'Fast rates for support vector machines using Gaussian kernels' - Ingo Steinwart & Clint Scovel

Statistical Learning Seminar 3. Vortrag

Ben Deitmar

02.06.2020

Albert-Ludwigs-Universität Freiburg

Gliederung

- Wiederholung der Grundlagen
- Hilberträume mit reproduzierendem Kern
- ERM mit Regularisierung
- Vorbereitende Definitionen
- Hauptaussage des Artikels
- Andere Betrachtungsweise des Klassifikators

Grundlegende Fragestellung

- P Verteilung auf $[0,1] \times \{-1,1\}$
- $T_n := ((X_1, Y_1), ..., (X_n, Y_n)) \sim \bigotimes_{i=1}^{n} P$ Trainingsdatensatz
- $(f_\alpha)_{\alpha \in \Lambda} \subset \{f : [0,1] \to \mathbb{R}\, mb.\}$ Funktionenfamilie

Finde abhängig von T_n ohne Kenntnis von P ein $\alpha \in \Lambda$, sodass f_α möglichst geringes Risiko besitzt.

Algorythmus: $\Phi : ([0,1] \times \{-1,1\})^n \to (f_\alpha)_{\alpha \in \Lambda}$

Risiko

- $g : [0, 1] \to \mathbb{R}$ messbar
- $L : \mathbb{R} \times \{-1, 1\} \to [0, \infty]$ Verlustfunktion

Risiko:

$$R_{L,P}(g) := \mathbb{E}_{(X,Y)\sim P}\left[L(g(X), Y)\right] \in [0, \infty]$$

Risiko

- $g : [0,1] \to \mathbb{R}$ messbar
- $L : \mathbb{R} \times \{-1,1\} \to [0,\infty]$ Verlustfunktion

Risiko:

$$R_{L,P}(g) := \mathbb{E}_{(X,Y)\sim P}\left[L(g(X),Y)\right] \in [0,\infty]$$

Aufteilung:

$$R_{L,P}(g) = \underbrace{R_{L,P}(g) - \inf_{\alpha\in\Lambda} R_{L,P}(f_\alpha)}_{\text{Algorithmus-Risiko (estimation error)}} + \underbrace{\inf_{\alpha\in\Lambda} R_{L,P}(f_\alpha) - \inf_{f:[0,1]\to\{-1,1\}\text{ mb.}} R_{L,P}(f)}_{\text{Familien-Risiko (approximation error)}} + \underbrace{\inf_{f:[0,1]\to\{-1,1\}\text{ mb.}} R_{L,P}(f)}_{\text{Grund-Risiko (statistical risk)}}$$

Empirical Risk Minimization

Algorythmus für beliebige Familie $(f_\alpha)_{\alpha \in \Lambda}$:

$$\Phi(T_n) := \arg\min_{\alpha \in \Lambda} R_{L,\hat{P}_n}(f_\alpha) = \arg\min_{\alpha \in \Lambda} \frac{1}{n} \sum_{i=1}^{n} L(f_\alpha(X_i), Y_i)$$

Empirical Risk Minimization

Algorythmus für beliebige Familie $(f_\alpha)_{\alpha\in\Lambda}$:

$$\Phi(T_n) := \arg\min_{\alpha\in\Lambda} R_{L,\hat{P}_n}(f_\alpha) = \arg\min_{\alpha\in\Lambda} \frac{1}{n}\sum_{i=1}^{n} L(f_\alpha(X_i), Y_i)$$

Problem:

- großes Algorythmus-Risiko (Overfitting) bei zu großen Familien $(f_\alpha)_{\alpha\in\Lambda}$
- großes Familien-Risiko bei zu kleinen Familien $(f_\alpha)_{\alpha\in\Lambda}$

Abschätzungen

Mögliche Formen:

- $\mathbb{E}_{P^{\otimes n}}\left[R_{L,P}(\Phi(T_n)) - \inf_{\alpha\in\Lambda} R_{L,P}(f_\alpha)\right] \leq C(n)$
- $\mathbb{P}_{P^{\otimes n}}\left(R_{L,P}(\Phi(T_n)) - \inf_{f:[0,1]\to\{-1,1\}\ mb.} R_{L,P}(f) > \varepsilon(n)\right) \leq C(n)$

Sollten am besten für ganze Klassen von Verteilungen gelten.

Hilberträume mit reproduzierendem Kern (RKHS.)

$(H, \langle \cdot, \cdot \rangle_H)$ ist ein Funktionen-Hilbertraum auf $[0, 1]$, falls:

- $H \subset \{f : [0,1] \to \mathbb{R}\, mb.\}$ Vektorraum
- $\langle \cdot, \cdot \rangle_H : H \times H \to \mathbb{R}$ Skalarprodukt
- H bezüglich $|| \cdot ||_H$ vollständig

Hilberträume mit reproduzierendem Kern (RKHS.)

$(H, \langle \cdot, \cdot \rangle_H)$ ist ein Funktionen-Hilbertraum auf $[0,1]$, falls:

- $H \subset \{f : [0,1] \to \mathbb{R}\, mb.\}$ Vektorraum
- $\langle \cdot, \cdot \rangle_H : H \times H \to \mathbb{R}$ Skalarprodukt
- H bezüglich $|| \cdot ||_H$ vollständig

Wir definieren für jedes $x \in [0,1]$:

$$\varphi_x : H \to \mathbb{R} \quad ; \quad \varphi_x(h) := h(x)$$

Hilberträume mit reproduzierendem Kern (RKHS.)

Definition:

$$(H, \langle \cdot, \cdot \rangle_H) \text{ ist } \textbf{RKHS.} \Leftrightarrow \forall x \in [0,1] : \varphi_x \text{ ist stetig}$$

Hilberträume mit reproduzierendem Kern (RKHS.)

Definition:

$$(H, \langle \cdot, \cdot \rangle_H) \text{ ist } \textbf{RKHS.} \Leftrightarrow \ \forall x \in [0,1] : \ \varphi_x \text{ ist stetig}$$

Darstellungssatz von Riesz (für Hilberträume):
Sei H ein Hilbertraum. Für jede stetige lineare Abbildung (Funktional) $\varphi : H \to \mathbb{R}$ existiert ein eindeutig bestimmtes $h_\varphi \in H$, sodass

$$\forall f \in H : \ \varphi(f) = \langle f, h_\varphi \rangle_H.$$

Hilberträume mit reproduzierendem Kern (RKHS.)

Definition:

$$(H, \langle \cdot, \cdot \rangle_H) \text{ ist } \textbf{RKHS.} \Leftrightarrow \ \forall x \in [0,1] : \ \varphi_x \text{ ist stetig}$$

Darstellungssatz von Riesz (für Hilberträume):
Sei H ein Hilbertraum. Für jede stetige lineare Abbildung (Funktional) $\varphi : H \to \mathbb{R}$ existiert ein eindeutig bestimmtes $h_\varphi \in H$, sodass

$$\forall f \in H : \ \varphi(f) = \langle f, h_\varphi \rangle_H.$$

Wir definieren:

$$K_x := h_{\varphi_x} \in H$$

Der reproduzierende Kern (RKHS.)

Eigenschaften der K_x:

- $\forall f \in H, \forall x \in [0,1]: \; f(x) = \langle f, K_x \rangle_H$
- $\forall x, y \in [0,1]: \; K_y(x) = \langle K_y, K_x \rangle_H$
- $\text{Span}_{\mathbb{R}}(K_x \mid x \in [0,1])$ ist dicht in H.

Der reproduzierende Kern (RKHS.)

Eigenschaften der K_x:

- $\forall f \in H, \forall x \in [0,1]: \; f(x) = \langle f, K_x \rangle_H$
- $\forall x, y \in [0,1]: \; K_y(x) = \langle K_y, K_x \rangle_H$
- $\mathrm{Span}_{\mathbb{R}}(K_x \mid x \in [0,1])$ ist dicht in H.

Wir definieren den **(reproduzierenden) Kern** K durch

$$K : [0,1] \times [0,1] \to \mathbb{R} \quad ; \quad K(x,y) := \langle K_y, K_x \rangle_H.$$

Der reproduzierende Kern (RKHS.)

Satz von Moore-Aronszajin:
Sei $K : [0,1] \times [0,1] \to \mathbb{R}$ positiv definit und symmetrisch, dann existiert genau ein RKHS. H, sodass K der reproduzierende Kern von H ist.

Der reproduzierende Kern (RKHS.)

Satz von Moore-Aronszajin:
Sei $K : [0,1] \times [0,1] \to \mathbb{R}$ positiv definit und symmetrisch, dann existiert genau ein RKHS. H, sodass K der reproduzierende Kern von H ist.

Wir definieren für $\sigma > 0$:

$$K_\sigma(x, y) := e^{-\frac{(x-y)^2}{\sigma^2}}$$

und den entsprechenden RKHS. als H_σ.

Skalarprodukt auf H_σ

Nach Definition:

$$\langle K_x, K_y \rangle_{H_\sigma} := K_\sigma(x, y) = e^{-\frac{(x-y)^2}{\sigma^2}}$$

H_σ wird als Vervollständigung von $(\mathrm{Span}_{\mathbb{R}}(K_x \mid x \in [0,1]), \langle K_x, K_y \rangle_{H_\sigma})$ Konstruiert. Daraus folgt:

- H_σ eindeutig bestimmt
- $\langle \cdot, \cdot \rangle_{H_\sigma}$ wohldefiniert

Eigenschaften von H_σ

Für jedes $h \in H_\sigma$ gilt:

- $||h||_{[0,1]} \leq ||h||_{H_\sigma}$
- $|h(x) - h(y)| \leq \sqrt{\frac{2}{\sigma^2}}\, ||h||_{H_\sigma}\, |x - y|$
- $||h||_{TV} := \lim_{m\to\infty} \sup_{x_1<\ldots<x_m} \sum_{i=1}^{m-1} |h(x_i) - h(x_{i+1})| \leq \sqrt{\frac{2}{\sigma^2}}\, ||h||_{H_\sigma}$

Eigenschaften von H_σ

Für jedes $h \in H_\sigma$ gilt:

- $||h||_{[0,1]} \leq ||h||_{H_\sigma}$
- $|h(x) - h(y)| \leq \sqrt{\frac{2}{\sigma^2}}\, ||h||_{H_\sigma}\, |x - y|$
- $||h||_{TV} := \lim_{m\to\infty} \sup_{x_1<\ldots<x_m} \sum_{i=1}^{m-1} |h(x_i) - h(x_{i+1})| \leq \sqrt{\frac{2}{\sigma^2}}\, ||h||_{H_\sigma}$

Beweis-Idee:

$$
\begin{aligned}
|h(x) - h(y)|^2 &= |\varphi_x(h) - \varphi_y(h)|^2 = |\langle h, K_x - K_y\rangle_{H_\sigma}|^2 \\
&\overset{CS.}{\leq} ||h||^2_{H_\sigma}\, ||K_x - K_y||^2_{H_\sigma} = 2||h||^2_{H_\sigma} \left(1 - e^{-\frac{(x-y)^2}{\sigma^2}}\right) \\
&\leq 2||h||^2_{H_\sigma} \frac{(x-y)^2}{\sigma^2} = \left(\sqrt{\frac{2}{\sigma^2}}\, ||h||_{H_\sigma}\, |x - y|\right)^2
\end{aligned}
$$

□

H_σ hat kein Familien-Risiko

Lemma 2.1:
Sei $\sigma > 0$ beliebig, dann ist H_σ eine dichte Teilmenge von $C([0,1])$.

Beweis in der Ausarbeitung.
(Arzelà-Acoli & Hahn-Banach & Riesz-Markov)

H_σ hat kein Familien-Risiko

Lemma 2.1:
Sei $\sigma > 0$ beliebig, dann ist H_σ eine dichte Teilmenge von $C([0,1])$.

Beweis in der Ausarbeitung.
(Arzelà-Acoli & Hahn-Banach & Riesz-Markov)

Ähnlich: 'Universal Approximation Theorem'
$\rho(x) := e^{-x^2}$ ist 'discrimminatory'.

H_σ hat kein Familien-Risiko

Falls die Verlustfunktion L im 1. Argument Lipschitz-stetig ist, dann verschwindet das Familienrisiko von H_σ.

H_σ hat kein Familien-Risiko

Falls die Verlustfunktion L im 1. Argument Lipschitz-stetig ist, dann verschwindet das Familienrisiko von H_σ.

Beweis-Idee:
$C([0,1])$ ist dicht in $L^1(P_X)$ und H_σ ist dicht in $C([0,1])$.
$\Rightarrow \forall f \in L^1(P_X)\, \exists (h_m)_{m\in\mathbb{N}} \subset H_\sigma : \mathbb{E}_{P_X}[|h_m - f|] \xrightarrow{m\to\infty} 0$

$$|R_{L,P}(h_m) - R_{L,P}(f)| = \left| \mathbb{E}_{P_X} \left[\sum_{i\in\{-1,1\}} P(Y = i \mid X)(L(h_m(X), i) - L(f(X), i)) \right] \right|$$

$$\leq \sum_{i\in\{-1,1\}} \mathbb{E}_{P_X}\left[|L(h_m(X), i) - L(f(X), i)|\right] \leq C\, \mathbb{E}_{P_X}\left[|h_m(X) - f(X)|\right] \to 0 \quad \square$$

ERM mit Regularisierung (ERMR)

Für Nullfolgen $(\sigma_n)_{n\in\mathbb{N}}, (\lambda_n)_{n\in\mathbb{N}}$ ist die **ERM mit Regularisierung** definiert durch:

$$h_{T_n} := \Phi(T_n) := \arg\min_{h\in H_{\sigma_n}} \left(\lambda_n ||h||^2_{H_{\sigma_n}} + R_{L,\hat{P}_n}(h)\right) \in H_{\sigma_n}$$

ERM mit Regularisierung (ERMR)

Für Nullfolgen $(\sigma_n)_{n\in\mathbb{N}}, (\lambda_n)_{n\in\mathbb{N}}$ ist die **ERM mit Regularisierung** definiert durch:

$$h_{T_n} := \Phi(T_n) := \arg\min_{h\in H_{\sigma_n}} \left(\lambda_n ||h||^2_{H_{\sigma_n}} + R_{L,\hat{P}_n}(h)\right) \in H_{\sigma_n}$$

Idee:
Die $||\cdot||_{H_\sigma}$-Norm enthält Information über die 'Komlpexität' der Funktion.
Hohe 'Komplexität' wird (für große n immer weniger) bestraft.

ERMR sucht nur aus $B^{H_{\sigma_n}}_{\sqrt{\frac{1}{\lambda_n}}}(0)$ aus.

Für die ERMR mit Verlustfunktion $L(a, y) := |a - y|$ gilt:

$$||\Phi(T_n)||_{H_{\sigma_n}} \leq \sqrt{\frac{1}{\lambda_n}}$$

ERMR sucht nur aus $B^{H_{\sigma_n}}_{\sqrt{\frac{1}{\lambda_n}}}(0)$ aus.

Für die ERMR mit Verlustfunktion $L(a, y) := |a - y|$ gilt:

$$||\Phi(T_n)||_{H_{\sigma_n}} \leq \sqrt{\frac{1}{\lambda_n}}$$

Beweis: Nach Def. gilt

$$\left(\lambda_n ||\Phi(T_n)||^2_{H_{\sigma_n}} + R_{L,\hat{P}_n}(\Phi(T_n))\right) \leq \left(\lambda_n ||0||^2_{H_{\sigma_n}} + R_{L,\hat{P}_n}(0)\right) = 0 + 1$$

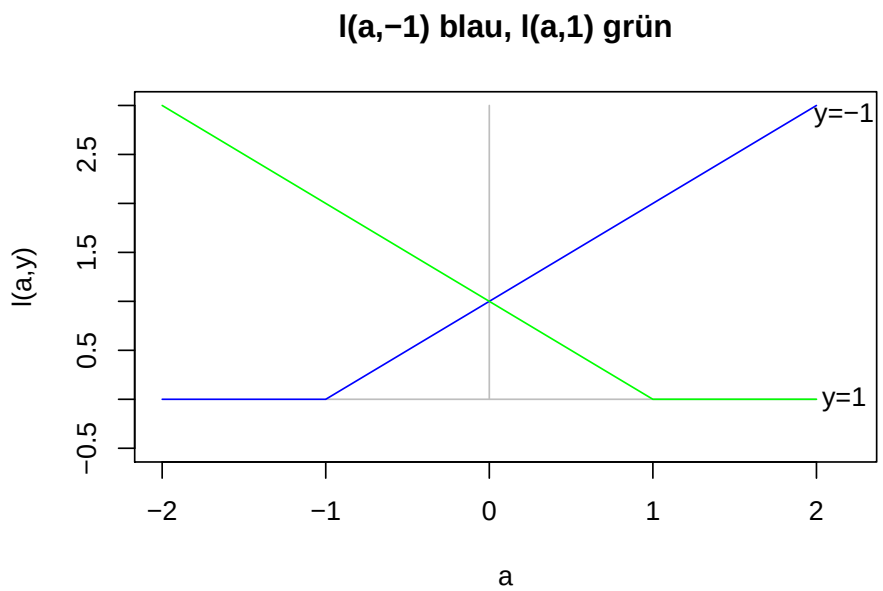
$$\Rightarrow \lambda_n ||\Phi(T_n)||^2_{H_{\sigma_n}} \leq 1 \Rightarrow ||\Phi(T_n)||_{H_{\sigma_n}} \leq \sqrt{\frac{1}{\lambda_n}}.$$ □

Analog gilt die Aussage mit der 'Hinge-Loss'-Verlustfunktion.

Hinge-Loss

Wir definieren die Verlustfunktion 'Hinge-Loss' durch:

$$l(a, y) := (1 - ay)^+$$



Tsybakov-Noise-Exponent (TNE)

Sei P eine Verteilung auf $[0,1] \times \{-1,1\}$, dann hat P den **Tsybakov-Noise-Exponent** $q \in [0,\infty]$, falls

$$\exists \varepsilon > 0,\ \exists C > 0,\ \forall t \in (0,\varepsilon):\ P_X\left(\left|\eta(X) - \frac{1}{2}\right| \leq t\right) \leq C\, t^q.$$

Für $\eta \overset{fs.}{=} P(Y = 1 \mid X = \cdot)$.

Tsybakov-Noise-Exponent (TNE)

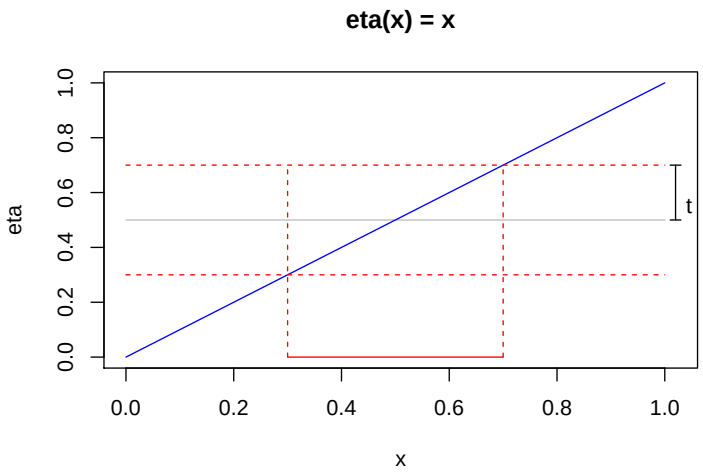
Sei P eine Verteilung auf $[0,1] \times \{-1,1\}$, dann hat P den **Tsybakov-Noise-Exponent** $q \in [0,\infty]$, falls

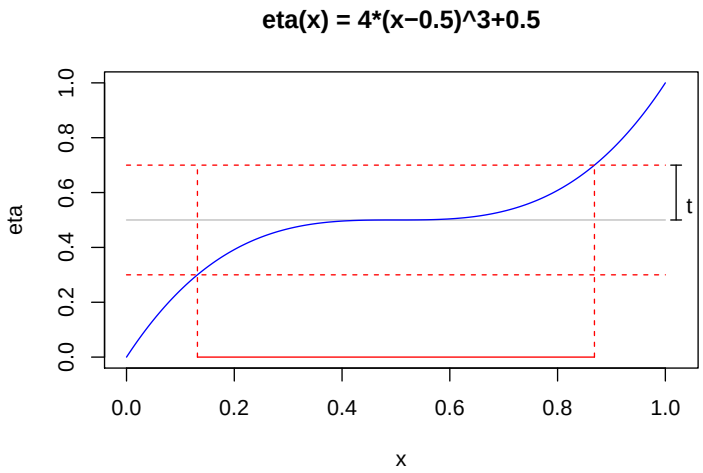
$$\exists \varepsilon > 0,\ \exists C > 0,\ \forall t \in (0,\varepsilon):\ P_X\left(\left|\eta(X) - \frac{1}{2}\right| \leq t\right) \leq C\, t^q.$$

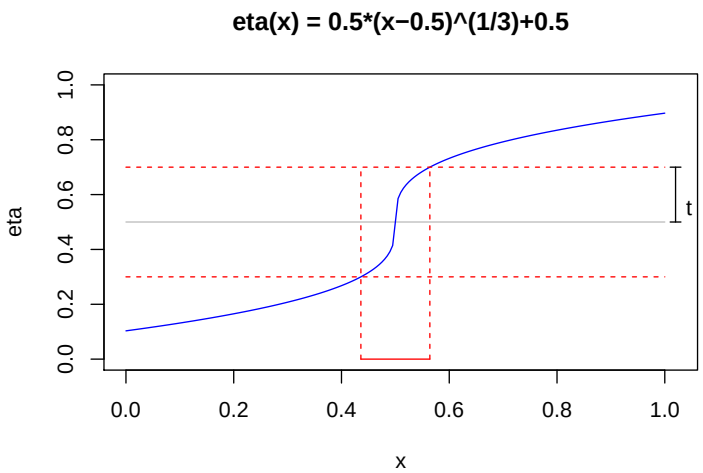
Für $\eta \overset{fs.}{=} P(Y = 1 \mid X = \cdot)$.

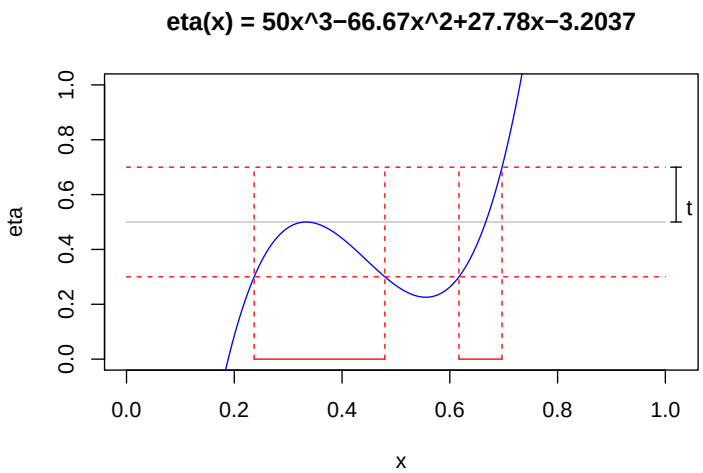
Bemerkung:
Die Definition kann man für $K \subset \mathbb{R}^d$ kompakt und P Verteilung auf $K \times \{-1,1\}$ analog betrachten.

TNE Beispiele

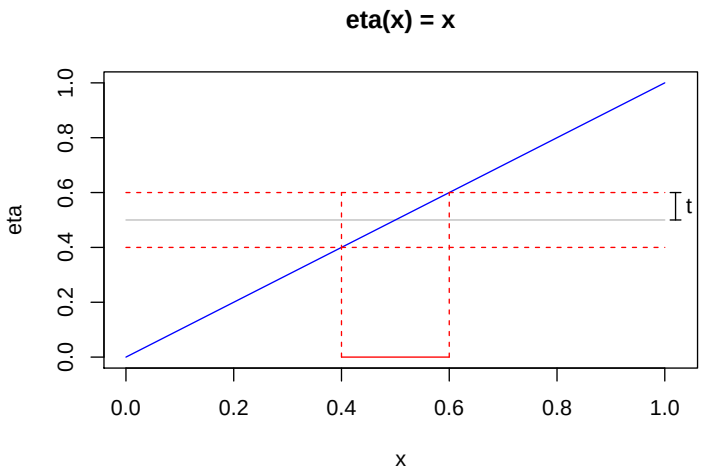


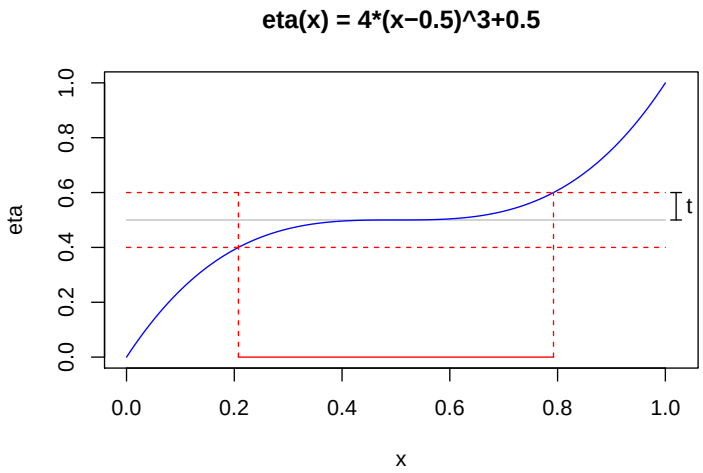


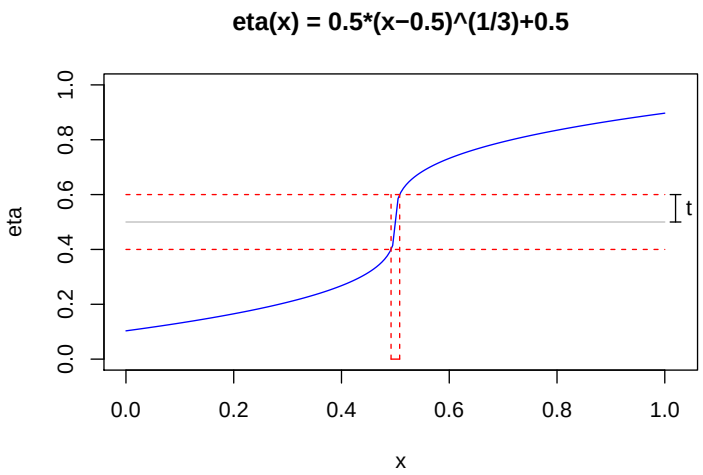


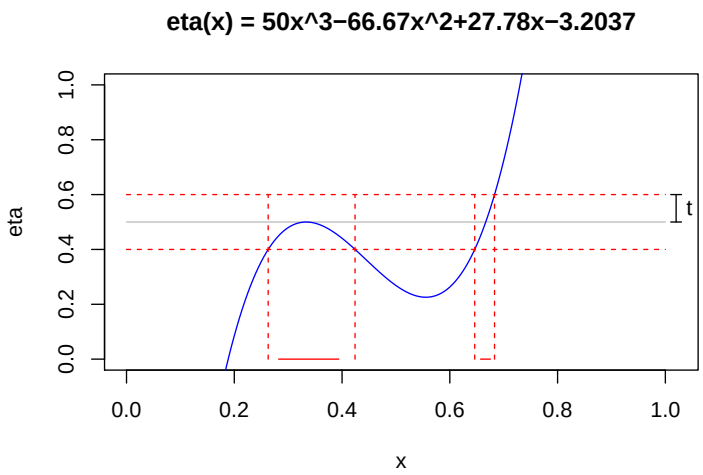


TNE Beispiele

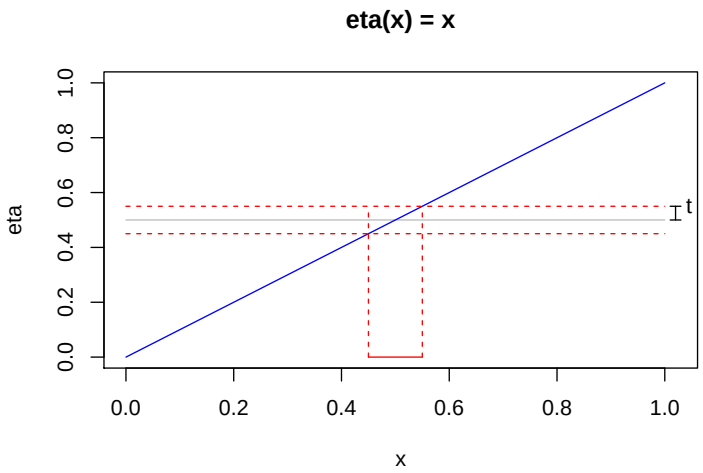


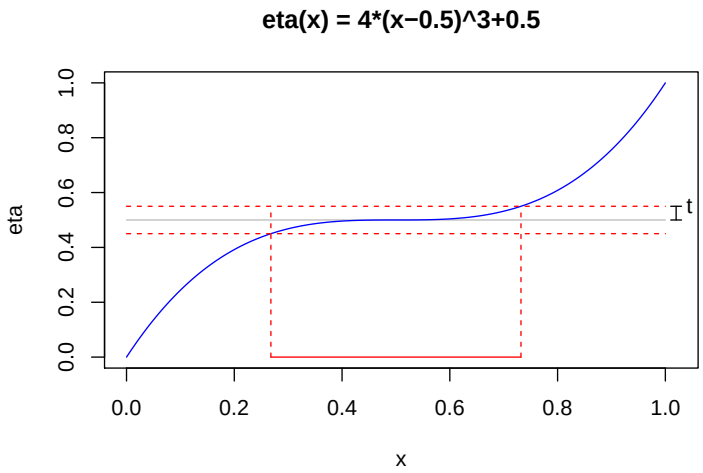


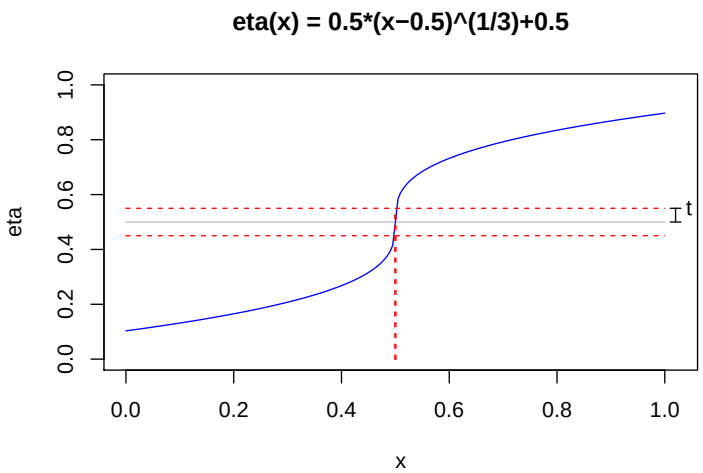


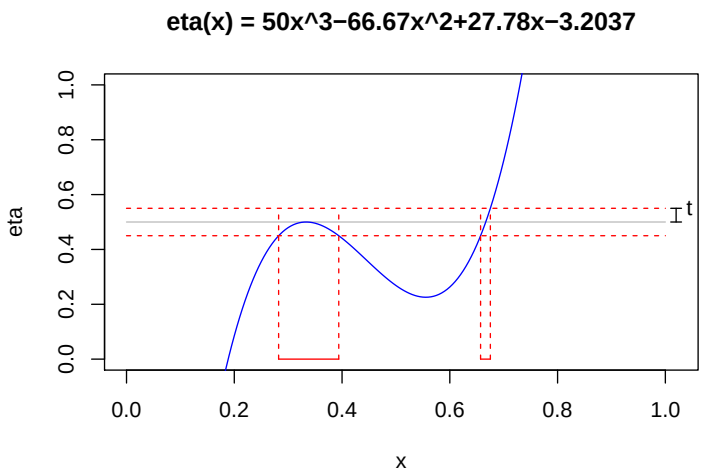


TNE Beispiele

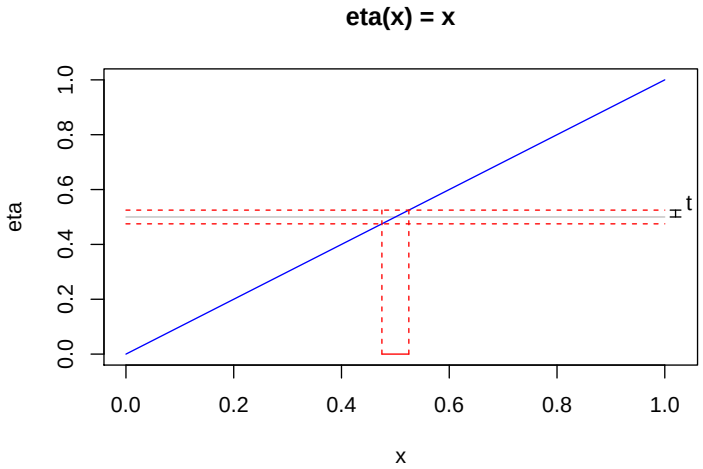


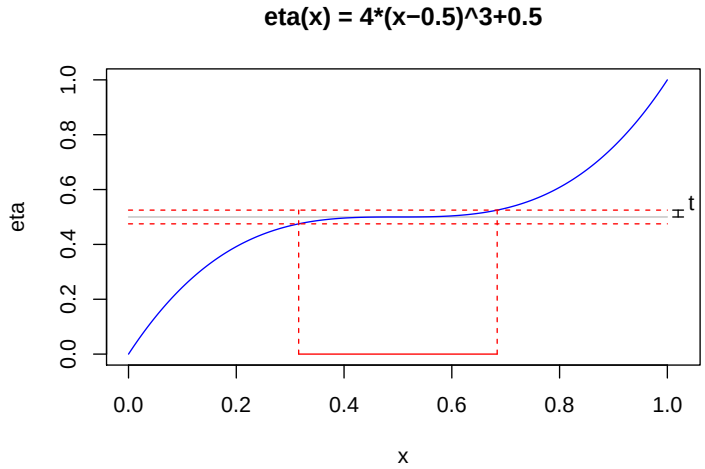


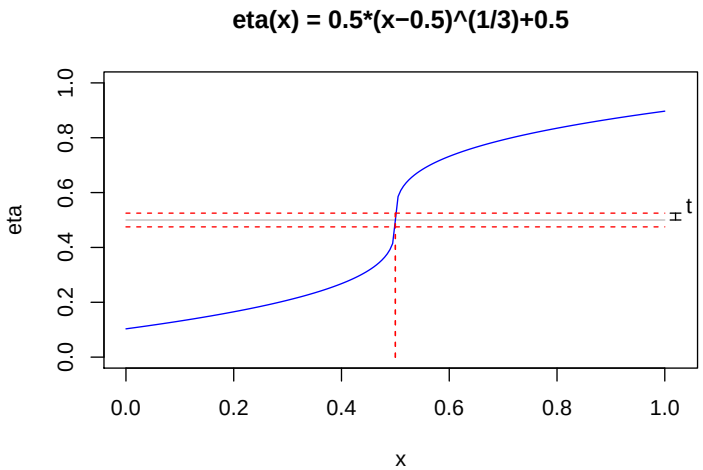


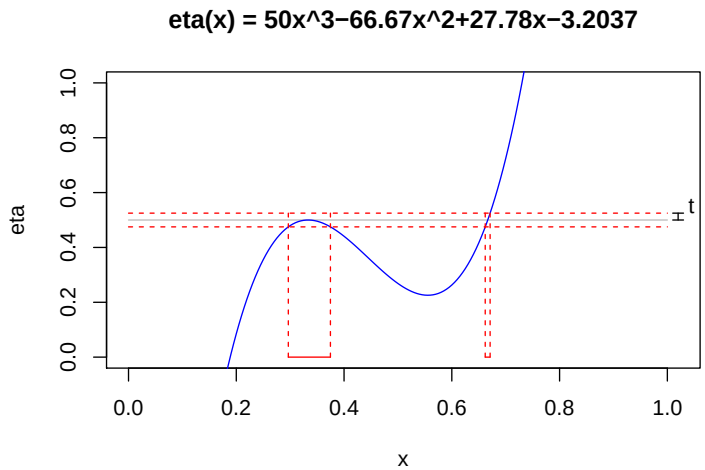


TNE Beispiele









Lemma 4.4 : TNE für η Poylnom

Falls:

- P_X hat beschränkte Lebesgue-Dichte
- $\exists p$ Polynom, sodass $P(Y = 1 \mid X = \cdot) - \frac{1}{2} \overset{fs.}{=} p$
- p hat Grad $m \geq 0$ und $p \neq 0$

Dann besitzt P (mindestens) jeden TNE $q \in [0, \frac{1}{m}]$.

(Beweis in der Ausarbeitung)

Geometric-Noise-Exponent

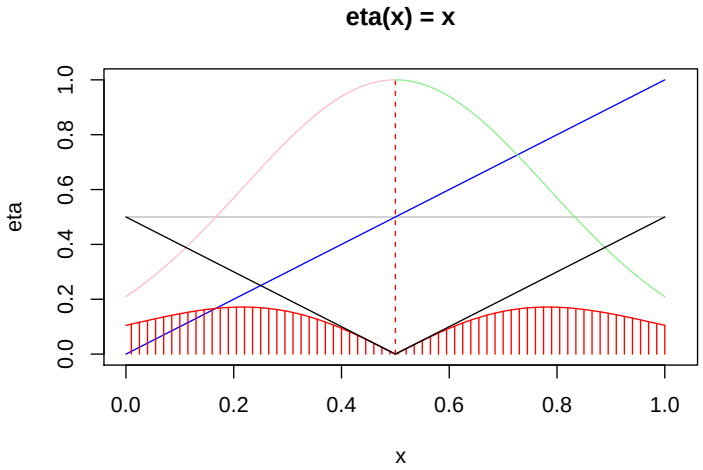
- P Verteilung auf $[0,1] \times \{-1,1\}$
- $\tau_P(x)$ der Abstand zur nächsten Stelle, an der $\eta - \frac{1}{2}$ das Vorzeichen wechselt

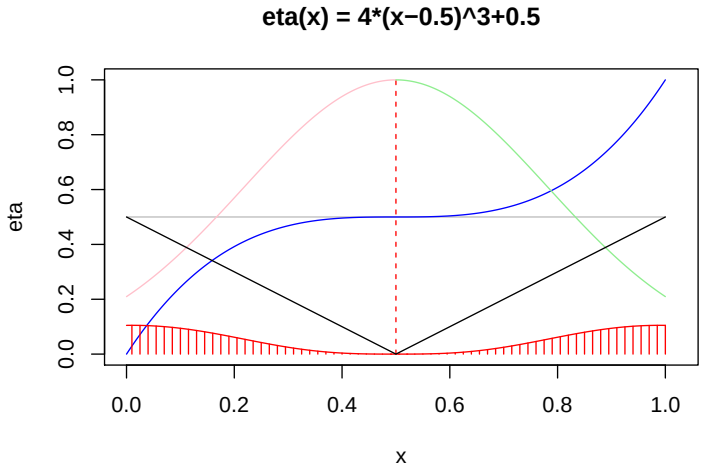
$$\tau_P(x) := 1_{\eta(x)>\frac{1}{2}}\, d\left(x, \left\{\eta \leq \frac{1}{2}\right\}\right) + 1_{\eta(x)<\frac{1}{2}}\, d\left(x, \left\{\eta \geq \frac{1}{2}\right\}\right).$$

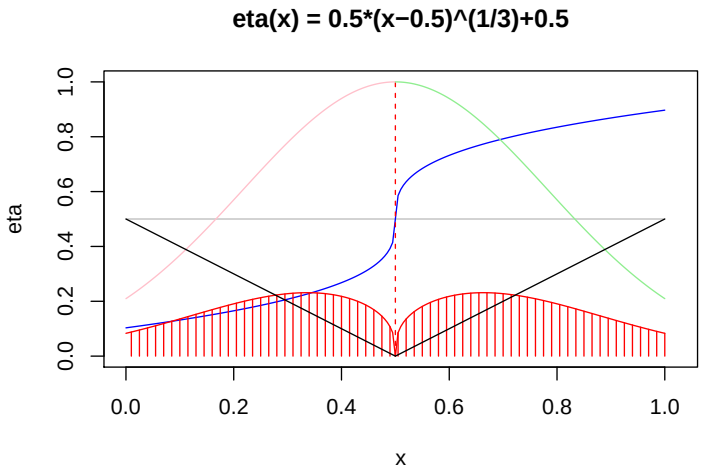
Die Verteilung P hat **Geometric-Noise-Exponent** $\alpha \in (0,\infty)$, falls

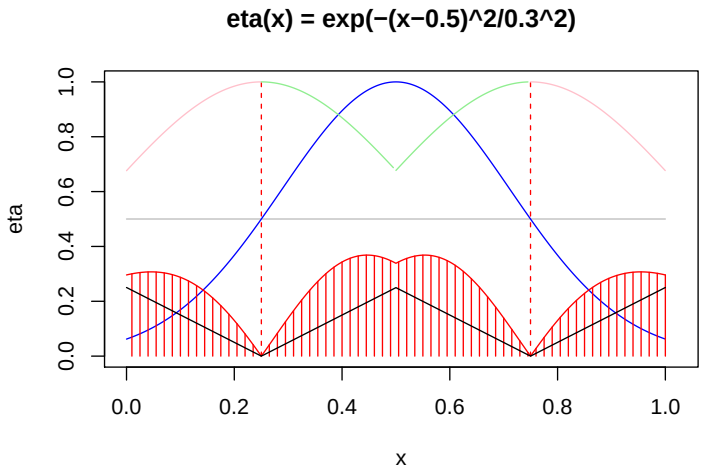
$$\exists C > 0, \forall t > 0 : \ \mathbb{E}_P\left[\left|\eta(X) - \frac{1}{2}\right| e^{-\frac{\tau_P(X)^2}{t^2}}\right] \leq C\, t^{\alpha}.$$

GNE Beispiele

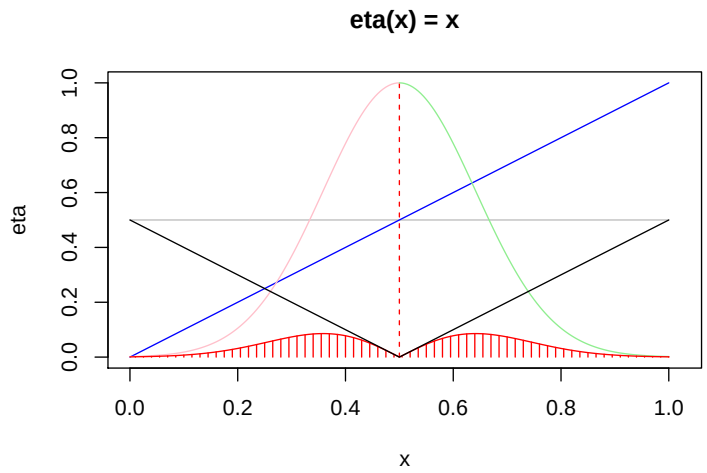


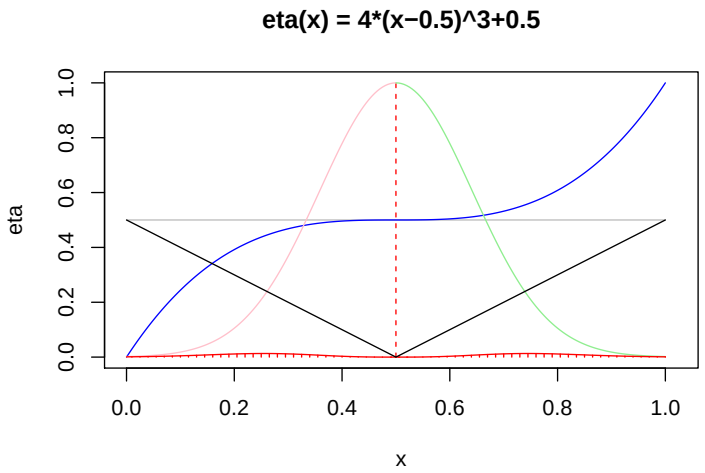


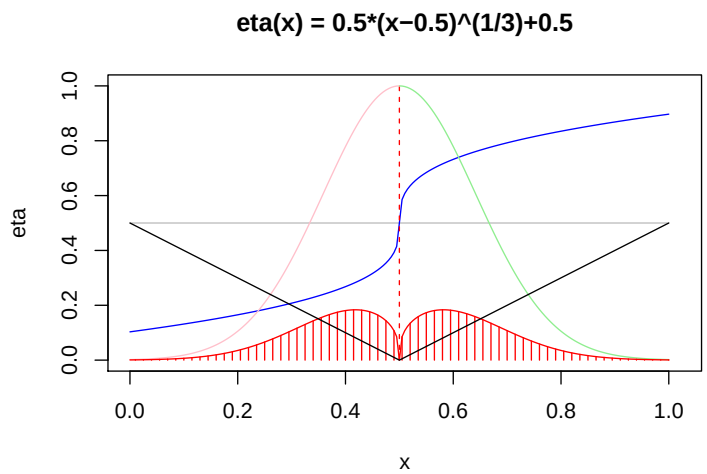


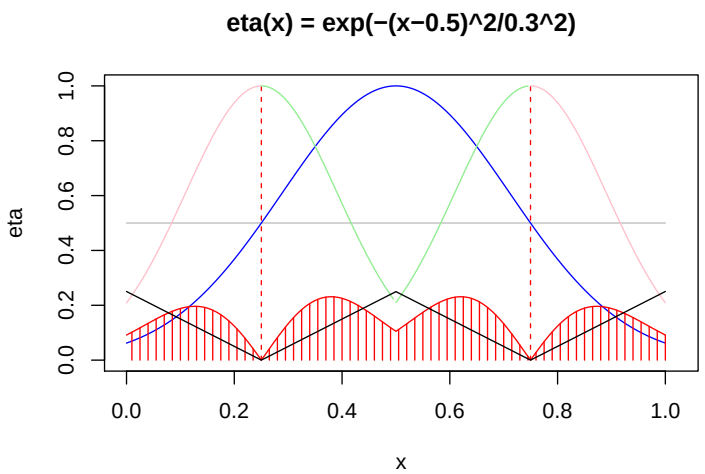


GNE Beispiele

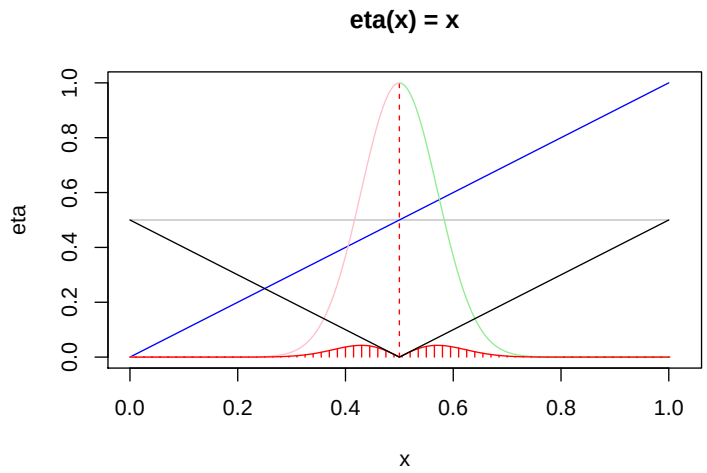


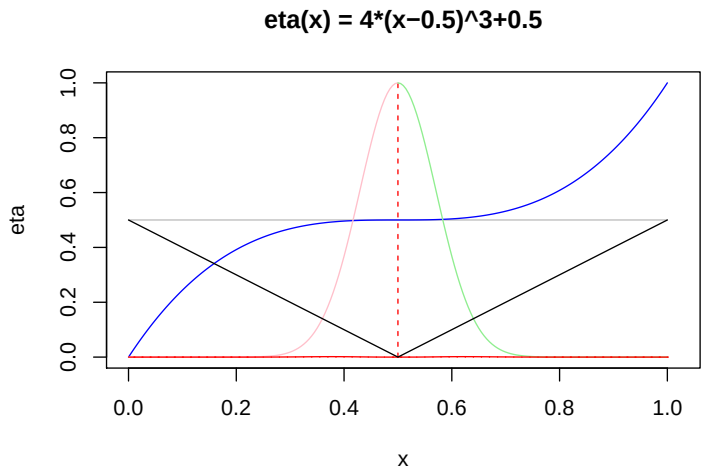


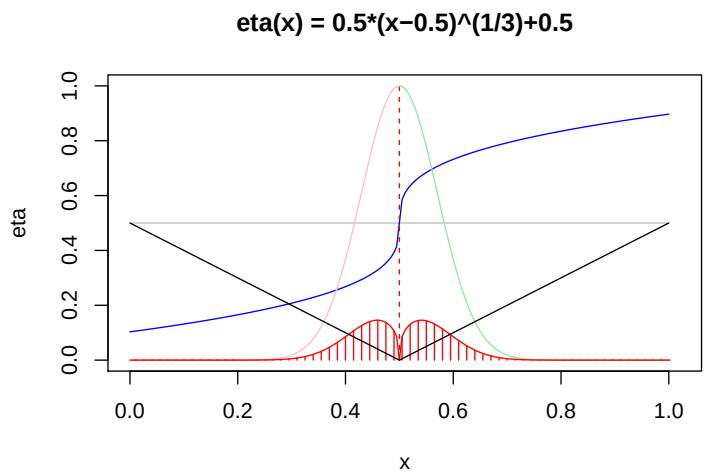


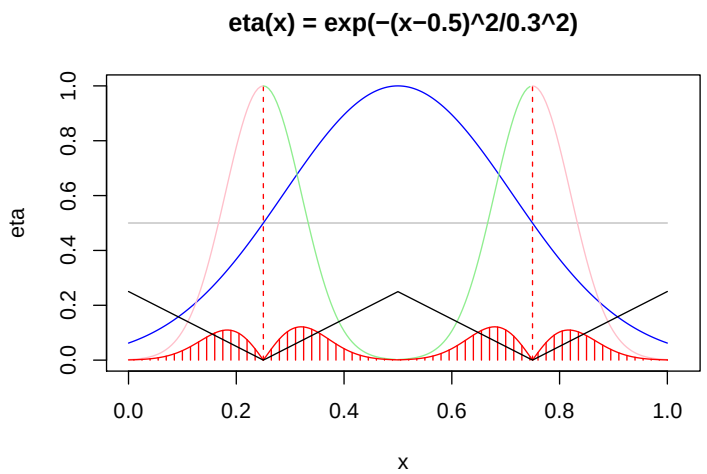


GNE Beispiele

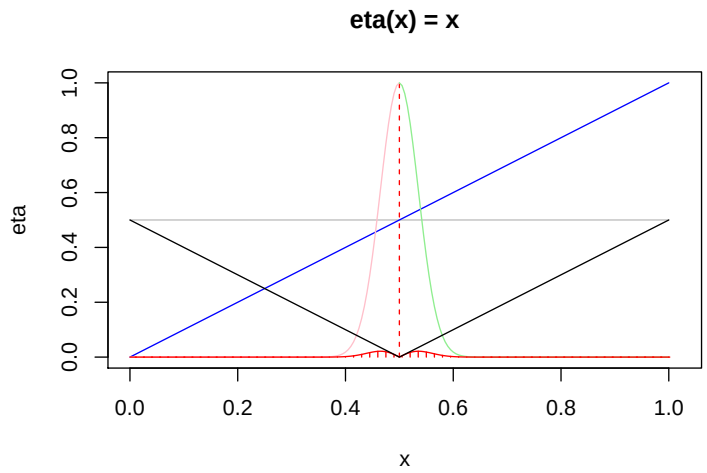


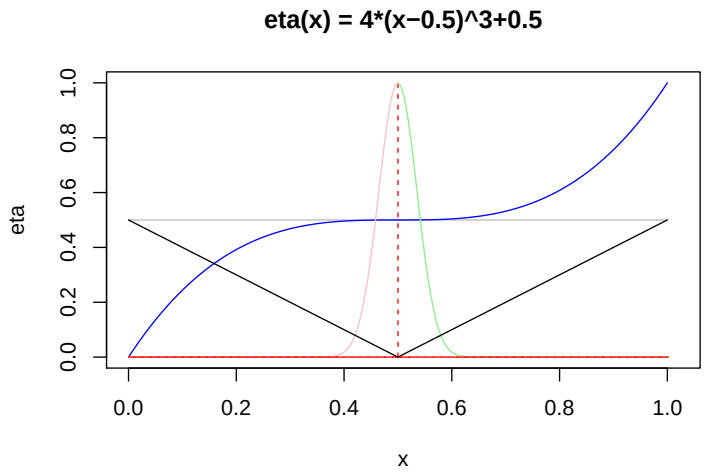


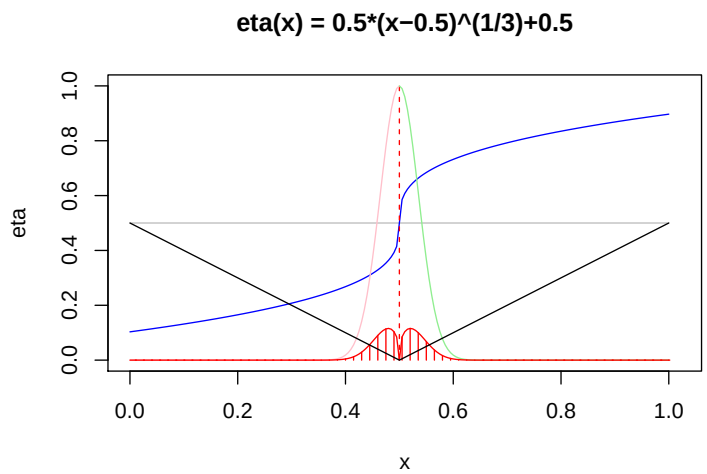


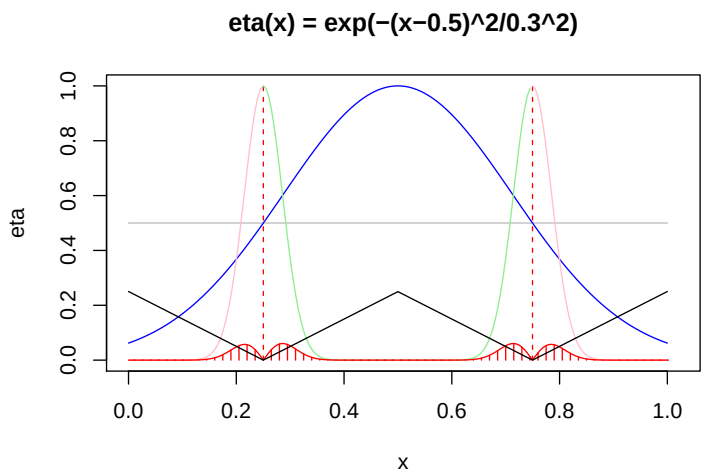


GNE Beispiele









Lemma 4.6 : GNE für η Hölderstetig

Falls gilt:

- P_X hat beschränkte Lebesgue-Dichte p
- $\exists f$ hölderstetige Funktion, sodass $P(Y = 1 \mid X = \cdot) - \frac{1}{2} \overset{fs.}{=} f$
- f hölderstetig bezgl. $\gamma > 0$, d.h.

$$\exists \widetilde{C} \forall x, y \in [0,1] : \ |f(x) - f(y)| \leq \widetilde{C} \, |x - y|^{\gamma}.$$

- f hat endlich viele Nullstellen

Dann besitzt P (mindestens) jeden GNE $\alpha \in [0, 1 + \gamma]$.

Lemma 4.6 : GNE für η Hölderstetig (Beweis)

Beweis-Idee:
Seien $x_1, ..., x_m$ die Nullstellen von f.

$$\mathbb{E}_P\left[\left|\eta(X) - \frac{1}{2}\right| e^{-\frac{\tau_P(X)^2}{t^2}}\right] = \int_0^1 p(x)\,|f(x)|\, e^{-\frac{\tau_P(x)^2}{t^2}}\, dx$$

$$\leq c \int_0^1 |f(x)|\, e^{-\frac{\tau_P(x)^2}{t^2}}\, dx$$

$$= c \sum_{i=1}^{m} \int_{x_i-\varepsilon}^{x_i+\varepsilon} |f(x)|\, e^{-\frac{\tau_P(x)^2}{t^2}}\, dx + c \int_{[0,1]\setminus\bigcup B_\varepsilon(x_i)} |f(x)|\, e^{-\frac{\tau_P(x)^2}{t^2}}\, dx$$

$$\leq c \sum_{i=1}^{m} \int_{x_i-\varepsilon}^{x_i+\varepsilon} |x - x_i|^\gamma\, e^{-\frac{(x-x_i)^2}{t^2}}\, dx + c \int_{[0,1]\setminus\bigcup B_\varepsilon(x_i)} |f(x)|\, e^{-\frac{\varepsilon^2}{t^2}}\, dx$$

Lemma 4.6 : GNE für η Hölderstetig (Beweis)

$$
\begin{aligned}
&= c \sum_{i=1}^{m} \int_{x_i-\varepsilon}^{x_i+\varepsilon} |x - x_i|^{\gamma} \, e^{-\frac{(x-x_i)^2}{t^2}} \, dx + c \int_{[0,1]\setminus\bigcup B_\varepsilon(x_i)} |f(x)| \, e^{-\frac{\varepsilon^2}{t^2}} \, dx \\
&\leq 2c \sum_{i=1}^{m} \int_{0}^{\varepsilon} x^{\gamma} \, e^{-\frac{x^2}{t^2}} \, dx + c \int_{[0,1]\setminus\bigcup B_\varepsilon(x_i)} e^{-\frac{\varepsilon^2}{t^2}} \, dx \\
&\leq 2mc \int_{0}^{\varepsilon} x^{\gamma} \, e^{-\frac{x^2}{t^2}} \, dx + ce^{-\frac{\varepsilon^2}{t^2}} \\
&= 2mc \int_{0}^{\frac{\varepsilon}{t}} (zt)^{\gamma} \, e^{-z^2} \, t \, dz + ce^{-\frac{\varepsilon^2}{t^2}} \\
&\leq t^{1+\gamma} \, 2mc \int_{0}^{\infty} z^{\gamma} \, e^{-z^2} \, dz + ce^{-\frac{\varepsilon^2}{t^2}} \qquad \square
\end{aligned}
$$

Satz 5.1 : Fehlerabschätzung für ERM mit Regularisierung

- P Verteilung auf $[0,1] \times \{-1,1\}$
- mit TNE $q \in [0,\infty]$
- mit GNE $\alpha \in (0,\infty)$

Definiere:

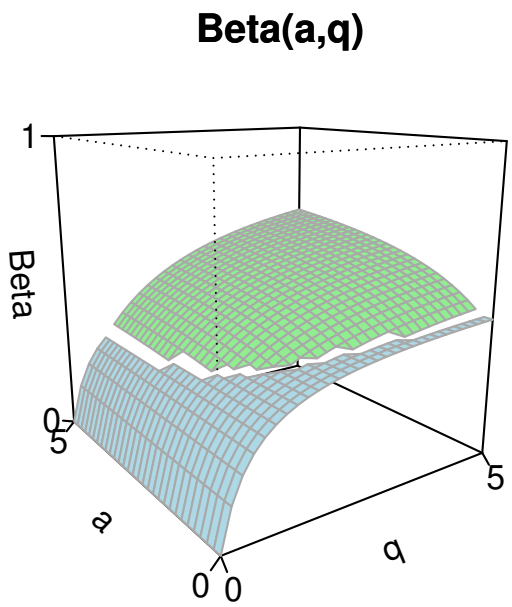
$$\beta = \max\left(\frac{\alpha}{2\alpha+1}, \frac{2\alpha(q+1)}{2\alpha(q+2)+3q+4}\right)$$

und $\lambda_n := n^{-\frac{\alpha+1}{\alpha\beta}}$ sowie $\sigma_n := n^{-\frac{\beta}{\alpha}}$, dann existiert für jedes $\varepsilon > 0$ ein $C > 0$, sodass für alle $n \in \mathbb{N}$ und $x \geq 1$ gilt:

$$\mathbb{P}\left(R_P\left(\operatorname{sign}(\Phi_{\lambda_n,\sigma_n}(T_n))\right) - \inf_{f:[0,1]\to\{-1,1\}mb.} R_P(f) \geq Cx^2 n^{-\beta+\varepsilon}\right) \leq e^{-x}$$

(Beweis-Idee in der Ausarbeitung)

$\beta(\alpha, q)$



$$\beta(\alpha, q) = \max\left(\frac{\alpha}{2\alpha+1}, \frac{2\alpha(q+1)}{2\alpha(q+2)+3q+4}\right)$$

- $\beta(\alpha, q)$ ist in α und in q monoton wachsend.
- $\beta(\alpha, q) > \frac{1}{2} \Leftrightarrow 2\alpha(q+1) > 2\alpha + 3q + 4$
  $\Leftrightarrow 2\alpha q > 3q + 4 \Leftrightarrow \alpha > \frac{3q+4}{2q}$

Satz 5.1 (Erwartetes Risiko)

Unter den Voraussetzungen von Satz 5.1 gilt:

$$
\begin{aligned}
&\mathbb{E}\left[R_P\left(\operatorname{sign}(\Phi_{\lambda_n,\sigma_n}(T_n))\right) - \inf_{f:[0,1]\to\{-1,1\}\,mb.} R_P(f)\right] \\
&= \int_0^\infty \mathbb{P}\left(R_P\left(\operatorname{sign}(\Phi_{\lambda_n,\sigma_n}(T_n))\right) - \inf_{f:[0,1]\to\{-1,1\}\,mb.} R_P(f) \geq t\right) dt \\
&= \int_0^\infty \mathbb{P}\left(R_P\left(\operatorname{sign}(\Phi_{\lambda_n,\sigma_n}(T_n))\right) - \inf_{f:[0,1]\to\{-1,1\}\,mb.} R_P(f) \geq Cx^2 n^{-\beta+\varepsilon}\right) 2Cxn^{-\beta+\varepsilon}\, dx \\
&\overset{5.1}{\leq} \int_0^\infty e^{-x}\, 2Cxn^{-\beta+\varepsilon}\, dx + \int_0^1 2Cxn^{-\beta+\varepsilon}\, dx = 2Cn^{-\beta+\varepsilon}\int_0^\infty e^{-x}\, x\, dx + Cn^{-\beta+\varepsilon} = 3Cn^{-\beta+\varepsilon}
\end{aligned}
$$

Verallgemeinerbarkeit

- $K \subset \mathbb{R}^d$ kompakt
- P Verteilung auf $K \times \{-1, 1\}$
- $\sigma_n := n^{-\frac{\beta}{\alpha d}}$ anstatt $\sigma_n := n^{-\frac{\beta}{\alpha}}$

Dann gilt Satz 5.1 analog:

$$\mathbb{P}\left(R_P\left(\text{sign}(\Phi_{\lambda_n,\sigma_n}(T_n))\right) - \inf_{f:K\to\{-1,1\}mb.} R_P(f) \geq Cx^2 n^{-\beta+\varepsilon} \right) \leq e^{-x}$$

Andere Betrachtungsweise von $\text{sign}(\Phi_{\lambda_n,\sigma_n}(T_n))$

Im Vortrag von Jens Nußberger haben wir gesehen, dass der Minimierer der ERMR. die Form

$$\Phi_{\lambda_n,\sigma_n}(T_n) = \sum_{i=1}^{n} \alpha_i \, Y_i K_{\sigma_n}(\cdot, X_i)$$

für bestimmte von T_n, σ_n, λ_n abhängige $\alpha_i \geq 0$ annimmt. Wir stellen um:

$$\Phi_{\lambda_n,\sigma_n}(T_n)(x) > 0 \;\Leftrightarrow\; \sum_{i\leq n, Y_i=1} \alpha_i K_{\sigma_n}(x, X_i) > \sum_{i\leq n, Y_i=-1} \alpha_i K_{\sigma_n}(x, X_i)$$

$$\Leftrightarrow \sum_{i\leq n, Y_i=1} \frac{\alpha_i}{\sum_{j=1}^{n} \alpha_j} \frac{K_{\sigma_n}(x, X_i)}{\sqrt{2\pi\sigma_n^2}} > \sum_{i\leq n, Y_i=-1} \frac{\alpha_i}{\sum_{j=1}^{n} \alpha_j} \frac{K_{\sigma_n}(x, X_i)}{\sqrt{2\pi\sigma_n^2}}$$

Andere Betrachtungsweise von $\text{sign}(\Phi_{\lambda_n,\sigma_n}(T_n))$

Naive Vorstellung: α_i alle gleich, dann wäre

$$\Phi^{Naiv}_{\lambda_n,\sigma_n}(T_n)(x) > 0 \Leftrightarrow \frac{1}{n}\sum_{i\leq n, Y_i=1}\frac{K_{\sigma_n}(x,X_i)}{\sqrt{2\pi\sigma_n^2}} > \frac{1}{n}\sum_{i\leq n, Y_i=-1}\frac{K_{\sigma_n}(x,X_i)}{\sqrt{2\pi\sigma_n^2}}$$

$$\Leftrightarrow \underbrace{\frac{n_1}{n}\frac{1}{n_1}\sum_{i\leq n, Y_i=1}\frac{K_{\sigma_n}(x,X_i)}{\sqrt{2\pi\sigma_n^2}}}_{=:\hat{p}_{1,\sigma_n}(x)} > \underbrace{\frac{n_{-1}}{n}\frac{1}{n_{-1}}\sum_{i\leq n, Y_i=-1}\frac{K_{\sigma_n}(x,X_i)}{\sqrt{2\pi\sigma_n^2}}}_{=:\hat{p}_{-1,\sigma_n}(x)}$$

für $n_y := \#\{i \leq n \mid Y_i = y\}$.

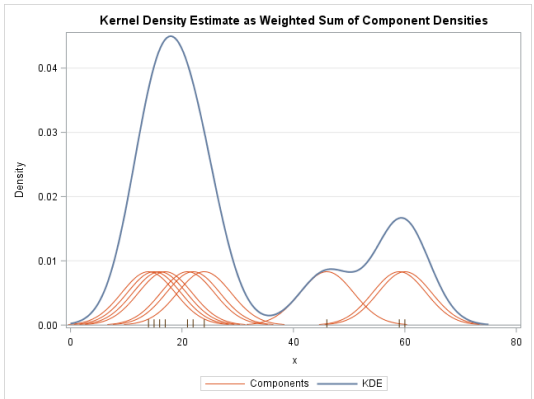
Kernel Density Estimation

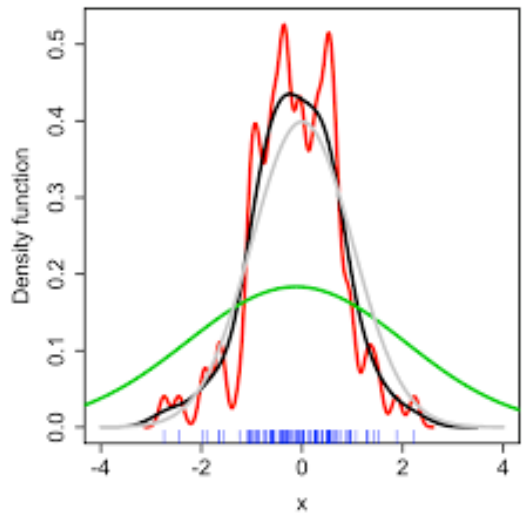
- $\widetilde{P}$ Verteilung auf $\mathbb{R}$ mit Lebesgue Dichte p
- $\widetilde{X}_{1,\widetilde{P}}, \widetilde{X}_{2,\widetilde{P}}, ...uiv. \sim \widetilde{P}$

Versuche die Dichte p abhängig von $\widetilde{X}_{1,\widetilde{P}}, ..., \widetilde{X}_{n,\widetilde{P}}$ zu schätzen:

Kerndichteschätzer mit Kern $\widetilde{K}(x) := \frac{1}{\sqrt{2\pi}} e^{-x^2}$

$$p_{n,\sigma_n}(x, \widetilde{X}_{1,\widetilde{P}}, ..., \widetilde{X}_{n,\widetilde{P}}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sigma_n} \widetilde{K} \left(\frac{x - \widetilde{X}_{i,\widetilde{P}}}{\sigma_n} \right)$$





KDE. und $\text{sign}(\Phi_{\lambda_n,\sigma_n}(T_n))$

- P Verteilung auf $\mathbb{R} \times \{-1, 1\}$
- P_1 und P_{-1} haben Dichten p_1, p_{-1}.
- $T_n = ((X_1, Y_1), ..., (X_n, Y_n)) \sim P^{\otimes n}$
- $T_{n_1}^1 := \{X_i \mid Y_i = 1\}$, $T_{n_{-1}}^{-1}$ analog

$$\hat{p}_{1,\sigma_n}(x) = \frac{1}{n_1} \sum_{i \leq n, Y_i = 1} \frac{K_{\sigma_n}(x, X_i)}{\sqrt{2\pi\sigma_n^2}} = \frac{1}{n_1} \sum_{i \leq n, Y_i = 1} \frac{1}{\sigma_n} \tilde{K}\left(\frac{x - X_i}{\sigma_n}\right)$$

$$= p_{n,\sigma_n}(x, T_{n_1}^1) \leftarrow \text{ Schätzer für } p_1(x)$$

analog $\hat{p}_{-1,\sigma_n}(x) = p_{n,\sigma_n}(x, T_{n_{-1}}^{-1})$

KDE. und $\text{sign}(\Phi_{\lambda_n,\sigma_n}(T_n))$

Es folgt

$$\Phi^{Naiv}_{\lambda_n,\sigma_n}(T_n)(x) > 0 \Leftrightarrow \frac{n_1}{n}\hat{p}_{1,\sigma_n}(x) > \frac{n_{-1}}{n}\hat{p}_{-1,\sigma_n}(x)$$
$$\Leftrightarrow \frac{n_1}{n}p_{n,\sigma_n}(x, T^1_{n_1}) > \frac{n_{-1}}{n}p_{n,\sigma_n}(x, T^{-1}_{n_{-1}})$$

Entscheidungsfunktion $\text{sign}(\Phi_{\lambda_n,\sigma_n}(T_n))$ ist also ein Vergleich von 'gewichteten' KDEs. (Gewichte α_i hängen vom Minimierungsproblem ab.)

Im Vergleich gilt für den Bayes-Klassifikator (fast sicher):

$$s^*(x) = 1 \Leftrightarrow \eta(x) > \frac{1}{2} \Leftrightarrow P(Y = 1 \mid X = x) > P(Y = -1 \mid X = x)$$
$$\Leftrightarrow \frac{p_1(x)P(Y = 1)}{c(x)} > \frac{p_{-1}(x)P(Y = -1)}{c(x)}$$
$$\Leftrightarrow P(Y = 1)p_1(x) > P(Y = -1)p_{-1}(x)$$

Quellen und Verweise

[1] Ingo Steinwart and Clint Scovel, *Fast rates for support vector machines using Gaussian kernels*, Ann. Statist. **35** (2007), no. 2, 575–607, DOI 10.1214/009053606000001226.

[2] Ingo Steinwart, *Consistency of Support Vector Machines and Other Regularized Kernel Classifiers*, IEEE TRANSACTIONS ON INFORMATION THEORY, **51** (2005).

[3] _____, *On the influence of the kernel on the consistency of support vector machines*, J. Mach. Learn. Res. **2** (2002), no. 1, DOI 10.1162/1532443027601852 52.

[4] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe, *Convexity, classification, and risk bounds*, J. Amer. Statist. Assoc. **101** (2006), no. 473, 138–156, DOI 10.1198/016214505000000907.

Vielen Dank für Ihre Aufmerksamkeit!

Albert-Ludwigs-Universität Freiburg