Asymptotic Normality of Support Vector Machine Variants and Other Regularized Kernel Methods

Robert Hable
Department of Mathematics
University of Bayreuth
Nonparametric Regression

\[ Y = f_0(X) + \varepsilon \]

with

- \( Y \): output variable (observable)
- \( X \): input variable (observable)
- \( f_0 \): regression function (unknown)
- \( \varepsilon \): error term (not observable)

**Goal:** Estimation of the unknown regression function \( f_0 \)
Support Vector Machines

\[ Y_i = f_0(X_i) + \varepsilon_i, \quad (X_i, Y_i) \sim P \quad \text{i.i.d.}, \quad i \in \{1, \ldots, n\} \]

**Goal:** Estimation of \( f_0 : \mathcal{X} \rightarrow \mathcal{Y} \subset \mathbb{R} \)
Support Vector Machines

\[ Y_i = f_0(X_i) + \varepsilon_i, \quad (X_i, Y_i) \sim P \quad \text{i.i.d.}, \quad i \in \{1, \ldots, n\} \]

**Goal:** Estimation of \( f_0 : \mathcal{X} \rightarrow \mathcal{Y} \subset \mathbb{R} \)

- Loss function
  \[ L : \mathcal{Y} \times \mathbb{R} \rightarrow [0, \infty) \]
  \[ L(y, t) : \text{loss caused by estimation } t = \hat{f}_n(x) \text{ if } y \text{ is true} \]
Support Vector Machines

\[ Y_i = f_0(X_i) + \varepsilon_i, \quad (X_i, Y_i) \sim P \quad \text{i.i.d.}, \quad i \in \{1, \ldots, n\} \]

**Goal:** Estimation of \( f_0 : \mathcal{X} \to \mathcal{Y} \subset \mathbb{R} \)

- Loss function
  \[ L : \mathcal{Y} \times \mathbb{R} \to [0, \infty) \]
  \( L(y, t) \): loss caused by estimation \( t = \hat{f}_n(x) \) if \( y \) is true
- Risk of an estimate \( \hat{f}_n : \mathcal{X} \to \mathbb{R} \)
  \[ \int L(y, \hat{f}_n(x)) \, P(d(x, y)) \]
Support Vector Machines

\[ Y_i = f_0(X_i) + \varepsilon_i, \quad (X_i, Y_i) \sim P \quad \text{i.i.d.,} \quad i \in \{1, \ldots, n\} \]

**Goal:** Estimation of \( f_0 : \mathcal{X} \to \mathcal{Y} \subset \mathbb{R} \)

- **Loss function**
  \[ L : \mathcal{Y} \times \mathbb{R} \to [0, \infty) \]
  \( L(y, t) \): loss caused by estimation \( t = \hat{f}_n(x) \) if \( y \) is true

- **empirical Risk of an estimate** \( \hat{f}_n : \mathcal{X} \to \mathbb{R} \)
  \[
  \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}_n(x_i))
  \]
Support Vector Machines

\[ Y_i = f_0(X_i) + \varepsilon_i, \quad (X_i, Y_i) \sim P \quad \text{i.i.d.}, \quad i \in \{1, \ldots, n\} \]

**Goal:** Estimation of \( f_0 : \mathcal{X} \to \mathcal{Y} \subset \mathbb{R} \)

- Loss function
  \[ L : \mathcal{Y} \times \mathbb{R} \to [0, \infty) \]
  \( L(y, t) \): loss caused by estimation \( t = \hat{f}_n(x) \) if \( y \) is true

- Empirical Risk of an estimate \( \hat{f}_n : \mathcal{X} \to \mathbb{R} \)
  \[ \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}_n(x_i)) \]

- RKHS \( H \) (certain Hilbert space of functions \( f : \mathcal{X} \to \mathbb{R} \))
Support Vector Machines

\[ Y_i = f_0(X_i) + \varepsilon_i, \quad (X_i, Y_i) \sim P \quad \text{i.i.d.}, \quad i \in \{1, \ldots, n\} \]

**Goal:** Estimation of \( f_0 : \mathcal{X} \to \mathcal{Y} \subset \mathbb{R} \)

- Loss function
  \[
  L : \mathcal{Y} \times \mathbb{R} \to [0, \infty) \]
  \( L(y, t) \): loss caused by estimation \( t = \hat{f}_n(x) \) if \( y \) is true

- Empirical Risk of an estimate \( \hat{f}_n : \mathcal{X} \to \mathbb{R} \)
  \[
  \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}_n(x_i))
  \]

- RKHS \( H \) (certain Hilbert space of functions \( f : \mathcal{X} \to \mathbb{R} \))

- Support vector machine
  \[
  S_n(\{(x_1, y_1), \ldots, (x_n, y_n)\}) = \operatorname{arg\ inf}_{f \in H} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i))
  \]
Overfitting
Overfitting
Support Vector Machines

\[ Y_i = f_0(X_i) + \varepsilon_i, \quad (X_i, Y_i) \sim P \quad \text{i.i.d.}, \quad i \in \{1, \ldots, n\} \]

**Goal:** Estimation of \( f_0 : \mathcal{X} \to \mathcal{Y} \subset \mathbb{R} \)

- **Loss function**
  \[ L : \mathcal{Y} \times \mathbb{R} \to [0, \infty) \]
  
  \( L(y, t) \): loss caused by prediction \( t \) if \( y \) is the true value

- **empirical Risk of an estimate** \( f : \mathcal{X} \to \mathbb{R} \)
  \[ \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) \]

- **RKHS** \( H \) (certain Hilbert space of functions \( f : \mathcal{X} \to \mathbb{R} \))

- **Support vector machine**
  \[ S_n((x_1, y_1), \ldots, (x_n, y_n)) = \arg \inf_{f \in H} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) \]
Support Vector Machines

\[ Y_i = f_0(X_i) + \varepsilon_i, \quad (X_i, Y_i) \sim P \quad \text{i.i.d.}, \quad i \in \{1, \ldots, n\} \]

**Goal:** Estimation of \( f_0 : \mathcal{X} \rightarrow \mathcal{Y} \subset \mathbb{R} \)

- Loss function
  \[
  L : \mathcal{Y} \times \mathbb{R} \rightarrow [0, \infty)
  \]
  \( L(y, t) \) : loss caused by prediction \( t \) if \( y \) is the true value

- Empirical Risk of an estimate \( f : \mathcal{X} \rightarrow \mathbb{R} \)
  \[
  \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i))
  \]

- RKHS \( H \) (certain Hilbert space of functions \( f : \mathcal{X} \rightarrow \mathbb{R} \))

- Support vector machine
  \[
  S_n((x_1, y_1), \ldots, (x_n, y_n)) = \arg \inf_{f \in H} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \|f\|^2_H
  \]
Overfitting
Overfitting
Reproducing Kernel Hilbert Space (RKHS)

Support Vector Machines

\[ S_n : (\mathcal{X} \times \mathcal{Y})^n \rightarrow H, \]
\[ ((x_1, y_1), \ldots, (x_n, y_n)) \mapsto \arg \inf_{f \in H} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \| f \|_H^2 \]

with \( H \) a reproducing kernel Hilbert space (RKHS)
Reproducing Kernel Hilbert Space (RKHS)

Support Vector Machines

\[ S_n : (\mathcal{X} \times \mathcal{Y})^n \rightarrow H, \]
\[ ((x_1, y_1), \ldots, (x_n, y_n)) \mapsto \text{arg inf}_{f \in H} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \|f\|^2_H \]

with \( H \) a reproducing kernel Hilbert space (RKHS)

Reproducing kernel Hilbert space \( H \)

- a Hilbert space of functions \( f : \mathcal{X} \rightarrow \mathbb{R} \)
- generated by a kernel function \( k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \)
- reproducing property

\[ \langle f, k(x, \cdot) \rangle_H = f(x) \quad \forall x \in \mathcal{X}, \quad \forall f \in H \]
Example: Gaussian Kernel

Gaussian Kernel \( \mathcal{X} = \mathbb{R} \)

\[ k : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}, \quad (x, x') \mapsto \exp \left( -\frac{1}{\gamma^2} |x - x'|^2 \right) \]

\[ H \subset L_p(P) \text{ dense} \]
Example: Polynomial Kernel

Polynomial Kernel \( \mathcal{X} = \mathbb{R} \)

\[
k : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}, \quad (x, x') \mapsto (x \cdot x' + c)^m
\]

\[
H = \{ f : \mathbb{R} \rightarrow \mathbb{R} \mid f \text{ a polynomial with degree } \leq m \} \cong \mathbb{R}^{m+1}
\]
Representer Theorem

How to calculate the SVM?

\[ D_n = ((x_1, y_1), \ldots, (x_n, y_n)) \]

\[
\text{SVM: } f_{D_n, \lambda} = \arg \inf_{f \in H} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \| f \|^2_H
\]
Representer Theorem

How to calculate the SVM?

\[ D_n = ((x_1, y_1), \ldots, (x_n, y_n)) \]

\[ \text{SVM: } f_{D_n,\lambda} = \arg \inf_{f \in H} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \|f\|_H^2 \]

Representer Theorem

There are \( \alpha_{D_n,1}, \ldots, \alpha_{D_n,n} \in \mathbb{R} \) such that

\[ f_{D_n,\lambda} = \sum_{i=1}^{n} \alpha_{D_n,i} k(x_i, \cdot) . \]
Representer Theorem

How to calculate the SVM?

\[ D_n = ((x_1, y_1), \ldots, (x_n, y_n)) \]

SVM:

\[
f_{D_n, \lambda} = \arg \inf_{f \in H} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \|f\|^2_{H}
\]

Represent Theorem

There are \( \alpha_{D_n, 1}, \ldots, \alpha_{D_n, n} \in \mathbb{R} \) such that

\[
f_{D_n, \lambda} = \sum_{i=1}^{n} \alpha_{D_n, i} k(x_i, \cdot).
\]

\[\rightarrow\] just solve a finite convex optimization problem
... and this really works?
...and this really works? Yes, quite good.
...and this really works? Yes, quite good.
... and this really works?  Yes, quite good.
Risk-Consistency

Risk of a predictor $f : \mathcal{X} \rightarrow \mathbb{R}$

$$R_P(f) = \int L(y, f(x)) \, P(d(x, y)) \quad \hat{=} \quad \text{Quality of } f$$

$D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$

SVM: $f_{D_n, \lambda_n} = \arg \inf_{f \in H} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i)) + \lambda_n \| f \|_H^2$
**Risk-Consistency**

Risk of a predictor $f : \mathcal{X} \rightarrow \mathbb{R}$

$$\mathcal{R}_P(f) = \int L(y, f(x)) \, P(d(x, y)) \quad \hat{=} \quad \text{Quality of } f$$

$\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$

**SVM:** 

$$f_{\mathcal{D}_n, \lambda_n} = \arg \inf_{f \in H} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i)) + \lambda_n\|f\|_H^2$$

Risk-consistency

$$\mathcal{R}_P(f_{\mathcal{D}_n, \lambda_n}) \xrightarrow{n \rightarrow \infty} \inf_{f : \mathcal{X} \rightarrow \mathbb{R}} \mathcal{R}_P(f) \quad \text{in probability}$$

essentially if

- $H \subset L_p(P)$ dense (e.g. Gaussian kernel)
- $\lambda_n \rightarrow 0$ not too fast (!)
Robustness – a non-parametric example

Regression:
Robustness – a non-parametric example

Regression: $k$-nearest neighbor
Robustness – a non-parametric example

Regression: $k$-nearest neighbor
Robustness – a non-parametric example

Regression:
Robustness – a non-parametric example

Regression: SVM
Robustness – a non-parametric example

Regression: SVM
Robustness

Loss function $L$

- $\varepsilon$-insensitive
- Pinball
- Least squares
Robustness

Loss function $L$ should be Lipschitz continuous

- $\varepsilon$-insensitive
- pinball
- least squares
Robustness

Loss function $L$ should be Lipschitz continuous

- $\varepsilon$-insensitive
- pinball
- least squares

bounded influenza function


bounds on the maxbias


qualitative robustness for $\lambda > 0$

Hable & Christmann (2011)
Rates of Convergence

Risk-consistency

\[
\mathcal{R}_P(f_{D_n,\lambda_n}) \xrightarrow{n \to \infty} \inf_{f: \mathcal{X} \to \mathbb{R}} \mathcal{R}_P(f) \quad \text{in probability}
\]
Rates of Convergence

Risk-consistency

$$\mathcal{R}_P(f_{D_n,\lambda_n}) \xrightarrow{n \to \infty} \inf_{f: \mathcal{X} \to \mathbb{R}} \mathcal{R}_P(f) \quad \text{in probability}$$

How fast is this convergence?

Is there a uniform rate $r_n$ such that

$$r_n \left( \mathcal{R}_P(f_{D_n,\lambda_n}) - \inf_{f: \mathcal{X} \to \mathbb{R}} \mathcal{R}_P(f) \right) \xrightarrow{n \to \infty} 0 \quad \text{in probability}$$

for every $P$?
Rates of Convergence

Risk-consistency

\[ \mathcal{R}_P(f_{D_n, \lambda_n}) \xrightarrow{n \to \infty} \inf_{f: \mathcal{X} \to \mathbb{R}} \mathcal{R}_P(f) \quad \text{in probability} \]

How fast is this convergence?

Is there a \textit{uniform} rate \( r_n \) such that

\[ r_n \left( \mathcal{R}_P(f_{D_n, \lambda_n}) - \inf_{f: \mathcal{X} \to \mathbb{R}} \mathcal{R}_P(f) \right) \xrightarrow{n \to \infty} 0 \quad \text{in probability} \]

for every \( P \)? \( \longrightarrow \) \textbf{No!} (no-free-lunch theorem)
Rates of Convergence

Risk-consistency

$$\mathcal{R}_P(f_{D_n,\lambda_n}) \xrightarrow{n \to \infty} \inf_{f: \mathcal{X} \to \mathbb{R}} \mathcal{R}_P(f) \quad \text{in probability}$$

How fast is this convergence?

Is there a uniform rate $r_n$ such that

$$r_n \left( \mathcal{R}_P(f_{D_n,\lambda_n}) - \inf_{f: \mathcal{X} \to \mathbb{R}} \mathcal{R}_P(f) \right) \xrightarrow{n \to \infty} 0 \quad \text{in probability}$$

for every $P$? $\rightarrow$ No! (no-free-lunch theorem)

Instead,

rates $r_n$ of convergence under assumptions on $P$

e.g. Steinwart and Scovel (2007), Caponnetto and De Vito (2007), Blanchard et al. (2008), Steinwart et al. (2009), Mendelson and Neeman (2010)
Smooth Approximation of the Regression Function

**Goal:** estimate a solution $f^* : \mathcal{X} \rightarrow \mathbb{R}$ of

$$\mathcal{R}_P(f) = \min! f : \mathcal{X} \rightarrow \mathbb{R}$$

or

$$\inf_{f \in H} \mathcal{R}_P(f) = \min! f \in H.$$
Smooth Approximation of the Regression Function

**Goal:** estimate a solution $f^* : \mathcal{X} \rightarrow \mathbb{R}$ of

$$R_P(f) = \min! \quad f : \mathcal{X} \rightarrow \mathbb{R}$$

or

$$\inf_{f \in H} R_P(f) = \min! \quad f \in H.$$  

However, these optimization problems

- there is no uniform rate of convergence to the solution (without substantial assumptions on $P$)
- either qualitatively robust or consistent (for SVMs $f_{D_n, \lambda_n}$ depends on the choice of $(\lambda_n)_{n \in \mathbb{N}}$)
Smooth Approximation of the Regression Function

**Goal:** estimate a solution $f^* : \mathcal{X} \to \mathbb{R}$ of

$$ \mathcal{R}_P(f) = \min! \quad f : \mathcal{X} \to \mathbb{R} $$

or

$$ \inf_{f \in H} \mathcal{R}_P(f) = \min! \quad f \in H. $$

However, these optimization problems

- there is no uniform rate of convergence to the solution (without substantial assumptions on $P$)
- either qualitatively robust or consistent (for SVMs $f_{D_n,\lambda_n}$ depends on the choice of $(\lambda_n)_{n \in \mathbb{N}}$)

Instead, consider the regularized problem

$$ \mathcal{R}_P(f) + \lambda_0 \|f\|_H^2 = \min! \quad f \in H. $$
Smooth Approximation of the Regression Function

Instead of estimating a solution $f^* : \mathcal{X} \rightarrow \mathbb{R}$ of

$$
\mathcal{R}_P(f) = \min! \quad f : \mathcal{X} \rightarrow \mathbb{R}
$$

we may estimate the solution $f_{P,\lambda_0}$ of the regularized problem

$$
\mathcal{R}_P(f) + \lambda_0 \| f \|_H^2 = \min! \quad f \in H.
$$

$f_{P,\lambda_0}$ serves as a “smoother approximation” of $f^*$. 
Smooth Approximation of the Regression Function

- Instead of estimating a solution \( f^* : \mathcal{X} \to \mathbb{R} \) of
  \[
  R_P(f) = \min! \quad f : \mathcal{X} \to \mathbb{R}
  \]
  we may estimate the solution \( f_{P,\lambda_0} \) of the regularized problem
  \[
  R_P(f) + \lambda_0 \|f\|_H^2 = \min! \quad f \in H.
  \]
  \( f_{P,\lambda_0} \) serves as a “smoother approximation” of \( f^* \).

- The regularized problem is equivalent to
  \[
  R_P(f) = \min! \quad f \in H, \quad \|f\|_H \leq r_0.
  \]
  \( r_0 \): bound on complexity of “smoother approximation”
Example
Example

\[ \lambda = 1 \]
Example

$\lambda = 0.1$
Example

\[ \lambda = 0.01 \]
Example

\[ \lambda = 0.001 \]
Example

\[ \lambda = 0.0001 \]
Example

\[ \lambda = 0.00001 \]
Example

$\lambda = 0.000001$
Asymptotic Normality of Regularized Problem

Under some
- assumptions on $\mathcal{X}$, $L$, and $k (\leftrightarrow H)$
- but (essentially) no assumptions on $P$,

we have

$$\sqrt{n}\left(\mathcal{R}(f_{D_n},\lambda_0) - \mathcal{R}(f_P,\lambda_0)\right) \sim \mathcal{N}(0, \sigma^2)$$
Asymptotic Normality of Regularized Problem

Under some

▶ assumptions on $\mathcal{X}$, $L$, and $k (\leftrightarrow H)$
▶ but (essentially) no assumptions on $P$,

we have

$$\sqrt{n} \left( \mathcal{R}(f_{D_n}, \lambda_0) - \mathcal{R}(f_P, \lambda_0) \right) \sim \mathcal{N}(0, \sigma^2)$$

and, even more,

$$\sqrt{n}(f_{D_n, \lambda_0} - f_P, \lambda_0) \sim \text{Gaussian process in } H$$
Asymptotic Normality of Regularized Problem

Under some

- assumptions on $\mathcal{X}$, $L$, and $k \leftrightarrow H$
- but (essentially) no assumptions on $P$,

we have

$$\sqrt{n}\left(\mathcal{R}(f_{D_n}, \lambda_0) - \mathcal{R}(f_P, \lambda_0)\right) \sim \mathcal{N}(0, \sigma^2)$$

and, even more,

$$\sqrt{n}(f_{D_n, \lambda_0} - f_P, \lambda_0) \sim \text{Gaussian process in } H$$
Asymptotic Normality of Regularized Problem

Under some

- assumptions on \( \mathcal{X}, L, k (\leftrightarrow H) \), and \( \lambda_n \xrightarrow{n \to \infty} \lambda_0 \)
- but (essentially) no assumptions on \( P \),

we have

\[
\sqrt{n} \left( \mathcal{R}(f_{D_n}, \lambda_n) - \mathcal{R}(f_P, \lambda_0) \right) \sim \mathcal{N}(0, \sigma^2)
\]

and, even more,

\[
\sqrt{n} \left( f_{D_n, \lambda_n} - f_{P, \lambda_0} \right) \sim \text{Gaussian process in } H
\]
Asymptotic Normality of Regularized Problem

Under some

- assumptions on $\mathcal{X}$, $L$, $k (\leftrightarrow H)$, and $\lambda_{D_n} \xrightarrow{n \to \infty} \lambda_0$
- but (essentially) no assumptions on $P$,

we have

$$\sqrt{n}\left(\mathcal{R}(f_{D_n},\lambda_{D_n}) - \mathcal{R}(f_P,\lambda_0)\right) \xrightarrow{\sim} \mathcal{N}(0, \sigma^2)$$

and, even more,

$$\sqrt{n}(f_{D_n,\lambda_{D_n}} - f_P,\lambda_0) \xrightarrow{\sim} \text{Gaussian process in } H$$
Asymptotic Normality of Regularized Problem

Corollary 1

In particular, we also have for every \( x_1, \ldots, x_m \in \mathcal{X} \)

\[
\sqrt{n} \begin{pmatrix}
  f_{D_n, \lambda_{D_n}}(x_1) - f_{P, \lambda_0}(x_1) \\
  \vdots \\
  f_{D_n, \lambda_{D_n}}(x_m) - f_{P, \lambda_0}(x_m)
\end{pmatrix} \sim \mathcal{N}_m(0, \Sigma)
\]

where \( \Sigma \) is a covariance matrix.

(follows from the reproducing property of the kernel \( k \))
Corollary 2

It follows that

$$\sqrt{n} \sup_{x \in \mathcal{X}} \left| f_{D_n, \lambda_n}(x) - f_{P, \lambda_0}(x) \right|$$

weakly converges to a random variable.
under some assumptions . . .

- $\mathcal{X} \subset \mathbb{R}^d$ compact
- $k$ more than $d/2$-times continuously differentiable
- $L$ smooth (2-times differentiable) and integrable

![Graphs](image-url)

- $\varepsilon$-insensitive
- pinball
- least squares
under some assumptions . . .

- $\mathcal{X} \subset \mathbb{R}^d$ compact
- $k$ more than $d/2$-times continuously differentiable
- $L$ smooth (2-times differentiable) and integrable

\(\varepsilon\)-insensitive | pinball | least squares
under some assumptions . . .

- $\mathcal{X} \subset \mathbb{R}^d$ compact
- $k$ more than $d/2$-times continuously differentiable
- $L$ smooth (2-times differentiable) and integrable

![smoothed $\varepsilon$-insensitive](image1.png)
![smoothed pinball](image2.png)
![least squares](image3.png)
under some assumptions . . .

and

\[ \sqrt{n}(\lambda_{D_n} - \lambda_0) \xrightarrow{n \to \infty} 0 \]
in probability.
under some assumptions . . .

and

\[ \sqrt{n}(\lambda_{D_n} - \lambda_0) \xrightarrow[n \to \infty]{} 0 \quad \text{in probability.} \]

For example:

Choose a (large) constant \( c > 0 \) and do a cross-validation in

\[ [\lambda_0, \lambda_0 + \frac{c}{\sqrt{n \ln(n)}}] . \]
Sketch of the Proof: Functional Delta-Method

Consider the SVM-functional

\[ S : \mathcal{M}_1 \rightarrow H, \quad P \mapsto f_{P,\lambda_0} \]

SVM-functional represents SVM-estimator:

\[ f_{D_n,\lambda_0} = S(P_{D_n}) \]
Sketch of the Proof: Functional Delta-Method

Consider the SVM-functional

\[ S : \mathcal{M}_1 \to H, \quad P \mapsto f_{P,\lambda_0} \]

SVM-functional represents SVM-estimator:

\[ f_{D_n,\lambda_0} = S(\mathbb{P}_{D_n}) \]

1. Show that \( \sqrt{n}(\mathbb{P}_{D_n} - P) \) converges weakly to a Gaussian process in a suitable space \( \ell_\infty(\mathcal{G}) \).
Sketch of the Proof: Functional Delta-Method

Consider the SVM-functional

\[ S : \mathcal{M}_1 \to H, \quad P \mapsto f_{P,\lambda_0} \]

SVM-functional represents SVM-estimator:

\[ f_{D_n,\lambda_0} = S(P_{D_n}) \]

1. Show that \( \sqrt{n}(P_{D_n} - P) \) converges weakly to a Gaussian process in a suitable space \( \ell_\infty(G) \).

2. Show that \( S \) is Hadamard-differentiable.
Sketch of the Proof: Functional Delta-Method

Consider the SVM-functional

\[ S : \mathcal{M}_1 \to H, \quad P \mapsto f_{P,\lambda_0} \]

SVM-functional represents SVM-estimator:

\[ f_{D_n,\lambda_0} = S(P_{D_n}) \]

1. Show that \( \sqrt{n}(P_{D_n} - P) \) converges weakly to a Gaussian process in a suitable space \( \ell_\infty(\mathcal{G}) \).
2. Show that \( S \) is Hadamard-differentiable.
3. Then, it follows from the functional delta-method that

\[ \sqrt{n}(f_{D_n,\lambda_0} - f_{P,\lambda_0}) = \sqrt{n}(S(P_{D_n}) - S(P)) \]

converges weakly to a Gaussian process.
Sketch of the Proof: Functional Delta-Method

Consider the SVM-functional

\[ S : \mathcal{M}_1 \rightarrow H, \quad P \mapsto f_{P,\lambda_0} \]

SVM-functional represents SVM-estimator:

\[ f_{D_n,\lambda_0} = S(P_{D_n}) \]

1. Show that \( \sqrt{n}(P_{D_n} - P) \) converges weakly to a Gaussian process in a suitable space \( \ell_\infty(G) \).
2. Show that \( S \) is Hadamard-differentiable:
3. Then, it follows from the functional delta-method that

\[ \sqrt{n}(f_{D_n,\lambda_0} - f_{P,\lambda_0}) = \sqrt{n}(S(P_{D_n}) - S(P)) \]

converges weakly to a Gaussian process.
Sketch of the Proof: Functional Delta-Method

How to deal with random parameters $\lambda_{D_n}$?
Sketch of the Proof: Functional Delta-Method

How to deal with random parameters $\lambda_{D_n}$?

Problem

$$\int L(y, f(x)) \, P(d(x, y)) + \lambda \| f \|^2_H = \min! \quad f \in H.$$ 

is equivalent to

$$\frac{\lambda_0}{\lambda} \left( \int L(y, f(x)) \, P(d(x, y)) + \lambda \| f \|^2_H \right) = \min! \quad f \in H.$$
Sketch of the Proof: Functional Delta-Method

How to deal with random parameters $\lambda_{D_n}$?

Problem

$$\int L(y, f(x)) P(d(x, y)) + \lambda \|f\|^2_H = \min! \quad f \in H.$$ 

is equivalent to

$$\left( \frac{\lambda_0}{\lambda} \right) \left( \frac{\lambda_0}{\lambda} \right) \left( \int L(y, f(x)) P(d(x, y)) + \lambda \|f\|^2_H \right) = \min! \quad f \in H$$

and

$$\int L(y, f(x)) \left( \frac{\lambda_0}{\lambda} P \right)(d(x, y)) + \lambda_0 \|f\|^2_H = \min! \quad f \in H.$$ 

Hence, $f_{P, \lambda} = f_{\frac{\lambda_0}{\lambda} P, \lambda_0} = S\left( \frac{\lambda_0}{\lambda} P \right)$
Sketch of the Proof: Functional Delta-Method

Consider the SVM-functional

\[ S : \mathcal{M}_1 \rightarrow H, \quad P \mapsto f_{P,\lambda_0} \]

SVM-functional represents SVM-estimator:

\[ f_{D_n,\lambda_{D_n}} = S \left( \frac{\lambda_0}{\lambda_{D_n}} \mathbb{P}_{D_n} \right) \]

1. Show that \( \sqrt{n}(\mathbb{P}_{D_n} - P) \) converges weakly to a Gaussian process in a suitable space \( \ell_\infty(G) \).
2. Show that \( S \) is Hadamard-differentiable:
3. Then, it follows from the functional delta-method that

\[ \sqrt{n}(f_{D_n,\lambda_0} - f_{P,\lambda_0}) = \sqrt{n}(S(\mathbb{P}_{D_n}) - S(P)) \]

converges weakly to a Gaussian process.
Sketch of the Proof: Functional Delta-Method

Consider the SVM-functional

\[ S : \mathcal{M}_1 \rightarrow H, \quad P \mapsto f_{P,\lambda_0} \]

SVM-functional represents SVM-estimator:

\[ f_{D_n,\lambda_{D_n}} = S\left( \frac{\lambda_0}{\lambda_{D_n}} P_{D_n} \right) \]

1. Show that \( \sqrt{n}(P_{D_n} - P) \) converges weakly to a Gaussian process in a suitable space \( \ell_\infty(\mathcal{G}) \).
2. Show that \( S \) is Hadamard-differentiable:
3. Then, it follows from the functional delta-method that

\[ \sqrt{n}(f_{D_n,\lambda_{D_n}} - f_{P,\lambda_0}) = \sqrt{n}\left( S\left( \frac{\lambda_0}{\lambda_{D_n}} P_{D_n} \right) - S(P) \right) \]

converges weakly to a Gaussian process.
Sketch of the Proof: Functional Delta-Method

Consider the SVM-functional

\[ S : \mathcal{M}_f \to H, \quad \mu \mapsto f_{\mu, \lambda_0} \]

SVM-functional represents SVM-estimator:

\[ f_{D_n, \lambda_{D_n}} = S \left( \frac{\lambda_0}{\lambda_{D_n}} P_{D_n} \right) \]

1. Show that \( \sqrt{n}(P_{D_n} - P) \) converges weakly to a Gaussian process in a suitable space \( \ell_\infty(G) \).

2. Show that \( S \) is Hadamard-differentiable:

3. Then, it follows from the functional delta-method that

\[ \sqrt{n}(f_{D_n, \lambda_{D_n}} - f_{P, \lambda_0}) = \sqrt{n} \left( S \left( \frac{\lambda_0}{\lambda_{D_n}} P_{D_n} \right) - S(P) \right) \]

converges weakly to a Gaussian process.
Example: Asymptotic Normality

Model: \( Y_i = \sin(X_i) + \varepsilon_i \) where

\[ \varepsilon_i \sim_{\text{i.i.d.}} \text{Unif}(-1, 1), \quad X_i \sim_{\text{i.i.d.}} \text{Unif}(-5, 5), \quad i \in \{1, \ldots, n\}. \]

- Sample size \( n = 100, \ n = 200 \) and \( n = 1000 \);
- 5000 runs

Estimation:
- Gaussian kernel
- Logistic loss
Example: Asymptotic Normality

Model: \( Y_i = \sin(X_i) + \varepsilon_i \) where

\[ \varepsilon_i \sim \text{i.i.d. Unif}(-1, 1), \quad X_i \sim \text{i.i.d. Unif}(-5, 5), \quad i \in \{1, \ldots, n\}. \]

\[
\begin{array}{ccc}
  n = 100 & n = 200 & n = 1000 \\
  \includegraphics[width=0.3\textwidth]{plot1} & \includegraphics[width=0.3\textwidth]{plot2} & \includegraphics[width=0.3\textwidth]{plot3}
\end{array}
\]
Example: Asymptotic Normality

**Model:** \( Y_i = \sin(X_i) + \varepsilon_i \) where

\[ \varepsilon_i \sim \text{i.i.d. Unif}(-1, 1), \quad X_i \sim \text{i.i.d. Unif}(-5, 5), \quad i \in \{1, \ldots, n\}. \]

\[ n = 100 \quad n = 200 \quad n = 1000 \]
Example: Asymptotic Normality

Model: \( Y_i = \sin(X_i) + \varepsilon_i \) where

\[ \varepsilon_i \sim \text{i.i.d. Unif}(-1,1), \quad X_i \sim \text{i.i.d. Unif}(-5,5), \quad i \in \{1, \ldots, n\}. \]
Example: Asymptotic Normality

Model: \( Y_i = \sin(X_i) + \varepsilon_i \) where

\[ \varepsilon_i \sim \text{i.i.d. Unif}(-1,1), \quad X_i \sim \text{i.i.d. Unif}(-5,5), \quad i \in \{1, \ldots, n\}. \]

\( n = 100 \) \hspace{2cm} \( n = 200 \) \hspace{2cm} \( n = 1000 \)
Example: Asymptotic Normality

Model: \( Y_i = \sin(X_i) + \varepsilon_i \) where

\[ \varepsilon_i \sim \text{i.i.d. Unif}(-1, 1), \quad X_i \sim \text{i.i.d. Unif}(-5, 5), \quad i \in \{1, \ldots, n\}. \]

\( n = 100 \)  
\( n = 200 \)  
\( n = 1000 \)
Example: Asymptotic Normality

Model: \[ Y_i = \sin(X_i) + \varepsilon_i \] where

\[ \varepsilon_i \sim \text{i.i.d. Unif}(-1, 1), \quad X_i \sim \text{i.i.d. Unif}(-5, 5), \quad i \in \{1, \ldots, n\}. \]
Example: Asymptotic Normality

Model: \( Y_i = \sin(X_i) + \varepsilon_i \) where

\[ \varepsilon_i \sim \text{i.i.d. } \text{Unif}(-1, 1), \quad X_i \sim \text{i.i.d. } \text{Unif}(-5, 5), \quad i \in \{1, \ldots, n\}. \]
Example: Asymptotic Normality

Model: $Y_i = \sin(X_i) + \varepsilon_i$ where

$\varepsilon_i \sim \text{i.i.d. Unif}(-1, 1), \quad X_i \sim \text{i.i.d. Unif}(-5, 5), \quad i \in \{1, \ldots, n\}.$

$n = 100, \quad n = 200, \quad n = 1000$
Example: Asymptotic Normality

Model: \( Y_i = \sin(X_i) + \varepsilon_i \) where

\[ \varepsilon_i \sim \text{i.i.d. } \text{Unif}(-1, 1), \quad X_i \sim \text{i.i.d. } \text{Unif}(-5, 5), \quad i \in \{1, \ldots, n\}. \]
Example: Asymptotic Normality

Model: \( Y_i = \sin(X_i) + \varepsilon_i \) where

\[ \varepsilon_i \sim_{\text{i.i.d.}} \text{Unif}(-1, 1), \quad X_i \sim_{\text{i.i.d.}} \text{Unif}(-5, 5), \quad i \in \{1, \ldots, n\}. \]

\( n = 100 \)  \hspace{1cm} \( n = 200 \)  \hspace{1cm} \( n = 1000 \)
Example: Asymptotic Normality

Model: \( Y_i = \sin(X_i) + \varepsilon_i \) where

\[ \varepsilon_i \sim \text{i.i.d. Unif}(-1, 1), \quad X_i \sim \text{i.i.d. Unif}(-5, 5), \quad i \in \{1, \ldots, n\}. \]

Asymptotic Normality:

\[
\sqrt{n} \left( f_{D_n,\lambda_{D_n}}(x_0) - f_{P,\lambda_0}(x_0) \right) \sim \mathcal{N}(0, \sigma^2_P), \quad x_0 = 0.
\]
Example: Asymptotic Normality

Model: \( Y_i = \sin(X_i) + \varepsilon_i \) where

\[ \varepsilon_i \sim_{\text{i.i.d.}} \text{Unif}(-1, 1), \quad X_i \sim_{\text{i.i.d.}} \text{Unif}(-5, 5), \quad i \in \{1, \ldots, n\}. \]

Asymptotic Normality:

\[ \sqrt{n} \left( f_{D_n, \lambda_{D_n}}(x_0) - f_{P, \lambda_0}(x_0) \right) \sim \mathcal{N}(0, \sigma_P^2), \quad x_0 = 0. \]

\[ n = 100 \quad n = 200 \quad n = 1000 \]
Example: Asymptotic Normality

Model: $Y_i = \sin(X_i) + \varepsilon_i$ where

$\varepsilon_i \sim_{i.i.d.} \text{Unif}(-1, 1), \quad X_i \sim_{i.i.d.} \text{Unif}(-5, 5), \quad i \in \{1, \ldots, n\}.$

Asymptotic Normality:

$$\sqrt{n} \left( f_{D_n, \lambda_n}(x_0) - f_{P, \lambda_0}(x_0) \right) \sim \mathcal{N}(0, \sigma_P^2), \quad x_0 = 0.$$
Inference

Asymptotic Normality:

\[ \sqrt{n} \left( f_{D_n, \lambda_{D_n}}(x_0) - f_{P, \lambda_0}(x_0) \right) \xrightarrow{\text{d}} N(0, \sigma^2_P), \]

Unknown variance: \( \sigma^2_P \)
Inference

Asymptotic Normality:

\[ \sqrt{n} \left( f_{D_n, \lambda_{D_n}}(x_0) - f_{P, \lambda_0}(x_0) \right) \sim \mathcal{N}(0, \sigma_P^2), \]

Unknown variance: \( \sigma_P^2 \)

Theorem

It follows that

\[ \sigma_P^2 = \text{Var}(g_{P, \lambda_0}(X_i, Y_i)), \]

where

\[ g_{P, \lambda_0}(x, y) = -L'(x, y, f_{P, \lambda_0}(x)) \left( K_P^{-1}(k(\cdot, x)) \right)(x_0) \]

and

\[ K_P : H \to H, \quad f \mapsto 2\lambda_0 f + \int L''(x, y, f_{P, \lambda_0}(x)) f(x) k(\cdot, x) P(d(x, y)). \]
Inference

Asymptotic Normality:

\[ \sqrt{n} \left( f_{D_n, \lambda_{D_n}}(x_0) - f_{P, \lambda_0}(x_0) \right) \overset{\mathcal{D}}{\sim} \mathcal{N}(0, \sigma_P^2), \]

Unknown variance: \( \sigma_P^2 \)

Estimator:

\[ \hat{\sigma}_{P, n}^2 = \text{SampleVariance}(g_{P, \lambda_0}(X_i, Y_i), \ i = 1, \ldots, n), \]

where

\[
 g_{P, \lambda_0}(x, y) = -L'(x, y, f_{P, \lambda_0}(x)) \left( K_P^{-1}(k(\cdot, x)) \right)(x_0)
\]

and

\[
 K_P : H \to H, \quad f \mapsto 2\lambda_0 f + \int L''(x, y, f_{P, \lambda_0}(x)) f(x) k(\cdot, x) P(d(x, y)) .
\]
Inference

Asymptotic Normality:

\[ \sqrt{n} \left( f_{D_n, \lambda_{D_n}}(x_0) - f_{P, \lambda_0}(x_0) \right) \sim \mathcal{N}(0, \sigma_P^2), \]

Unknown variance: \( \sigma_P^2 \)

Estimator:

\[ \hat{\sigma}_{P, n}^2 = \frac{\text{SampleVariance}(g_{P, \lambda_0}(X_i, Y_i), i = 1, \ldots, n)}{n}, \]

where

\[ g_{P, \lambda_0}(x, y) = -L'(x, y, f_{P, \lambda_0}(x)) \left( K_P^{-1}(k(\cdot, x)) \right)(x_0) \]

and

\[ K_P : H \to H, \quad f \mapsto 2\lambda_0 f + \int L''(x, y, f_{P, \lambda_0}(x)) f(x) k(\cdot, x) P(d(x, y)). \]
Inference

Asymptotic Normality:

$$\sqrt{n} \left( f_{D_n, \lambda_{D_n}}(x_0) - f_{P, \lambda_0}(x_0) \right) \sim \mathcal{N}(0, \sigma_P^2),$$

Unknown variance: $\sigma_P^2$

Estimator:

$$\hat{\sigma}_{D_n}^2 = \text{SampleVariance} \left( g_{D_n, \lambda_{D_n}}(X_i, Y_i), \ i = 1, \ldots, n \right),$$

where

$$g_{D_n, \lambda_{D_n}}(x, y) = -L'(x, y, f_{D_n, \lambda_{D_n}}(x)) \left( K_{D_n, \lambda_{D_n}}^{-1}(k(\cdot, x)) \right)(x_0)$$

and

$$K_{D_n, \lambda_{D_n}} : f \mapsto 2\lambda_{D_n} f + \int L''(x, y, f_{D_n, \lambda_{D_n}}(x)) f(x) k(\cdot, x) \mathbb{P}_{D_n}(d(x, y)).$$
Example: Asymptotic Confidence Interval

Model: \( Y_i = \sin(X_i) + \varepsilon_i \) where
\[
\varepsilon_i \sim_{\text{i.i.d.}} \text{Unif}(-1, 1), \quad X_i \sim_{\text{i.i.d.}} \text{Unif}(-5, 5), \quad i \in \{1, \ldots, n\}.
\]

Asymptotic Normality:
\[
\sqrt{n} \left( f_{D_n, \lambda_n}(x_0) - f_{P, \lambda_0}(x_0) \right) \sim \mathcal{N}(0, \sigma_P^2), \quad x_0 = 0.
\]
Example: Asymptotic Confidence Interval

**Model:** \( Y_i = \sin(X_i) + \varepsilon_i \) where
\[ \varepsilon_i \sim \text{i.i.d. Unif}(-1, 1), \quad X_i \sim \text{i.i.d. Unif}(-5, 5), \quad i \in \{1, \ldots, n\}. \]

**Asymptotic Normality:**
\[ \sqrt{n} \left( f_{D_n, \lambda_{D_n}}(x_0) - f_{P, \lambda_0}(x_0) \right) \sim \mathcal{N}(0, \sigma_P^2), \quad x_0 = 0. \]

**Estimation of \( \sigma_P^2 \)**

**Confidence Interval**

<table>
<thead>
<tr>
<th>( n )</th>
<th>confidence level</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.897</td>
</tr>
<tr>
<td>200</td>
<td>0.923</td>
</tr>
<tr>
<td>1000</td>
<td>0.941</td>
</tr>
<tr>
<td>theor.</td>
<td>0.95</td>
</tr>
</tbody>
</table>
References


