WIAS Berlin 9. November 2010

Asymptotic Normality of Support Vector Machines

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)
- ε : error term (not observable)

Goal: Estimation of the unknown regression function f_0

Support Vector Machines

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

Goal: Estimation of $f_0 : \mathcal{X} \to \mathcal{Y} \subset \mathbb{R}$

Support Vector Machines

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

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

Support Vector Machines

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

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 $\hat{f}_n : \mathcal{X} \to \mathbb{R}$

$$\int L\big(y, \hat{f}_n(x)\big)\, P\big(d(x, y)\big)$$

Support Vector Machines

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

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 $\hat{f}_n : \mathcal{X} \to \mathbb{R}$

$$\frac{1}{n} \sum_{i=1}^{n} L\big(y_i, \hat{f}_n(x_i)\big)$$

Support Vector Machines

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

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 $\hat{f}_n : \mathcal{X} \to \mathbb{R}$

$$\frac{1}{n} \sum_{i=1}^{n} L\big(y_i, \hat{f}_n(x_i)\big)$$

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

Support Vector Machines

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

Goal: Estimation of $f_0 : \mathcal{X} \to \mathcal{Y} \subset \mathbb{R}$

- Loss function

$$L \;:\; \mathcal{Y} \times \mathbb{R} \;\to\; [0, \infty)$$

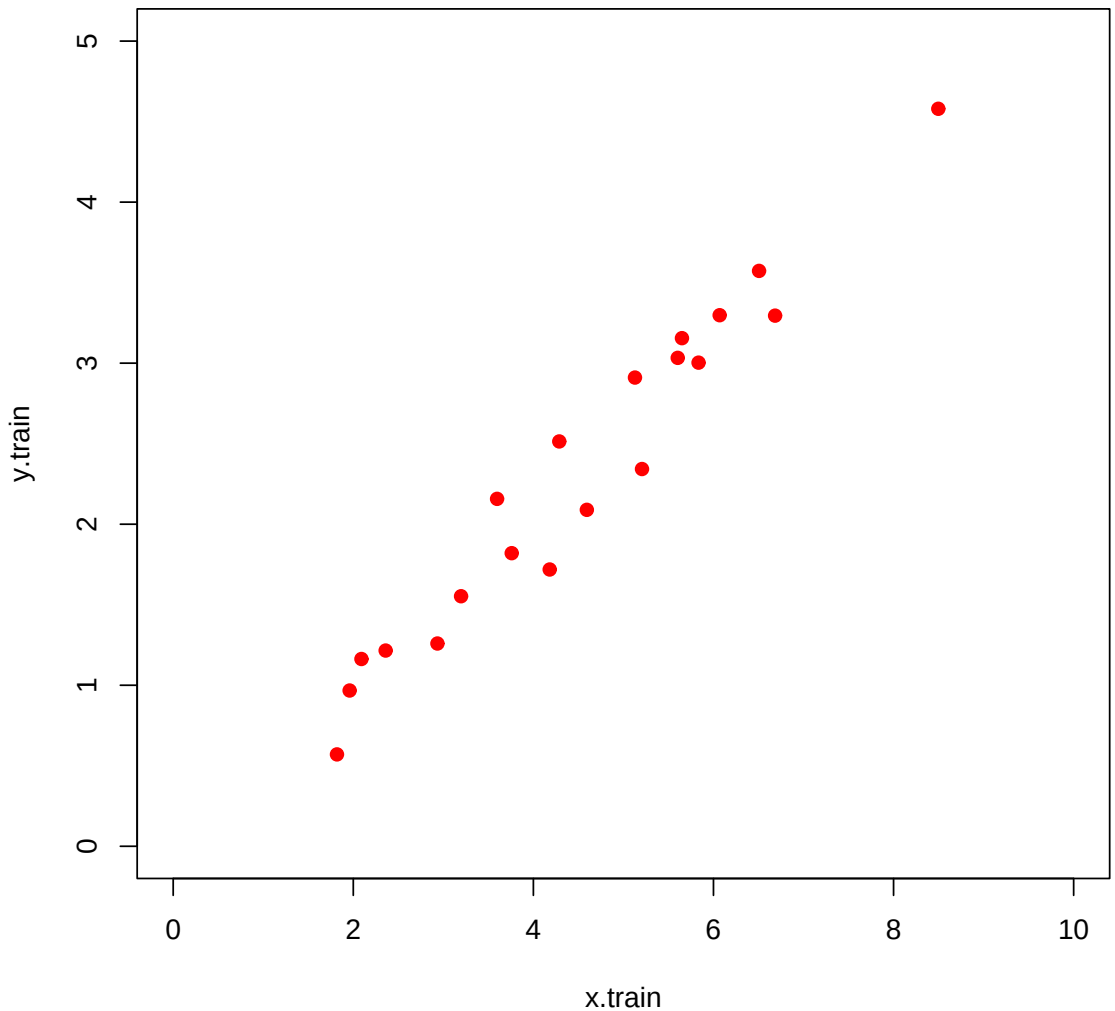
  $L(y, t)$: loss caused by estimation $t = \hat{f}_n(x)$ if y is true
- empirical Risk of an estimate $\hat{f}_n : \mathcal{X} \to \mathbb{R}$

$$\frac{1}{n} \sum_{i=1}^{n} L\big(y_i, \hat{f}_n(x_i)\big)$$

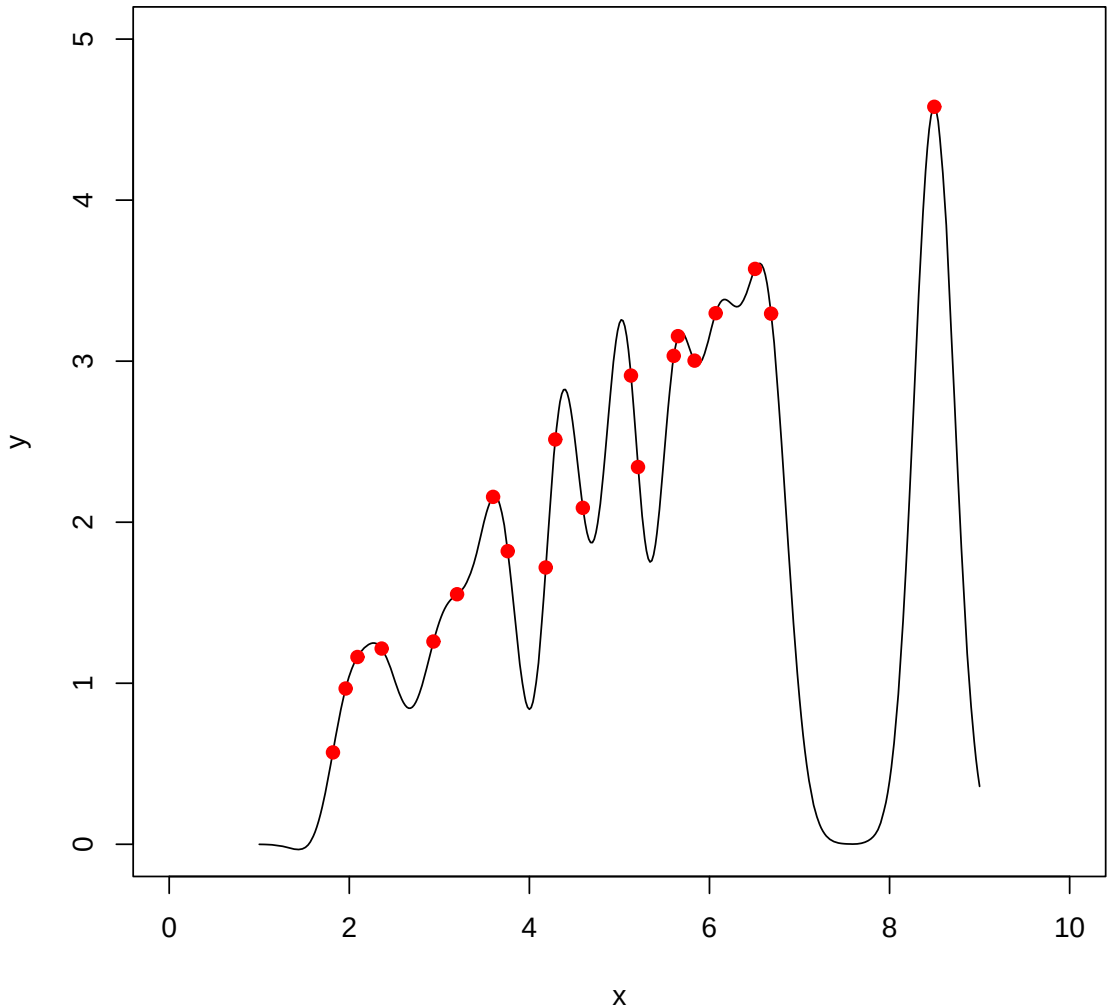
- RKHS H (certain Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$)
- Support vector machine

$$S_n\big((x_1, y_1), \ldots, (x_n, y_n)\big) \;=\; \arg\inf_{f \in H} \; \frac{1}{n} \sum_{i=1}^{n} L\big(y_i, f(x_i)\big)$$

Overfitting



Overfitting



Support Vector Machines

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

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 $f : \mathcal{X} \to \mathbb{R}$

$$\frac{1}{n} \sum_{i=1}^{n} L\big(y_i, f(x_i)\big)$$

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

$$S_n\big((x_1, y_1), \ldots, (x_n, y_n)\big) \;=\; \arg\inf_{f \in H} \; \frac{1}{n} \sum_{i=1}^{n} L\big(y_i, f(x_i)\big)$$

Support Vector Machines

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

Goal: Estimation of $f_0 : \mathcal{X} \to \mathcal{Y} \subset \mathbb{R}$

- Loss function

$$L \;:\; \mathcal{Y} \times \mathbb{R} \;\to\; [0, \infty)$$

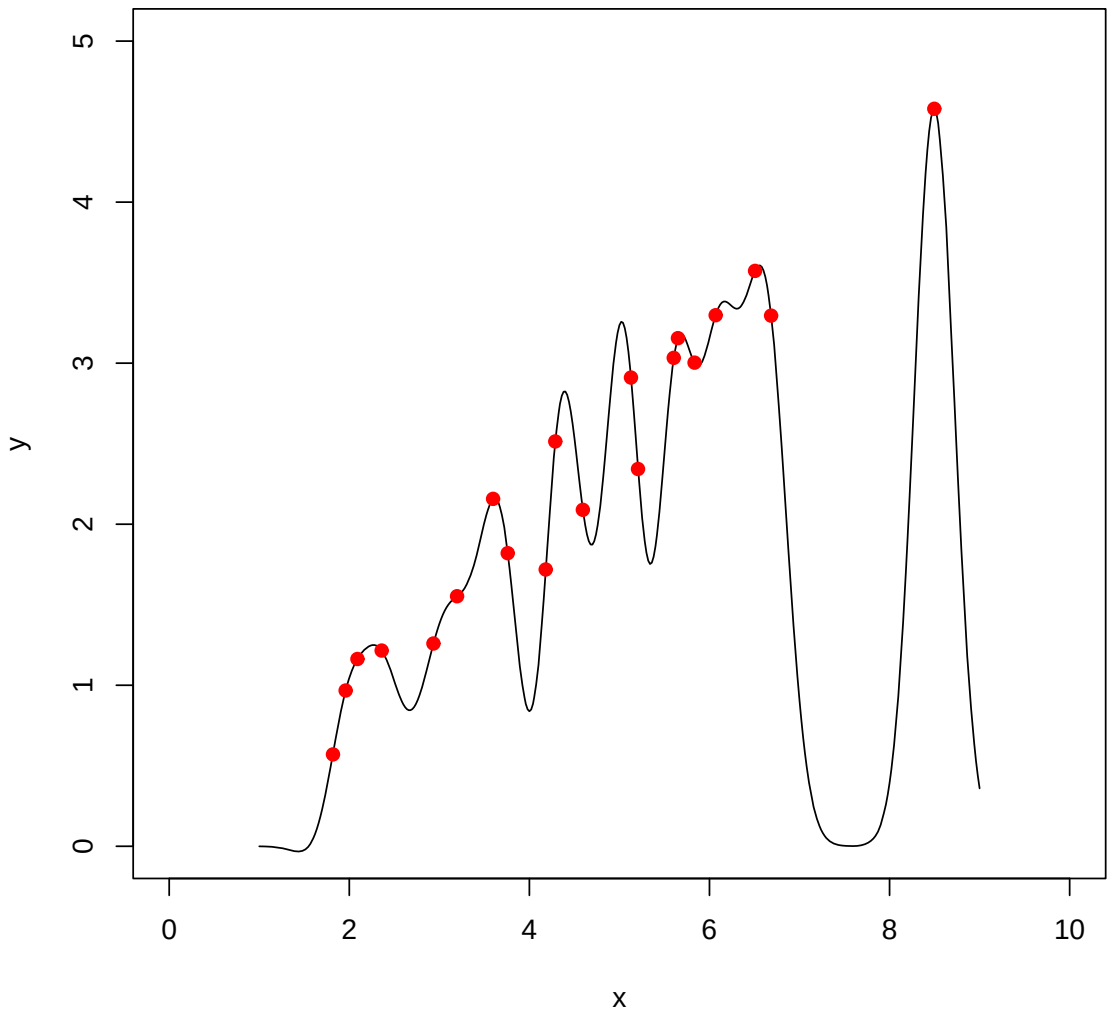
  $L(y, t)$: loss caused by prediction t if y is the true value
- empirical Risk of an estimate $f : \mathcal{X} \to \mathbb{R}$

$$\frac{1}{n} \sum_{i=1}^{n} L\big(y_i, f(x_i)\big)$$

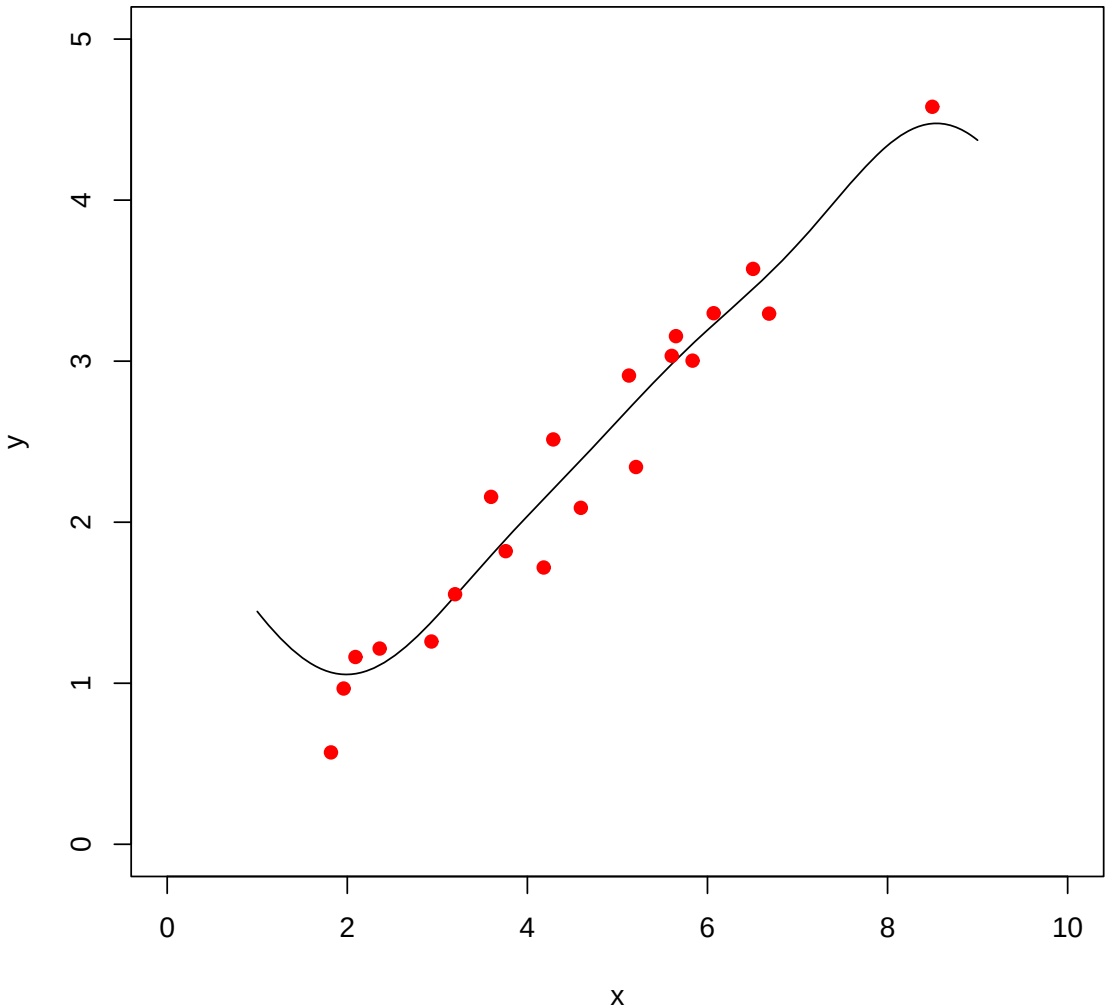
- RKHS H (certain Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$)
- Support vector machine

$$S_n\big((x_1, y_1), \ldots, (x_n, y_n)\big) \;=\; \arg\inf_{f \in H} \; \frac{1}{n} \sum_{i=1}^{n} L\big(y_i, f(x_i)\big) + \lambda \|f\|_H^2$$

Overfitting



Overfitting



Reproducing Kernel Hilbert Space (RKHS)

Support Vector Machines

$$
\begin{array}{rccl}
S_n : & (\mathcal{X} \times \mathcal{Y})^n & \to & H\,, \\
\big((x_1, y_1), \ldots, (x_n, y_n)\big) & & \mapsto & \arg\inf\limits_{f \in H} \; \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} L\big(y_i, f(x_i)\big) \; + \; \lambda \|f\|_H^2
\end{array}
$$

with H a reproducing kernel Hilbert space (RKHS)

Reproducing Kernel Hilbert Space (RKHS)

Support Vector Machines

$$\begin{aligned} S_n \;:\quad (\mathcal{X}\times\mathcal{Y})^n \quad &\to \quad H\,, \\ \big((x_1,y_1),\ldots,(x_n,y_n)\big) \quad &\mapsto \quad \arg\inf_{f\in H} \; \frac{1}{n}\sum_{i=1}^{n} L\big(y_i, f(x_i)\big) \;+\; \lambda\|f\|_H^2 \end{aligned}$$

with H a reproducing kernel Hilbert space (RKHS)

Reproducing kernel Hilbert space H

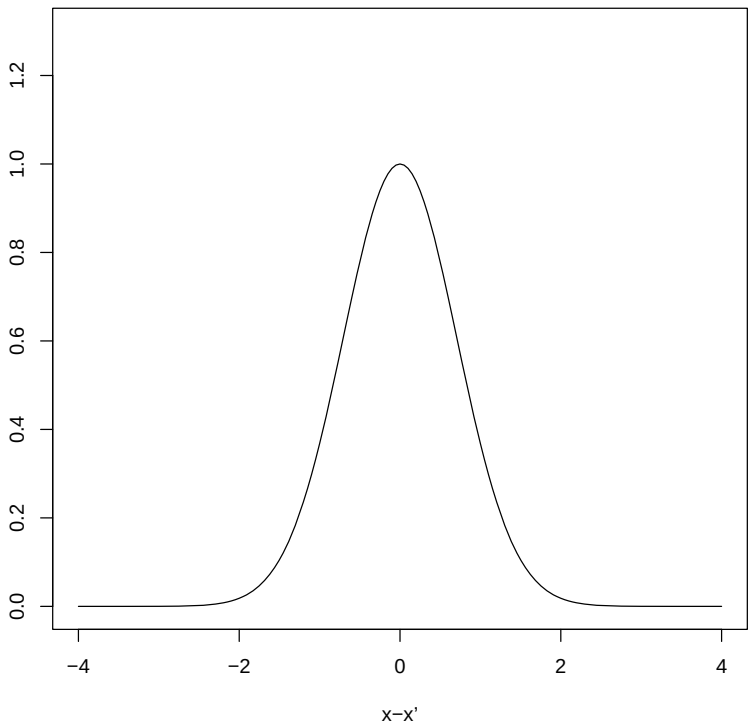
- a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$
- generated by a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
- reproducing property

$$\big\langle f\,,\, k(x,\cdot)\big\rangle_H \;=\; f(x) \qquad \forall\, x\in\mathcal{X}, \quad \forall\, f\in H$$

Example: Gaussian Kernel

Gaussian Kernel $\mathcal{X} = \mathbb{R}$

$$k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \qquad (x, x') \mapsto \exp\left(-\tfrac{1}{\gamma^2}|x - x'|^2\right)$$

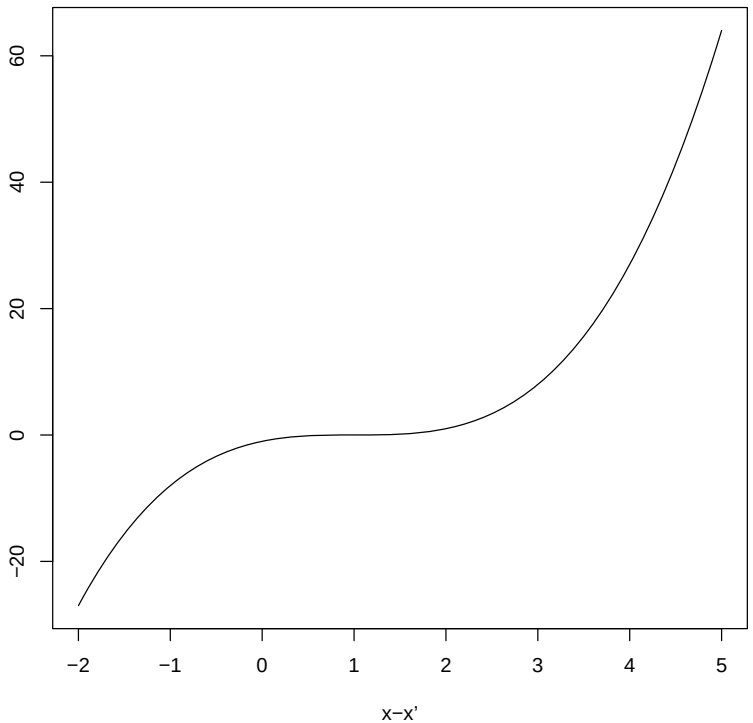


$H \subset L_p(P)$ dense

Example: Polynomial Kernel

Polynomial Kernel $\mathcal{X} = \mathbb{R}$

$$k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \qquad (x, x') \mapsto \left(x \cdot x' + c\right)^m$$



$$H = \left\{ f : \mathbb{R} \to \mathbb{R} \,\middle|\, f \text{ a polynomial with degree } \leq m \right\} \cong \mathbb{R}^{m+1}$$

Representer Theorem

How to calculate the SVM?

$$D_n = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$$

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

Representer Theorem

How to calculate the SVM?

$$D_n = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$$

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

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?

$$D_n = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$$

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

Representer Theorem

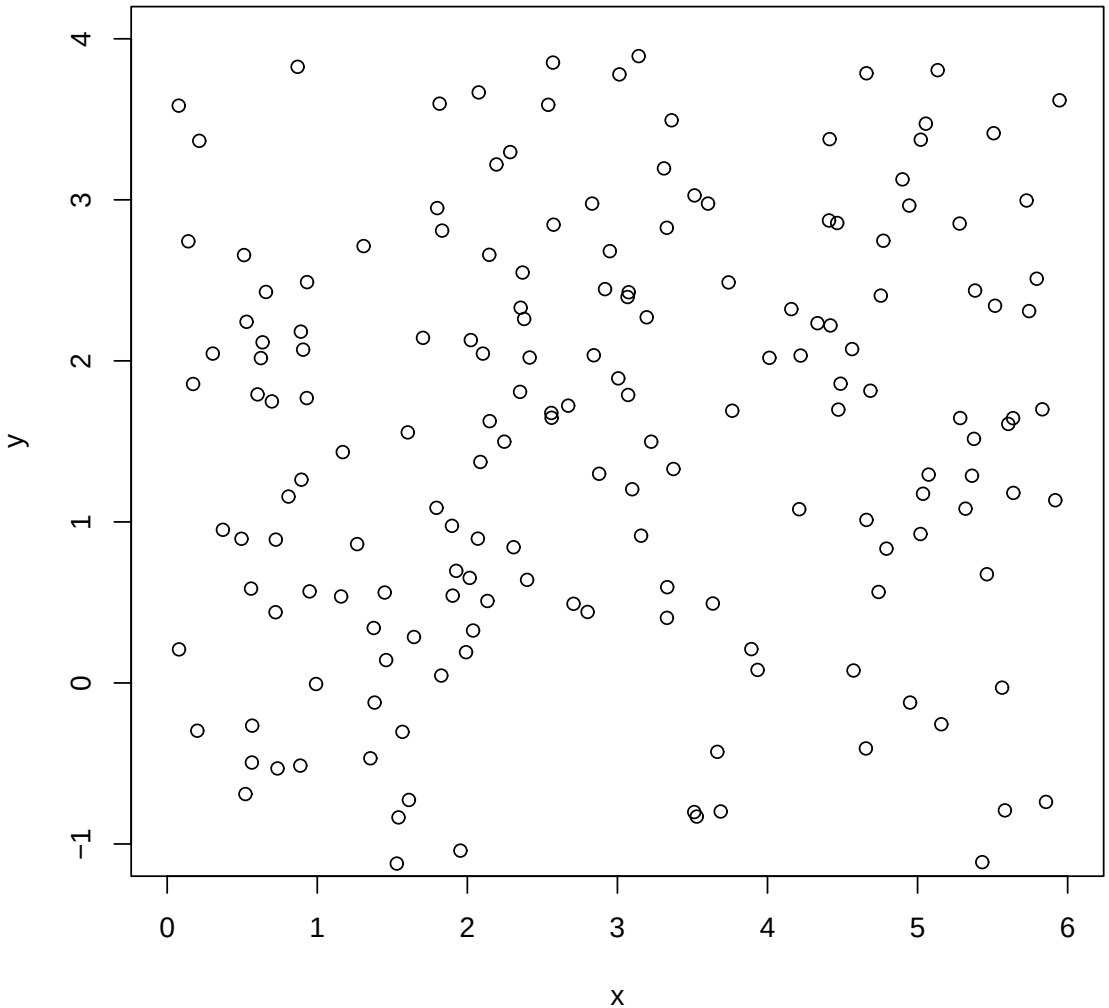
There are $\alpha_{D_n,1}, \ldots, \alpha_{D_n,n} \in \mathbb{R}$ such that

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

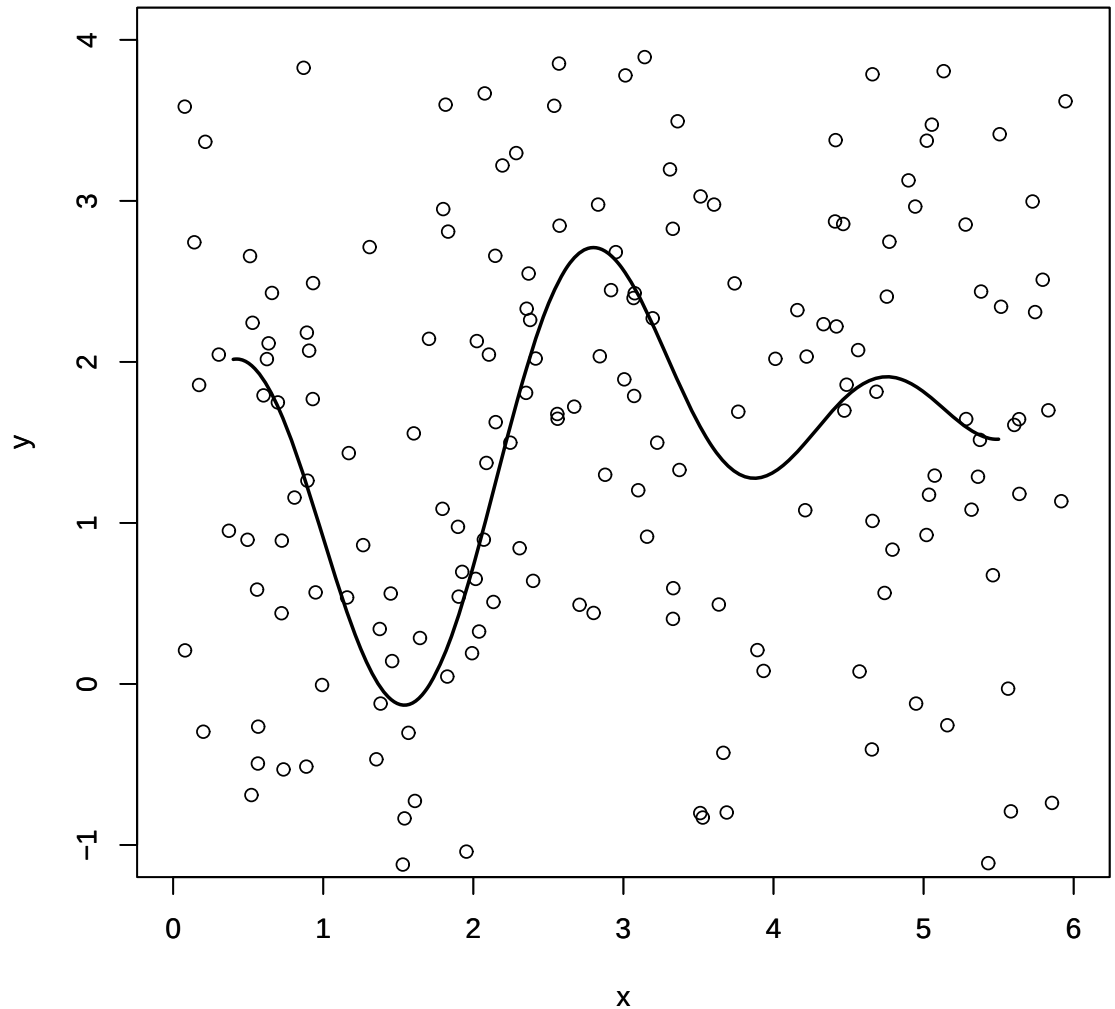
$\longrightarrow$ **just solve a finite convex optimization problem**

. . . and this really works?

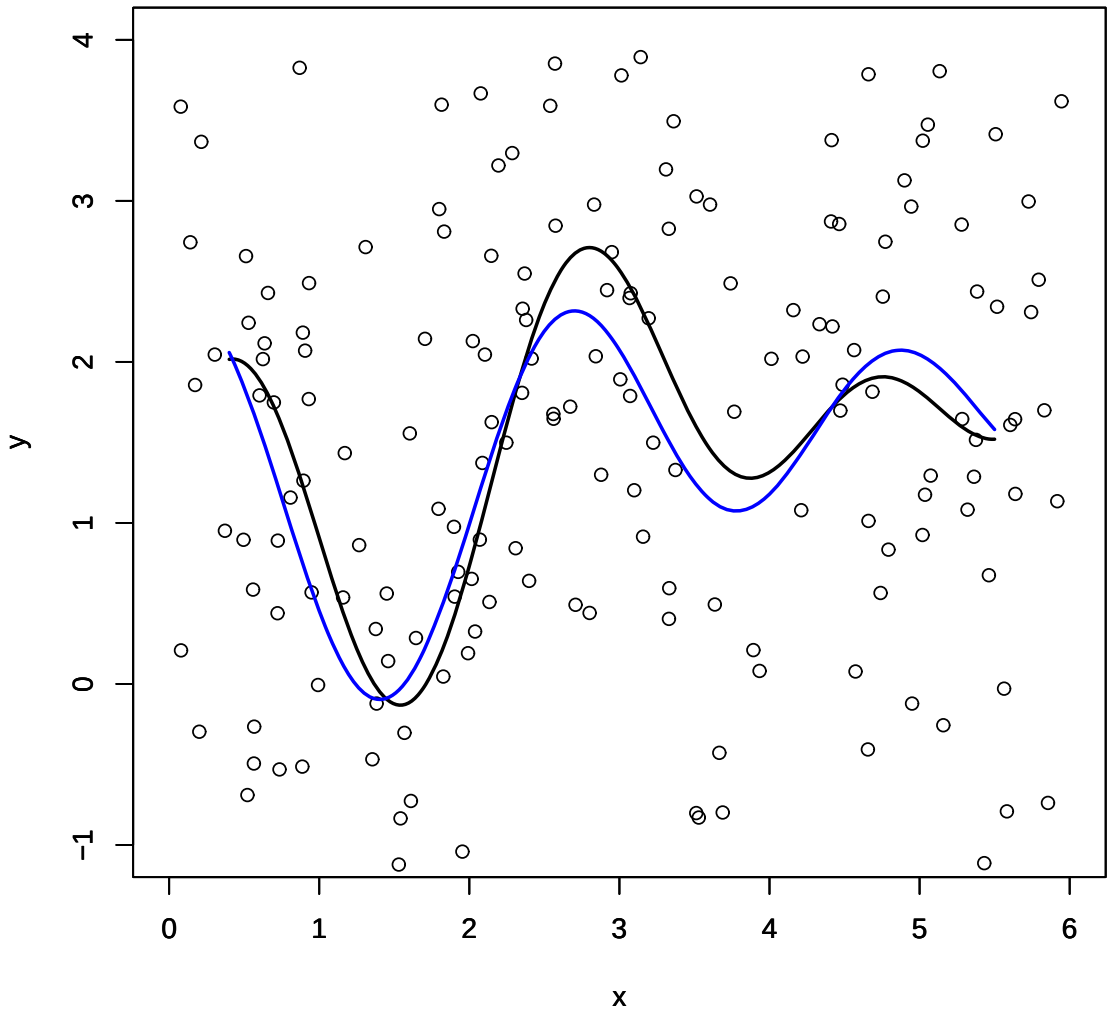
. . . and this really works? Yes, quite good.



. . . and this really works? Yes, quite good.



. . . and this really works? Yes, quite good.



Risk-Consistency

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

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

$$\mathbf{D}_n \;=\; \big((X_1, Y_1), \ldots, (X_n, Y_n)\big)$$

$$\text{SVM:} \qquad f_{\mathbf{D}_n, \lambda_n} \;=\; \arg\inf_{f \in H} \; \frac{1}{n} \sum_{i=1}^{n} L\big(Y_i, f(X_i)\big) \;+\; \lambda_n \|f\|_H^2$$

Risk-Consistency

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

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

$$\mathbf{D}_n \;=\; \big((X_1, Y_1), \ldots, (X_n, Y_n)\big)$$

$$\text{SVM:} \qquad f_{\mathbf{D}_n,\lambda_n} \;=\; \arg\inf_{f \in H} \; \frac{1}{n} \sum_{i=1}^{n} L\big(Y_i, f(X_i)\big) \;+\; \lambda_n \|f\|_H^2$$

Risk-consistency

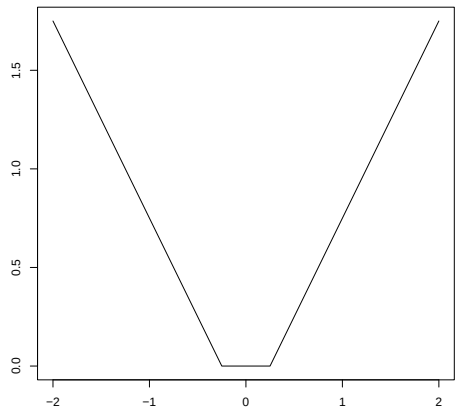
$$\mathcal{R}_P(f_{\mathbf{D}_n,\lambda_n}) \;\xrightarrow[n\to\infty]{}\; \inf_{f:\mathcal{X}\to\mathbb{R}} \mathcal{R}_P(f) \qquad \text{in probability}$$

essentially if

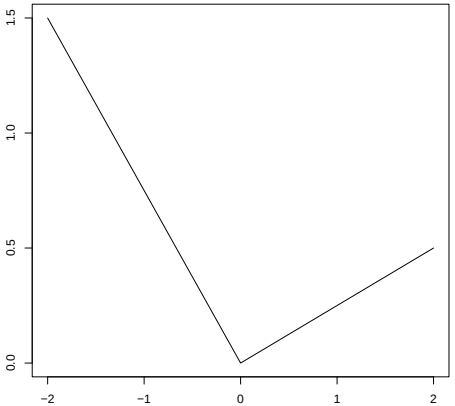
- $H \subset L_p(P)$ dense (e.g. Gaussian kernel)
- $\lambda_n \to 0$ not too fast (!)

Robustness

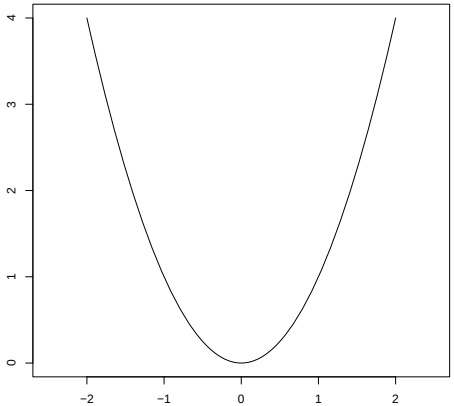
Loss function L



ε-insensitive



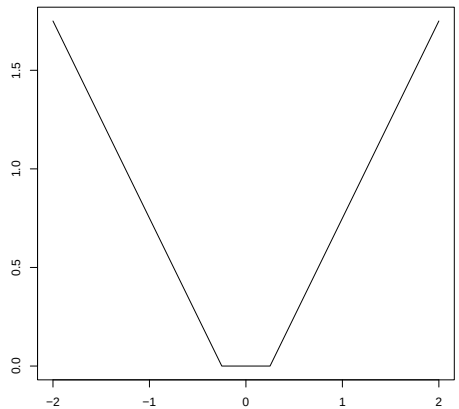
pinball



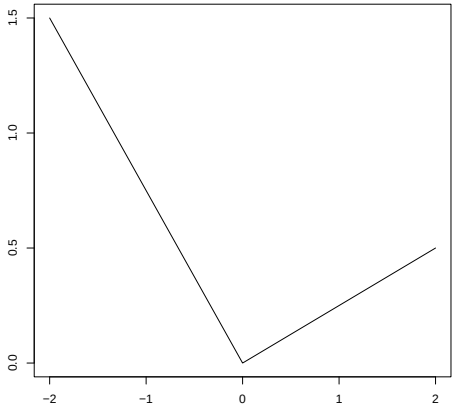
least squares

Robustness

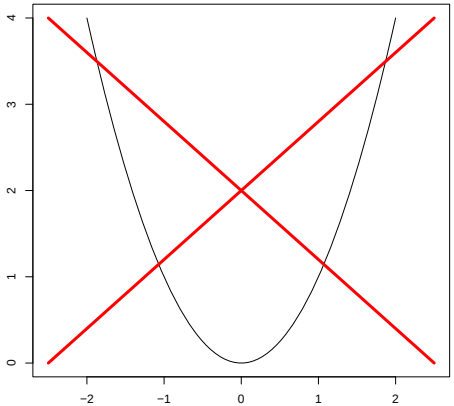
Loss function L should be Lipschitz continuous



ε-insensitive



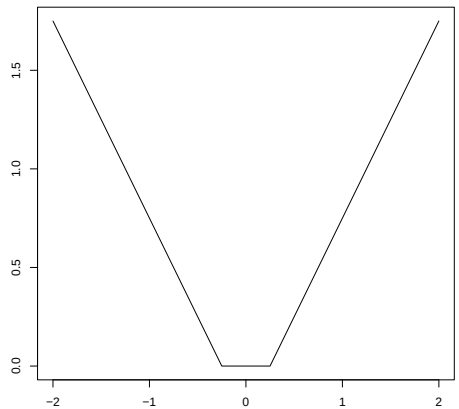
pinball



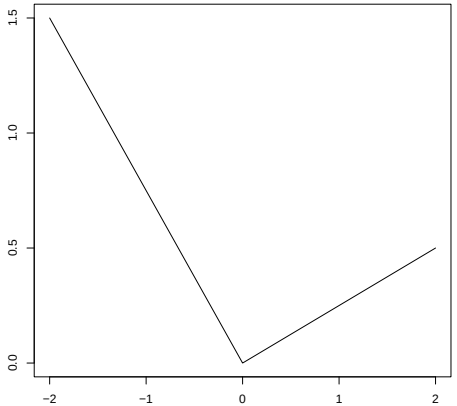
least squares

Robustness

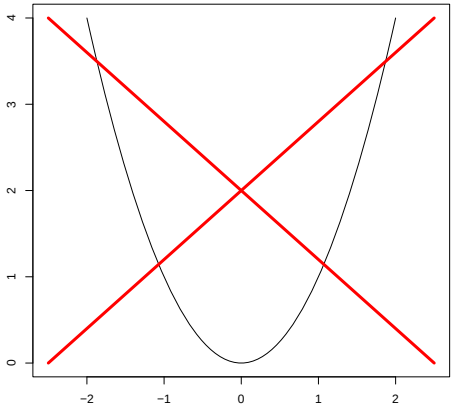
Loss function L should be Lipschitz continuous



ε-insensitive



pinball



least squares

- bounded influence function (Christmann & Van Messem, 2008)
- bounds for maxbias (Christmann & Van Messem, 2008)
- qualitative robustness (Hable & Christmann, 2010)

Rates of Convergence

Risk-consistency

$$\mathcal{R}_P(f_{\mathbf{D}_n,\lambda_n}) \xrightarrow[n\to\infty]{} \inf_{f:\mathcal{X}\to\mathbb{R}} \mathcal{R}_P(f) \qquad \text{in probability}$$

Rates of Convergence

Risk-consistency

$$\mathcal{R}_P(f_{\mathbf{D}_n,\lambda_n}) \xrightarrow[n\to\infty]{} \inf_{f:\mathcal{X}\to\mathbb{R}} \mathcal{R}_P(f) \qquad \text{in probability}$$

How fast is this convergence?

Is there a uniform rate r_n such that

$$r_n\Big(\mathcal{R}_P(f_{\mathbf{D}_n,\lambda_n}) - \inf_{f:\mathcal{X}\to\mathbb{R}} \mathcal{R}_P(f)\Big) \xrightarrow[n\to\infty]{} 0 \quad \text{in probability}$$

for every P?

Rates of Convergence

Risk-consistency

$$\mathcal{R}_P(f_{\mathbf{D}_n,\lambda_n}) \xrightarrow[n\to\infty]{} \inf_{f:\mathcal{X}\to\mathbb{R}} \mathcal{R}_P(f) \qquad \text{in probability}$$

How fast is this convergence?

Is there a uniform rate r_n such that

$$r_n\Big(\mathcal{R}_P(f_{\mathbf{D}_n,\lambda_n}) - \inf_{f:\mathcal{X}\to\mathbb{R}} \mathcal{R}_P(f)\Big) \xrightarrow[n\to\infty]{} 0 \quad \text{in probability}$$

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

Rates of Convergence

Risk-consistency

$$\mathcal{R}_P(f_{\mathbf{D}_n,\lambda_n}) \xrightarrow[n\to\infty]{} \inf_{f:\mathcal{X}\to\mathbb{R}} \mathcal{R}_P(f) \qquad \text{in probability}$$

How fast is this convergence?

Is there a uniform rate r_n such that

$$r_n\Big(\mathcal{R}_P(f_{\mathbf{D}_n,\lambda_n}) - \inf_{f:\mathcal{X}\to\mathbb{R}} \mathcal{R}_P(f)\Big) \xrightarrow[n\to\infty]{} 0 \quad \text{in probability}$$

for every P? $\longrightarrow$ **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 $f^* : \mathcal{X} \to \mathbb{R}$ of

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

or

$$\inf_{f \in H} \mathcal{R}_P(f) \;=\; \text{min!} \qquad f \in H\,.$$

Smooth Approximation of the Regression Function

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

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

or

$$\inf_{f \in H} \mathcal{R}_P(f) \;=\; \text{min!} \qquad f \in H\,.$$

However, these optimization problems

- are ill-posed and
- there is no uniform rate of convergence to the solution
  (without substantial assumptions on P)

Smooth Approximation of the Regression Function

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

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

or

$$\inf_{f \in H} \mathcal{R}_P(f) \;=\; \text{min!} \qquad f \in H\,.$$

However, these optimization problems

- are ill-posed and
- there is no uniform rate of convergence to the solution (without substantial assumptions on P)

Instead, consider the regularized problem

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

Smooth Approximation of the Regression Function

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

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

we may estimate the solution f_{P,λ_0} of the regularized problem

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

f_{P,λ_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

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

  we may estimate the solution f_{P,λ_0} of the regularized problem

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

  f_{P,λ_0} serves as a "smoother approximation" of f^*.

- The regularized problem is equivalent to

$$\mathcal{R}_P(f) \;=\; \text{min!} \qquad f \in H\,, \quad \|f\|_H \le r_0\,.$$

  r_0: bound on complexity of "smoother approximation"

Asymptotic Normality of Regularized Problem

Under some

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

we have

$$\sqrt{n}\Big(\mathcal{R}(f_{\mathbf{D}_n,\lambda_0}) - \mathcal{R}(f_{P,\lambda_0})\Big) \;\rightsquigarrow\; \mathcal{N}(0,\sigma^2)$$

Asymptotic Normality of Regularized Problem

Under some

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

we have

$$\sqrt{n}\Big(\mathcal{R}(f_{\mathbf{D}_n,\lambda_0}) - \mathcal{R}(f_{P,\lambda_0})\Big) \;\rightsquigarrow\; \mathcal{N}(0,\sigma^2)$$

and, even more,

$$\sqrt{n}\big(f_{\mathbf{D}_n,\lambda_0} - f_{P,\lambda_0}\big) \;\rightsquigarrow\; \text{Gaussian process in } H$$

Asymptotic Normality of Regularized Problem

Under some

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

we have

$$\sqrt{n}\Big(\mathcal{R}(f_{\mathbf{D}_n,\lambda_0}) - \mathcal{R}(f_{P,\lambda_0})\Big) \;\rightsquigarrow\; \mathcal{N}(0,\sigma^2)$$

and, even more,

$$\sqrt{n}\big(f_{\mathbf{D}_n,\lambda_0} - f_{P,\lambda_0}\big) \;\rightsquigarrow\; \text{Gaussian process in } H$$

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

$$\sqrt{n}\Big(\mathcal{R}(f_{\mathbf{D}_n,\lambda_n}) - \mathcal{R}(f_{P,\lambda_0})\Big) \;\rightsquigarrow\; \mathcal{N}(0,\sigma^2)$$

and, even more,

$$\sqrt{n}\big(f_{\mathbf{D}_n,\lambda_n} - f_{P,\lambda_0}\big) \;\rightsquigarrow\; \text{Gaussian process in } H$$

Asymptotic Normality of Regularized Problem

Under some

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

we have

$$\sqrt{n}\Big(\mathcal{R}(f_{\mathbf{D}_n,\lambda_{\mathbf{D}_n}}) - \mathcal{R}(f_{P,\lambda_0})\Big) \;\rightsquigarrow\; \mathcal{N}(0,\sigma^2)$$

and, even more,

$$\sqrt{n}\big(f_{\mathbf{D}_n,\lambda_{\mathbf{D}_n}} - f_{P,\lambda_0}\big) \;\rightsquigarrow\; \text{Gaussian process in } H$$

Asymptotic Normality of Regularized Problem

Corollary

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

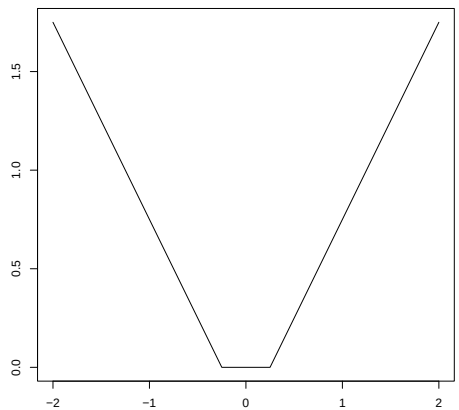
$$\sqrt{n} \begin{pmatrix} f_{\mathbf{D}_n, \lambda_{\mathbf{D}_n}}(x_1) - f_{P,\lambda_0}(x_1) \\ \vdots \\ f_{\mathbf{D}_n, \lambda_{\mathbf{D}_n}}(x_m) - f_{P,\lambda_0}(x_m) \end{pmatrix} \rightsquigarrow \mathcal{N}_m(0, \Sigma)$$

where Σ is a covariance matrix.

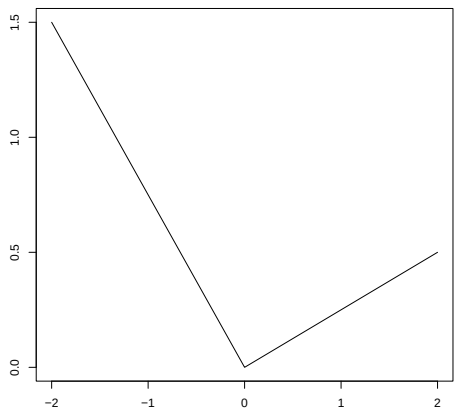
(follows from the reproducing property of the kernel k)

under some assumptions . . .

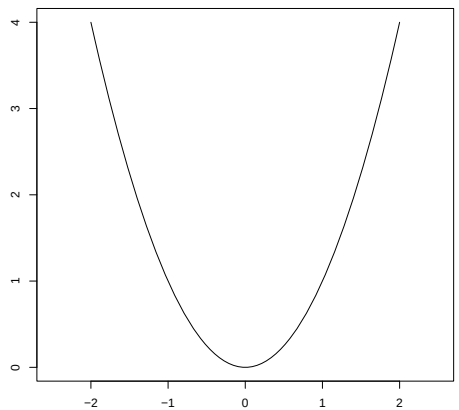
- $\mathcal{X} \subset \mathbb{R}^d$ compact
- k more than $d/2$ - times continuously differentiable
- L smooth (2-times differentiable) and integrable



ε-insensitive



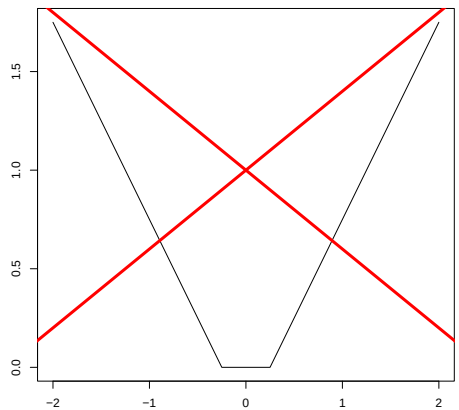
pinball



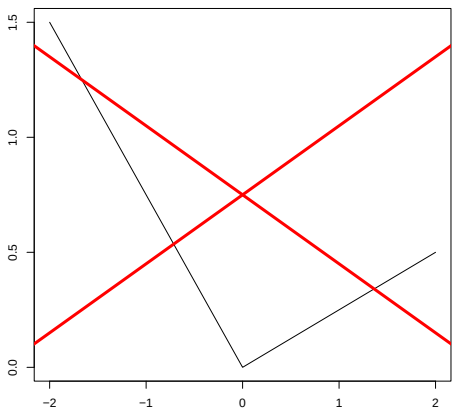
least squares

under some assumptions . . .

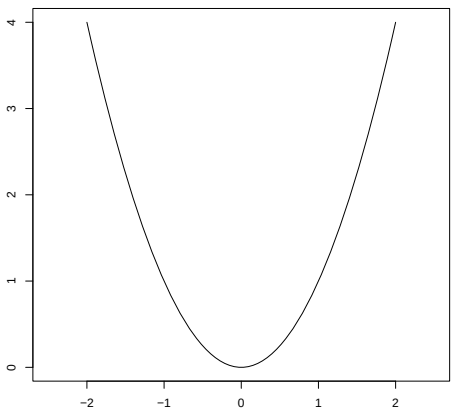
- $\mathcal{X} \subset \mathbb{R}^d$ compact
- k more than $d/2$ - times continuously differentiable
- L smooth (2-times differentiable) and integrable



ε-insensitive



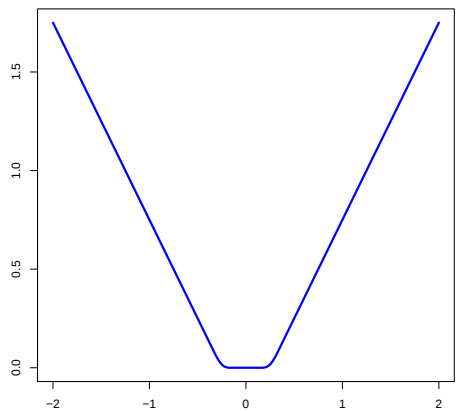
pinball



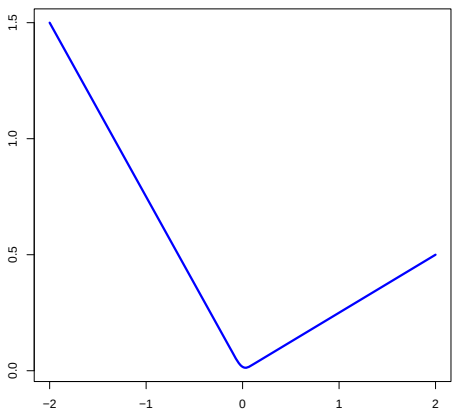
least squares

under some assumptions . . .

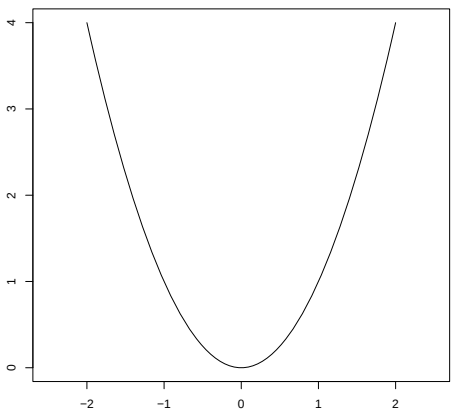
- $\mathcal{X} \subset \mathbb{R}^d$ compact
- k more than $d/2$ - times continuously differentiable
- L smooth (2-times differentiable) and integrable



smoothed ε-insensitive



smoothed pinball



least squares

under some assumptions . . .

and

- $\sqrt{n}\big(\lambda_{\mathbf{D}_n} - \lambda_0\big) \xrightarrow[n\to\infty]{} 0$ in probability.

under some assumptions . . .

and

- $\sqrt{n}\big(\lambda_{\mathbf{D}_n} - \lambda_0\big) \xrightarrow[n\to\infty]{} 0$ in probability.

For example:

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

$$\left[\lambda_0 \ , \ \lambda_0 + \frac{c}{\sqrt{n \ln(n)}}\right] .$$

Sketch of the Proof: Functional Delta-Method

Consider the SVM-functional

$$S : \mathcal{M}_1 \to H, \qquad P \mapsto f_{P,\lambda_0}$$

SVM-functional represents SVM-estimator:

$$f_{\mathbf{D}_n,\lambda_0} = S(\mathbb{P}_{\mathbf{D}_n})$$

Sketch of the Proof: Functional Delta-Method

Consider the SVM-functional

$$S : \mathcal{M}_1 \to H, \qquad P \mapsto f_{P,\lambda_0}$$

SVM-functional represents SVM-estimator:

$$f_{\mathbf{D}_n,\lambda_0} = S(\mathbb{P}_{\mathbf{D}_n})$$

1. Show that $\sqrt{n}(\mathbb{P}_{\mathbf{D}_n} - P)$ 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

$$S : \mathcal{M}_1 \to H, \qquad P \mapsto f_{P,\lambda_0}$$

SVM-functional represents SVM-estimator:

$$f_{\mathbf{D}_n,\lambda_0} = S(\mathbb{P}_{\mathbf{D}_n})$$

1. Show that $\sqrt{n}(\mathbb{P}_{\mathbf{D}_n} - P)$ converges weakly to a Gaussian process in a suitable space $\ell_\infty(\mathcal{G})$.
2. Show that S is Hadamard-differentiable.

Sketch of the Proof: Functional Delta-Method

Consider the SVM-functional

$$S : \mathcal{M}_1 \to H, \qquad P \mapsto f_{P,\lambda_0}$$

SVM-functional represents SVM-estimator:

$$f_{\mathbf{D}_n,\lambda_0} = S(\mathbb{P}_{\mathbf{D}_n})$$

1. Show that $\sqrt{n}(\mathbb{P}_{\mathbf{D}_n} - P)$ 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

$$\sqrt{n}(f_{\mathbf{D}_n,\lambda_0} - f_{P,\lambda_0}) = \sqrt{n}\big(S(\mathbb{P}_{\mathbf{D}_n}) - S(P)\big)$$

converges weakly to a Gaussian process.

Sketch of the Proof: Functional Delta-Method

Consider the SVM-functional

$$S : \mathcal{M}_1 \to H, \qquad P \mapsto f_{P,\lambda_0}$$

SVM-functional represents SVM-estimator:

$$f_{\mathbf{D}_n,\lambda_0} = S(\mathbb{P}_{\mathbf{D}_n})$$

1. Show that $\sqrt{n}(\mathbb{P}_{\mathbf{D}_n} - P)$ 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

$$\sqrt{n}(f_{\mathbf{D}_n,\lambda_0} - f_{P,\lambda_0}) = \sqrt{n}\big(S(\mathbb{P}_{\mathbf{D}_n}) - S(P)\big)$$

converges weakly to a Gaussian process.

Sketch of the Proof: Functional Delta-Method

How to deal with random parameters $\lambda_{\mathbf{D}_n}$?

Sketch of the Proof: Functional Delta-Method

How to deal with random parameters $\lambda_{\mathbf{D}_n}$?

Problem

$$\int L\big(y, f(x)\big)\, P\big(d(x,y)\big) + \lambda \|f\|_H^2 \;=\; \text{min!} \qquad f \in H\,.$$

is equivalent to

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

Sketch of the Proof: Functional Delta-Method

How to deal with random parameters $\lambda_{\mathbf{D}_n}$?

Problem

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

is equivalent to

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

and

$$\int L\big(y, f(x)\big)\, \big(\tfrac{\lambda_0}{\lambda} P\big)\big(d(x,y)\big) + \lambda_0 \|f\|_H^2 \;=\; \min! \qquad f \in H\,.$$

Hence, $\quad f_{P,\lambda} \;=\; f_{\frac{\lambda_0}{\lambda} P, \lambda_0} \;=\; S\big(\tfrac{\lambda_0}{\lambda} P\big)$

Sketch of the Proof: Functional Delta-Method

Consider the SVM-functional

$$S \;:\; \mathcal{M}_1 \;\to\; H\,, \qquad P \;\mapsto\; f_{P,\lambda_0}$$

SVM-functional represents SVM-estimator:

$$f_{\mathbf{D}_n,\lambda_{\mathbf{D}_n}} = S\Big(\tfrac{\lambda_0}{\lambda_{\mathbf{D}_n}}\mathbb{P}_{\mathbf{D}_n}\Big)$$

1. Show that $\sqrt{n}(\mathbb{P}_{\mathbf{D}_n} - P)$ 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

$$\sqrt{n}(f_{\mathbf{D}_n,\lambda_0} - f_{P,\lambda_0}) \;=\; \sqrt{n}\big(S(\mathbb{P}_{\mathbf{D}_n}) - S(P)\big)$$

converges weakly to a Gaussian process.

Sketch of the Proof: Functional Delta-Method

Consider the SVM-functional

$$S \;:\; \mathcal{M}_1 \;\to\; H\,, \qquad P \;\mapsto\; f_{P,\lambda_0}$$

SVM-functional represents SVM-estimator:

$$f_{\mathbf{D}_n,\lambda_{\mathbf{D}_n}} = S\left(\tfrac{\lambda_0}{\lambda_{\mathbf{D}_n}}\mathbb{P}_{\mathbf{D}_n}\right)$$

1. Show that $\sqrt{n}(\mathbb{P}_{\mathbf{D}_n} - P)$ 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

$$\sqrt{n}(f_{\mathbf{D}_n,\lambda_{\mathbf{D}_n}} - f_{P,\lambda_0}) \;=\; \sqrt{n}\left(S\left(\tfrac{\lambda_0}{\lambda_{\mathbf{D}_n}}\mathbb{P}_{\mathbf{D}_n}\right) - S(P)\right)$$

converges weakly to a Gaussian process.

Sketch of the Proof: Functional Delta-Method

Consider the SVM-functional

$$S \;:\; \mathcal{M}_f \;\to\; H\,, \qquad \mu \;\mapsto\; f_{\mu,\lambda_0}$$

SVM-functional represents SVM-estimator:

$$f_{\mathbf{D}_n,\lambda_{\mathbf{D}_n}} = S\left(\tfrac{\lambda_0}{\lambda_{\mathbf{D}_n}}\mathbb{P}_{\mathbf{D}_n}\right)$$

1. Show that $\sqrt{n}(\mathbb{P}_{\mathbf{D}_n} - P)$ 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

$$\sqrt{n}(f_{\mathbf{D}_n,\lambda_{\mathbf{D}_n}} - f_{P,\lambda_0}) \;=\; \sqrt{n}\left(S\left(\tfrac{\lambda_0}{\lambda_{\mathbf{D}_n}}\mathbb{P}_{\mathbf{D}_n}\right) - S(P)\right)$$

converges weakly to a Gaussian process.

References

- **R. Hable (2010)**: Asymptotic Normality of Support Vector Machines for Classification and Regression. *arXiv:1010.0535v2*
- **I. Steinwart and A. Christmann (2008)**: *Support vector machines*. Springer, New York.
- **V. N. Vapnik (1998)**: *Statistical learning theory*. John Wiley & Sons, New York.

The handout to this talk is also available on my homepage

http://www.staff.uni-bayreuth.de/~btms04/index.html