) := \lim_{\varepsilon \downarrow 0} \frac{S((1 - \varepsilon)P + \varepsilon \delta_z) - S(P)}{\varepsilon} = \mathbb{E}_P \nabla^F_3 L^*(X, Y, f_{L^*,P,\lambda}(X)) T^{-1} \Phi(X) - \nabla^F_{3,3} L^*(x, y, f_{L^*,P,\lambda}(x)) T^{-1} \Phi(x),
\]

where \( T : \mathcal{H} \rightarrow \mathcal{H} \) with

\[
T(\cdot) := 2\lambda \text{id}_{\mathcal{H}}(\cdot) + \mathbb{E}_P \nabla^F_{3,3} L^*(X, Y, f_{L^*,P,\lambda}(X)) \langle \Phi(X), \cdot \rangle_{\mathcal{H}} \Phi(X).
\]
Bouligand Influence Function of SVM

- There is a similar result for the **Bouligand influence function** of SVMs for non-Fréchet differentiable $L$ (CHR & Van Messem '08)
- Special cases
  - for regression: $\epsilon$-insensitive loss, Huber’s loss
  - for quantile regression: pinball loss
Bias of SVM increases at most linearly

Bounds for Maxbias

For all $\lambda > 0$, all $\varepsilon \in [0, 1]$, and all $P, Q \in M_1(X \times Y)$:

$$\|f_{L^*(1-\varepsilon)P+\varepsilon Q} - f_{L^*,P,\lambda}\|_H \leq c_{P,Q} \cdot \varepsilon,$$

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

CHR & Van Messem '08
Computation of SVMs

Let $D$ be any data set. Then
- there exists $\alpha := (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$ such that
  \[ f_{L,D,\lambda}(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i). \]
  “weighted sum of kernel terms”
- $\alpha$ is solution of convex problem with side conditions; for some $L$ even quadratic problem (use Lagrangian approach)
- efficient algorithms for $\alpha$ and hence for $f_{L,D,\lambda}$ exist
  - SVM$_{\text{light}}$, SVM$_{\text{struct}}$: Joachims (1999)
  - LIBSVM: Chang & Lin (2004), Fan et al. (2005)
  - mySVM, myKLR: Rüping (2000, 2005)
  - in R: function svm in package e1071
- if $\alpha_i \neq 0$ then $(x_i, y_i)$ is called support vector (SV)
- number of SVs is crucial for fast numerical evaluation
Stochastic bounds for number of SVs

Thm

Steinwart & CHR '09

- $L_\epsilon(x, y, t) := \max \{0, |y - t| - \epsilon\}$ $\epsilon$-insensitive loss, $\epsilon > 0$,
- $\mathcal{R}_{L_\epsilon, P}(0) < \infty$,
- $k$ bounded kernel with $\|k\|_\infty \leq 1$,
- $\mathcal{H}$ separable RKHS, dense in $L_1(P_X)$,
- $(\lambda_n) \in (0, \infty)$ with $\lambda_n \to 0$ and $\lambda_n^2 n \to \infty$.

Then, for all $\rho > 0$,

$$\lim_{n \to \infty} P^n \left( D \in (\mathcal{X} \times \mathbb{R})^n : \frac{\#SV(f_{L,D,\lambda_n})}{n} \in [a - \rho, b + \rho] \right) = 1,$$

$$a := \liminf_{m \to \infty} P\left( |f_{L,P,\lambda_m}(X) - Y| > \epsilon \right), \quad b := \limsup_{m \to \infty} P\left( |f_{L,P,\lambda_m}(X) - Y| \geq \epsilon \right).$$

Sometimes it is possible to show that $a = b$. 
**Conclusions**

SVMs are consistent and robust solutions of a well-posed problem in Hadamard’s sense

- for convex, Lipschitzian loss function and bounded kernel
- can “learn” from any totally unknown distribution
- can efficiently solve even non-standard real-life problems

**Main References:**

- my homepage: [http://www.stoch.uni-bayreuth.de](http://www.stoch.uni-bayreuth.de)
References

Thm.: Weak convergence of Bochner integrals

Let $\mathcal{X}$ be a Polish space and let $\mathcal{H}$ be a separable Hilbert space. Let $(P_n)_{n \in \mathbb{N}}$ and $P$ be Borel probability measures on $\mathcal{X}$ such that $P_n \xrightarrow{w} P$. Then

$$\lim_{n \to \infty} \left\| \int_{\mathcal{X}} f \, dP_n - \int_{\mathcal{X}} f \, dP \right\|_{\mathcal{H}} = 0$$

for every continuous function $f : \mathcal{X} \to \mathcal{H}$ such that

$$\sup_{x \in \mathcal{X}} \| f(x) \|_{\mathcal{H}} < \infty.$$

**Remark:** Nielsen (2007, Thm. 2.3) stated this theorem for any metric space $\mathcal{X}$, but a Polish space is needed.
Let \((\mathcal{X}_1, d_1)\) and \((\mathcal{X}_2, d_2)\) be complete, separable metric spaces, endowed their Borel \(\sigma\)-algebras.

Denote the **Prohorov metric** by
\[
d_{\text{Pro}}(P, Q) := \inf \{\varepsilon > 0 : P(A) < Q(A^\varepsilon) + \varepsilon, \ \forall A \in \mathcal{B}(\mathcal{X}_1)\}.
\]

Same for \(d_{\text{Pro}}\) on \(\mathcal{M}_1(\mathcal{X}_2)\).

**Def. Qualitative robustness**

A sequence \((S_n)_{n \in \mathbb{N}}\) of \(\mathcal{X}_2\)-valued estimates is called qualitatively robust for \(P \in \mathcal{M}_1(\mathcal{X}_1)\), if, \(\forall \varepsilon > 0, \ \exists \delta > 0\)

\[
d_{\text{Pro}}(P, Q) < \delta \implies d_{\text{Pro}}(\mathcal{L}_P(S_n), \mathcal{L}_Q(S_n)) < \varepsilon, \ \forall n \in \mathbb{N}.
\]
Thm. Qualitative robustness

Let \((S_n)_{n \in \mathbb{N}}\) be a sequence of estimates such that there exists a function \(S : \mathcal{M}_1(X_1) \rightarrow X_2\) with \(S_n = S(P_n)\) with \(P_n\) the empirical distribution.

If \(S\) is \textbf{continuous} on \(\mathcal{M}_1(X_1)\), then the sequence \((S_n)_{n \in \mathbb{N}}\) is \textbf{qualitatively robust} at \(P\), for all \(P \in \mathcal{M}_1(X_1)\).

Cuevas (1988, Thm. 2)
Derivatives and Influence Functions

Hadamard $\rightarrow$ Gâteaux

Fréchet $\rightarrow$ Bouligand $\rightarrow$ BIF

CHR & Van Messem (2008)

Notation: $\nabla^F, \nabla^G, \nabla^B, \nabla^B_3,$ etc.

Property: $\nabla^F_3 L^* = \nabla^F_3 L$, $\nabla^B_3 L^* = \nabla^B_3 L$
Thm. Weak convergence

Let $(\Omega, d_\Omega)$ be a complete separable metric space and let $\mathcal{X} := \mathcal{M}_1(\Omega)$ be the set of all probability measures on $\Omega$ enveloped with its Borel $\sigma$-algebra $\mathcal{B}(\Omega)$. Denote the Prohorov metric on $\mathcal{X}$ by $d_{\text{Pro}}$.

Then weak convergence of probability measures is equivalent to $d_{\text{Pro}}$-convergence, $\mathcal{X}$ is a complete separable metric space, and a set in $\mathcal{X}$ is relatively compact if and only if its $d_{\text{Pro}}$-closure is $d_{\text{Pro}}$-compact.

Billingsley (1999, Thm. 6.8)
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 $Q \neq P$ is the special Bouligand-derivative

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

(if it exists).

If BIF exists and $Q = \delta_z$: IF exists and $BIF = IF$

**Goal:** Bounded BIF