Inhaltsverzeichnis

1 Wiederholung der Grundlagen

- **Grundlegende Fragestellung:**
  Es sei P ein W-Maß auf $[0,1]\times\{-1,1\}$ (mit induzierter Borel-Algebra) und $T_n = ((X_1,Y_1),...,(X_n,Y_n))$ ein Trainingsdatensatz. Suche abhängig von T_n eine Funktion aus einer gegebenen Funktionenfamilie $(f_\alpha : [-1,1] \to \mathbb{R})_{\alpha\in\Lambda}$ aus, die möglichst geringes Risiko $R_P(f)$ besitzt.

- **Risiko:**
  Das P-Risiko einer Funktion $f : [0,1] \to \mathbb{R}$ bezüglich einer Verlust-Funktion $L : \{-1,1\} \times \mathbb{R} \to [0,\infty]$ ist definiert durch

$$R_{L,P}(f) := \mathbb{E}_P\left[L(Y, f(X))\right].$$

  Normalerweise wird die Funktionenfamilie so gewählt, dass

$$f_\alpha : [0,1] \to \{-1,1\} \text{ messbar}, \forall\alpha \in \Lambda.$$

  Dann kann $L(y,a) = 1_{y\neq a} = |y-a|$ gewählt werden. In diesem Fall schreiben wir für das Risiko auch nur R_P.
  Um das Risiko zu minimieren will man die folgenden Terme/Risiken möglichst gering halten.

$$\begin{aligned} R_P(\Phi(T_n)) = {} & \underbrace{R_P(\Phi(T_n)) - \inf_{\alpha\in\Lambda} R_P(f_\alpha)}_{\text{Algorithmus-Risiko (estimation error)}} \\ & + \underbrace{\inf_{\alpha\in\Lambda} R_P(f_\alpha) - \inf_{f:[0,1]\to\{-1,1\}\text{ mb.}} R_P(f)}_{\text{Familien-Risiko (approximation error)}} \\ & + \underbrace{\inf_{f:[0,1]\to\{-1,1\}\text{ mb.}} R_P(f)}_{\text{Grund-Risiko (statistical risk)}} \end{aligned}$$

- **Konsistenz:**
  Den Algorithmus können wir als Funktion

$$\Phi : \bigcup_{n=1}^{\infty} ([0,1]\times\{-1,1\})^n \to \{f_\alpha \mid \alpha\in\Lambda\}$$

  darstellen. Ein solcher Algorithmus Φ heißt **konsistent** für das W-Maß P, falls gilt, dass für $T_n \sim \bigotimes_{i=1}^{n} P$, $n\in\mathbb{N}$ die Konvergenz

$$R_P(\Phi(T_n)) \xrightarrow{n\to\infty}_p \inf_{\alpha\in\Lambda} R_P(f_\alpha)$$

  in Wahrscheinlichkeit erfüllt ist.

- **Empirical Risk Minimization:**
  Da die ursprüngliche Verteilung P nicht bekannt ist, kann man stattdessen versuchen das Risiko für die empirische Verteilung $\hat{P}_n$ zu minimieren. Dabei kann es aber leicht passieren, dass ein Algorithmus sich nur um die genauen Positionen der Trainingsdaten $X_1, ..., X_n$ herum anpasst und etwas weiter entfernte Punkte beliebig zuordnet. Es kommt somit zu einer zu starken Anpassung der Funktion $\Phi(T_n)$ an die Positionen der Daten (Overfitting = Über-Anpassung).
  Dies lässt sich durch eine passende ('kleinere') Wahl der Familie $(f_\alpha)_{\alpha\in\Lambda}$ verhindern, womit jedoch das Familien-Risiko wächst.

- **Abschätzungen**
  Für kleine Trainingsdatensätze ist eine reine Konvergenz-Aussage für das Risiko nicht sehr hilfreich, es wäre besser das Risiko abhängig von n abschätzen zu können. Solche Abschätzungen können z.B. die Form

$$\mathbb{E}_{P^{\otimes n}}\left[R_P(\Phi(T_n)) - \inf_{f:[0,1]\to[-1,1]\, mb.} R_P(f)\right] \leq C(n)$$

  oder

$$\mathbb{P}\left(R_P(\Phi(T_n)) - \inf_{f:[0,1]\to[-1,1]\, mb.} R_P(f) > \varepsilon(n)\right) \leq C(n)$$

  annehmen und sollten am besten gleich für ganze Klassen von Verteilungen P gelten, da P in der Praxis nicht bekannt ist.

2 Hilberträume mit reproduzierendem Kern

Sei $H \subset \{f : [0,1] \to \mathbb{R} \mid f \text{ messbar}\}$ und $\langle \cdot, \cdot \rangle_H : H \times H \to \mathbb{R}$, sodass $(H, \langle \cdot, \cdot \rangle_H)$ ein Hilbertraum (d.h. H ist Vektorraum, $\langle \cdot, \cdot \rangle_H$ ist Skalarprodukt und $(H, ||\cdot||_H)$ ist vollständig), dann heißt $(H, \langle \cdot, \cdot \rangle_H$ ein **Hilbertraum mit reproduzierendem Kern** auf $[0,1]$, falls für jedes $x \in [0,1]$ die Auswertungsfunktion

$$\varphi_x : (H, ||\cdot||_H) \to \mathbb{R} \; ; \; f \mapsto f(x)$$

stetig ist. Man nennt φ_x dann auch ein Funktional auf H. (Achtung: φ_x ist zwar linear, aber nicht unbedingt stetig, da H auch unendlichdimensional sein kann.)

Nun besagt der Satz von Riesz aus der Funktionalanalysis, dass zu jedem Funktional φ auf H auch ein Element $h \in H$ existiert, sodass gilt

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

Das zu φ_x korrespondierende Element nennen wir K_x, also

$$\forall f \in H : \; \langle f, K_x \rangle_H = \varphi_x(f) = f(x).$$

Nun ist der **reproduziernde Kern** K von H definiert durch

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

Tatsächlich enthält K sämtliche Informationen über H und man kann zu jedem symmetrischen positiv definiten Kern K einen eindeutigen Hilbertraum H konstruieren, dessen reproduzierender Kern K ist. Es ist klar, dass mit K bereits alle K_x für $x \in [0,1]$ bekannt sind und man stellt fest, dass H die Vervollständigung von $\text{Span}_{\mathbb{R}}(\{K_x \mid x \in [0,1]\})$ mit

$$\langle \sum_{i=1}^{n} a_i K_{x_i}, \sum_{j=1}^{m} b_j K_{y_j} \rangle_{H_0} = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j K(x_i, y_j)$$

ist. (Genauer: Satz von Moore–Aronszajn)

Wir werden in Zukunft den Hilbertraum H_σ zu dem reproduzierenden Kern $K_\sigma(x,y) = e^{-\frac{(x-y)^2}{\sigma^2}}$ betrachten. Das ist möglich, da nach dem Satz von Bochner K_σ positiv definit ist.

Lemma 2.1.
Sei $\sigma > 0$ fest, dann ist H_σ eine dichte Teilmenge von $(C([0,1]), ||\cdot||_{[0,1]})$.

Beweis. [eigenständig bewiesen]
Stetigkeit:
Nach Konstruktion existiert für jedes $f \in H_\sigma$ eine Folge $(f_m)_{m\in\mathbb{N}} \subset \mathrm{Span}_{\mathbb{R}}(\{K_x \mid x \in [0,1]\})$, sodass $||f - f_m||_{H_\sigma} \xrightarrow{m\to\infty} 0$, womit auch für jedes $x \in [0,1]$ gilt, dass

$$|f(x) - f_m(x)| = |\varphi_x(f - f_m)| \xrightarrow{m\to\infty} 0, \text{ wegen der Stetigkeit von } \varphi_x.$$

Nach der Konvergenz-Eigenschaft muss auch ein $C > 0$ existieren, sodass $\forall m \in \mathbb{N} : ||f_m||_{H_\sigma} \leq C$. Damit berechnen wir nun, dass $(f_m)_{m\in\mathbb{N}}$ gleichmäßig beschränkt und gleichgradig stetig ist, womit der Satz von Arzelà-Ascoli bereits die Stetigkeit von f liefert.
-gleichmäßig beschränkt:

$$|f_m(x)| = |\varphi_x(f_m)| = |\langle f_m, K_x\rangle_{H_\sigma}| \overset{CS.}{\leq} ||f_m||_{H_\sigma}\, ||K_x||_{H_\sigma} \leq C \cdot 1$$

-gleichgradig stetig:

$$\begin{aligned}
&|f_m(x_1) - f_m(x_2)|^2 = |\langle f_m, K_{x_1} - K_{x_2}\rangle_{H_\sigma}|^2 \overset{CS.}{\leq} ||f_m||^2_{H_\sigma}\, ||K_{x_1} - K_{x_2}||^2_{H_\sigma} \\
&\leq C^2 ||K_{x_1} - K_{x_2}||^2_{H_\sigma} = C^2 \left(||K_{x_1}||^2_{H_\sigma} - 2\langle K_{x_1}, K_{x_2}\rangle_{H_\sigma} + ||K_{x_2}||^2_{H_\sigma}\right) \\
&= 2C^2 \left(1 - e^{-\frac{(x_1-x_2)^2}{\sigma^2}}\right) =: l(|x_1 - x_2|)^2
\end{aligned}$$

Somit existiert für jedes $\varepsilon > 0$ ein $\delta := l^{-1}(\varepsilon) > 0$ sodass

$$\forall m \in N\, \forall x_1, x_2 \in [0,1]; |x_1 - x_2| < \delta : \; |f_m(x_1) - f_m(x_2)| \leq l(|x_1 - x_2|) < l(\delta) = \varepsilon.$$

Dichtheit:
Sei $\overline{H}$ der Abschluss von H in $(C([0,1], ||\cdot||_{[0,1]}))$. Angenommen es existiere eine Funktion $g \in C([0,1] =$, die nicht in $\overline{H}$ liegt, dann definieren wir die lineare Abbildung

$$\widetilde{\varphi} : \overline{H} \oplus \mathbb{R} \cdot g \to \mathbb{R} \; ; \; h + c \cdot g \mapsto c$$

und stellen, dass $\widetilde{\varphi}$ eine stetige Linearform auf ihrem Definitionsbereich ist. Nun kann man den Satz von Hahn-Banach verwenden um eine Erweiterung

$$\varphi : C([0,1]) \to \mathbb{R} \text{ mit } \varphi|_{\overline{H}\oplus(\mathbb{R}\cdot g)} = \widetilde{\varphi}$$

zu erhalten. Nun besagt des Satz von Riesz-Markov, dass für jedes stetige Funktional auf $C([0,1])$ ein (endliches) signiertes Radonmaß μ auf $\mathcal{B}([0,1])$ existiert, sodass

$$\forall f \in C([0,1]); \; \varphi(f) = \int_0^1 f \, d\mu.$$

Jetzt wollen wir zeigen, dass hier bereits $\mu = 0$ gelten muss, womit wir einen Widerspruch zu $\varphi(g) > 0$ gefunden hätten.

Dazu definieren wir zuerst die Polynome $p_0, p_1, ...$ durch $\frac{\partial^k}{\partial x^k} e^{-\frac{x^2}{\sigma^2}} = p_k(x)\, e^{-\frac{x^2}{\sigma^2}}$. Man sieht leicht, dass dann das Polynom p_k den Grad k haben wird, womit die $p_0, p_1, ...$ eine Basis des Polynom-Raums bilden.
Nach Annahme ist $\varphi|_{\overline{H}} = 0$, womit auch für jedes $x \in [0,1]$ gilt $0 = \varphi(K_x)$ und somit auch

$$
\begin{aligned}
&\forall x \in (0,1),\, \forall k \in \mathbb{N}: \\
&0 = \frac{\partial^k}{\partial x^k}\varphi(K_x) = \frac{\partial^k}{\partial x^k}\left(\int_0^1 K_x\, d\mu\right) \overset{\text{dom.K.}}{=} \int_0^1 \frac{\partial^k}{\partial x^k} K_x\, d\mu \\
&= \int_0^1 \frac{\partial^k}{\partial x^k} e^{-\frac{(x-t)^2}{\sigma^2}}\, d\mu(t) = \int_0^1 p_k(x-t) e^{-\frac{(x-t)^2}{\sigma^2}}\, d\mu(t) \\
&\Rightarrow\ \forall p \text{ Polynom}:\ 0 = \int_0^1 p(x-t) e^{-\frac{(x-t)^2}{\sigma^2}}\, d\mu(t).
\end{aligned}
$$

Nach Stone-Weierstraß sind die Polynome dicht in $C([0,1])$ und wir finden für jedes $f \in C([0,1])$ und festes $x \in (0,1)$ eine folge von Polynomen $(q_m)_{m\in\mathbb{N}}$, sodass $||q_m(x-\cdot) - f(\cdot)e^{\frac{(x-\cdot)^2}{\sigma^2}}||_{[0,1]} \xrightarrow{m\to\infty} 0$ somit wird auch das Integral approximiert und wir erhalten

$$0 = \int_0^1 f(t)\, d\mu(t)\ ,\ \forall f \in C([0,1]).$$

Nun kann man unter anderem die Charakteristische Funktion (Fourier-Transformierte) vom μ für beliebiges $t > 0$ approximieren und erhält so, dass $\mu = 0$, da seine Charakteristische Funktion immer Null ist. □

3 ERM. mit Regularisierung

Die normale ERM. wird durch

$$\Phi(T_n) := \arg\min_{\alpha\in\Lambda} \left(R_{L,\hat{P}_n}(f_\alpha) \right)$$

für eine feste Funktionenfamilie

$$(f_\alpha)_{\alpha\in\Lambda} \subset \{f : [0,1] \to \mathbb{R} \mid f \text{ mb. }, \ \ R_{P.L}(f) \text{ definiert}\}$$

bestimmt.

Für eine zu große Familie $(f_\alpha)_{\alpha\in\Lambda}$ kann es schnell zu Overfitting kommen, für eine zu kleine ist das Familien-Risiko

$$\inf_{\alpha\in\Lambda} R_{L,P}(f_\alpha) - \inf_{f:[0,1]\to\{-1,1\} \text{ mb.}} R_{L,P}(f)$$

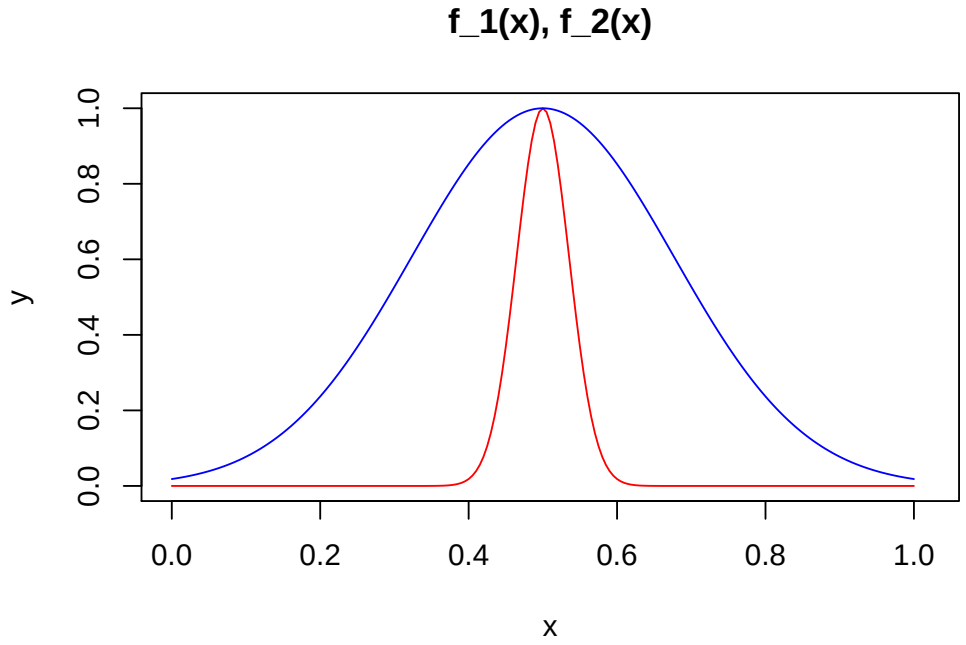
so groß, dass mit mehr Daten eine Verkleinerung des Algorithmus-Risiko nur noch wenig an dem Gesamtrisiko $R_{L,P}(\Phi(T_n))$ ändert.

Eine Möglichkeit dieses Problem zu umgehen wäre die Funktionenfamilie mit zunehmender Datenmenge zu vergrößern. Eine andere Möglichkeit ist es 'komplizierte' Funktionen zu vermeiden und 'einfachere' Funktionen aus der Familie zu bevorzugen. Dies wird durch einen Algorithmus der Form

$$\Phi(T_n) := \arg\min_{\alpha\in\Lambda} \left(\lambda_n \cdot W(f)^2 + R_{L,\hat{P}_n}(f_\alpha) \right)$$

dargestellt, wobei $W : (f_\alpha)_{\alpha\in\Lambda} \to [0,\infty)$ die 'Komplexität' einer Funktion quantifizieren soll und $(\lambda_n)_{n\in\mathbb{N}}$ eine Nullfolge ist.

Um so einen Algorithmus verwenden zu können, sollten wir zunächst klarstellen was mit 'Komplexität' gemeint ist. Ein generelles Problem der ERM. ist das Overfittig, wobei aus der Funktionenfamilie $(f_\alpha)_{\alpha\in\Lambda}$ eine Funktion gewählt wird, die nur in direkter Umgebung der Datenpunkte $X_1, ..., X_n$ an die Klassifizierungen $Y_1, ..., Y_n$ angepasst ist. Wir wollen Funktionen, für die dies für mehr Verteilungen P auftreten kann als 'Komplexer' betrachten. Grob gesagt ist eine Funktion $f : [0,1] \to \mathbb{R}$ mit stärkeren Höhenunterschieden auf kleinem Raum komplexer.
Man könnte versuchen die Funktionenfamilie $C^1([0,1])$ zu wählen und die Komplexität einer Funktion $f \in C^1([0,1])$ durch $||f'||_{L^p(\lambda)}$ zu bestimmen, allerdings könnte diese Norm z.B. für $p = 1$ nicht gut zwischen den hier gezeigten f_1 und f_2 unterscheiden:



Wir verwenden stattdessen den RKHS. H_σ als Funktionenfamilie $(f_\alpha)_{alpha\in\Lambda}$, da die Hilbertraum-Norm $||\cdot||_{H_\sigma}$ gute Eigenschaften besitzt um die Komplexität einer Funktion aus H_σ zu bestimmen.
Bei der so entstandenen Regularisierung

$$\Phi_\sigma(T_n) = \arg\min_{f\in H_\sigma} \left(\lambda_n \cdot ||f||^2_{H_\sigma} + R_{L,\hat{P}_n}(f)\right)$$

wirkt der Faktor σ sich darauf aus in welcher Entfernung eines Datenpunktes (X_i, Y_i) dieser noch signifikanten Einfluss auf die Entscheidungsfunktion $f_{T_n} = \Phi_\sigma(T_n)$ hat. Die Regularisierung lässt sich also noch verbessern, wenn man σ von der Anzahl der Daten (und von der Dimension des zugrunde liegenden Raumes der X_i) abhängig immer kleiner wählt. Die Regularisierung hat dann die Form

$$\Phi_{\lambda_n,\sigma_n}(T_n) = \arg\min_{f\in H_{\sigma_n}} \left(\lambda_n \cdot ||f||^2_{H_{\sigma_n}} + R_{L,\hat{P}_n}(f)\right)$$

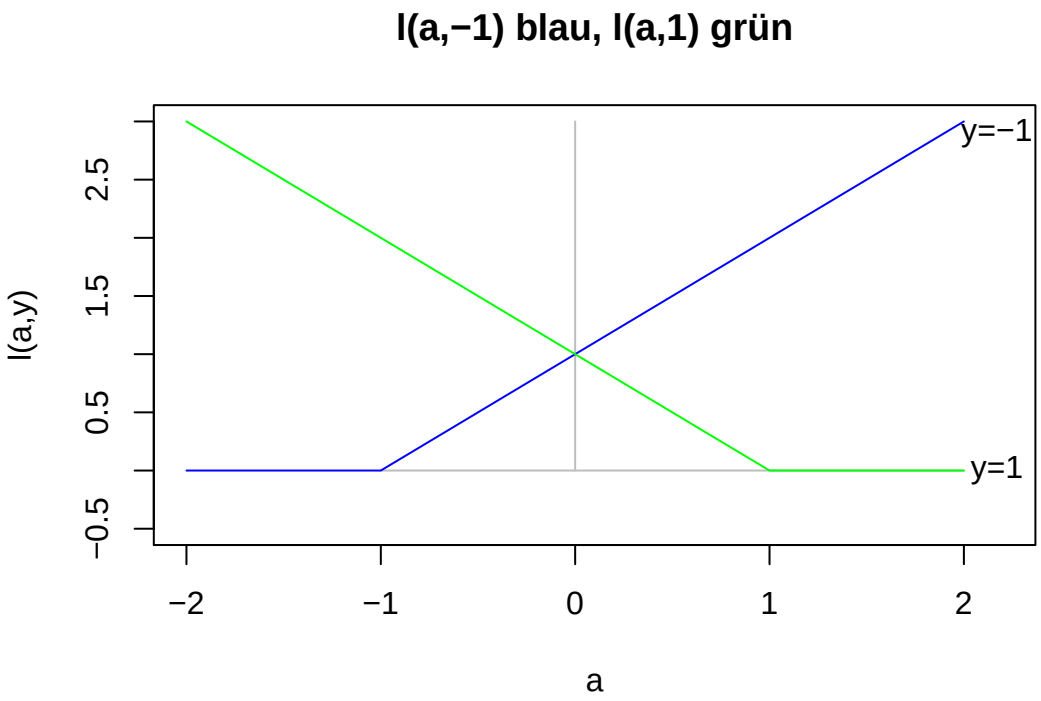
für Nullfolgen $(\sigma_n)_{n\in\mathbb{N}}$ und $(\lambda_n)_{n\in\mathbb{N}}$.

4 Vorbereitung

Definition 4.1 (Hinge-Loss)**.**
Hinge-Loss ist eine Verlustfunktion $l : \mathbb{R} \times \{-1, 1\} \to [0, \infty)$*, die durch*

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

definiert ist. Falls $a \in [-1, 1]$ *stimmt sie mit der Funktion* $|a - y|$ *überein.*



In Zukunft wird $R_P(f)$ für $R_{l,P}(f) = \mathbb{E}_P[l(f(X), Y)]$ stehen.

Definition 4.2 (Tsybakov-Noise-Exponent)**.**
Sei P *ein W-Maß auf* $[0, 1] \times \{-1, 1\}$ *und* η *beschreibe eine reguläre Version der Bedingten Wahrscheinlichkeit* $P(Y = 1 \mid X = \cdot)$*, dann hat* P *den* ***Tsybakov-Noise-Exponent*** $q \in [0, \infty]$*, falls*

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

Damit ist q *nicht eindeutig. Falls* P *den TNE.* q *besitzt, dann besitzt es auch* q' *für jedes* $q' \leq q$*.*

Anschaulich bedeutet ein größerer Tsybakov-Noise-Exponent, dass die Verteilung P besser zwischen den Kategorien $Y = -1$ und $Y = 1$ unterscheidet.

Beispiel 4.3.

- *Sei* $P^{X|Y=-1} = P^{X|Y=1} = U([0,1])$ *und* P_Y *beliebig, dann gilt*

$$\eta = \frac{1}{2}, \lambda - fs.$$

  womit P *den TNE.* 0 *besitzt.*

- *Sei* $P^{X|Y=1} = \delta_1$, $P^{X|Y=-1} = \delta_0$ *und* P_Y *beliebig, dann gilt* $supp(P_X) = \{0,1\}$ *und*

$$\forall x \in \{0,1\} : \ \eta(x) = x.$$

  Damit gilt für $t < \frac{1}{2}$, *dass*

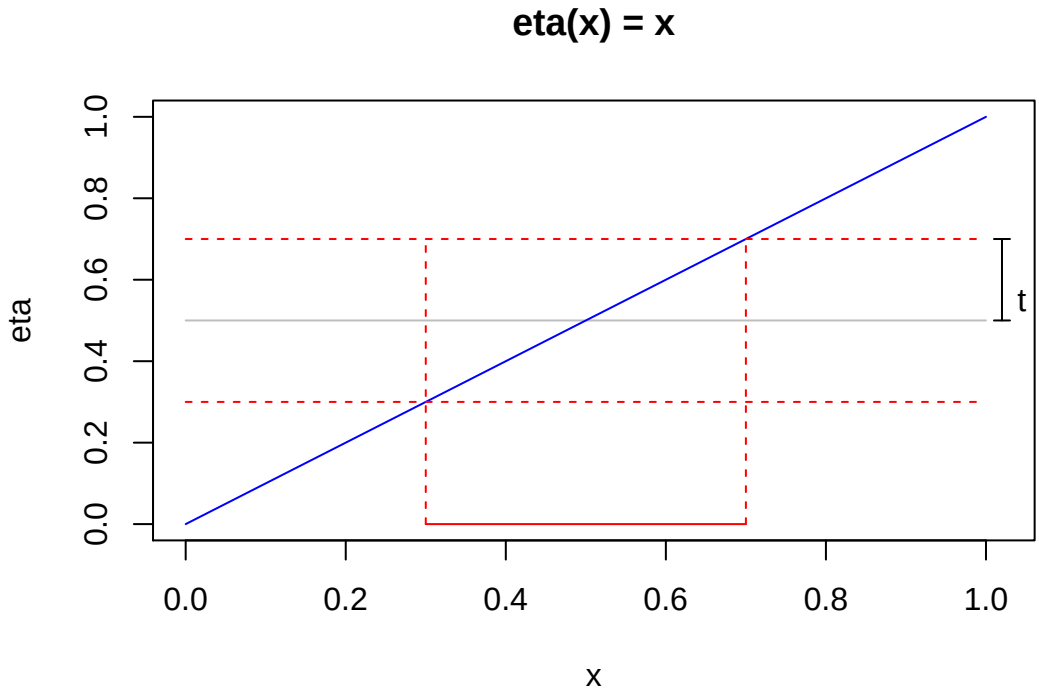
$$P_X\left(\left|\eta(X) - \frac{1}{2}\right| \leq t\right) \leq P_X\left(\left|X - \frac{1}{2}\right| < \frac{1}{2}\right) = 0 = t^{\infty}$$

  und P *besitzt den TNE.* ∞.

- *Sei* $P_X = U([0,1])$ *und* $\eta(x) = f_i(x)$ *mit*

$$f_1(x) = x \ ; \quad f_2(x) = 4\left(x - \frac{1}{2}\right)^3 + \frac{1}{2} \ ; \quad f_3(x) = \frac{1}{2}\left(x - \frac{1}{2}\right)^{\frac{1}{3}} + \frac{1}{2},$$

  dann hat P_1 *den TNE.* 1, P_2 *den TNE.* $\frac{1}{3}$ *und* P_3 *den TNE.* 3.



Lemma 4.4. (TNE. für η Polynom)
Sei P ein W-Maß auf $[0,1] \times \{-1,1\}$, sodass P_X eine beschränkte Dichte bezüglich des Lebesgue-Maßes besitzt. Falls $0 \neq P(Y=1 \mid X=x) - \frac{1}{2}$ sich fs. durch ein Polynom $\eta(x) = \widetilde{p}(x) = a_m x^m + ... + a_0$ darstellen lässt, dann besitzt P (mindestens) jeden TNE. $q \in [0, \frac{1}{m}]$.

Beweis.
Seien $x_1, ..., x_{m'} \in [0,1]$ die Nullstellen von p, dann existiert zu jedem x_j ein $k_j \leq m$, sodass $p^{(k_j)}(x_j) \neq 0$, denn sonst wäre $p = 0$.
Die Taylorentwicklung um x_j ergibt:

$$p(x_j + y) - \underbrace{p(x_j)}_{=0} = \sum_{k=1}^{m} \frac{p^{(k)}(x_j)}{k!}(y - x_j)^k = \sum_{k=k_j}^{m} \frac{p^{(k)}(x_j)}{k!}(y - x_j)^k$$

Wir folgern, dass ein $\varepsilon_j > 0$ existiert, sodass für alle $y \in B_{\varepsilon_j}$ gilt:

$$|p(x_j + y)| = \left| \sum_{k=k_j}^{m} \frac{p^{(k)}(x_j)}{k!}(y - x_j)^k \right| = |y - x_j|^{k_j} \left| \sum_{k=k_j}^{m} \frac{p^{(k)}(x_j)}{k!}(y - x_j)^{k-k_j} \right|$$

$$= |y - x_j|^{k_j} \underbrace{\left| \frac{p^{(k_j)}(x_j)}{k_j!} + \sum_{k=k_j+1}^{m} \frac{p^{(k)}(x_j)}{k!} (y - x_j)^{k-k_j} \right|}_{\geq \delta > 0 \text{ in einer kleinen Umgebung } B_{\varepsilon_j} \text{ um } x_j} \geq \delta \, |y - x_j|^{k_j}$$

Damit berechnet man

$$P_X \left(\left| \eta(X) - \frac{1}{2} \right| \leq t \right) \leq c \, \lambda \left(\{ |p| \leq t \} \right) \overset{t<<1}{\leq} \sum_{j=1}^{m'} \lambda \left(\left\{ \delta \, |x_j - \cdot|^{k_j} \leq t \right\} \right)$$
$$= \sum_{j=1}^{m'} \lambda \left(B_{\left(\frac{t}{\delta}\right)^{\frac{1}{k_j}}}(x_j) \right) = \sum_{j=1}^{m'} 2 \left(\frac{t}{\delta} \right)^{\frac{1}{k_j}} \leq \left(\sum_{j=1}^{m'} 2 \frac{1}{\delta^{\frac{1}{k_j}}} \right) t^{\frac{1}{m}}$$

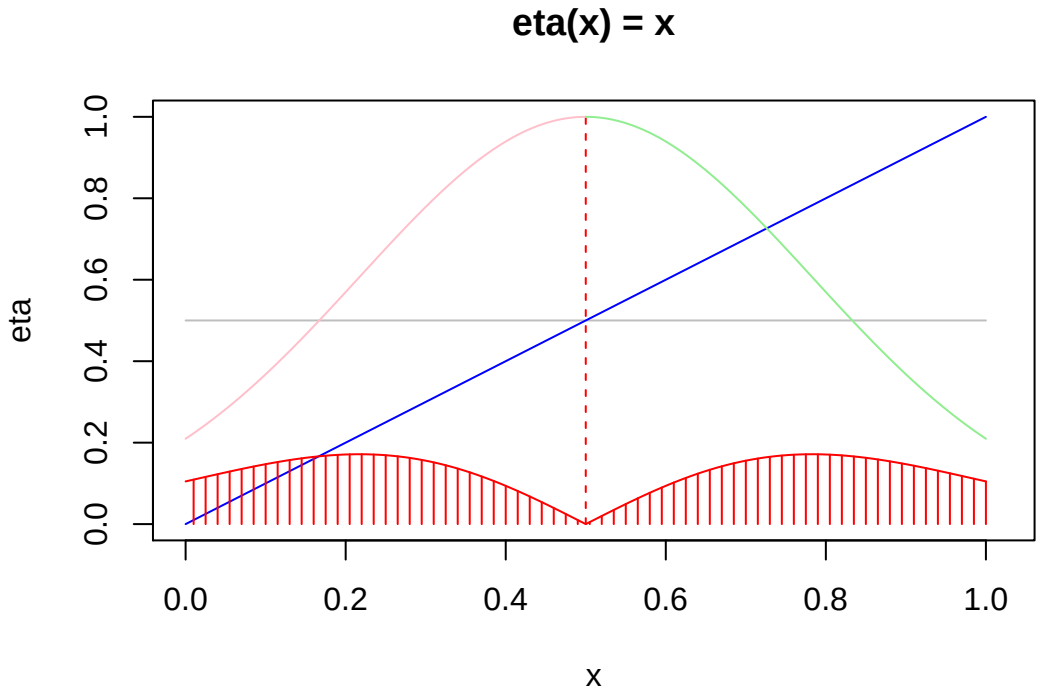
□

Definition 4.5 (Geometric-Noise-Exponent)**.**
Sei P ein W-Maß auf $[0,1] \times \{0,1\}$. Zu einem $x \in [0,1]$ beschreibe $\tau_P(x)$ den Abstand zu nächsten Stelle $\widetilde{x} \in [0,1]$, an der $\eta - \frac{1}{2}$ das Vorzeichen wechselt, also

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

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

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



Lemma 4.6. (GNE. für $\eta - \frac{1}{2}$ Hölder-stetig mit endl. vielen Nullstellen)
Sei P ein W-Maß auf $[0,1] \times \{-1,1\}$, sodass P_X eine beschränkte Dichte bezüglich des Lebesgue-Maßes besitzt. Falls für eine reguläre Version der Bedingten Verteilung η gilt, dass $\eta - \frac{1}{2}$ Hölder-stetig zur Konstante γ ist und nur endlich viele Nullstellen besitzt, dann hat P einen GNE. $\alpha \geq 1 + \gamma$.

Beweis. [eigenständig bewiesen]
Sei $p : [0,1] \to [0,M]$ die Dichtefunktion von P_X. Aus der Stetigkeit von $P(Y \mid X = \cdot) - \frac{1}{2}$ folgt, dass τ_P stückweise differenzierbar mit Ableitung in $\{-1,0,1\}$ ist. Seien $(a_1,b_1),...,(a_{m_+},b_{m_+})$ die (maximal großen) Intervalle mit $\tau_P' = 1$ und $(c_1,d_1),...,(c_{m_-},d_{m_-})$ die Intervalle mit $\tau_P' = -1$. Es gibt jeweils nur endlich viele, da gefordert ist, dass $\eta - \frac{1}{2}$ endlich viele Nullstellen hat. Wir berechnen:

$$\begin{aligned}
&\mathbb{E}_P\left[\left|\eta(X) - \frac{1}{2}\right| e^{-\frac{\tau_P(X)^2}{t^2}}\right] = \int_0^1 p(x)\,|f(x)|\, e^{-\frac{\tau_P(x)^2}{t^2}}\, dP_X(x) \\
&= \int_0^1 p(x)\,|f(x) - \underbrace{f(x \pm \tau_P(x))}_{=0}|\, e^{-\frac{\tau_P(x)^2}{t^2}}\, dP_X(x) \\
&\leq M\widetilde{C} \int_0^1 \tau_P(x)^\gamma\, e^{-\frac{\tau_P(x)^2}{t^2}}\, dx
\end{aligned}$$

$$
\begin{aligned}
&= M\widetilde{C} \sum_{i=1}^{m_+} \int_{a_i}^{b_i} (x-a_i)^\gamma \, e^{-\frac{(x-a_i)^2}{t^2}} \, dx + M\widetilde{C} \sum_{i=1}^{m_-} \int_{c_i}^{d_i} (d_i-x)^\gamma \, e^{-\frac{(d_i-x)^2}{t^2}} \, dx \\
&= M\widetilde{C} \sum_{i=1}^{m_+} \int_0^{b_i-a_i} x^\gamma \, e^{-\frac{x^2}{t^2}} \, dx + M\widetilde{C} \sum_{i=1}^{m_-} \int_0^{d_i-c_i} x^\gamma \, e^{-\frac{x^2}{t^2}} \, dx \\
&= M\widetilde{C} \sum_{i=1}^{m_+} \int_0^{\frac{b_i-a_i}{t}} (zt)^\gamma \, e^{-z^2} \, t \, dz + M\widetilde{C} \sum_{i=1}^{m_-} \int_0^{\frac{d_i-c_i}{t}} (zt)^\gamma \, e^{-z^2} \, t \, dz \\
&\leq M\widetilde{C} \, t^{1+\gamma} \sum_{i=1}^{m_+} \int_0^{\infty} z^\gamma \, e^{-z^2} \, dz + M\widetilde{C} \, t^{1+\gamma} \sum_{i=1}^{m_-} \int_0^{\infty} z^\gamma \, e^{-z^2} \, dz \\
&= t^{1+\gamma} \, M\widetilde{C} \, (m_+ + m_-) \int_0^{\infty} z^\gamma \, e^{-z^2} \, dz
\end{aligned}
$$

□

Satz 4.7 (GNE. für $\eta - \frac{1}{2}$ Hölder-stetig und bekanntem TNE.)**.**
Sei P ein W-Maß auf $[0,1] \times \{-1,1\}$, sodass ein $c > 0$ existiert mit

$$\left| \eta - \frac{1}{2} \right| \leq c\tau(x)^\gamma \quad P_X \text{ fast sicher,}$$

dann gilt für jeden TNE. $q \in [0,\infty]$, dass $\alpha := (q+1)\gamma$ auch eine GNE. von P ist.

(vgl. [1] Thm. 2.6)

In höheren Dimensionen, also P W-Maß auf $K \times \{-1,1\}$ für $K \subset \mathbb{R}^d$ kompakt, muss $\alpha := \frac{(q+1)\gamma}{d}$ gewählt werden.

5 Aussagen des Artikels

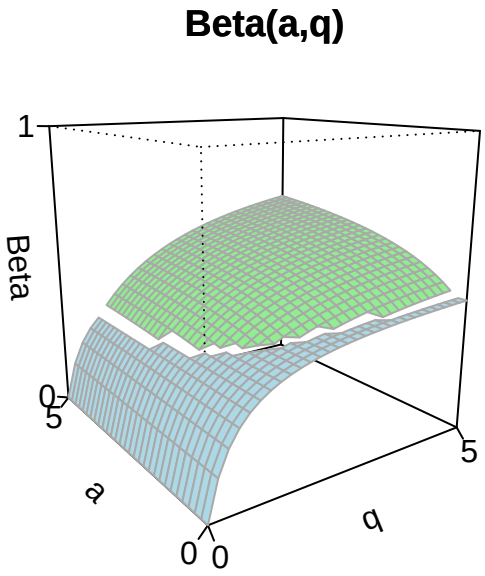
Satz 5.1.
Sei P ein W-Maß auf $[0,1] \times \{-1,1\}$ mit TNE. $q \in [0,\infty]$ und GNE. $\alpha \in (0,\infty)$. Wir definieren β durch

$$\beta = \begin{cases} \frac{\alpha}{2\alpha+1} & , \alpha \leq \frac{q+2}{2q} \\ \frac{2\alpha(q+1)}{2\alpha(q+2)+3q+4} & , \alpha > \frac{q+2}{2q} \end{cases} = \max\left(\frac{\alpha}{2\alpha+1}, \frac{2\alpha(q+1)}{2\alpha(q+2)+3q+4}\right)$$

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

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

(vgl. [1] Thm.2.8.)



Beweis.
Der Großteil des Artikels [1] dreht sich um den Beweis dieses Satzes, deswegen hier nur die Beweis-Idee.
Zuerst wird ein Satz aus [4] leicht modifiziert um folgende Aussage zu erhalten.

Sei $Z \subset \mathbb{R}^d$ kompakt und $\mathcal{F} \subset \{f : Z \to \mathbb{R}\, mb.\}$ konvex. Sei $L : \mathcal{F} \times Z \to [0, \infty)$ eine im ersten Argument konvexe Verlustfunktion mit der Eigenschaft, dass $t \mapsto L(tf_1 + (1-t)f_2, z)$ stetig ist.
Für ein W-Maß P auf Z, für welches ein Minimierer $f_{L,P} = \min_{f \in \mathcal{F}} R_{L,P}(f) = \min_{f \in \mathcal{F}} \mathbb{E}_{z \sim P}[L(f, z)]$ existiert, definieren wir

$$\mathcal{G}_{\mathcal{F},P} := \{g_f := L(f, \cdot) - L(f_{L,P}, \cdot) \mid f \in \mathcal{F}\}.$$

Falls konstanten $c, \delta, B > 0; \alpha \in (0, 1]$ existieren, sodass

$$\forall g \in \mathcal{G}_{\mathcal{F},P} : \ \mathbb{E}_P[g^2] \leq c\mathbb{E}_P[g]^\alpha + \delta \ \text{ und } \ ||g||_\infty \leq B,$$

dann gilt für alle $n, x \in \mathbb{N}$ und $\varepsilon > 0$ mit

$$\varepsilon \geq 10 \max\left(\mathbb{E}\left[\sup_{g \in \mathcal{G}_{L,P}\,;\, \mathbb{E}_P[g^2] < c\varepsilon^\alpha + \delta} \left|\hat{P}_n(g) - P(g)\right|\right], \sqrt{\frac{\delta x}{n}}, \left(\frac{4cx}{n}\right)^{\frac{1}{2-\alpha}}, \frac{Bx}{n}\right),$$

dass

$$\mathbb{P}^*\left(R_{L,P}(f_{n,L}) - R_{L,P}(f_{L,P}) \leq \varepsilon\right) \leq e^{-x}.$$

(vgl. [1] Thm. 5.1)
Wir wollen diesen Satz mit $Z = [0, 1] \times \{-1, 1\}$, $\mathcal{F}_m = B_{r_m}^{H_{\sigma_m}}(0)$ und $L_m(f, (x, y)) = \lambda_m ||f||_{H_{\sigma_m}} + l(f(x), y)$ anwenden.

- Konvexität und Stetigkeit von L_m:
  Aus der Dreiecksungleichung und der Konvexität von l folgt sofort

  $$\begin{aligned} &L_m(tf_1 + (1-t)f_2, (x, y)) \\ &= \lambda_m ||tf_1 + (1-t)f_2||_{H_{\sigma_m}} + l(tf_1(x) + (1-t)f_2(x), y) \\ &\leq \lambda_m \left(t||f_1||_{H_{\sigma_m}} + (1-t)||f_2||_{H_{\sigma_m}}\right) + t\, l(f_1(x), y) + (1-t)\, l(f_2(x), y) \\ &= t\, L_m(f_1, (x, y)) + (1-t)\, L_m(f_2, (x, y)). \end{aligned}$$

  Auch die Stetigkeit von $t \mapsto L(tf_1 + (1-t)f_2, z)$ folgt aus der Halblinearität von $|| \cdot ||_{H_{\sigma_m}}$ und der Stetigkeit von l.

- Existenz der Konstanten c, δ, B, α:
  Für jedes $f \in \mathcal{F}_m$ gilt

  $$g_f(x, y) = \lambda_m ||f||_{H_{\sigma_m}} + l(f(x), y) - \lambda_m ||f_{L,P,m}||_{H_{\sigma_m}} - l(f_{L,P,m}(x), y)$$

$$= \lambda_m \underbrace{\left(||f||_{H_{\sigma_m}} - ||f_{L,P,m}||_{H_{\sigma_m}}\right)}_{\leq r_m} + l(\underbrace{f(x)}_{|\cdot|\leq r_m}, y) - l(\underbrace{f_{L,P,m}(x)}_{|\cdot|\leq r_m}, y)$$
$$\leq (\lambda_m + 2)\, r_m =: B.$$

Weiter wird in [1] Thm. 6.1 gezeigt, dass:

> Falls P den TNE. $q \in [0, \infty]$ besitzt, dann existiert ein Minimierer $f_{l,P}$, sodass für jede messbare Funktion $f : [0,1] \to \mathbb{R}$ gilt
>
> $$\mathbb{E}\left[(l(f(X), Y) - l(f_{l,P}(x), Y))^2\right]$$
> $$\leq \widetilde{C}\ (||f||_\infty + 1)^{\frac{q+2}{q+1}}\ \mathbb{E}\left[l(f(X), Y) - l(f_{l,P}(x), Y)\right]^{\frac{q}{q+1}}$$
>
> für
>
> $$\widetilde{C} = \left(\sup_{t>0} t^q P_X\left(t\left|\mathbb{E}[Y \mid X] - \frac{1}{2}\right| > 2\right)\right)^{\frac{1}{q}} + 2.$$
>
> (Falls $q = 0$, kann $\widetilde{C} = 1$ gewählt werden.)

(vgl. [1] Lemma 6.1)
Dies wird in [1] Prop. 6.3 verwendet um für $f \in B_\gamma^{H_{\sigma_m}}(0)$ zu zeigen:

$$\mathbb{E}[g_f^2] \leq 8\widetilde{C}(K\gamma + 1)^{\frac{q+2}{q+1}}\, \mathbb{E}[g_f]^{\frac{q}{q+1}} + 16\widetilde{C}(K\gamma + 1)^{\frac{q+2}{q+1}} a(\gamma)^{\frac{q}{q+1}},$$

wobei $K > 0$ eine von f und γ unabhängige Konstante ist und

$$a(\lambda) = \inf_{h \in H_{\sigma_m}} \left(\lambda ||h||_{H_{\sigma_m}} + R_{l,P}(h) - \inf_{f:[0,1]\to[0,1]\, mb.} R_{l,P}(f)\right).$$

Wir erhalten also $\alpha = \frac{q}{q+1}$, $c = 8\widetilde{C}(K\gamma + 1)^{\frac{q+2}{q+1}}$ und $\delta = 2c\,(a(r_m))^{\frac{q}{q+1}}$.

- Abschätzung von $\omega_n\left(\mathcal{G}_{L,P}, c\varepsilon^\alpha + \delta\right) := \sup_{g \in \mathcal{G}_{L,P}\,;\, \mathbb{E}_P[g^2] < c\varepsilon^\alpha + \delta} \left|\hat{P}_n(g) - P(g)\right|$:
  In [1] Sektion 5.2 wird die Abschätzung durch den lokalen Rademacher-Durchschnitt

  $$\omega_n\left(\mathcal{G}_{L,P}, c\varepsilon^\alpha + \delta\right) \leq 2\,\mathcal{R}_n(\mathcal{G}_{L,P}, c\varepsilon^\alpha + \delta)$$

  gezeigt, vergleiche [5] Lemma 2.3.1. (Mehr zum Rademacher-Durchschnitt im zweiten Vortrag.)
  Dieser wird im Rest von [1] Sektion 5 weiter abgeschätzt.

- Abschätzung von $\sqrt{\frac{\delta x}{n}}$ und $\frac{Bx}{n}$:
  Wir haben bereits festgestellt, dass

  $$\delta = 2c(a(r_m))^{\frac{q}{q+1}} \text{ und } B = r_m(\lambda_m + 2).$$

  Durch eine passende Wahl der aufsteigenden Folge $(r_m)_{m\in\mathbb{N}}$ und einer späteren Gleichsetzung von n und m, lässt sich die Fallgeschwindigkeit der beiden Terme kontrollieren.

- Abschätzung von $\left(\frac{4cx}{n}\right)^{\frac{1}{2-\alpha}}$:
  Dieser Term lässt sich nur begrenzt abschätzen, weshalb er so großen Einfluss auf das oben beschriebene Beta hat. Genaueres findet man in [1] auf Seite 598.

□

Bemerkung 5.2 (Einfache Anwendung von Satz 5.1)**.**
Sei P ein W-Maß auf $[0,1] \times \{-1,1\}$, sodass P_X eine beschränkte Dichte bezüglich des Lebesgue-Maßes besitzt.

1) Falls $\eta - \frac{1}{2}$ auf $[0,1]$ stetig differenzierbar ist und endlich viele Nullstellen besitzt, dann ist $\beta \geq \frac{1}{5}$.

2) Falls $\eta - \frac{1}{2}$ auf $[0,1]$ stetig differenzierbar ist, endlich viele Nullstellen besitzt und für jede Nullstelle x_0 gilt $g'(x_0) \neq 0$, dann ist $\beta \geq \frac{8}{19}$.

3) Falls $\eta - \frac{1}{2}$ Hölder-stetig zur Konstante γ ist und endlich viele Nullstellen besitzt, dann ist $\beta \geq \frac{1+\gamma}{3+2\gamma}$.

Beweis. [eigenständig bewiesen]

1) Aus Lemma 4.6 folgt $\alpha \geq 2$, da $\widetilde{p}$ auf dem Kompakten Raum $[0,1]$ Lipschitzstetig ist. Man berechnet leicht

$$\beta \geq \frac{\alpha}{2\alpha + 1} \geq \frac{1}{5}.$$

2) Mit Hilfe des Mittelwertsatzes ist leicht gezeigt, dass für den oben beschriebenen Fall $q = 1$ gilt. Genau wie im 1. Fall ist g hier auch Lipschitzstetig, womit $\alpha \geq 2$ gilt. Man berechnet

$$\beta \geq \frac{2\alpha(q+1)}{2\alpha(q+2) + 3q + 4} \geq \frac{8}{19}.$$

3) Aus Lemma 4.6 wissen wir, dass $\alpha \geq 1 + \gamma$. Wir berechnen

$$\beta \geq \frac{2\alpha(q+1)}{2\alpha(q+2)+3q+4} \geq \frac{1+\gamma}{2(1+\gamma)+1} = \frac{1+\gamma}{3+2\gamma}.$$

□

Bemerkung 5.3.
Um $\beta > \frac{1}{2}$ zu erhalten muss $\alpha > \frac{3q+4}{2q}$ gelten.

6 Sonstiges

Wir definieren die Variations-Semi-Norm $||\cdot||_{TV}$ durch

$$||f||_{TV} := \lim_{m\to\infty} \sup_{x_1<\ldots<x_m} \sum_{i=1}^{m-1} |f(x_i) - f(x_{i+1})|.$$

Die folgenden Aussagen werden für den Vergleich mit dem 2. Vortrag benötigt.

Lemma 6.1.
Für jedes $f \in H_\sigma$ gilt

$$||f||_{TV} \leq \sqrt{\frac{2}{\sigma^2}}||f||_{H_\sigma}.$$

Beweis. [eigenständig bewiesen]

Jedes $h \in \mathrm{Span}_{\mathbb{R}}\{K_x \mid x \in [0,1]\}$ ist stetig differenzierbar, da $K_x = e^{-\frac{(x-\cdot)^2}{\sigma^2}} \in C^1([0,1])$. Seien (a_i, b_i) die (maximalen) Intervalle auf denen $h' > 0$ gilt und (c_i, d_i) die (maximalen) Intervalle auf denen $h' < 0$ gilt.
Es folgt

$$\begin{aligned}
||h||_{TV} &= \int_0^1 |h'(y)|\,dy = \sum_{i=1}^{\infty}\int_{a_i}^{b_i} h'(y)\,dy - \sum_{i=1}^{\infty}\int_{c_i}^{d_i} h'(y)\,dy \\
&= \sum_{i=1}^{\infty}\big(h(b_i) - h(a_i)\big) + \sum_{i=1}^{\infty}\big(h(c_i) - h(d_i)\big) \\
&\leq \sqrt{2}||h||_{H_\sigma}\left(\sum_{i=1}^{\infty}\sqrt{1 - e^{-\frac{(a_i-b_i)^2}{\sigma^2}}} + \sum_{i=1}^{\infty}\sqrt{1 - e^{-\frac{(c_i-d_i)^2}{\sigma^2}}}\right) \\
&\leq \sqrt{2}||h||_{H_\sigma}\left(\sum_{i=1}^{\infty}\frac{(a_i-b_i)}{\sigma} + \sum_{i=1}^{\infty}\frac{(c_i-d_i)}{\sigma}\right) \leq \sqrt{\frac{2}{\sigma^2}}||h||_{H_\sigma}.
\end{aligned}$$

Damit ist die lineare Funktion

$$Id : (\mathrm{Span}_{\mathbb{R}}(K_x \mid x \in [0,1]), ||\cdot||_{H_\sigma}) \to (E_{TV}, ||\cdot||_{TV})$$

für

$$E_{TV} := \{f : [0,1] \to \mathbb{R}\, mb. \mid ||f||_{TV} < \infty\}$$

stetig und besitzt eine eindeutige stetige Fortsetzung

$$Id : (H_\sigma, ||\cdot||_{H_\sigma}) \to \overline{(E_{TV}, ||\cdot||_{TV})}$$

mit derselben Operatornorm.

□

Literatur

[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 10.1162/153244302760185252.

[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.

[5] Aad W. van der Vaart and Jon A. Wellner, *Weak convergence and empirical processes*, Springer Series in Statistics, Springer-Verlag, New York, 1996. With applications to statistics.