Jan-Cornelius Molnar

Nichtparametrische Statistik und statistische Lerntheorie

Aufbauend auf einer Vorlesung von Prof. Dr. I. Steinwart

Stuttgart, Sommersemester 2010

Version vom 1. März 2012

Diese Rohfassung enthält sicher noch zahlreiche Druckfehler. Für Hinweise auf solche und Kommentare jeder Art bin ich stets dankbar.[1] Viel Spaß!

[1] Jan-Cornelius Molnar jan.molnar@studentpartners.de

Inhaltsverzeichnis

0 Grundlagen

0-1 Statistische Lerntheorie (SLT)

Die statistische Lerntheorie ist Teil der nichtparametrischen Statistik sowie der Informatik. Sie bildet den mathematischen Arm des maschinellen Lernens und befasst sich sowohl mit den theoretischen Grundlagen wie auch Fragen der Implementierung und besitzt zahlreiche Anwendungen.

0-1-1 Inhalt der Vorlesung

- Was ist Lernen?
- Klassische Verfahren
- Empirische Risikominimierung (ERM)
- Regularisierte ERMs
- Kernbasierte Verfahren (support vector machines, SVMs)
- Verlustfunktionen

0-2 Lernszenarien und -ziele

0-2-1 Überwachtes Lernen

Beim überwachten Lernen wird der Lernprozess durch einen Lehrer geführt. Den Eingaberaum bezeichnen wir mit X, den Ausgaberaum mit Y. Weiterhin sind Trainingsdaten in der From

$$((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$$

gegeben. Das Ziel ist es eine Funktion

$$f_D : X \to Y$$

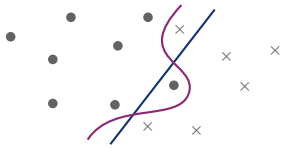
zu so finden, dass $f_D(x) \approx y$ für neue Paare (x, y). Es geht also nicht darum, die Daten einfach auswendig zu lernen, sondern y für zukünftiges x vorherzusagen. Was "$\approx$" genau bedeutet werden wir später erklären.

Beim überwachten Lernen, erhält der Lehrer ein x_i und ordnet diesem ein Label y_i zu. Der Lernende erhält dann vom Lehrer das Tupel (x_i, y_i) und soll darauf f_D ableiten.

Klassifikation

Gegeben sei der Ausgaberaum $Y = \{-1, 1\}$ und Ziel ist f_D so zu bestimmen, dass $f_D(x) = y$ „möglichst häufig".

Es ist nun grundsätzlich zu entscheiden, ob man den Trainingsdaten traut oder eher ein Rauschen vermutet.



0.1 **Zwei Möglichkeiten zur Klassifikation von Trainingsdaten.**

Geht man davon aus, dass auf den Trainingsdaten keine Fehler sind, besteht die Gefahr des sogenannten overfitting, d.h. man traut den Daten zu sehr. Vermutet man andererseits ein Rauschen besteht die Gefahr des underfitting, d.h. die Annahmen an die Zielfunktion sind unzureichend.

Over- und underfitting konkurrieren miteinander. Gute Lernverfahren finden einen guten bzw. den "besten" Kompromiss.

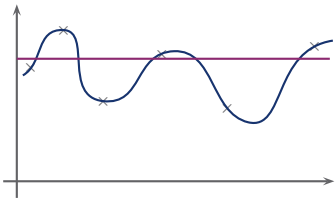
Anwendungen der Klassifikation

- Spamfilter
- Fraud Detection (Betrugsvorfälle bei Kreditkartenunternehmen)

- Schriftenerkennung
- Diagnoseverfahren

Regression

Sei $Y = [-M, M]$ oder $Y = \mathbb{R}$. Auch hier ist das Ziel $f_D(x) \approx y$ möglichst häufig und es besteht ebenfalls die Gefahr des over- bzw. underfittings.



0.2 **Zwei Möglichkeiten zur Regression (datentreu oder linear).**

Was man genau unter "$\approx$" versteht, fällt unter den Terminus Verlustfunktion. Je nachdem wie man "$\approx$" mathematisch präzesiert, erreicht man unterschiedliche Lernziele.

Anwendung der Regression

- Finanzindustrie: Kursvorhersage von Aktien, Ausfallrisiken von Krediten, ...
- Data Mining
- Klassifikation $P(Y = y \mid x) = ?$
- Schätzungen von Restlebenszeiten von Teilen

0-2-2 Halbüberwachtes Lernen

Hier sind X und Y wie beim überwachten Lernen definiert. Die Trainingsdaten sind jedoch nur noch zum Teil mit einem Label versehen. Sie bestehen also aus

einer „kleinen“ Menge an

$$((x_1, y_1), \ldots, (x_n, y_n)) \in (X, Y)^n,$$

wobei $n = 2$ den Extremfall bildet, sowie einer „großen“ Menge an

$$(x_{n+1}, \ldots, x_m) \in X^{m-n}, \qquad m >> 1.$$

Nun kann man ebenfalls Klassifikation oder Regression betreiben.

Die Motivation hierbei ist, dass die Beschaffung der y_i entweder “teuer” oder gar unmöglich ist. Beispiele dafür sind medizinische Studien oder Diagnoseverfahren zur Überwachung von Brücken.

Die Methoden für das halbüberwachte Lernen sind noch weitgehend unausgereift. Es gibt noch sehr viele Fragestellungen an denen man arbeiten kann.

0-2-3 Unüberwachtes Lernen

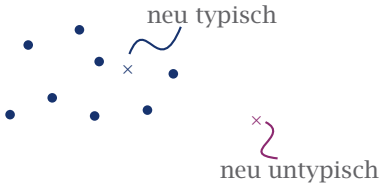
Beim unüberwachten Lernen sind nur Trainingsdaten ohne Label gegeben

$$x_1, \ldots, x_n \in X.$$

Es lassen sich nun verschiedene Lernziele definieren. Wir wollen hier nur einen kurzen Überblick geben; ausführlich werden wir uns lediglich mit der Ausreißeridentifikation beschäftigen.

Ausreißeridentifikation

Es geht darum “untypische” Daten zu identifizieren und von „typischen“ zu unterscheiden.

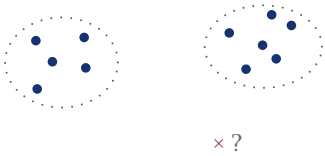


0.3 **Zwei Möglichkeiten zur Klassifikation von Trainingsdaten.**

Die Ausreißeridentifikation hat viele Ähnlichkeiten mit der Klassifikation und bereits gut verstanden.

Clusteranalyse

Hier ist zunächst festzustellen, ob die Daten Cluster bilden, und diese gegebenenfalls zu ermitteln.



0.4 **Clusterbildung.**

Im Gegensatz zur Ausreißeridentifikation bietet die Clusteranalyse noch sehr viele offene Fragen.

Dimensionsreduzierung

Viele Modelle erzeugen sehr hochdimensionale Daten. Hochdimensionale Daten sind schlecht numerisch zu handhaben und außerdem schlecht zu visualisieren.

0.5 **Dimensionsreduktion durch Approximation der Daten durch eine Funktion.**

Bsp Eine Kaffeekanne wird unter Rotation fotografiert, um so ein dreidimensionales Bild zu erhalten. Jedes Foto besteht aus n Pixeln, die man als Punkte im $\mathbb{R}^n$ ansehen kann. Diese gewaltige Datenmenge lässt sich jedoch oft auf den $\mathbb{R}^3$ reduzieren. ■

Auch die Dimensionsreduzierung ist noch weitgehend unerforscht.

Dichteschätzung

Die Dichteschätzung ist mit klassischen Verfahren gut handhabbar und bereits sehr ausgereift.

0.6 **Zur Dichteschätzung.**

0-2-4 Weitere Lernszenarien

- *Reinforcement Learning*

  Reinforcement Learning hat das Lernen von Handlungen zum Ziel. Beispielsweise misst ein Roboter Umwelteinflüsse und entscheidet sich dannach für Handlungen. Welche Entscheidung "gut" war ist jedoch nur mit einer gewissen Verzögerung feststellbar.

- *Multi-armed Bandit Learning*

  Man hat n Optionen, wovon jede einen Gewinn ausschüttet. Aufgrund beschränkter Ressourcen muss man eine Priorisierung vornehmen, um den größtmöglichen Gewinn zu erzielen.

0-2-5 Datengenerierung

- *Batch Learning*

  Hier liegen alle Trainingsdaten vor Trainingsbeginn vor.

- *Online Learning*

  Das Online Learning ist bei den Informatikern sehr beliebt, da sich damit sehr gut "worst case"-Szenarien realisieren lassen.

  Ein möglicher Algorithmus wäre
  a) Habe bereites $(x_1, y_1), \ldots, (x_i, y_i)$
  b) Finde $f_i : X \to Y$
  c) Erhalte x_{i+1}
  d) Mache Vorhersage $f_i(x_{i+1})$
  e) Bekomme y_{i+1}
  f) Zurück zu a)

Eine “billige"Generierung von f_i ist wesentlich für den Erfolg des Online Learning.

- *Active Learning*

  Bestimme $x_1, \dots, x_n$ selbst und erhalte anschließend die zugehörigen Labels $y_1, \dots, y_n$.

Natur der Daten

- *Zufällige Daten.* Dies ist typisch für batch learning. Wir werden uns später damit ausführlich beschäftigen.
- *Pseudozufällige Daten.* Die Daten werden eigentlich von einem deterministischen System generiert, dieses ist aber chaotisch.
- *“worst case” Daten.* Typisch für online learning.

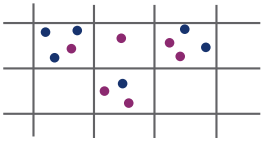
0-3 Klassische Verfahren zur Klassifikation

Sei $X = \mathbb{R}^d$ und $Y = \{-1, 1\}$.

- *Histogrammregel.*

  Sei $(Q_j)_{j \geqslant 1}$ eine Zerlegung von X in Würfel. Es gilt dann offenbar

  $$\forall\, x \in X\ \exists\, j \geqslant 1 : x \in Q_j.$$



0.7 **Zerlegung von X.**

Wir definieren nun

$$f_D(x) \coloneqq \operatorname{sign}\left(\sum_{x_i \in Q_i(x)} y_i \right).$$

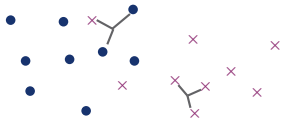
- *Nearest Neighbor.*

  Sei (X, d) ein metrischer Raum (z. B. $X = \mathbb{R}^d$). Fixiere die Anzahl k der betrachteten Nachbarn. Für $x \in X$ suche die k Beispiele

  $$(x_{i_1}, y_{i_1}), \ldots, (x_{i_k}, y_{i_k})$$

  aus D, die am nächsten liegen.

  $$f_D(x) = \operatorname{sign}\left(\sum_{j=1}^{k} y_{i_j}\right).$$



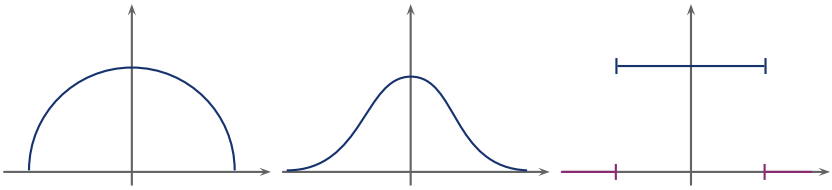
0.8 **Nearest Neighbor Identifikation.**

- *Moving window / kernel rules.*

  Sei $X = \mathbb{R}^d$ ein euklidischer Raum und

  $$K : \mathbb{R} \to [0, \infty)$$

  symmetrisch und auf $[0, \infty)$ monoton fallend gegeben.



0.9 **Beispiele für Funktionen $K : \mathbb{R} \to [0, \infty)$.**

Fixiere $h > 0$ und setze

$$f_D(x) = \operatorname{sign}\left(\sum_{i=1}^{k} y_i \, K(h^{-1} \|x - x_i\|)\right)$$

0.10 **Identifikation mittels moving window.**

1 Grundkonzepte der statistischen Lerntheorie

1-1 Grundannahmen

Folgende Notationen werden durchgehend verwendet.

- $(X, \mathcal{A})$ sei ein metrischer Raum.
- $Y \subset \mathbb{R}$ sei abgeschlossen mit $Y = \{-1, 1\}$, $Y = [-M, M]$ oder $Y = \mathbb{R}$. Y sei stets mit der Borel-σ-Algebra versehen.
- P sei ein W-Maß auf $X \times Y$.
- $D = ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$.
- $\mathcal{L}_0(X) := \{f : X \to \mathbb{R} : f \text{ messbar}\}$.

Wir benötigen folgenden Satz zu regulären bedingten Wahrscheinlichkeiten.

Satz *Es existiert eine Abbildung*

$$P(\cdot \mid \cdot) : \mathcal{B} \times X \to [0, 1]$$

mit folgenden Eigenschaften.

1) $P(\cdot \mid x) : \mathcal{B} \to [0, 1]$ *ist ein W-Maß auf* $\mathcal{B}$ *für alle* $x \in X$.
2) $x \mapsto P(B \mid x)$ *ist messbar für alle* $B \in \mathcal{B}$.

3) Für alle messbaren $A \in \mathcal{A}$ und $B \in \mathcal{B}$ gilt

$$P(A \times B) = \int_A P(B \mid x)\, \mathrm{d}P_X(x) \qquad (*)$$

wobei P_X ein W-Maß auf $(X, \mathcal{A})$ ist, das durch $P_X(A) := P(A \times Y)$ für $A \in \mathcal{A}$ definiert ist.

4) $P(\cdot \mid \cdot)$ ist P_X-f.s. eindeutig. ⋊

Interpretation. Falls $P = P_Y \otimes P_X$, so ist $P(B \mid x) = P_Y(B)$ und daher

$$(*) = P(A \times B) = P_Y(B) P_X(A).$$

$P(\cdot \mid \cdot)$ ermöglicht es, P im Sinne von (*) aufzuspalten ("desintegrate"). ⊸

Als Folgerung erhalten wir eine Verallgemeinerung des Satzes von Fubini von Produktmaßen auf solche P.

Korollar *Sei $f : X \times Y \to \mathbb{R}$ P-integrierbar, so gilt*

$$\mathbf{E}_P f := \int_{X \times Y} f \,\mathrm{d}P = \int_X \int_Y f(x, y) P(\mathrm{d}y \mid x)\, \mathrm{d}P_X(x). \quad ⋊$$

Bemerkung. Für allgemeine messbare Räume $(Y, \mathcal{A}')$ ist dies *nicht* möglich. (Wesentlich ist hier, dass $Y \subset \mathbb{R}$ abgeschlossen). ⊸

Weitere Grundannahmen sind.

- P ist uns völlig unbekannt. Wir wissen lediglich, dass P existiert.
- Die Trainingsdaten D sind Beobachtungen einer Folge $Z = (Z_i)_{i=1}^n$ mit $Z_i = (X_i, Y_i)$ von Zufallsvariablen, die $X \times Y$-wertig und unabhängig sind und alle die Verteilung P haben.

  Interpretation.
  $X_i \sim P_X$, dies gibt eine Beobachtung $x_i \in X$.
  $Y_i \sim P(\cdot \mid x_i)$, dies gibt eine Beobachtung $y_i \in Y$.

  Datengenerierungsalgorithmus.

- $i = 1$.
- "Würfle" x gemäß der Verteilung P_X und unabhängig von (x_1, y_1), $\dots$, (x_{i-1}, y_{i-1}).
- "Würfle" y gemäß der Verteilung $P(\cdot \mid x_i)$.
- Wiederhole bis $i = n$.

- Alle zukünftigen Daten (x, y) sind Beobachtungen von einer Zufallsvariable $\tilde{Z} = (\tilde{X}, \tilde{Y})$, die $X \times Y$-wertig ist, Verteilung P hat und unabhängig von der Folge $(Z_i)_{i=1}^n$ ist.

- Für eine große Anzahl zukünftiger Beobachtungen, betrachte die Folge $(Z_i)_{i=n+1}^\infty$ von $(X \times Y)$-wertigen Zufallsvariablen, die untereinander und von $Z_1, \dots, Z_n$ unabhängig sind und Verteilung P haben.

1-1.1 **Definition** *Eine Lernmethode oder auch Lernverfahren ist eine Folge $(L_n)_{n \geqslant 1}$ von Abbildungen*

$$L_n : (X \times Y)^n \to \mathcal{L}_0(X), \qquad D \mapsto f_D. \quad \bowtie$$

Bsp a) Histogrammregel.

b) Nearest Neighbor.

c) Moving windows, kernel rules. ■

1-1.2 *Bemerkung.* Wir werden später benötigen, dass die Abbildungen

$$(X \times Y)^n \times X \to \mathbb{R}, \qquad (D, x) \mapsto f_D(x)$$

messbar sind. ⊸

1-2 Verlustfunktionen und Risiken

Ziel dieses Abschnitts ist es, genauer zu beschreiben, was $f_D(x) \approx y$ bedeutet.

1-2.1 **Definition** *Eine messbare Abbildung*

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

heißt Verlustfunktion. Selbiges gilt für Funktionen

$$L : X \times \mathbb{R} \to [0, \infty), \quad \textit{oder} \quad L : Y \times \mathbb{R} \to [0, \infty). \quad \bowtie$$

Interpretation. $L(x, y, f(x)) \in [0, \infty)$ beschreibt den "Verlust" bei der Entscheidung $f_D(x)$, falls (x, y) beobachtet wird. Generell sind kleine Verluste besser als große! $\multimap$

1-2.2 **Definition** *Sei L eine Verlustfunktion. Dann ist das Risiko einer messbaren Funktion $f : X \to \mathbb{R}$ durch*

$$\begin{aligned} \mathcal{R}_{L,P}(f) &= \int_{X \times Y} L(x, y, f(y)) \, \mathrm{d}P(x, y) \\ &= \int_X \int_Y L(x, y, f(x)) P(\, \mathrm{d}y \mid x) \, \mathrm{d}P_X(x) \end{aligned}$$

definiert. $\bowtie$

Interpretation. Ist $Z = (Z_i)_{i=1}^n$ eine Folge von $X \times Y$-wertigen, unabhängigen und P-verteilten Zufallsvariablen und $\mathcal{R}_{L,P}(f) < \infty$, so besagt das starke Gesetz der großen Zahlen, dass

$$\mathcal{R}_{L,P}(f) = \lim_{m \to \infty} \frac{1}{m - n - 1} \sum_{i=n+1}^{\infty} L(x_i, y_i, f(x_i)) \; P\text{-f.s.}$$

$\mathcal{R}_{L,P}(f)$ beschreibt daher den mittleren zukünftigen Verlust. $\multimap$

Wir wollen jetzt genauer spezifizieren, was wir unter "kleinen" Risiken zu verstehen ist.

1-2.3 **Definition** *Sei L eine Verlustfunktion, dann heißt das kleinstmögliche Risiko*

$$\mathcal{R}^*_{L,P} := \inf \left\{ \mathcal{R}_{L,P}(f) \, : \, f : X \to \mathbb{R} \textit{ messbar} \right\}$$

Bayes-Risiko. Eine messbare Funktion $f^*_{L,P} : X \to \mathbb{R}$ für die gilt

$$\mathcal{R}_{L,P}(f^*_{L,P}) = \mathcal{R}^*_{L,P}$$

heißt Bayes-Entscheidungsfunktion. $\bowtie$

$f^*_{L,P}$ ist im Allgemeinen nicht P-f.s. eindeutig.

Unser informelles Lernziel besteht nun darin, ein Lernverfahren $\mathcal{L}$ zu finden, für das das Überschussrisiko

$$\mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P}, \qquad f_D \in \mathcal{L}$$

mit hoher Wahrscheinlichkeit klein ist. In Abschnitt 1-3 werden wir präzisieren was wir unter "hoher Wahrscheinlichkeit" und "klein" verstehen.

Bsp Klassifikation Hier ist unser Ausgaberaum $Y = \{-1, 1\}$ und unser Ziel, y "richtig vorauszusagen". Mithilfe der vorangegangen Definitionen lässt sich dies nun mathematisch formulieren.

Wir definieren eine Verlustfunktion wie folgt

$$L_{\text{class}} : Y \times \mathbb{R} \to [0, \infty), \quad (y, t) \mapsto \mathbf{I}_{(-\infty,0]}(y \operatorname{sign} t),$$

$$L_{\text{class}}(y, t) = \begin{cases} 0, & y = -1,\ t < 0, \\ 1, & y = -1,\ t \geqslant 0, \\ 1, & y = 1,\ t < 0, \\ 0, & y = 1,\ t \geqslant 0. \end{cases}$$

L_{class} bestraft somit Vorhersagen, für die $y \neq \operatorname{sign} t$. Wir können nun das Risiko angeben,

$$\begin{aligned} \mathcal{R}_{L_{\text{class}},P}(f) &= \int_{X \times Y} \mathbf{I}_{(-\infty,0]}(y \operatorname{sign} f(x)) \, \mathrm{d}P(x, y) \\ &= P\left\{(x, y) : y \neq \operatorname{sign} f(x)\right\} \\ &= \int_X \int_Y L_{\text{class}}(y, f(x)) P(\mathrm{d}y \mid x) \, \mathrm{d}P_X(x). \end{aligned}$$

Setzen wir nun $\eta(x) = P(y = 1 \mid x)$ so können wir das Integral schreiben als

$$\int_X \int_Y \eta(x) \mathbf{I}_{(-\infty,0)}(f(x)) + (1 - \eta(x)) \mathbf{I}_{[0,\infty)}(f(x)) \, \mathrm{d}P_X(x).$$

Damit der Integrand minimal wird muss gelten,

$$\eta(x) > \frac{1}{2} \Rightarrow f(x) \geqslant 0,$$
$$\eta(x) < \frac{1}{2} \Rightarrow f(x) < 0.$$

Daher minimiert f das Risiko genau dann, wenn

$$f(x) \geqslant 0, \qquad \text{auf } [\eta > \frac{1}{2}],$$
$$f(x) < 0, \qquad \text{auf } [\eta < \frac{1}{2}].$$

In diesem Fall ist $f = f^*_{L_{\text{class}},P}$ und folglich

$$\mathcal{R}^*_{L,P} = \int_X \min\{\eta, 1-\eta\} \, \mathrm{d}P_X.$$

Wir wollen noch zeigen, dass

$$\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P} = \int_X |2\eta - 1| \, \mathbf{I}_{(-\infty,0]}((2\eta-1)\operatorname{sign} f(x)) \, \mathrm{d}P_X.$$

» *Beweisskizze.* Dazu betrachtet man einfach die 6 möglichen Kombinationen von

$\eta(x) > \frac{1}{2}$	$\eta(x) < \frac{1}{2}$	$\eta(x) = \frac{1}{2}$
$f(x) \geqslant 0$	$f(x) < 0$.	

z.B. $\eta(x) > \frac{1}{2}$ und $f(x) \geqslant 0$, so ist

$$|2\eta(x) - 1| \, \mathbf{I}_{(-\infty,0]}(\underbrace{(2\eta-1)\operatorname{sign} f(x)}_{>0}) = 0$$

und

$$\eta(x) \underbrace{\mathbf{I}_{(-\infty,0)}(f(x))}_{=0} + (1-\eta(x)) \underbrace{\mathbf{I}_{[0,\infty)}(f(x))}_{=1} - \underbrace{\min\{\eta(x), 1-\eta(x)\}}_{1-\eta(x)} = 0.$$

oder $\eta(x) > \frac{1}{2}$ und $f(x) < 0$, so ist

$$|2\eta(x) - 1| \, \mathbf{I}_{(-\infty,0]}(\underbrace{(2\eta-1)\operatorname{sign} f(x)}_{<0}) = 2\eta(x) - 1$$

und

$$\eta(x)\underbrace{\mathbf{I}_{(-\infty,0)}(f(x))}_{=1}+(1-\eta(x))\underbrace{\mathbf{I}_{[0,\infty)}(f(x))}_{=0}-\underbrace{\min\{\eta(x),1-\eta(x)\}}_{1-\eta(x)}$$
$$=2\eta(x)-1.$$

Die übrigen Fälle folgen analog. «

In unserer Definition von L_{class} haben wir alle Fehler gleich gewichtet. Oft nimmt man eine unterschiedliche Gewichtung vor (Übungsaufgabe). ■

REGRESSION MIT KLEINSTEN FEHLERQUADRATEN Sei $Y \subset \mathbb{R}$ ein Intervall oder ganz $\mathbb{R}$. Unser Ziel ist es f zu finden mit $f(x) \approx y$. Dies wollen wir nun mathematisch fassen.

Dazu definieren wir folgende Verlustfunktion

$$L_{\text{LS}} : Y \times \mathbb{R} \to [0,\infty), \qquad (y,t) \mapsto (y-t)^2.$$

LS steht hier für least squares. Das Risiko ist dann gegeben durch

$$\begin{aligned}\mathcal{R}_{L,P}(f) &= \int_X \int_Y (y-f(x))^2 P(\mathrm{d}y \mid x)\,\mathrm{d}P_X(x)\\ &= \int_X \int_Y y^2 - 2yf(x) + f^2(x) P(\mathrm{d}y \mid x)\,\mathrm{d}P_X(x)\\ &= \int_X \int_Y y^2 P(\mathrm{d}y \mid x) - 2f(x)\int_Y yP(\mathrm{d}y \mid x)\\ &\qquad + f^2(x)\,\mathrm{d}P_X(x).\end{aligned}$$

Setzen wir $t = f(x)$ und $h(t) = -2t\int_Y yP(\mathrm{d}y \mid x) + t^2$, so erhalten wir ein Minimum durch

$$0 = h'(t) = -2\int_Y yP(\mathrm{d}y \mid x) + 2t \Rightarrow t = \int_Y yP(\mathrm{d}y \mid x).$$

Somit ist $f : X \to \mathbb{R}$ eine Bayes-Entscheidungsfunktion, wenn

$$f(x) = \int_Y yP(\mathrm{d}y \mid x) =: \mathbf{E}_P(Y \mid x)\ P\text{-f.s.}$$

f ist in diesem Fall tatsächlich eindeutig.

Das alles funktioniert jedoch nur, wenn $\mathcal{R}^*_{L,S} < \infty$, denn sonst ist jede Funktion eine Bayes-Entscheidungsfunktion.

Setzen wir nun in $\mathcal{R}_{L,S}(\cdot)$ ein, erhalten wir

$$\begin{aligned}\mathcal{R}_{L,S}(f^*_{L,S}) &= \int_X \int_Y y^2 P(\mathrm{d}y \mid x) - 2(\mathbf{E}_P(Y \mid x))^2 + (\mathbf{E}_P(Y \mid x))^2 \,\mathrm{d}P_X(x) \\ &= \int_X \int_Y y^2 P(\mathrm{d}y \mid x) - (\mathbf{E}_P(Y \mid x))^2 \,\mathrm{d}P_X(x) \\ &= \text{mittlere Varianz der Label } y.\end{aligned}$$

Somit ist das Überschussrisiko

$$\begin{aligned}\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P} &= \int_X \int_Y y^2 - 2f(x)y + f^2 - y^2 \\ &\qquad - (\mathbf{E}_P(Y \mid x))^2 P(\mathrm{d}y \mid x)\,\mathrm{d}P_X(x) \\ &= \int_X (f(x) - f^*_{L,P}(x))^2 \,\mathrm{d}P_X(x) = \left\| f - f^*_{L,S} \right\|^2_{L^2(P_X)}.\end{aligned}$$

Minimierung des Risikos ist also äquivalent zur L^2-Approximation von $f^*_{L,S}$. ■

Bis jetzt waren alle konkreten Verlustfunktionen von x unabhängig.

1-2.4 **Definition** *Eine messbare Funktion $L : Y \times \mathbb{R} \to [0, \infty)$ heißt strikt überwachte Verlustfunktion.* ⋊

1-2.5 **Definition** *Eine messbare Funktion $L : X \times \mathbb{R} \to [0, \infty)$ heißt unüberwachte Verlustfunktion.* ⋊

Es liegt die Vermutung nahe, dass unüberwachte Verlustfunktionen nutzlos für überwachtes Lernen ist. Dem ist aber nicht so, wie folgendes Beispiel zeigen wird.

Bsp Medianregression $Y \subset \mathbb{R}$ ist hier typischerweise ein Intervall oder $\mathbb{R}$.

Für $x \in X$ ist der bedingte Median (Median von der bedingten Erwartung in x $P(\cdot \mid x)$) gegeben durch

$$F^*_{1/2} := \left\{ t^* \in \mathbb{R} : \frac{1}{2} \leqslant P((-\infty, t^*] \mid x) \text{ und } \frac{1}{2} \leqslant P([t^*, \infty) \mid x) \right\}.$$

Im Allgemeinen ist dieser nicht eindeutig, wir nehmen nun aber an, dass dem so ist.

Das Ziel ist es nun f^* bei bekanntem $D = ((x_1, y_1), \dots, (x_n, y_n))$ abzuschätzen. Wie genau das gemacht wird behandeln wir später, jetzt soll es nur darum gehen, die Verluste zu bewerten.

Betrachte die Verlustfunktion

$$L : X \times \mathbb{R} \to [0, \infty), \qquad (x, t) \mapsto \left| f^*(x) - t \right|.$$

Das Risiko ist dann

$$\mathcal{R}_{L,P}(f) \coloneqq \int_X \left| f^*(x) - f(x) \right| \, \mathrm{d}P_X(x),$$
$$\mathcal{R}^*_{L,P} = \mathcal{R}_{L,P}(f^*) = 0.$$

Somit ist das Überschussrisiko

$$\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P} = \left\| f^* - f \right\|_{L^1(P_X)}.$$

Die Verlustfunktion ist also tatsächlich unabhängig von den Daten, obwohl wir überwachte Labels haben.

Problematisch hierbei ist, dass L auf dem bedingten Median beruht, den wir nicht kennen und so L überhaupt nicht berechnen können. Man muss also einen anderen Ansatz wählen.

Dies ist auch ein gutes Beispiel dafür, dass ein Unterschied darin besteht, eine Verlustfunktion zu finden, die das Lernziel gut beschreibt, und eine, die in der Praxis gut verwendbar ist. ■

1-3 Universelle Konsistenz

Ziel dieses Abschnittes ist es die Formulierung

$$\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}$$

ist mit hoher Wahrscheinlichkeit klein, mathematisch zu präzisieren.

1-3.1 **Definition** *Sei $L : X \times Y \times \mathbb{R} \to [0, \infty)$ eine Verlustfunktion. Eine Lernmethode $\mathcal{L}$ heißt P-konsistent bezüglich L, falls für alle $\varepsilon > 0$ gilt*

$$\lim_{n\to\infty} P^n \left\{ D \in (X \times Y)^n : \mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P} < \varepsilon \right\} = 1.$$

$\mathcal{L}$ heißt universell konsistent bezüglich L, falls $\mathcal{L}$ P-konsistent bezüglich L für W-Maße P auf $X \times Y$ für die $\mathcal{R}^*_{L,P} < \infty$. ⋈

Die universelle Konsistenz ist einerseits eine schwache Forderung, da sie eine rein asymptotische Aussage macht. Sie ist andererseits aber auch eine sehr starke Forderung wie das folgende Beispiel zeigen wird.

Bsp Wir betrachten Klassifikation auf $X = [0, 1]$. Seien dazu

$$X_1 = C \subset [0, 1], \qquad \text{“Cantor Menge”},$$
$$X_{-1} = (\mathbb{Q} \cap [0, 1]) \setminus C$$

Nun existiert ein W-Maß μ auf X_1 (z.B. das Hausdorffmaß) und ein W-Maß ν auf X_{-1} (z.B. Zählmaß+Dichte), so dass für $P_X := \frac{1}{2}(\mu + \nu)$ gilt

$$P(Y = 1 \mid x) = \begin{cases} 0.51, & x \in X_1, \\ 0.49 & x \in X_{-1}. \end{cases}$$ ■

1-3-1 Plug-in-rules für Klassifikation

Wir wollen nun für ein spezielles Verfahren nachweisen, dass es universell konsistent ist.

Sei $Y = \{-1, 1\}$ und $\eta(x) = P(Y = 1 \mid x)$ die Wahrscheinlichkeit dafür, dass x ein positives Label besitzt. Gegeben sei eine Lernmethode $\mathcal{L}$, die die Entscheidungsfunktion f_D konstruiert und $f_D \approx \eta$.

Wir haben bereits gezeigt, dass für die Bayes-Klassifikationsentscheidungsfunktion gilt

$$f^*_{L_{\text{class}},P}(x) = \begin{cases} 1, & \eta(x) > \frac{1}{2}, \\ -1, & \eta(x) < \frac{1}{2}. \end{cases}$$

1-3.2 **Definition** *Das zu $\mathcal{L}$ gehörige plug-in-Verfahren ist durch*

$$\hat{f}_D(x) := \begin{cases} 1, & f_D(x) > \frac{1}{2}, \\ -1, & f_D(x) < \frac{1}{2}, \end{cases}$$

gegeben. ⋈

Wie gut das Verfahren arbeitet, beschreibt das folgende

1-3.3 **Lemma** *Sei P ein W-Maß auf $X \times Y$, $Y = \{-1, 1\}$, $\eta(x) = P(Y = 1 \mid x)$ und $x \in X$. Für $h : X \to \mathbb{R}$ gilt dann*

$$\mathcal{R}_{L_{\text{class}}}(2h - 1) - \mathcal{R}^*_{L_{\text{class}},P} \leqslant 2 \int_X |\eta - h| \, \mathrm{d}P_X. \quad ⋈$$

Interpretation. Ist h eine "gute" Schätzung von η, dann ist $2h - 1$ (bzw. $\operatorname{sign}(2h - 1)$) eine "gute" Klassifikationsentscheidungsfunktion.

» *Beweis des Lemmas.* Sei $f := 2h - 1$. Wir haben bereits gezeigt, dass

$$\mathcal{R}_{L_{\text{class}}}(f) - \mathcal{R}^*_{L_{\text{class}},P} = \int_X \underbrace{|2\eta - 1| \, \mathbf{I}_{(-\infty,0]}((2\eta - 1) \operatorname{sign} f)}_{(1)} \, \mathrm{d}P_X.$$

Falls $(2\eta - 1) \operatorname{sign} f > 0$, so ist der Integrand Null und daher

$$(1) \leqslant 2 \, |\eta - h| \, .$$

Falls $(2\eta - 1) \operatorname{sign} f < 0$, so wertet die Indikatorfunktion zu 1 aus. Ist nun $\eta > \frac{1}{2}$, so ist $f < 0$ und somit $-f - 1 > -1$, also

$$\begin{aligned} |2\eta - 1| &= 2\eta - 1 \leqslant 2\eta - 1 - f = 2\eta - 1 - (2h - 1) = 2(\eta - h) \\ &\leqslant 2 \, |\eta - h| \, . \end{aligned}$$

Ist dagegen $\eta < \frac{1}{2}$, so ist $f \geqslant 0$ und somit $f + 1 \geqslant 1$, also

$$|2\eta - 1| = 1 - 2\eta \leqslant f + 1 - 2\eta = 2 \, |\eta - h| \, .$$

Ist $\eta = \frac{1}{2}$, so ist (1)= 0. «

Betrachten wir die Histogrammregel mit $Y = \{-1, 1\}$, $X = \mathbb{R}^d$ und

$$D = ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n.$$

Sei nun $\mathcal{A} = (A_i)_{i \geqslant 1}$ mit A_i messbar eine Partition von X, d.h.

$$X = \bigcup_{i \geqslant 1} A_i, \qquad A_i \cap A_j = \emptyset, \quad i \neq j.$$

Für jedes $x \in X$ existiert ein eindeutiger Index $j \geqslant 1$, so dass $x \in A_j$. Setzen wir

$$N_D(x) := \sum_{i=1}^{n} \mathbf{I}_{A_i(x)}(x_i) = \#\left\{\text{Samples } (x_i, y_i) \text{ für die } x_i \in A(x)\right\},$$

so erhalten wir

$$\hat{n}_D(x) = \begin{cases} \frac{1}{N_D(x)} \sum_{\{i : y_i = 1\}} \mathbf{I}_{A(x)}(x_i), & N_D(x) \neq 0, \\ 0, & \text{sonst.} \end{cases}$$

Damit können wir nun eine plug-in-Regel definieren.

$$f_D^{\mathcal{A}}(x) := \begin{cases} -1, & \hat{n}_D(x) < \frac{1}{2}, \\ 1 & \hat{n}_D(x) \geqslant \frac{1}{2}. \end{cases}$$

Für $A \subset \mathbb{R}^d$ ist der Durchmesser definiert als

$$\operatorname{diam} A := \sup_{x, x' \in A} \left\| x - x' \right\|.$$

Für $n \geqslant 1$ betrachten wir nun eine weitere Partition $\mathcal{A}_n = (A_{n,j})_{j \geqslant 1}$. Das Lernverfahren $\mathcal{L}$ sei nun durch

$$(X \times Y) \ni D \mapsto f_D^{\mathcal{A}_n} =: f_D \tag{*}$$

gegeben.

1-3.4 **Satz** *Sei P ein W-Maß auf $X \times Y$, $Y = \{-1, 1\}$ und $X = \mathbb{R}^d$. Ferner gelten*

1) $\lim_{n \to \infty} \sup_j \operatorname{diam} A_{n,j} = 0$.

2) $\lim_{n\to\infty} P^n \otimes P_X(\{(D,x) : N_D(x) \leqslant l\}) = 0$ *f.s. für* $l \geqslant 1$.

Dann ist die Lernmethode (*) P-konsistent bezüglich L_{class}. ⋊

» Nach Lemma 1-3.3 gilt für $h = \hat{\eta}_D$ und $f_D = \operatorname{sign}(2\hat{\eta}_D - 1)$,

$$\mathcal{R}_{L_{\text{class}}}(f_D) - \mathcal{R}^*_{L_{\text{class}},P} \leqslant 2 \int_X |\eta - \hat{\eta}_D| \, \mathrm{d}P_X.$$

Definiere nun $\overline{\eta} : X \to [0,1]$,

$$\overline{\eta}(x) = \begin{cases} \frac{1}{P_X(A(x))} \int_{A(x)} \eta \, \mathrm{d}P_X, & P_X(A(x)) > 0, \\ 0, & \text{sonst}, \end{cases}$$

als "mittleres η auf $A(x)$". $\overline{\eta}$ ist somit auf jedem $A(x)$ konstant. Dann gilt

$$\int |\eta - \hat{\eta}_D| \, \mathrm{d}P_X \leqslant \underbrace{\int |\eta - \overline{\eta}| \, \mathrm{d}P_X}_{(1)} + \underbrace{\int |\overline{\eta} - \hat{\eta}_D| \, \mathrm{d}P_X}_{(2)}.$$

Wir betrachten zunächst den von D abhängigen Term (2),

$$\begin{aligned} &\int_{X\times Y} \int_X |\overline{\eta}(x) - \hat{\eta}_D(x)| \, \mathrm{d}P_X(x) \, \mathrm{d}P^n(D) \\ &\quad = \int_X \int_{X\times Y} |\overline{\eta}(x) - \hat{\eta}_D(x)| \, \mathrm{d}P^n(D) \, \mathrm{d}P_X(x) \\ &\quad = \int_X \sum_{k=0}^{n} \int_{\{D : N_D(x) = k\}} |\overline{\eta}(x) - \hat{\eta}_D(x)| \, \mathrm{d}P^n(D) \, \mathrm{d}P_X(x) \end{aligned}$$

Sei $N_D(x) = k$, dann existieren eindeutige $x_{i_1}, \ldots, x_{i_k} \in A(x)$. Die zugehörigen $y_{i_1}, \ldots, y_{i_k}$ sind unabhängige Zufallsvariablen, wovon jede einzelne mit der Wahrscheinlichkeit $\overline{\eta}(x)$ positiv ist.

Somit ist $\sum_{\{j : Y_{i_j} = 1\}} y_{i_j}$ binomialverteilt, genauer $B(k, \overline{\eta}(x))$-verteilt und

$$\sum_{\{i : y_i = 1\}} \mathbf{I}_{A(x)}(x_i) = N_D(x)\hat{\eta}_D(x) = k \cdot \hat{\eta}_D(x).$$

Sei $\xi : \Omega \to \mathbb{R}$ eine $B(k, \overline{\eta}(x))$-verteilte Zufallsvariable. Dann gilt $\xi/k \sim \hat{\eta}_D(x)$ für $k \geqslant 1$ und folglich

$$\begin{aligned}
&\frac{1}{P^n(\{D : N_D(x) = k\})} \int_{\{D : N_D(x) = k\}} |\overline{\eta}(x) - \hat{\eta}_D(x)| \, \mathrm{d}P^n(D) \\
&\quad = \mathbf{E}\left|\overline{\eta}(x) - \frac{1}{k}\xi\right| \\
&\quad \leqslant \frac{1}{k}\left(\mathbf{E}\left((k\overline{\eta}(x) - \xi)^2\right)\right)^{1/2} \\
&\quad = \frac{1}{k}\left(\mathbf{E}(\mathbf{E}\xi - \xi)^2\right)^{1/2},
\end{aligned}$$

nach Hölders Ungleichung. Die Varianz einer $B(k, \overline{\eta}(x))$-verteilten Zufallsvariablen ist $k\overline{\eta}(x)(1 - \overline{\eta}(x)) \leqslant k/4$ und damit erhalten wir für obigen Ausdruck die Abschätzung

$$\frac{1}{k}\left(\mathbf{E}(\mathbf{E}\xi - \xi)^2\right)^{1/2} = \frac{1}{k}\left(k\overline{\eta}(x)(1 - \overline{\eta}(x))\right)^{1/2} \leqslant \frac{1}{2\sqrt{k}}.$$

Für $k = 0$ gilt auf $\{D : N_D(x) = k\}$ nach Definition $\hat{\eta} \equiv 0$ sowie $|\overline{\eta}(x)| \leqslant 1$. Deshalb ist

$$\int_{\{D : N_D(x) = k\}} |\overline{\eta}(x) - \hat{\eta}_D(x)| \, \mathrm{d}P^n(x) \leqslant P^n(\{D : N_D(x) = k\}).$$

Damit folgt

$$\begin{aligned}
&\int_{(X\times Y)^n} \int_X |\overline{\eta}(x) - \hat{\eta}_D(x)| \, \mathrm{d}P_X(x) \, \mathrm{d}P^n(D) \\
&\quad \leqslant \int_X \sum_{k=1}^n \frac{1}{2\sqrt{k}} P^n(\{D : N_D(x) = k\}) + P^n(\{D : N_D(x) = 0\}) \, \mathrm{d}P_X(x) \\
&\quad \leqslant \int_X \sum_{k=1}^n \int_{\{D : N_D(x) = k\}} \frac{1}{2\sqrt{k}} \, \mathrm{d}P^n(D) + P^n(\{D : N_D(x) = 0\}) \, \mathrm{d}P_X(x) \\
&\quad \leqslant \frac{1}{2} \int_X \int_{\{D : N_D(x) > 0\}} \frac{\mathrm{d}P^n(D)}{\sqrt{N_D(x)}} \, \mathrm{d}P_X(x) + P^n \otimes P(\{(D, x) : N_D(x) = 0\})
\end{aligned} \tag{**}$$

Für $l \geqslant 1$ gilt dann

$$\begin{aligned}
&\int_X \int_{\{D : N_D(x) > 0\}} \frac{\mathrm{d}P^n(D)}{\sqrt{N_D(x)}}\, \mathrm{d}P_X(x) \\
&\quad \leqslant \int_X \int_{\{D : 0 < N_D(x) \leqslant l\}} \underbrace{\frac{\mathrm{d}P^n(D)}{\sqrt{N_D(x)}}}_{\leqslant 1}\, \mathrm{d}P_X(x) + \int_X \int_{\{D : l < N_D(x)\}} \underbrace{\frac{\mathrm{d}P^n(D)}{\sqrt{N_D(x)}}}_{\leqslant \sqrt{l}^{-1}}\, \mathrm{d}P_X(x) \\
&\quad \leqslant P^n \otimes P(\{(D,x) : N_D(x) \leqslant l\}) + \frac{1}{\sqrt{l}}. \qquad (***)
\end{aligned}$$

Insgesamt gilt also

$$(**) \leqslant \frac{3}{2} P^n \otimes P(\{(D,x) : N_D(x) \leqslant l\}) + \frac{1}{2\sqrt{l}}.$$

Sei nun $\varepsilon > 0$ und $l > \varepsilon^{-2}$, so ist $\frac{1}{\sqrt{l}} < \varepsilon$ und nach Voraussetzung (ii) gilt für hinreichend große n,

$$P^n \otimes P(\{(D,x) : N_D(x) \leqslant l\}) \leqslant \varepsilon$$

und somit $(***) \leqslant \frac{5}{2}\varepsilon$. «

1-3.5 **Satz** *Sei μ ein endliches Borelmaß auf $\mathbb{R}^d$ und*

$$f : \mathbb{R}^d \to \mathbb{R}, \qquad \mu\text{-integrierbar.}$$

Dann gibt es für jedes $\varepsilon > 0$ eine gleichmäßig stetige Abbildung $h : \mathbb{R}^d \to \mathbb{R}$, die μ-integrierbar ist und für die gilt

$$\int |f - h| \, \mathrm{d}\mu \leqslant \varepsilon. \quad \rtimes$$

» Der Beweis findet sich in [7], oder ergibt sich direkt, wenn das Lebesgue-Integral nach der Methode von Daniel Stone eingeführt wird. «

P_X ist als W-Maß insbesondere endlich und $\eta = P(Y = 1 \mid \cdot) \in [0,1]$ ist beschränkt und damit P_X-integrierbar. Für $\varepsilon > 0$ gibt es nun eine gleichmäßig stetige Abbildung $\eta_\varepsilon : \mathbb{R}^d \to \mathbb{R}$, die ebenfalls P_X-integrierbar ist und

$$\int |\eta - \eta_\varepsilon| \, \mathrm{d}P_X < \varepsilon$$

erfüllt. Setzen wir also

$$\overline{\eta}_\varepsilon(x) := \begin{cases} \frac{1}{P_X(A(x))} \int_{A(x)} \eta_\varepsilon \, \mathrm{d}P_X, & P_X(A(x)) > 0, \\ 0, & \text{sonst}, \end{cases}$$

so gilt

$$\int |\eta - \overline{\eta}| \, \mathrm{d}P_X \leqslant \underbrace{\int |\eta - \eta_\varepsilon| \, \mathrm{d}P_X}_{\text{(I)}} + \underbrace{\int \left|\eta_\varepsilon - \overline{\eta}_\varepsilon\right| \, \mathrm{d}P_X}_{\text{(II)}} + \underbrace{\int \left|\overline{\eta}_\varepsilon - \overline{\eta}\right| \, \mathrm{d}P_X}_{\text{(III)}}$$

(I) ist nach Konstruktion $< \varepsilon$. Analog folgt

$$\begin{aligned} \text{(III)} &= \int \left|\overline{\eta}_\varepsilon - \overline{\eta}\right| \, \mathrm{d}P_X \\ &\overset{\frac{0}{0}:=0}{=} \int_X \frac{1}{P_X(A(x))} \int_{A(x)} \left|\eta_\varepsilon(x') - \eta(x')\right| \, \mathrm{d}P_X(x') \, \mathrm{d}P_X(x) \\ &= \int_X \int_X \frac{1}{P_X(A(x))} \mathbf{I}_{A(x)}(x) \left|\eta_\varepsilon(x') - \eta(x')\right| \, \mathrm{d}P_X(x) \, \mathrm{d}P_X(x') \\ &= \int_X \left|\eta_\varepsilon(x') - \eta(x')\right| \, \mathrm{d}P_X(x') \leqslant \varepsilon. \end{aligned}$$

Da η_ε gleichmäßig stetig, gibt es ein $\delta > 0$, so dass

$$\left\|x - x'\right\| < \delta \Rightarrow \left|\eta_\varepsilon(x) - \eta_\varepsilon(x')\right| < \varepsilon$$

und nach Voraussetzung (i) des Satzes gilt für große Zahlen n,

$$\sup_{j \geqslant 1} \operatorname{diam} A_{n,j} \leqslant \delta.$$

Außerdem erhalten wir

$$\begin{aligned} \text{(II)} &= \int_X \left|\eta_\varepsilon - \overline{\eta}_\varepsilon\right| \, \mathrm{d}P_X \\ &= \int_X \left| \eta_\varepsilon(x) - \frac{1}{P_X(A(x))} \int_{A(x)} \eta_\varepsilon(x') \, \mathrm{d}P_X(x') \right| \, \mathrm{d}P_X(x) \\ &\leqslant \int_X \frac{1}{P_X(A(x))} \int_{A(x)} \underbrace{\left|\eta_\varepsilon(x) - \eta_\varepsilon(x')\right|}_{<\varepsilon} \, \mathrm{d}P_X(x') \, \mathrm{d}P_X(x) \leqslant \varepsilon. \end{aligned}$$

Insgesamt haben wir somit gezeigt,

$$\forall\ \varepsilon > 0\ \exists\ N_0 \in \mathbb{N}\ \forall\ n \geqslant N : \int |\eta - \overline{\eta}|\ \mathrm{d}P_X < \varepsilon.$$

Damit haben wir gezeigt, dass

$$\mathbf{E}_{D\sim P^n} \int |\eta - \hat{\eta}_D|\ \mathrm{d}P_X \to 0, \qquad n \to \infty.$$

Jetzt müssen wir uns noch davon überzeugen, dass dies auch für die universelle Konsistenz genügt.

Schreibe $h_D \coloneqq \mathcal{R}_{L_{\text{class}},P}(f_D) - \mathcal{R}^*_{L_{\text{class}},P} \geqslant 0$, so gilt nach Lemma 1-3.3 und dem soeben gezeigten, dass

$$\mathbf{E}_{D\sim P^n} h_D \to 0, \qquad n \to \infty.$$

Ferner gilt

$$\mathbf{E}_{D\sim P^n} h_D = \mathbf{E}_{D\sim P^n} |h_D| = \int_{[h_D<\varepsilon]} |h_D|\ \mathrm{d}P^n(D) + \underbrace{\int_{[h_D\geqslant\varepsilon]} |h_D|\ \mathrm{d}P^n(D)}_{\geqslant\varepsilon P^n(\{D\,:\,|h_D|\geqslant\varepsilon\})}.$$

Das heißt

$$\lim_{n\to\infty} P^n\left(\left\{D\,:\,\mathcal{R}_{L_{\text{class}},P}(f_D) - \mathcal{R}^*_{L_{\text{class}},P} \geqslant \varepsilon\right\}\right) = 0,$$

da $\mathbf{E}_{D\sim P^n} h_D \to 0$.

Wir wollen jetzt zeigen, dass es Lernmethoden gibt, die universell konsistent sind.

1-3.6 **Universelle Konsistenz der Histogrammregel** *Sei $\mathcal{A}_n \coloneqq (A_{n,j})_{j\geqslant 1}$ eine Partition von $\mathbb{R}^d$, so dass jedes $A_{n,j}$ ein Hyperwürfel der Kantenlänge h_n ist, d.h. wir haben zu $x_{n,j} \in \mathbb{R}^d$,*

$$A_{n,j} = x_{n,j} + [0, h_n]^d.$$

Es gelten

1) $\lim_{n\to\infty} h_n = 0$,

2) $\lim_{n\to\infty} n \cdot h_n = \infty$.

Dann ist die zugehörige Histogrammregel universell konsistent. ⋊

» Wir weisen die Voraussetzungen von Satz 1-3.4 nach.
"1)": $\lim_{n\to\infty} \sup_{j\geqslant 1} \operatorname{diam} A_{n,j} = 0$ ist erfüllt.
"2)": Dazu führen wir $S > 0$ ein. Dann gibt es höchstens $c_1 + c_2 h_n^{-d}$ Hyperwürfel $A_{n,j}$ mit $A_{n,j} \cap B_S(0) \neq \emptyset$ und somit

$$\begin{aligned} &P^n \otimes P_X(\{(D,x) : N_D(x) \leqslant l\}) \leqslant P^n \otimes P_X(\{(D,x) : x \notin B_S(0)\}) \\ &\leqslant \sum_{j: A_{n,j}\cap B_S(0)\neq\emptyset} P^n \otimes P_X\left(\{(D,x) : N_D(x) \leqslant l\} \cap (X\times Y)^n \times A_{n,j}\right) \end{aligned}$$

Dann ist $P^n \otimes P_X(\{(D,x) : x \notin B_S(0)\}) = P_X(\mathbb{R}^d \setminus B_S(0))$ also

$$\begin{aligned} &\sum_{j: A_{n,j}\cap B_S(0)\neq\emptyset} P^n \otimes P_X\left(\left\{(D,x) : N_D(x) \leqslant l,\ x \in A_{n,j}\right\}\right) \\ &\leqslant \sum_{\substack{j: A_{n,j}\cap B_S(0)\neq\emptyset \\ P_X(A_{n,j}) \leqslant \frac{2l}{n}}} P^n \otimes P_X\left(\left\{(D,x) : N_D(x) \leqslant l,\ x \in A_{n,j}\right\}\right) \qquad (1) \\ &+ \sum_{\substack{j: A_{n,j}\cap B_S(0)\neq\emptyset \\ P_X(A_{n,j}) > \frac{2l}{n}}} \left(\left\{(D,x) : N_D(x) \leqslant l,\ x \in A_{n,j}\right\}\right) \qquad (2) \end{aligned}$$

Zu (1): $\left\{(D,x) : N_D(x) \leqslant l,\ x \in A_{n,j}\right\} \subset (X\times Y)^n \times A_{n,j}$ und daher

$$\begin{aligned} &P^n \otimes P_X\left(\left\{(D,x) : N_D(x) \leqslant l,\ x \in A_{n,j}\right\}\right) \leqslant P^n \otimes P_X((X\times Y)^n \times A_{n,j}) \\ &= P_X(A_{n,j}) \end{aligned}$$

Folglich ist

$$(1) \leqslant \sum_{\substack{j: A_{n,j}\cap B_S(0)\neq\emptyset \\ P_X(A_{n,j}) \leqslant \frac{2l}{n}}} P_X(A_{n,j}) \leqslant \frac{2l}{n}(c_1 + c_2 h_n^{-d}) \to 0,$$

nach Voraussetzung 2).

Zu (2): Setze $\mu_{D_X} := n^{-1} \sum_{i=1}^{n} \delta_{\{x_i\}}$, wobei

$$\delta_{\{x_i\}}(A) = \begin{cases} 1, & x_i \in A, \\ 0, & \text{sonst.} \end{cases}$$

μ_{D_X} heißt empirisches Maß bezüglich des Datensatzes $D_X = (x_1, \ldots, x_n)$.

Für $x \in A_{n,j}$ gilt

$$N_D(x) = \sum_{i=1}^{n} \mathbf{I}_{A(x)}(x_i) = n \cdot \mu_{D_X}(A(x)).$$

Daher folgt

$$\begin{aligned} &P^n \otimes P_X \left(\left\{ (D, x) : N_D(x) \leqslant l,\ x \in A_{n,j} \right\} \right) \\ &\quad \leqslant P_X(A_{n,j}) P^n \left(\left\{ D : n \cdot \mu_{D_X}(A_{n,j}) \leqslant l \right\} \right) \end{aligned}$$

Es folgt wegen $P_X(A_{n,j}) > \frac{2l}{n}$

$$\begin{aligned} &P^n \left(\left\{ D : \mu_{D_X}(A_{n,j}) \leqslant \frac{l}{n} \right\} \right) \\ &\quad = P^n \left(\left\{ D : \mu_{D_X}(A_{n,j}) - P_X(A_{n,j}) \leqslant \frac{l}{n} - P_X(A_{n,j}) \right\} \right) \\ &\quad \leqslant P^n \left(\left\{ D : \mu_{D_X}(A_{n,j}) - P_X(A_{n,j}) \leqslant -\frac{1}{2} P_X(A_{n,j}) \right\} \right). \end{aligned}$$

Zur Fortsetzung des Beweises benötigen wir folgendes Lemma. «

1-3.7 **Lemma** *Sei $(\Omega, \mathcal{A}, P)$ ein W-Raum und $h : \Omega \to \mathbb{R}$ messbar mit $\int h^2 \, \mathrm{d}P < \infty$. Dann gilt für alle $n \geqslant 1$ und $t > 0$,*

$$P^n \left(\left\{ \omega \in \Omega : \frac{1}{n} \sum_{i=1}^{n} h(\omega_i) - \mathbf{E}_P h \geqslant t \right\} \right) \leqslant \frac{\mathbf{E}_P h^2}{t^2 \cdot n}. \quad \rtimes$$

» Die Tschebyscheff-Ungleichung besagt für quadratintegrierbares f,

$$P\left[\left| f \right| \geqslant t \right] \leqslant \frac{\mathbf{E}_P f^2}{t^2}.$$

Wenden wir dies auf $f = \frac{1}{n}\sum_{i=1}^n h \circ \pi_i - \mathbf{E}_P h$, wobei

$$\pi_i : \Omega^n \to \Omega$$

die i-te Projektion.

$$P^n\left(\left\{\omega \in \Omega^n : \frac{1}{n}\sum_{i=1}^n h(\omega_i) - \mathbf{E}_P h \geqslant t\right\}\right) \leqslant \frac{\mathbf{E}_{P^n}\left(\frac{1}{n}\sum_{i=1}^n h \circ \pi_i - \mathbf{E}_P h\right)^2}{t^2}$$

und mit $\eta_i \coloneqq h \circ \pi_i - \mathbf{E}_P h$ gilt

$$\begin{aligned}\mathbf{E}_{P^n}\left(\frac{1}{n}\sum_{i=1}^n (h \circ \pi_i - \mathbf{E}_P h)\right)^2 &= \frac{1}{n^2}\mathbf{E}_{P^n}\left(\sum_{i=1}^n \eta_i\right)^2 \\ &= \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}_{P^n} h^2 \circ \pi_i + \frac{1}{n^2}\sum_{i\neq j}\underbrace{\mathbf{E}_{P^n}\eta_i\eta_j}_{=0} \\ &= \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}_{P^n} h^2 \circ \pi_i = \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}_P h^2 = \frac{1}{n}\mathbf{E}_P h^2. \quad «\end{aligned}$$

» *Fortsetzung des Beweises von Satz 1-3.6.* Setze nun in Lemma 1-3.5, $h = \mathbf{I}_{A_{n,j}}$ und $t = \frac{1}{2}P_X(A_{n,j})$, so gilt

$$\begin{aligned}P^n\left(\left\{D : \mu_{D_X}\right\}(A_{n,j}) - P(A_{n,j}) \leqslant -\frac{1}{2}P_X(A_{n,j})\right) &\leqslant \frac{\mathbf{E}_P \mathbf{I}^2_{A_{n,j}}}{4P_X^2(A_{n,j})n} \\ &= \frac{1}{4n \cdot P_X(A_{n,j})}.\end{aligned}$$

Somit gilt schließlich

$$\begin{aligned}(2) &= \sum_{\substack{j: A_{n,j}\cap B_S(0)\neq\varnothing \\ P_X(A_{n,j})>\frac{2l}{n}}} \left(\left\{(D,x) : N_D(x) \leqslant l,\ x \in A_{n,j}\right\}\right) \\ &\leqslant \sum_{\substack{j: A_{n,j}\cap B_S(0)\neq\varnothing \\ P_X(A_{n,j})>\frac{2l}{n}}} \frac{1}{4n} \leqslant \frac{c_1 + c_2 h_n^{-d}}{4n} \xrightarrow{n\to\infty} 0.\end{aligned}$$

Zum Abschluss des Beweises, müssen wir noch S passend wählen. Sei also $\varepsilon > 0$, dann existiert ein $\delta > 0$, so dass $P_X(\mathbb{R}^d \setminus B_S(0)) \leqslant \varepsilon$ und für hinreichend große n gilt ferner

$$\frac{2l}{n}\left(c_1 + c_2 h_n^{-d}\right) \leqslant \varepsilon,$$
$$\left(c_1 + c_2 h_n^{-d}\right)\frac{1}{4n} \leqslant \varepsilon,$$

und letztlich

$$P^n \otimes P_X\left(\{(D,x) : N_D(x) \leqslant \varepsilon\}\right) \leqslant 3\varepsilon. \quad «$$

Schlussbemerkungen A. Es gibt universell konsistente Klassifizierungsmethoden. Stone hat 1977 gezeigt, dass das Nearest-Neighbour-Verfahren universell konsistent ist, falls $K_n \to \infty$ und $\frac{K_n}{n} \to 0$.

B. Gibt es ein $S > 0$, so dass $P_X(\mathbb{R}^d \setminus B_S(0)) = 0$, dann ist die Konvergenz von

$$P^n \otimes P_X\left(\{(D,x) : N_D(x) \leqslant l\}\right) \tag{*}$$

für $n \to \infty$ von P unabhängig. Der wesentliche Grund dafür ist, dass Lemma **??** von P unabhängig ist.

C. Die Konvergenzgeschwindigkeit von

$$\int_{(X\times Y)^n}\int_X |\overline{\eta} - \hat{\eta}_D| \,\mathrm{d}P_X \,\mathrm{d}P^n \overset{n\to\infty}{\to} 0$$

hängt lediglich von (*) ab und ist daher von P unabhängig.

D. In Satz 1-3.4 benötigen wir eine gleichmäßig stetige Approximation von η und $\overline{\eta}$ und können daher keine Aussage zur Konvergenzrate von

$$\mathcal{R}_{L_{\text{class}},P}(f_D) - \mathcal{R}^*_{L_{\text{class}},P} \leqslant \int |\eta - \overline{\eta}| \,\mathrm{d}P_X \to 0$$

machen.

E. Ist η lipschitz stetig mit Konstante $\leqslant 1$, so können wir eine Konvergenzrate von

$$\mathcal{R}_{L_{\text{class}},P}(f_D) - \mathcal{R}^*_{L_{\text{class}},P} \leqslant \int |\eta - \overline{\eta}| \,\mathrm{d}P_X \to 0$$

berechnen. Diese ist unabhängig von P, in dem Sinn, dass man nur die P betrachtet für die η lipschitz ist. ⊸

1-4 Lernraten

Ziel dieses Abschnitts ist es, die Konvergenzraten von

$$\mathcal{R}_{L_{\text{class}},P}(f_D) - \mathcal{R}^*_{L_{\text{class}},P} \to 0$$

zu untersuchen.

1-4.1 **Definition** *Sei $L : X \times Y \times \mathbb{R} \to [0,\infty)$ eine Verlustfunktion, P ein W-Maß auf $X \times Y$, $\mathcal{L}$ eine Lernmethode und (a_n) eine Folge echt positiver Zahlen mit $a_n \downarrow 0$.*

Wir sagen, dass $\mathcal{L}$ im Mittel mit Rate (a_n) lernt, falls es ein $c > 0$ gibt, so dass

$$\mathbf{E}_{D\sim P^n}\left(\mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P}\right) \leqslant c a_n, \qquad n \geqslant 1.$$

Ferner sagen wir, dass $\mathcal{L}$ in Verteilung mit Rate (a_n) lernt, falls es eine Familie $(c_\tau)_\tau \subset [1,\infty)$ und $\tau \in [0,1]$ gibt, so dass gilt

$$P^n\left(\left\{D : \mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P} \leqslant c_\tau a_n\right\}\right) \geqslant 1-\tau, \qquad n \geqslant 1,\ \tau \in [0,1]. \quad \rtimes$$

Bevor wir klären, ob es überhaupt Lernraten gibt, die für alle P gelten, wollen wir erstmal beide Begriffe vergleichen.

Bemerkungen. A. Ist $\mathcal{L}$ konsistent, so kann man zeigen, dass $\mathcal{L}$ im Mittel mit einer von P abhängigen Rate im Mittel lernt.

B. Im Beweis von Satz 1-3.6 haben wir benutzt,

$$\mathbf{E}_{D\sim P^n}\left(\mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P}\right) \geqslant \varepsilon P^n\left(\left\{D : \mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P} > \varepsilon\right\}\right)$$

Ist nun $\mathbf{E}_{D\sim P^n}\left(\mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P}\right) \leqslant c a_n$, dann ist

$$P^n\left(\left\{D : \mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P} > \varepsilon\right\}\right) \geqslant \frac{c a_n}{\varepsilon}.$$

D.h. für $\tau = \frac{c a_n}{\varepsilon} \Leftrightarrow \varepsilon = \frac{c}{\tau} a_n$ und $c_\tau := c\tau^{-1}$ ist

$$P^n\left(\left\{D : \mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P} \leqslant c_\tau a_n\right\}\right) \leqslant 1-\tau.$$

Lernraten im Mittel implizieren somit Lernraten in Verteilung.

c. Später werden wir häufig Lernraten in Verteilung beweisen, die dann aufgrund der Struktur der c_τ auch im Mittel gelten. ⊸

Offene Frage *Gibt es Verfahren $\mathcal{L}$, die mit Rate (a_n) für alle P lernen?* ⋈

Wir werden im Folgenden sehen, dass dies "normalerweise" nicht so ist.

1-4.2 **Definition** *Sei $(X, \mathcal{A}, \mu)$ ein endlicher Maßraum. Ein $A \in \mathcal{A}$ heißt Atom, wenn*

$$\forall\, B \subset A, \quad B \in \mathcal{A} \text{ gilt entweder } \mu(B) = 0 \text{ oder } \mu(A \setminus B) = 0.$$

$(X, \mathcal{A}, \mu)$ heißt atomfrei, falls kein $A \in \mathcal{A}$ ein Atom ist. ⋈

Bsp a) Versehen wir die natürlichen Zahlen versehen mit dem Zählmaß ν mit $\nu(\{n\}) = 1$, so ist jede einelementige Menge $\{n\}$ ein Atom.

b) Das Intervall $[0, 1]$ versehen mit dem Lebesgue-Maß ist atomfrei. ■

1-4.3 **Satz von Lyapunov** *Ist $(X, \mathcal{A}, \mu)$ ein endlicher atomfreier Maßraum, so ist*

$$\{\mu(A) : A \in \mathcal{A}\} = [0, \mu(X)]. \quad ⋈$$

» Ein Beweis findet sich in [8]. «

1-4.4 **No-free-lunch Theorem (Devroye 1982)** *Sei $(a_n) \subset [0, 1/32]$ eine fallende Nullfolge, $(X, \mathcal{A}, \mu)$ ein atomfreier W-Raum und $Y := \{-1, 1\}$. Dann gibt es zu jeder Lernmethode $\mathcal{L}$ ein W-Maß P auf $X \times Y$ mit*

1) $P_X = \mu$.

2) $\mathcal{R}^*_{L_{\text{class}},P} = 0$.

3) $\mathbf{E}_{D\sim P^n} \mathcal{R}_{L_{\text{class}},P}(f_D) - \mathcal{R}^*_{L_{\text{class}},P} \geqslant a_n$ *für* $n \geqslant 1$. ⋈

Bemerkungen. A. Das NFL-Theorem besagt, dass es im Allgemeinen keine universellen mittleren Raten gibt. Dies gilt auch für Raten in Verteilung.

B. Das NFL-Theorem gilt für *praktisch jede* Verlustfunktion.

c. Für $X = \mathbb{N}$ gilt das NFL-Theorem, falls $\mu = P_X$ nicht vorher festgelegt wird.

d. Für *endliche* X ist das NFL-Theorem falsch. ⊸

1-4.5 **Lemma** *Sei $(a_n) \subset (0, 1/16]$ mit $a_n \downarrow 0$, dann existiert eine Folge $(p_n) \subset (0,1)$ mit $p_1 \geqslant p_2 \geqslant \ldots$ und $\sum_{n=1}^{\infty} p_n = 1$ und*

$$\sum_{i=n+1}^{\infty} p_i \geqslant \max\left\{8a_n, 32np_{n+1}\right\}. \quad \bowtie$$

» Da $p_n \geqslant p_{n+1}$ gezeigt wird, genügt es zu zeigen, dass

$$\sum_{i=n+1}^{\infty} p_i \geqslant \max\left\{8a_n, 32np_n\right\}. \quad \bowtie$$

Für $l \leqslant m$ definiere

$$H(m,l) \coloneqq \sum_{i=l}^{m-1} \frac{1}{i}.$$

Sei nun $n_1 \coloneqq 1$, so gilt

$$8a_{n_1} = 8a_1 \leqslant 8\frac{1}{16} \leqslant 2^{-1}.$$

Ferner gibt es ein $n_2 > n_1$ mit $H(n_2, n_1) \geqslant 32$, da $\sum_{i \geqslant 1} i^{-1} = \infty$. Außerdem ist $H(\cdot, n_1)$ wachsend und (a_n) fallend nach Voraussetzung, also können wir $8a_{n_2} \leqslant 2^{-2}$ für n_2 hinreichend groß garantieren. Sukzessive finden wir n_k für $k \geqslant 3$, so dass

$$n_k > n_{k-1}, \qquad H(n_k, n_{k-1}) \geqslant H(n_{k-1}, n_{k-2}), \qquad 8a_{n_k} \leqslant 2^{-k}.$$

Definiere nun

$$c_k \coloneqq \frac{32}{2^k H(n_{k+1}, n_k)}, \qquad k \geqslant 1,$$

so ist c_k fallend und

$$\frac{1}{32} \sum_{k \geqslant 1} c_k H(n_{k+1}, n_k) = \sum_{k \geqslant 1} \frac{1}{2^k} = 1.$$

Für $n \in [n_k, n_{k+1})$ definiere $p_n := \frac{c_k}{32n}$, dann fällt p_n monoton und

$$\sum_{n\geqslant 1} p_n = \sum_{k\geqslant 1} \sum_{i=n_k}^{n_{k+1}-1} \frac{c_k}{32i} = \sum_{k\geqslant 1} \frac{c_k}{32} \underbrace{\sum_{i=n_k}^{n_{k+1}-1} \frac{1}{i}}_{H(n_{k+1},n_k)} = 1.$$

Ferner gilt für $n \in [n_k, n_{k+1})$

$$\begin{aligned}\sum_{i\geqslant n+1} p_i &\geqslant \sum_{i\geqslant n_{k+1}} p_i = \sum_{i\geqslant k+1} \sum_{j=n_i}^{n_{i+1}-1} \frac{c_i}{32j} = \sum_{i\geqslant k+1} \frac{c_i}{32} H(n_{i+1}, n_i) \\ &= \sum_{i\geqslant k+1} 2^{-i} = 2^{-k}.\end{aligned}$$

Da außerdem $H(n_{k+1}, n_k) \geqslant H(n_2, n_1) \geqslant 32$ folgt

$$\frac{1}{2^k} \geqslant \frac{32}{2^k H(n_{k+1}, n_k)} = c_k = 32np_n.$$

Zudem ist

$$\frac{1}{2^k} \geqslant 8a_{n_k} \geqslant 8a_n. \quad «$$

» *Beweis des No-free-lunch Theorems.* Ohne Einschränkung sei $f_D \in \{-1, 1\}$, ansonsten betrachte $\operatorname{sign} f_D$. Fixiere eine Folge $(p_i) \downarrow$ gemäß dem vorangegangen Lemma und zerlege X in Partitionen $(A_j)_{j\geqslant 1}$ mit $\mu(A_j) = p_j$ (Anwendung des Satzes von Lyapunov). Sei nun $\overline{\nu}$ ein W-Maß auf $\{0, 1\}$ mit $\overline{\nu}(\{0\}) = 1/2$, so ist

$$\nu := \bigotimes_{i=1}^{\infty} \overline{\nu},$$

ein W-Maß auf $\Omega := \{0, 1\}^\infty$. Für $\omega = (\omega_j) \in \Omega$ schreibe

$$\eta_\omega(x) := \sum_{j=1}^{\infty} \omega_j \mathbf{I}_{A_j}(x), \quad \in \{0, 1\}.$$

Nun sei P_ω das W-Maß auf $X \times Y$, das durch

$$\begin{aligned}&P_X = \mu, \\ &P(Y = 1 \mid x) = \eta_\omega(x), \quad P\text{-f.s.}\end{aligned}$$

charakterisiert wird, so gilt

$$\mathcal{R}^*_{L,P_\omega} = \int_X \underbrace{\min\{\eta_\omega, 1-\eta_\omega\}}_{=0} \, \mathrm{d}P_X = 0.$$

Wir zeigen nun, dass

$$\int_\Omega \inf_{n\geqslant 1} \frac{1}{a_n} \int_{(X\times Y)^n} \mathcal{R}_{L,P_\omega}(f_D) \, \mathrm{d}P^n_\omega(D) \, \mathrm{d}\nu(\omega) \geqslant \frac{1}{2}, \tag{*}$$

denn dann existiert ein ω, so dass auch

$$\inf_{n\geqslant 1} \frac{1}{a_n} \int_{(X\times Y)^n} \mathcal{R}_{L,P_\omega}(f_D) \, \mathrm{d}P^n_\omega(D) \geqslant \frac{1}{2},$$

und für die Folge $(2a_n)_{n\geqslant 1}$ folgt die Behauptung.

Idee: Im besten Fall genügt ein Punkt, um alle Informationen auf dem zugehörigen A_j zu liefern, denn η_ω ist dort konstant. Haben wir beispielsweise 4 Datenpunkte mit je einem in $A_1, \dots, A_4$, dann können wir auf diesen "richtig" entscheiden, jedoch haben wir keine Möglichkeit über A_4 hinaus noch "richtig" zu entscheiden, denn egal wie entschieden wird, existieren immer ω, für die die Entscheidung schlecht ist. Wir können dann also nur noch raten…

Der folgende Beweis ist jedoch sehr technisch und daher ist es schwer, die anschauliche Idee dort wieder zu finden.

Sei $D = (X \times Y)^\infty$ und schreibe $D_n := ((x_1, y_1), \dots, (x_n, y_n))$, so gilt

$$\begin{aligned}(*) &= \int_\Omega \inf_{n\geqslant 1} \frac{1}{a_n} \int_{(X\times Y)^n} \mathcal{R}_{L,P_\omega}(f_D) \, \mathrm{d}P^n_\omega(D) \, \mathrm{d}\nu(\omega) \\ &\geqslant \int_\Omega \int_{(X\times Y)^n} \inf_{n\geqslant 1} \frac{1}{a_n} \mathcal{R}_{L,P_\omega}(f_D) \, \mathrm{d}P^n_\omega(D) \, \mathrm{d}\nu(\omega).\end{aligned}$$

Ist $\mathcal{R}_{L,P_\omega}(f_D) \geqslant a_n$, so ist auch $\frac{1}{a_n}\mathcal{R}_{L,P_\omega}(f_D) \geqslant 1$ und daher

$$\dots \geqslant \int_\Omega \int_{(X\times Y)^n} \mathbf{I}_{\bigcap_{n=1}^\infty \{\mathcal{R}_{L,P_\omega}(f_D)\geqslant a_n\}} \, \mathrm{d}P^n_\omega(D) \, \mathrm{d}\nu(\omega).$$

Setzen wir $B_n := \left\{\mathcal{R}_{L,P_\omega}(f_D) \geqslant a_n\right\}$, so ist $\bigcap_{n\geqslant 1} B_n = X \setminus \left(\bigcup_{n\geqslant 1} X \setminus B_n\right)$ und für

das Maß $Q = P_\omega^\infty$ gilt folglich

$$\begin{aligned} Q\left(\bigcap_{n\geqslant 1} B_n\right) &= 1 - Q\left(\bigcup_{n\geqslant 1} X \setminus B_n\right) \geqslant 1 - \sum_{n\geqslant 1} Q(X \setminus B_n) \\ &\geqslant 1 - \underbrace{\sum_{n\geqslant 1} \int_\Omega \int_{(X\times Y)^\infty} \mathbf{I}_{\{\mathcal{R}_{L,P}(f_D)\leqslant a_n\}} \,\mathrm{d}P_\omega^n(D)\,\mathrm{d}\nu(\omega)}_{(**)}. \end{aligned}$$

Im Folgenden wollen wir (**) nach oben abschätzen. Für $n \geqslant 1$ definiere

$$\overline{f}_{D_n} := \begin{cases} 1, & \mu\left(\left\{f_D = 1\right\} \cap A_j\right) \geqslant \mu\left(\left\{f_D = -1\right\} \cap A_j\right), \\ 0, & \text{sonst}, \end{cases}$$

und schreibe ferner

$$E_{\omega,j}(f_{D_n}) := A_j \cap \left\{f_D \neq 2\eta_\omega - 1\right\}$$

für die Menge der Fehlentscheidungen von f_{D_n} auf A_j für das Maß P_ω. Auf A_j ist $\eta_\omega(x) \equiv \omega_j$ und daher gilt

$$\overline{f}_{D_n}(j) \neq 2\omega - 1 \Rightarrow \mu(E_{\omega,j}(f_{D_n})) \geqslant \frac{p_j}{2}.$$

Um dies einzusehen sei z.B. $\omega_j = 1$, dann ist $\overline{f}_{D_n}(j) = -1$ und daher nach Definition von $\overline{f}_{D_n}$

$$\mu\left(\left\{f_{D_n} = 1\right\} \cap A_j\right) = \mu\left(\left\{f_{D_n} = -1\right\} \cap A_j\right)$$

also

$$\begin{aligned} p_j = \mu(A_j) &= \mu\left(\left\{f_{D_n} = 1\right\} \cap A_j\right) + \mu\left(\left\{f_{D_n} = -1\right\} \cap A_j\right) \\ &\leqslant 2\mu\left(\underbrace{\left\{f_{D_n} = -1\right\} \cap A_j}_{E_{\omega,j}(f_{D_n})}\right) \end{aligned}$$

Damit folgt $\mathbf{I}_{\left\{\mu(E_{\omega,j}(f_{D_n}))\geqslant \frac{p_j}{2}\right\}} \geqslant \mathbf{I}_{\left\{\overline{f}_{D_n}\neq 2\omega_j - 1\right\}}$ und wegen

$$\mu\left(E_{\omega,j}(f_{D_n})\right) \geqslant \frac{p_j}{2}\mathbf{I}_{\left\{\mu(E_{\omega,j}(f_{D_n}))\geqslant \frac{p_j}{2}\right\}}$$

ist

$$\begin{aligned}\mathcal{R}_{L,P_\omega}(f_D) &= \mathcal{R}_{L,P_\omega}(f_D) - \underbrace{\mathcal{R}^*_{L,P_\omega}}_{=0} \\ &= \int_X |2\eta_\omega - 1|\, \mathbf{I}_{(-\infty,0]}\left((2\eta_\omega - 1)\operatorname{sign} f_{D_n}\right) \mathrm{d}\mu \\ &= \mu\left(\left\{f_{D_n} \neq (2\eta_\omega - 1)\right\}\right) = \sum_{j=1}^{\infty} \mu\left(E_{\omega,j}(f_{D_n})\right) \\ &\geqslant \frac{1}{2}\sum_{j=1}^{\infty} p_j \mathbf{I}_{\{f_{D_n} \neq 2\omega_j - 1\}} \geqslant \frac{1}{2}\sum_{\substack{j:j\leqslant n \\ x_i \neq A_j}} p_j \mathbf{I}_{\left\{\overline{f}_{D_n}(j) \neq 2\omega_j - 1\right\}}\end{aligned}$$

Daraus folgt

$$\begin{aligned}&\int_\Omega \int_{(X\times Y)^\infty} \mathbf{I}_{\{\mathcal{R}_{L,P_\omega}(f_{D_n}) \leqslant a_n\}} \,\mathrm{d}P_\omega^\infty(D)\,\mathrm{d}\nu(\omega) \\ &\leqslant \int_\Omega \int_{(X\times Y)^\infty} \mathbf{I}_{\left\{\sum_{\substack{j:j\leqslant n \\ x_i\in A_j}} p_j \mathbf{I}_{\{\overline{f}_{D_n} \neq 2\omega_j - 1\}} \leqslant 2a_n\right\}} \,\mathrm{d}P_\omega^\infty(D)\,\mathrm{d}\nu(\omega).\end{aligned}$$

Für alles Weitere fixieren wir $\mathcal{J} \coloneqq \left\{j : j \leqslant n,\ x_i \in A_j\right\}$.

Zu jedem $x \in X$ existiert genau ein $j_x \geqslant 1$ mit $x \in A_j$, wir schreiben dafür $j(x) \coloneqq j_x$. Nach Konstruktion sind die x_i von ω unabhängig und es gilt

$$y_i = 2\omega_j(x_i) - 1.$$

Die Labels y_i lassen sich somit explizit berechnen und daher hängt $\overline{f}_D$ nur von $D_X = (x_1, \ldots, x_n)$ und $\omega_{D_X} = (\omega_j(x_1), \ldots, \omega_j(x_n))$ ab. Wir schreiben daher $\overline{f}_{D_X,\omega}$ für $\overline{f}_D$ und erhalten,

$$\begin{aligned}&\int_\Omega \int_{(X\times Y)^\infty} \mathbf{I}_{\left\{\sum_{j\in\mathcal{J}} p_j \mathbf{I}_{\{\overline{f}_D(j) \neq 2\omega_j - 1\}} \leqslant 2a_n\right\}} \,\mathrm{d}P_\omega^n(D)\,\mathrm{d}\nu(\omega) \\ &= \int_{X^n} \int_\Omega \mathbf{I}_{\left\{\sum_{j\in\mathcal{J}} p_j \mathbf{I}_{\{\overline{f}_{D_X,\omega}(j) \neq 2\omega_j - 1\}} \leqslant 2a_n\right\}} \,\mathrm{d}\nu(\omega)\,\mathrm{d}\mu^n(D_X).\end{aligned}$$

Sei weiterhin Ω_{D_X} das Kreuzprodukt, das durch die Koordinaten $j(x_1), \ldots, j(x_n)$ beschrieben wird, und $\Omega_{\not D_X}$ stehe für die "übrigen" Koordinaten. So ist $\Omega =$

$\Omega_{D_X} \times \Omega_{\not D_X}$, in selber Weise definieren wir ν_{D_X} und $\nu_{\not D_X}$. $\overline{f}_{D_X,\omega}$ ändert sich nur in den Koordinaten von ω, die zu Ω_{D_X} gehören, also ist

$$\begin{aligned}
&\int_\Omega \mathbf{I}_{\left\{\sum_{j\in J} p_j \mathbf{I}_{\{\overline{f}_{D_X,\omega}(j)\neq 2\omega_j-1\}} \leqslant 2a_n\right\}} \,\mathrm{d}\nu(\omega) \\
&= \int_{\Omega_{D_X}} \int_{\Omega_{\not D_X}} \mathbf{I}_{\left\{\sum_{j\in J} p_j \mathbf{I}_{\{\overline{f}_{D_X,\omega}(j)\neq 2\omega_j-1\}} \leqslant 2a_n\right\}} \,\mathrm{d}\nu_{\not D_X}(\omega_{\not D_X})\,\mathrm{d}\nu_{D_X}(\omega_{D_X}).
\end{aligned}$$

Außer dem ist $\overline{f}_{D_X,\omega}(j) \equiv 1$ oder -1 für alle $\omega_{\not D_X}$ bei festem D_X und ω_{D_X}, wobei $w_j = 0$ oder 1 jeweils mit Wahrscheinlichkeit $\frac{1}{2}$, also

$$\begin{aligned}
\ldots &= \int_{\Omega_{D_X}} \int_{\Omega_{\not D_X}} \mathbf{I}_{\left\{\sum_{j\in J} p_j \underbrace{\mathbf{I}_{\{\omega_j=1\}}}_{=\omega_j} \leqslant 2a_n\right\}} \,\mathrm{d}\nu_{\not D_X}(\omega_{\not D_X})\,\mathrm{d}\nu_{D_X}(\omega_{D_X}) \\
&\geqslant \int_\Omega \mathbf{I}_{\left\{\sum_{j=n+1}^\infty p_j\omega_j \leqslant 2a_n\right\}} \,\mathrm{d}\nu(\omega). \qquad (**)
\end{aligned}$$

Da p_i monoton fällt und (**) unabhängig von D_X, ist

$$\begin{aligned}
(**) &= \nu\left(\left\{\omega : -\sum_{j=n+1}^\infty p_j\omega_j > -2a_n\right\}\right) \\
&= \nu\left(\left\{\omega : \exp\left(2sa_n - s\sum_{j=n+1}^\infty p_j\omega_j\right) > 1\right\}\right)
\end{aligned}$$

für ein $s > 0$, das wir später wählen. Anwendung der Markov-Ungleichung ergibt,

$$\begin{aligned}
\ldots &\leqslant \mathbf{E}_{\omega\sim\nu} \exp\left(2sa_n - s\sum_{j=n+1}^\infty p_j\omega_j\right) = \exp(2sa_n) \prod_{j=n+1}^\infty \mathbf{E}_{\omega\sim\nu} \exp(-sp_j\omega_j) \\
&= \exp(2sa_n) \prod_{j=n+1}^\infty \left(\frac{1}{2} + \frac{1}{2}\exp(-sp_j)\right)
\end{aligned}$$

Verwenden wir nun, dass $e^{-t} \leqslant 1 - t + \frac{t^2}{2}$ für $t \geqslant 0$, so ist

$$\ldots \leqslant \exp(2sa_n) \prod_{j=n+1}^\infty \frac{1}{2}\left(2 - sp_j + \frac{s^2p_j^2}{2}\right)$$

und, da $1 - t \leqslant e^{-t}$ ist

$$\begin{aligned}
\ldots &\leqslant \exp(2sa_n) \prod_{j=n+1}^{\infty} \left(\exp\left(\frac{-sp_j}{2} + \frac{s^2 p_j^2}{4} \right) \right) \\
&= \exp\left(2sa_n - s \sum_{j=n+1}^{\infty} p_j + \frac{s^2}{4} \sum_{j=n+1}^{\infty} p_j^2 \right) \\
&\leqslant \exp\left(2sa_n - s \sum_{j=n+1}^{\infty} p_j + \frac{s^2 p_{n+1}}{4} \sum_{j=n+1}^{\infty} p_j \right) \\
&= \exp\left(2sa_n + \left(\frac{s^2 p_{n+1}}{4} - s \right) \sum_{j=n+1}^{\infty} p_j \right)
\end{aligned}$$

Setzen wir nun $A \coloneqq \sum_{j=n+1}^{\infty} p_j$ und

$$s = \frac{\sum_{j=n+1}^{\infty} p_j - 4a_n}{p_{n+1} \sum_{j=n+1}^{\infty} p_j} = \frac{A - 4a_n}{Ap_{n+1}},$$

so ist $s > 0$ nach Konstruktion der p_j. Eine längere aber elementare Rechnung zeigt,

$$\exp\left(2sa_n + \left(\frac{s^2 p_{n+1}}{4} - s \right) \sum_{j=n+1}^{\infty} p_j \right) = \exp\left(-\frac{1}{4} \frac{(A - 4a_n)^2}{Ap_{n+1}} \right)$$

Die Funktion $x \mapsto (A - 4x)^2$ hat ein globales Minimum bei $x = \frac{A}{4}$ und ist auf $(-\infty, \frac{A}{4}]$ monoton fallend. Folglich ist $(A - 4x)^2 \geqslant \left(A - \frac{A}{2}\right)^2 = \frac{A^2}{4}$ für $x \leqslant \frac{A}{8}$ und daher

$$\ldots \leqslant \exp\left(-\frac{A}{16p_{n+1}} \right) \leqslant \exp(-2n),$$

denn $A \geqslant 32p_{n+1}n$.

Schließlich folgt

$$\int_\Omega \inf_{n\geqslant 1} \frac{1}{a_n} \int_{(X\times Y)^n} \mathcal{R}_{L,P}(f_D)\,\mathrm{d}P_\omega^n(D)\,\mathrm{d}\nu(\omega)$$
$$\geqslant 1 - \sum_{n=1}^\infty \int_\Omega \int_{(X\times Y)^\infty} \mathbf{I}_{\{\mathcal{R}_{L,P}(f_{D_n})\leqslant a_n\}}\,\mathrm{d}P_\omega^\infty(D)\,\mathrm{d}\nu(\omega)$$
$$\geqslant 1 - \sum_{n=1}^\infty e^{-2n} = \frac{e^2-2}{e^2-1} \geqslant \frac{1}{2}. \quad «$$

Abschlussbemerkungen. A. Das NFL gilt auch in den folgenden Situationen:

- $X = \mathbb{R}^d$ und $\eta \in C^\infty([0,1))$.
- $X = \mathbb{R}^2$ und η ist unimodal, d.h. es gibt ein $x_0 \in X$ mit $\lambda \mapsto \eta(\lambda x_0)$ fallend für wachsendes $\lambda > 0$.
- $\eta \in \{0,1\}$, $X \subset \mathbb{R}^2$ und $\{\eta = 1\}$ ist kompakt, konvex und $0 \in \{\eta = 1\}$.

B. Es gibt keine Super-Klassifikationsmethoden. Ist $\mathcal{L}$ eine Klassifikationsmethode, so gibt es eine universell konsistente Klassifikationsmethode $\mathcal{L}'$ und ein W-Maß P auf $X \times Y$ mit

$$\mathbf{E}_{D\sim P^n} \mathcal{R}_{L_{\text{class}},P}(f_D) > \mathbf{E}_{D\sim P^n} \mathcal{R}_{L_{\text{class}},P}(f'_D), \qquad n \geqslant 1.$$

Dies erscheint auf den ersten Blick natürlich negativ, es hat jedoch auch positive Auswirkungen, denn dadurch wird das Feld stetig neu belebt…

C. Für jede Methode $\mathcal{L}$, die $\mathcal{R}^*_{L_{\text{class}},P}$ abschätzt und jedes $n \geqslant 1$, $\varepsilon > 0$ gibt es ein P mit

$$\mathbf{E}_{D\sim P^n} \left|\mathcal{R}_{L_{\text{class}},P}(f_D) - \mathcal{R}^*_{L_{\text{class}},P}\right| \geqslant \frac{1}{4} - \varepsilon. \quad \multimap$$

Wir wollen dieses Kapitel mit einer offenen Frage beenden. Dazu die folgende

Definition *Eine Klassifikationsmethode $\mathcal{L}$ heißt smart, wenn $\mathbf{E}_{D\sim P^n} \mathcal{R}_{L_{\text{class}},P}(f_D)$ monoton fallend für $n \to \infty$ und alle P.* ⋊

Offene Frage *Gibt es eine universell konsistente und smarte Klassifikationsmethode?* ⋊

Bis jetzt ist noch keine solche Lernmethode bekannt. Dabei ist die universelle Konsistenz entscheidend. Lässt man diese Voraussetzung fallen, so gibt es natürlich zahllose smarte Lernmethoden.

2 Verlustfunktionen

Sei L eine Verlustfunktion, dann ist

$$\mathcal{R}_{L,D}(f) \coloneqq \frac{1}{n}\sum_{i=1}^{n} L(x_i, y_i, f(x_i))$$

sinnvoll, wenn man D mit dem empirischen Maß

$$\mu = n^{-1}\sum_{i=1}^{n} \delta_{\{(x_i,y_i)\}}$$

identifiziert. Sei weiterhin $\mathcal{F}$ eine Menge von Funktionen $f : X \to \mathbb{R}$, dann suchen wir $(f, D) \in \mathcal{F} \times (X \times Y)^n$, so dass das Infimum

$$\inf_{f\in\mathcal{F}} \mathcal{R}_{L,D}(f) + \gamma(f)$$

für einen Strafterm γ angenommen wird.

Um diese Frage zu klären, wollen wir nach Eigenschaften der Verlustfunktion suchen, die sich auf das Risiko übertragen.

2-1 Eigenschaften von Verlustfunktionen

Wir betrachten im Folgenden Verlustfunktionen

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

Während X und Y relativ "strukturlos" sind, hat $\mathbb{R}$ sehr viel "Struktur". Wir suchen nun nach Eigenschaften von $\mathbb{R}$ wie Topologie, Metrik, Konvexität, ... die sich auf L übertragen.

2-1.1 **Definition** *Eine Verlustfunktion $L : X \times Y \times \mathbb{R} \to [0,\infty)$ heißt (strikt) konvex, falls $L(x,y,\cdot) : \mathbb{R} \to [0,\infty)$ (strikt) konvex.* ⋊

Erinnerung. $\mathcal{L}_0(X) \coloneqq \{f : X \to \mathbb{R} : f \text{ messbar}\}$. ⊸

2-1.2 **Lemma** *Sei L eine (strikt) konvexe Verlustfunktion, dann ist auch*

$$\mathcal{R}_{L,P}(\cdot) : \mathcal{L}_0(X) \to [0,\infty]$$

(strikt) konvex. ⋊

» Übung. «

2-1.3 **Definition** *Eine Verlustfunktion $L : X \times Y \times \mathbb{R} \to [0,\infty)$ heißt stetig, falls $L(x,y,\cdot) : \mathbb{R} \to [0,\infty)$ stetig für alle $(x,y) \in (X \times Y)$. L heißt lokal lipschitz-stetig, falls*

$$\forall\ a > 0\ \exists\ c_a : \left|L(x,y,t) - L(x,y,t')\right| \leqslant c_a \left|t - t'\right|$$

für alle $(x,y) \in (X \times Y)$ und $t,t' \in [-a,a]$. Die kleinste Konstante c_a wird mit $|L|_{a,1}$ bezeichnet.

L heißt lipschitz-stetig, falls $|L|_1 \coloneqq \sup_{a \geqslant 0} |L|_{a,1} < \infty$. ⋊

2-1.4 **Lemma** *Sei L eine stetige Verlustfunktion und P ein W-Maß auf $X \times Y$, so gilt für alle Folgen (f_n) in $\mathcal{L}_0(X)$ und $f \in \mathcal{L}_0(X)$ mit $f_n \xrightarrow{P} f$,*

$$\mathcal{R}_{L,P}(f) \leqslant \liminf_{n\to\infty} \mathcal{R}_{L,P}(f_n).$$ ⋊

Das Risiko ist also "halbstetig von unten".

» Da $f_n \xrightarrow{P} f$ existiert eine Teilfolge mit $f_{n_k} \to f$ P-f.s. und daher existiert auch eine weitere Teilfolge $(f_{n_{k_l}})$ mit

$$\lim_{l\to\infty} \mathcal{R}_{L,P}(f_{n_{k_l}}) = \liminf_{n\to\infty} \mathcal{R}_{L,P}(f_n).$$

Wir schreiben nun kürzer (f_{n_k}) für die Teilfolge mit beiden Eigenschaften. Da L stetig gilt $L(x,y,f_{n_k}(x)) \to L(x,y,f(x))$ P_X-f.s. und mit dem Lemma von Fatou folgt,

$$\begin{aligned}
\mathcal{R}_{L,P}(f) &= \int_{X\times Y} L(x,y,f(x))\,\mathrm{d}P(x,y) = \int_{X\times Y} \lim_{k\to\infty} L(x,y,f_{n_k}(x))\,\mathrm{d}P(x,y) \\
&\leqslant \liminf_{k\to\infty} \int_{X\times Y} L(x,y,f_{n_k}(x))\,\mathrm{d}P(x,y) = \lim_{k\to\infty} \mathcal{R}_{L,P}(f_{n_k}) \\
&= \liminf_{n\to\infty} \mathcal{R}_{L,P}(f_n). \quad «
\end{aligned}$$

Optimal für unsere Zwecke wäre "=" anstat "$\leqslant$" im Lemma 2-1.4. Es stellt sich jedoch heraus, dass dies nicht ohne zusätzliche Voraussetzungen an L möglich ist.

2-1.5 **Definition** *Eine Verlustfunktion $L : X \times Y \times \mathbb{R} \to [0, \infty)$ heißt Nemitski-Verlustfunktion (NVF), falls eine messbare Funktion $b : X \times Y \to [0, \infty)$ und eine messbare und monoton wachsende Funktion $h : \mathbb{R} \to [0, \infty)$ existiert, so dass*

$$L(x, y, t) \leqslant b(x, y) + h(|t|), \qquad \forall\, x, y, t.$$

L heißt NVF der Ordnung $p \in (0, \infty)$, falls ein $c > 0$ existiert, so dass

$$L(x, y, t) \leqslant b(x, y) + c \cdot |t|^p, \qquad \forall\, x, y, t.$$

Ist P ein W-Maß auf $(X \times Y)$ und L eine NVF, so heißt L P-integrierbar, falls b P-integrierbar. ⋈

2-1.6 **Lemma** *Sei P ein W-Maß auf $X \times Y$ und L eine stetige, P-integrierbare NVF, dann gelten.*

1) *Sei (f_n) eine gleichmäßig beschränkte Folge in $\mathcal{L}_0(P_X)$ und $f \in \mathcal{L}_\infty(P_X)$ mit $f_n \to f$ P-f.s. Dann folgt*

$$\mathcal{R}_{L,P}(f) = \lim_{n\to\infty} \mathcal{R}_{L,P}(f_n).$$

2) *$\mathcal{R}_{L,P}(\cdot) : \mathcal{L}_\infty(P_X) \to [0, \infty)$ ist stetig.*

3) *Ist L außerdem von der Ordnung $p \in [1, \infty)$, so ist*

$$\mathcal{R}_{L,P}(f) : \mathcal{L}_p(P_X) \to [0, \infty)$$

wohldefiniert und stetig. ⋈

» "1)": f ist beschränkt mit $\left\|f\right\|_\infty \leqslant B$. Ferner gilt

$$\lim_{n\to\infty} \underbrace{\left|L(x, y, f_n(x)) - L(x, y, f(x))\right|}_{g_n(x)} = 0 \; P\text{-f.s.}$$

und

$$\begin{aligned} g_n(x) &\leqslant L(x,y,f_n(x)) + L(x,y,f(x)) \\ &\leqslant b(x,y) + h(\left|f_n(x)\right|) + b(x,y) + h(\left|f(x)\right|) \\ &\leqslant 2(b(x,y) + h(B))\ P\text{-f.s.} \end{aligned}$$

wobei die rechte Seite als Funktion in (x,y) P-integrierbar ist. Mit dem Satz von Lebesgue folgt nun,

$$\left|\mathcal{R}_{L,P}(f_n) - \mathcal{R}_{L,P}(f)\right| \leqslant \int_{X\times Y} g_n(x)\,\mathrm{d}P(x,y) \to 0, \qquad n\to\infty.$$

"2)": Sei (f_n) Folge in $\mathcal{L}_\infty(X)$ und $f\in\mathcal{L}_\infty(X)$ mit $\left\|f_n - f\right\|_\infty \to 0$. Dann ist f_n gleichmäßig beschränkt und $f_n \to f$ P_X-f.s., also $\mathcal{R}_{L,P}(f_n)\to\mathcal{R}_{L,P}(f)$.

"3)": Es gilt für $f\in\mathcal{L}_p(P_X)$,

$$R_{L,P}(f) = \int L(x,y,f(x))\,\mathrm{d}P(x,y) \leqslant \int b(x,y) + c\left|f(x)\right|^p\,\mathrm{d}P(x,y) < \infty.$$

Somit ist $\mathcal{R}_{L,P}(\cdot):\mathcal{L}_p(P_X)\to[0,\infty)$ wohldefiniert.

Sei nun $(f_n)\subset\mathcal{L}_p(P_X)$ und $f\in\mathcal{L}_p(P_X)$ mit $\left\|f_n - f\right\|_p \to 0$, dann gilt auch $f_n \xrightarrow{P} f$. Mit Lemma 2-1.4 folgt

$$\mathcal{R}_{L,P}(f) \leqslant \liminf_{n\to\infty} \mathcal{R}_{L,P}(f_n). \tag{*}$$

Definiere nun $\overline{L}(x,y,t) := b(x,y) + c\,|t|^p - L(x,y,t) \geqslant 0$, so ist $\overline{L}: X\times Y\times\mathbb{R}\to[0,\infty)$ eine stetige Verlustfunktion. Erneute Anwendung von 2-1.4 ergibt

$$\begin{aligned} \|b\|_1 + c\left\|f\right\|_p^p - \mathcal{R}_{L,P}(f) &= \mathcal{R}_{\overline{L},P}(f) \leqslant \liminf_{n\to\infty} \mathcal{R}_{\overline{L},P}(f_n) \\ &= \|b\|_1 + \liminf_{n\to\infty} c\left\|f_n\right\|_p^p - \mathcal{R}_{L,P}(f_n). \end{aligned}$$

Da $\left\|f_n\right\|_p \to \left\|f\right\|_p$ ist $-\mathcal{R}_{L,P}(f) \leqslant -\limsup_{n\to\infty} \mathcal{R}_{L,P}(f_n)$, d.h. mit (*) folgt

$$\limsup_{n\to\infty} \mathcal{R}_{L,P}(f_n) \leqslant \liminf_{n\to\infty} \mathcal{R}_{L,P}(f_n)$$

also $\lim_{n\to\infty} \mathcal{R}_{L,P}(f_n) = \mathcal{R}_{L,P}(f)$. «

2-1.7 **Lemma** *Ist $L : Y \times \mathbb{R} \to [0,\infty)$ strikt überwacht und konvex und Y endlich. Dann ist L lokal lipschitz.* ⋊

» Wir benutzen, dass jede konvexe Abbildung $g : [-a,a] \to [0,\infty)$ lokal lipschitz ist. Nach dieser Aussage ist $L(y,\cdot) : \mathbb{R} \to [0,\infty)$ lokal lipschitz für alle $y \in Y$. Da Y endlich, folgt die Aussage. «

2-1.8 **Lemma** *1) Ist L lokal lipschitz, so ist L NVF.*

2) Ist L lokal lipschitz und $\mathcal{R}_{L,P}(0) < \infty$, so ist L P-integrierbare NVF.

3) Ist L lipschitz stetig, so ist L NVF der Ordnung $p = 1$. ⋊

» 1) $\big|L(x,y,t) - L(x,y,0)\big| \leqslant |L|_{|t|,1}\,|t|$. Somit ist

$$L(x,y,t) \leqslant \underbrace{|L|_{|t|,1}\,|t|}_{h(t)} + \underbrace{L(x,y,0)}_{b(x,y)}.$$

2) Falls $\mathcal{R}_{L,P}(0) < \infty$ ist $b(x,y)$ P-integrierbar.

3) Falls L lipschitz, ist $|L|_{|t|,1} \leqslant |L|_1$ und damit

$$L(x,y,t) \leqslant L(x,y,0) + |L|_1\,|t|\,. \quad «$$

2-1.9 **Lemma** *Sei L lokal lipschitz, $B \geqslant 0$ und $f,g \in \mathcal{L}_\infty(P_X)$ mit $\|f\|_\infty, \|g\|_\infty \leqslant B$. Dann gilt*

$$\big|\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,P}(g)\big| \leqslant |L|_{B,1}\,\big\|f-g\big\|_{\mathcal{L}_1(P_X)}\,. \quad ⋊$$

» Übung. «

2-1.10 **Definition** *Eine Verlustfunktion L kann bei $M > 0$ abgeschnitten werden, wenn für alle $(x,y) \in X \times Y$ und $t \in \mathbb{R}$ gilt*

$$L(x,y,\widehat{t}) \leqslant L(x,y,t),$$

wobei

$$\widehat{t} = \begin{cases} -M, & t \leqslant -M, \\ t, & t \in (-M,M) \\ M, & t \geqslant M. \end{cases} \quad ⋊$$

Man kann dies so interpretieren, dass Abschneiden den Verlust *nicht* erhöht.

2-1.11 **Lemma** *Sei L eine konvexe Verlustfunktion und $M > 0$. Dann sind folgende Aussagen äquivalent:*

(i) L kann bei M abgeschnitten werden.

(ii) Für alle $(x, y) \in X \times Y$ hat die Funktion $L(x, y, \cdot) : \mathbb{R} \to [0, \infty)$ mindestens ein globales Minimum in $[-M, M]$. ⋊

» Schreibe $M_{x,y} := \left\{ t^* \in \mathbb{R} : L(x, y, t^*) = \inf_{t \in \mathbb{R}} L(x, y, t) \right\}$. Da L konvex ist, ist $M_{x,y}$ ein Intervall.

"(i)⇒(ii)": Angenommen es gibt ein $(x, y) \in X \times Y$ mit $M_{x,y} \cap [-M, M] = \emptyset$.

1. Fall $M_{x,y} = \emptyset$. L ist konvex, also ist $L(x, y, \cdot)$ strikt monoton, denn falls $L(x, y, \cdot) \in C^2(\mathbb{R})$, so ist aufgrund der Konvexität $L''(x, y, \cdot) \geqslant 0$. Da aber $M_{x,y} = \emptyset$ folgt $L'(x, y, t) \neq 0$ für alle t. Für allgemeines $L(x, y, \cdot)$ folgt die strikte Monotonie aus der Betrachtung von Subdifferenzialen.

Aber da $L(x, y, \cdot)$ strikt monoton, kann L nicht abgeschnitten werden. ↯

2. Fall $M_{x,y} \neq \emptyset$. Da $M_{x,y}$ ein abgeschlossenes Intervall, folgt ohne Einschränkung $t := \inf M_{x,y}$ erfüllt $M < t < \infty$, d.h. $M_{x,y}$ liegt rechts von $[-M, M]$.

Somit ist $L(x, y, \widehat{t}) = L(x, y, M) > L(x, y, t)$ da $t > M$ und $L(x, y, \cdot)$ aufgrund der Konvexität strikt fallend links von $M_{x,y}$. ↯

"(ii)⇒(i)": Es gilt $M_{x,y} \cap [-M, M] \neq \emptyset$ und daher ist $\inf M_{x,y} \leqslant M$ und $\sup M_{x,y} \geqslant M$.

$L(x, y, \cdot)$ ist strikt konvex, also ist $L(x, y, \cdot)$ auf $[\sup M_{x,y}, \infty)$ wachsend und auf $(-\infty, \inf M_{x,y}]$ fallend. Somit kann L abgeschnitten werden. «

2-2 Margin basierte Verlustfunktionen

Sei $Y = \{-1, 1\}$ und $\eta(x) = P(Y = 1 \mid x)$.

2-2.1 **Definition** *Eine strikt überwachte Verlustfunktion $L : Y \times \mathbb{R} \to [0, \infty)$ heißt margin-basiert, falls eine repräsentative Funktion $\varphi : \mathbb{R} \to [0, \infty)$ existiert, d.h. $L(y, t) = \varphi(y \cdot t)$.* ⋊

2-2.2 **Lemma** *Sei L margin-basiert und φ die repräsentative Funktion. Dann gelten:*

1) *L ist genau dann (strikt) konvex, wenn φ (strikt) konvex.*

2) *L ist genau dann stetig, wenn φ stetig.*

3) *L ist genau dann (lokal) lipschitz, wenn φ (lokal) lipschitz.*

4) *Ist L konvex, so ist L (lokal) lipschitz.*

5) *L ist P-integrierbare NVF.*

6) *Ist L lipschitz stetig, so ist L P-integrierbare NVF der Ordnung $p = 1$.* ⋈

» Übung. «

2-2.3 **Bsp** *Kleinste Quadrate.* Scharfes Hinsehen ergibt,

$$L_{\mathrm{LS}}(y,t) = (y-t)^2 = (1-yt)^2, \quad \Rightarrow \quad \varphi(t) = (1-t)^2.$$

Damit ist L_{LS} strikt konvex, da φ strikt konvex. $|L_{\mathrm{LS}}|_{a,1} = 2a + 2$ für $a \geqslant 0$. L_{LS} kann bei $M = 1$ abgeschnitten werden (φ hat ein globales Minimum bei $t = 1$ und damit $L(1,\cdot)$ bei $t = 1$ und $L(-1,\cdot)$ bei $t = -1$). ■

2-2.4 **Bsp** *Hinge loss.*

$$L_{\mathrm{Hinge}}(y,t) = \max\{0, 1-yt\}, \quad \Rightarrow \quad \varphi(t) = \max\{0, 1-t\}.$$

L_{Hinge} ist konvex, lipschitz und $|L_{\mathrm{Hinge}}|_1 = 1$. Sie ist *nicht* strikt konvex und kann bei $M = 1$ abgeschnitten werden. ■

2-2.5 **Bsp** *Quadrierter Hinge-Loss.*

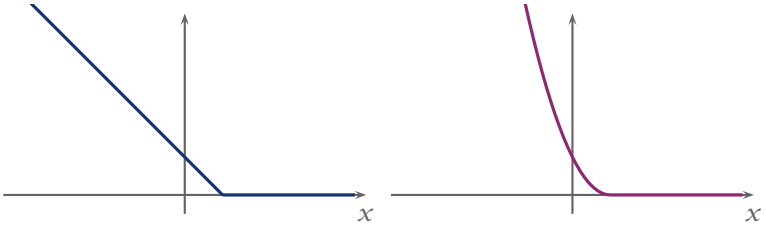
$$L(y,t) = \left(\max\{0, 1-yt\}\right)^2, \quad \Rightarrow \quad \varphi(t) = (\max\{0, 1-t\})^2.$$

Sie stellt eine Mischung aus LS und HL dar. L ist konvex (nicht strikt), lokal lipschitz, $|L|_{a,1} = 2a + 2$ und kann bei $M = 1$ abgeschnitten werden. ■

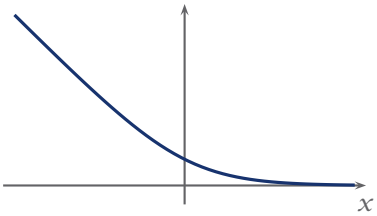
2-2.6 **Bsp** *Logistische Verlustfunktion für Klassifikation.*

$$L_{\log}(y,t) = \log(1+\exp(-yt)), \quad \Rightarrow \quad \varphi(t) = \log(1+\exp(-yt)).$$

$L_{\log}$ ist strikt konvex, lipschitz $|L_{\log}|_1 = 1$, kann aber *nicht* abgeschnitten werden. ■



2.1 **Hinge loss und quadrierte hinge loss Verlustfunktion**



2.2 **Logistische Verlustfunktion**

In Übung 6 wurde gezeigt

$$\mathcal{R}_{L_{\text{class}},P}(f) - \mathcal{R}^*_{L_{\text{class}},P} \leqslant \sqrt{\mathcal{R}_{L_{\text{LS}},P}(f) - \mathcal{R}^*_{L_{\text{LS}},P}}$$

2-2.7 **Lemma** *Für $\eta \in [0,1]$ und $t \in [-1,1]$ gilt*

$$|2\eta - 1|\, \mathbf{I}_{(-\infty,0]}((2\eta-1)\operatorname{sign} t) \leqslant |2\eta - 1| \left|t - \operatorname{sign}(2\eta - 1)\right|. \quad \bowtie$$

» *1. Fall* $\eta = \frac{1}{2}$ ist klar.
2. Fall $\eta < \frac{1}{2}$. Für $t \in [-1,0)$ ist

$$\underbrace{(2\eta - 1)}_{<0} \underbrace{\operatorname{sign} t}_{<0} > 0,$$

somit verschwindet die linke Seite und die rechte ist $\geqslant 0$.
Für $t \in [0,1)$ gilt umgekehrt,

$$\underbrace{(2\eta - 1)}_{<0} \underbrace{\operatorname{sign} t}_{=1} < 0,$$

somit ist die linke Seite gleich $|2\eta - 1| < 1$ und $\left|t - \operatorname{sign}(2\eta - 1)\right| = |t + 1| \geqslant 1$.
«

2-2.8 **Zhang's (Un-)Gleichung** *Sei* $f^*_{L_{\text{class}}}(x) \coloneqq \operatorname{sign}(2\eta(x) - 1)$ *für* $x \in X$. *Dann gilt für* $f : X \to [-1, 1]$*:*

$$\mathcal{R}_{L_{\text{hinge}},P}(f) - \mathcal{R}^*_{L_{\text{hinge}},P} = \int_X \left|f(x) - f^*_{L_{\text{class}},P}(x)\right| |2\eta(x) - 1| \, \mathrm{d}P_X(x)$$

„Überschuss-L_{hinge}-Risiko = gewichtete L^1-Norm von $f - f^*_{L_{\text{class}},P}$".
Und für $h : X \to \mathbb{R}$ gilt

$$\mathcal{R}_{L_{\text{class}},P}(f) - \mathcal{R}^*_{L_{\text{class}},P} \leqslant \mathcal{R}_{L_{\text{hinge}},P}(f) - \mathcal{R}^*_{L_{\text{hinge}},P}. \quad \bowtie$$

» $L_{\text{hinge}}(y, t) = \max\{0, 1 - yt\} \overset{t\in[-1,1]}{=} 1 - yt$. Für $f : X \to [-1, 1]$ gilt daher

$$\begin{aligned}\mathcal{R}_{L_{\text{hinge}},P}(f) &= \int_X \eta(x)(1 - f(x)) + (1 - \eta(x))(1 + f(x)) \, \mathrm{d}P_X \\ &= \int_X 1 + f(x)(1 - 2\eta(x)) \, \mathrm{d}P_X.\end{aligned}$$

$f(x)(1 - 2\eta(x))$ ist minimal genau dann, wenn

$$\left.\begin{aligned} f = 1 \text{ auf } \left\{\eta > \tfrac{1}{2}\right\} \\ f = -1 \text{ auf } \left\{\eta < \tfrac{1}{2}\right\}\end{aligned}\right\} \text{ d.h. } f = f^*_{L_{\text{class}},P} \text{ auf } \left\{\eta \neq \frac{1}{2}\right\}$$

Da L_{hinge} bei $M = 1$ abgeschnitten werden kann folgt somit

$$\mathcal{R}^*_{L_{\text{hinge}},P} = \inf_{f:X\to\mathbb{R}} \mathcal{R}_{L_{\text{hinge}},P}(f) = \inf_{f:X\to[-1,1]} \mathcal{R}_{L_{\text{hinge}},P}(f).$$

Damit ist

$$\begin{aligned}\mathcal{R}_{L_{\text{hinge}},P}(f) - \mathcal{R}^*_{L_{\text{hinge}},P} &= \int_X 1 + f \cdot (1 - 2\eta) - 1 - f^*_{L_{\text{class}},P} \cdot (1 - 2\eta) \, \mathrm{d}P_X \\ &= \int_X \underbrace{(f - f^*_{L_{\text{class}},P}) \cdot (1 - 2\eta)}_{\geqslant 0} \, \mathrm{d}P_X,\end{aligned}$$

wobei der Integrand positiv ist, da $f^*_{L_{\text{class}},P}$ punktweise das Überschussrisiko minimiert und dieses ist positiv, also

$$\mathcal{R}_{L_{\text{hinge}},P(f)} - \mathcal{R}^*_{L_{\text{hinge}},P} = \int_X \left|f - f^*_{L_{\text{class}},P}\right| |1-2\eta| \, \mathrm{d}P_X.$$

Ungleichung. Da L_{hinge} bei $M = 1$ abgeschnitten werden kann, folgt

$$\mathcal{R}_{L_{\text{hinge}},P}(\widehat{f}) - \mathcal{R}^*_{L_{\text{hinge}},P} \leqslant \mathcal{R}_{L_{\text{hinge}},P}(f) - \mathcal{R}^*_{L_{\text{hinge}},P}$$

und

$$\mathcal{R}_{L_{\text{class}},P}(\widehat{f}) - \mathcal{R}^*_{L_{\text{class}},P} = \mathcal{R}_{L_{\text{class}},P}(f) - \mathcal{R}^*_{L_{\text{class}},P}.$$

Daher ist ohne Einschränkung $f : X \to [-1,1]$. In Kapitel 1-2 haben wir gezeigt

$$\begin{aligned}
\mathcal{R}_{L_{\text{class}},P}(f) - \mathcal{R}^*_{L_{\text{class}},P} &= \int_X |2\eta - 1| \, \mathbf{I}_{(-\infty,0]}((2\eta-1)\operatorname{sign} f) \, \mathrm{d}P_X \\
&\overset{2\text{–}2.7}{\leqslant} \int_X |2\eta - 1| \left| f - \underbrace{\operatorname{sign}(2\eta-1)}_{=f^*_{L_{\text{class}},P}} \right| \mathrm{d}P_X \\
&\overset{\text{Zhang}}{=} \mathcal{R}_{L_{\text{hinge}}}(f) - \mathcal{R}^*_{L_{\text{hinge}},P}. \quad \text{«}
\end{aligned}$$

2-3 Distanzbasierte Verlustfunktionen

2-3.1 **Definition** *Eine Verlustfunktion $L : Y \times \mathbb{R} \to [0,\infty)$ heiß distanzbasiert, falls eine Funktion $\psi : \mathbb{R} \to [0,\infty)$ existiert mit*

$$L(y,t) = \psi(y-t), \qquad y \in Y, \quad t \in \mathbb{R}.$$

Eine distanzbasierte Verlustfunktion heißt symmetrisch, falls $\psi(r) = \psi(-r)$ für alle $r \in \mathbb{R}$ heißt . ⋊

2-3.2 **Lemma** *Sei L distanzbasiert, so gelten:*

1) *L ist genau dann (strikt) konvex, wenn ψ (strikt) konvex.*

2) *L ist genau dann stetig, wenn ψ stetig.*

3) *L ist genau dann lipschitz, wenn ψ lipschitz. (Gilt im Allgemeinen nicht für lokal lipschitz, siehe L_{LS})*

4) *Ist ferner $Y \subset [-M, M]$, so ist L genau dann lokal lipschitz, wenn ψ lokal lipschitz.*

5) *Ist L konvex, so ist ψ lokal lipschitz.*

6) *L ist eine P-integrierbare NVF.* ⋊

» 4) „⇐“: Sei $t \in [-a, a]$ und $a > 0$, so gilt

$$\left|L(y,t) - L(y,t')\right| = \left|\psi(y-t) - \psi(y-t')\right| \leqslant |\psi|_{1+M,1} \left|t - t'\right|.$$

„⇒“: Analog folgt $|L|_{a,1} \leqslant |\psi|_{a+M,1}$.

5) Sei L konvex, dann folgt für $y = 0$, dass auch

$$t \mapsto \psi(-t) = \psi(0-t) = L(0,t)$$

konvex und daher ist ψ konvex und folglich auch lokal lipschitz. Mit 4) folgt nun, dass L lokal lipschitz.

6) folgt aus Lemma 2-1.9. «

2-3.3 **Bsp** *Kleinste Quadrate*

$$L_{\mathrm{LS}}(y,t) \coloneqq (y-t)^2 \Rightarrow \psi(r) = r^2.$$

ψ ist strikt konvex, nicht lipschitz aber lokal lipschitz und symmetrisch. ■

2-3.4 **Bsp** *Betragsfunktion.*

$$L_{\mathrm{abs}}(y,t) \coloneqq \left|y-t\right| \Rightarrow \psi(r) = |r|.$$

ψ ist konvex (nicht strikt), symmetrisch und Lipschitz mit $|L|_1 = 1$.

2-3.5 **Bsp** *Pinball für $\varepsilon \in (0,1)$.*

$$\psi(r) \coloneqq \begin{cases} -1(1-\varepsilon)r, & r < 0, \\ \varepsilon r, & r \geqslant 0 \end{cases}$$

ψ ist konvex (nicht strikt), symmetrisch genau dann, wenn $\varepsilon = \frac{1}{2}$ und lipschitz mit $|L|_1 = \min\{\varepsilon, 1-\varepsilon\}$. ■

3 Konzentrationsungleichungen

Im Lemma 1-3.7 hatten wir gezeigt:

$$\mu^n\left(\left\{\omega\in\Omega^n\,:\,\frac{1}{n}\sum_{i=1}^{n}h(\omega_i)-\mathbf{E}_\mu h\geqslant t\right\}\right)\leqslant\frac{\mathbf{E}_\mu h^2}{t^2 n}$$

für alle Maße μ, $n\geqslant 1$, $t>0$ und $h:\Omega\to\mathbb{R}$.

3-0.6 **Satz** *Sei $(\Omega,\mathcal{A},\mu)$ ein Wahrscheinlichkeitsraum, dann gilt für alle $f:\Omega\to\mathbb{R}$ mit $\mathbf{E}_\mu\left|f\right|<\infty$ und $t>0$,*

$$\mu\left(\left\{\omega\in\Omega\,:\,f(\omega)\geqslant t\right\}\right)\leqslant\frac{\mathbf{E}_\mu\left|f\right|}{t}.\quad\bowtie$$

3-0.7 **Lemma** *Für $x>-1$ gilt*

$$(1+x)\ln(1+x)-x\geqslant\frac{3}{2}\frac{x^2}{x+3}.\quad\bowtie$$

» Seien $f(x)=(1+x)\ln(1+x)-x$ und $g(x)=\frac{3}{2}\frac{x^2}{x+3}$. Dann folgt

$$f'(x)=\ln(1+x),\qquad g'(x)=\frac{3}{2}\frac{x^2+6x}{(x+3)^2},$$
$$f''(x)=\frac{1}{1+x},\qquad g''(x)=\frac{27}{(x+3)^3},$$

und es gelten

$$f'(0)=0,\quad f(0)=0,$$
$$f'(0)=0,\quad g(0)=0.$$

Wir zeigen nun $f''(x) \geqslant g''(x)$ für alle $x > -1$. Es ist

$$\begin{aligned} & 0 \leqslant x^2(x+9) = x^3 + 9x^2 \\ \Rightarrow\ & x^3 + 9x^2 + 27x + 27 = (x+3)^3 \geqslant 27(1+x) = 27x + 27 \\ \Rightarrow\ & \frac{27}{(x+3)^3} \leqslant \frac{1}{1+x}. \end{aligned}$$

Für $x \geqslant 0$ folgt mit dem Hauptsatz der Differential- und Integralrechnung,

$$f'(x) = f'(x) - f'(0) = \int_0^x f''(t)\,\mathrm{d}t \geqslant \int_0^x g''(t)\,\mathrm{d}t = g'(x) - g'(0) = g'(x).$$

Das selbe Argument angewandt auf f liefert, $f(x) \geqslant g(x)$ und für $x \in (-1, 0]$ folgt $-f(x) \geqslant -g(x)$. «

3-0.8 **Bernsteins Ungleichung** *Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $B > 0$, $\sigma > 0$ und $n \geqslant 1$. Ferner seien $\xi_1, \dots, \xi_n : \Omega \to \mathbb{R}$ i.i.d. (unabhängige und identisch verteilte) Zufallsvariablen, so dass gilt*

$$\left.\begin{aligned} \mathbf{E}_P \xi_i &= 0, \\ \left\|\xi_i\right\|_\infty &\leqslant B, \\ \mathbf{E}_P \xi^2 &\leqslant \sigma^2 \end{aligned}\right\} \text{ für alle } i = 1, \dots, n.$$

Dann gilt

$$P\left(\frac{1}{n}\sum_{i=1}^n \xi_i \geqslant \sqrt{\frac{2\sigma^2\tau}{n}} + \frac{2B\tau}{3n}\right) \leqslant e^{-\tau}, \quad \text{für alle } \tau > 0. \quad \bowtie$$

» Fixiere $t \geqslant 0$ und $\varepsilon > 0$, dann gilt

$$\begin{aligned} P\left(\sum_{i=1}^n \xi_i \geqslant \varepsilon n\right) &= P\left(\exp\left(t\sum_{i=1}^n \xi_i\right) \geqslant \exp(t\varepsilon n)\right) \\ &\overset{\text{Markov}}{\leqslant} e^{-t\varepsilon n} \mathbf{E}_P \exp\left(t\sum_{i=1}^n \xi_i\right) \\ &\overset{\text{Unabh.}}{=} e^{-t\varepsilon n} \prod_{i=1}^n \mathbf{E}_P e^{t\xi_i}. \end{aligned}$$

Folglich ist

$$\mathbf{E}_P e^{t\xi_i} = \mathbf{E}_P \sum_{k=0}^{\infty} \frac{t^k}{k!} \xi_i^k = \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathbf{E}_P \xi^k$$

und da $\mathbf{E}_P \xi_i = 0$ und $\mathbf{E}_P \xi^k \leqslant \sigma^2 B^{k-2}$ folgt

$$\begin{aligned}\mathbf{E}_P e^{t\xi} &= 1 + \sum_{k=2}^{\infty} \frac{t^k}{k!} \mathbf{E}_P \xi_i^k \leqslant 1 + \sum_{k=2}^{\infty} \frac{t^k}{k!} \sigma^k B^{k-2} = 1 + \frac{\sigma^2}{B^2} \sum_{k=2}^{\infty} \frac{t^k}{k!} B^k \\ &= 1 + \frac{\sigma^2}{B^2} \left(e^{tB} - tB - 1 \right).\end{aligned}$$

Damit gilt

$$\begin{aligned}P\left(\sum_{i=1}^{n} \xi_i \geqslant \varepsilon n \right) &\leqslant e^{-tn\varepsilon} \left(1 + \frac{\sigma^2}{B^2} \left(e^{tB} - tB - 1 \right) \right)^n \\ &\leqslant e^{-tn\varepsilon} \exp\left(n \frac{\sigma^2}{B^2} \left(e^{tB} - tB - 1 \right) \right)\end{aligned}$$

Sei $h(t) \coloneqq -tn\varepsilon + \frac{n\sigma^2}{B^2} \left(e^{tB} - tB - 1 \right)$, dann suchen wir kritische Punkte

$$h'(t) = -n\varepsilon + \frac{n\sigma^2}{B^2} \left(Be^{tB} - B \right) = -n\varepsilon + \frac{n\sigma^2}{B} \left(e^{tB} - 1 \right) \stackrel{!}{=} 0.$$

t ist genau dann kritisch, wenn

$$\frac{\sigma^2 n}{B} e^{tB} = \varepsilon n + \frac{\sigma^2 n}{B} \Leftrightarrow e^{tB} = \frac{\varepsilon B}{\sigma^2} + 1 \Leftrightarrow t^* = \frac{1}{B} \log\left(1 + \frac{\varepsilon B}{\sigma^2} \right).$$

t^* ist der einzige Kandidat für ein Optimum. Weiterhin ist

$$\lim_{t \to \pm\infty} h(t) = \infty$$

und folglich hat h ein globales Minimum bei t^*. Setzen wir $\gamma = \frac{\varepsilon B}{\sigma^2}$, dann gilt

$t^* = \frac{1}{B}\log(1+\gamma)$ und somit

$$\begin{aligned}
&-t^*\varepsilon n + \frac{\sigma^2 n}{B^2}\left(e^{t^*B} - t^*B - 1\right)\\
=&-\frac{\varepsilon n}{B}\log(1+\gamma) + \frac{\sigma^2 n}{B^2}(1+\gamma-\log(1+\gamma)-1)\\
=&\frac{\sigma^2 n}{B^2}\left(-\gamma\log(1+\gamma)+\gamma-\log(1+\gamma)\right)\\
=&-\frac{\sigma^2 n}{B^2}\left((1+\gamma)\log(1+\gamma)-\gamma\right)\\
\overset{3\text{–}0.7}{\leqslant}&-\frac{\sigma^2 n}{B^2}\frac{3}{2}\frac{\gamma^2}{\gamma+3} \overset{\text{Rechnung}}{=} -\frac{3n\varepsilon^2}{2\varepsilon B + 6\sigma^2}.
\end{aligned}$$

Damit ist

$$P\left(\frac{1}{n}\sum_{i=1}^n \xi_i \geqslant \varepsilon\right) \leqslant \exp\left(-\frac{3n\varepsilon^2}{2\varepsilon B + 6\sigma^2}\right).$$

Wähle nun

$$\tau := \frac{3n\varepsilon^2}{2\varepsilon B + 6\sigma^2},$$

dann folgt

$$\varepsilon = \sqrt{\frac{2\sigma^2\tau}{n} + \frac{B^2\tau^2}{9n^2}} + \frac{B\tau}{3n} \leqslant \sqrt{\frac{2\sigma^2\tau}{n}} + \frac{2B\tau}{3n}. \quad «$$

3-0.9 **Hoeffdings Ungleichung** *Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, $a < b$, $1 \leqslant n \in \mathbb{N}$ und $\xi_1, \dots, \xi_n : \Omega \to [a,b]$ i.i.d. Dann gilt für $\tau > 0$,*

$$P\left(\frac{1}{n}\sum_{i=1}^n (\xi_i - \mathbf{E}_P\xi_i) \geqslant (b-a)\sqrt{\frac{\tau}{2n}}\right) \leqslant e^{-\tau}. \quad \bowtie$$

Beobachtung. Für $a = -b$ betrachte $\eta_i = \xi_i - \mathbf{E}_P\xi_i$.
Dann ist $\mathbf{E}_P\eta_i = 0$, $\mathbf{E}_P\eta_i^2 = \mathbf{E}_P\xi_i^2 - (\mathbf{E}_P\xi_i)^2 \leqslant b^2 =: \sigma^2$ und $\|\eta_i\|_\infty \leqslant 2b =: B$. Mit Bernstein folgt,

$$P\left(\frac{1}{n}\sum_{i=1}^n \eta_i \geqslant \sqrt{\frac{2b^2\tau}{n}} + \frac{4b\tau}{3n}\right) \leqslant e^{-\tau},$$

wobei $\sqrt{\frac{2b^2\tau}{n}} = 2b\sqrt{\frac{\tau}{2n}} = (b-a)\sqrt{\frac{\tau}{2n}}$. $\multimap$

3-0.10 **Lemma (Union bound)** *Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, $f_1, \dots, f_n : \Omega \to \mathbb{R}$ messbar. Dann gilt für alle $t \in \mathbb{R}$*

$$\mu\left(\sup_{i=1,\dots,n} f_i \geqslant t\right) \leqslant \sum_{i=1}^{n} \mu(f_i \geqslant t). \quad \bowtie$$

» $\left\{\sup_{i=1,\dots,n} f_i \geqslant t\right\} = \bigcup_{i=1}^{n} \left\{f_i \geqslant t\right\}$. «

Bemerkung. Betrachte $\xi_1, \dots, \xi_n, -\xi_1, \dots, -\xi_n$ und das union bound für die Summe über 1. Block $(=: f_1)$ + 2. Block $(=: f_2)$. Somit ist

$$\left|\frac{1}{n}\sum_{i=1}^{n} \xi_i\right| = \sup\left\{f_1, f_2\right\}$$

und folglich gilt

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} \xi_i\right| \geqslant \sqrt{\frac{2\sigma^2\tau}{n}} + \frac{2B\tau}{3n}\right) \leqslant 2e^{-\tau}. \quad \multimap$$

„Zweiseitige Bernstein-Ungleichung“ (Hoeffdings analog!)

4 Empirische Risikominimierung (ERM)

Motivation. Sei D das empirische Wahrscheinlichkeitsmaß,

$$\mathcal{R}_{L,D}(f) \coloneqq \frac{1}{n}\sum_{i=1}^{n} L(x_i, y_i, f(x_i)) \overset{n\to\infty}{\longrightarrow} \mathcal{R}_{L,P}(f).$$

Ziel des Lernens: Finde f mit $\mathcal{R}_{L,P}(f)$ klein.
Heuristik von ERM: Minimiere Schätzung $\mathcal{R}_{L,D}(\cdot)$ von $\mathcal{R}_{L,P}(\cdot)$. ⊸

4-1 ERM über endliche Funktionenmengen

4-1.1 **Definition** *Sei L eine Verlustfunktion, $\emptyset \neq \mathcal{F} \subset \mathcal{L}_0(X)$. Dann heißt eine Lernmethode $\mathcal{L}$ ERM über $\mathcal{F}$, wenn*

$$\mathcal{R}_{L,D}(f_D) = \inf_{f\in\mathcal{F}} \mathcal{R}_{L,D}(f), \qquad \forall\ n \geqslant 1,\ D \in (X\times Y)^n \text{ und } f_D \in \mathcal{F}. \quad \bowtie$$

Bemerkungen. A. Im Allgemeinen gibt es keine ERM.

B. Im Allgemeinen ist ERM nicht eindeutig.

C. Für $\mathcal{F} = \mathcal{L}_0(X)$ oder $\mathcal{F} = \mathcal{L}_\infty(X)$ erhält man im Allgemeinen overfitted ERM.

D. Für „kleine“ $\mathcal{F}$ erhält man im Allgemeinen underfitted ERM.

Im Folgenden sei $\mathcal{F}$ endlich.

4-1.2 **Orakelungleichung I** *Sei L eine Verlustfunktion und $\mathcal{F} \subset \mathcal{L}_0(X)$ nicht leer und endlich. Ferner sei $B > 0$ mit*

$$L(x, y, f(x)) \leqslant B, \qquad \forall\ (x, y) \in X\times Y, \quad f \in \mathcal{F}.$$

Für den Wert des Orakels

$$\mathcal{R}^*_{L,P,\mathcal{F}} := \inf_{f\in\mathcal{F}} \mathcal{R}_{L,P}(f)$$

und ERM gilt für $n \geqslant 1$ und $\tau > 0$,

$$P^n\left(\left\{D \in (X\times Y)^n : \mathcal{R}_{L,P}(f_D) \leqslant \mathcal{R}^*_{L,P,\mathcal{F}} + B\sqrt{\frac{2\tau + 2\log(2\left|\mathcal{F}\right|)}{n}}\right\}\right) \geqslant 1 - e^{-\tau}. \quad \rightthreetimes$$

» Für $\delta > 0$ existiert ein $f_\delta \in \mathcal{F}$ mit $\mathcal{R}_{L,P}(f_\delta) \leqslant \mathcal{R}^*_{L,P,\mathcal{F}} + \delta$. Damit ist

$$\begin{aligned}\mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P,\mathcal{F}} &\leqslant \mathcal{R}_{L,P}(f_D) - \mathcal{R}_{L,D}(f_D) + \mathcal{R}_{L,D}(f_D) - \mathcal{R}_{L,P}(f_\delta) + \delta\\ &\leqslant \left|\mathcal{R}_{L,P}(f_D) - \mathcal{R}_{L,D}(f_D)\right| + \left|\mathcal{R}_{L,P}(f_\delta) - \mathcal{R}_{L,D}(f_\delta)\right| + \delta\\ &\leqslant 2\sup_{f\in\mathcal{F}}\left|\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)\right| + \delta.\end{aligned}$$

Für „$\delta \to 0$“ folgt

$$\mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P,\mathcal{F}} \leqslant 2\sup_{f\in\mathcal{F}}\left|\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)\right|$$

Damit folgt

$$\begin{aligned}&P^n\left(\left\{D : \mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P,\mathcal{F}} \geqslant B\sqrt{\frac{2\tau}{n}}\right\}\right)\\ \leqslant\ &P^n\left(\left\{D : \sup_{f\in\mathcal{F}}\left|\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}\right| \geqslant B\sqrt{\frac{\tau}{2n}}\right\}\right)\\ \overset{\text{union bound}}{\leqslant}\ &\sum_{f\in\mathcal{F}} P^n\left(\left\{D : \left|\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}\right| \geqslant B\sqrt{\frac{\tau}{2n}}\right\}\right).\end{aligned}$$

Setzen wir $\mathbf{E}_P\xi_1 = \mathcal{R}_{L,P}(f)$ und $\xi_i = L(x_i, y_i, f(x_i))$, so ist $\mathcal{R}_{L,D}(f) = \frac{1}{n}\sum_{i=1}^n \xi_i$ und mit der 2-seitigen Hoeffding Ungleichung folgt

$$\ldots \leqslant \sum_{f\in\mathcal{F}} 2e^{-\tau} = 2\left|\mathcal{F}\right| e^{-\tau} = e^{\log(2|\mathcal{F}|)-\tau}.$$

Mit $-t' = \log(2\left|\mathcal{F}\right|) - \tau$ folgt die Behauptung. «

Idee: Falls $\left|\mathcal{F}\right| = \infty$ finde $\mathcal{F}_\varepsilon \subset \mathcal{F}$ endlich, das $\mathcal{F}$ gut „approximiert“.

4-2 ERM über unendlichen Funktionenmengen

4-2.1 **Definition** *Sei (T,d) ein metrischer Raum (MR) und $\varepsilon > 0$.*

(a) *$S \subset T$ heißt ε-Netz falls*

$$\forall\ t \in T\ \exists\ s \in S : d(t,s) \leqslant \varepsilon.$$

(b) *Die Überdeckungszahl (covering number) ist definiert als*

$$\mathcal{N}(T,d,\varepsilon) \coloneqq \inf\left\{ n \geqslant 1 :\ \exists\ s_1,\dots,s_n \in T : T \subset \bigcup_{i=1}^{n} B_d(s_i,\varepsilon) \right\}.$$

D.h. $S = \{s_1,\dots,s_n\}$ ist ε-Netz. Damit entspricht $\mathcal{N}(T,d,\varepsilon)$ der Größe des kleinsten ε-Netzes von T.

(c) *Sei $T \subset E$, E normierter Raum. Dann ist*

$$\mathcal{N}(T, \|\cdot\|_E, \varepsilon) \coloneqq \mathcal{N}(T, d_E, \varepsilon), \qquad d_E(x,x') = \left\|x - x'\right\|_E. \quad \bowtie$$

Interpretation. Ist T kompakt, dann ist $\mathcal{N}(T,d,\varepsilon) < \infty$ für jedes $\varepsilon > 0$. Genauer ist T genau dann präkompakt, wenn $\mathcal{N}(T,d,\varepsilon) < \infty$ für jedes $\varepsilon > 0$.

Da für $\varepsilon < \varepsilon'$ gilt $\mathcal{N}(T,d,\varepsilon') < \mathcal{N}(T,d,\varepsilon)$, stellt das Wachstumsverhalten von $\mathcal{N}(T,d,\varepsilon)$ für $\varepsilon \downarrow 0$ ein quantitatives Maß für die Kompaktheit dar.

$\mathcal{N}(T,d,\varepsilon) \to \infty$ für $\varepsilon \to 0$ genau dann, wenn T unendlich. $\multimap$

4-2.2 **Definition** *Sei (T,d) ein metrischer Raum und $n \geqslant 1$.*

(a) *Die n-te (b-a-dische) Entropiezahl von T ist*

$$e_n(T,d) = \inf\left\{ \varepsilon > 0 :\ \exists\ s_1,\dots,s_{2^{n-1}} \in T : T \subset \bigcup_{i=1}^{2^{n-1}} B_d(s_i,\varepsilon) \right\}.$$

(b) *$e_n(T, \|\cdot\|_E)$ für $T \subset (E, \|\cdot\|_E)$ ist wie üblich definiert.*

(c) *Sei $S : E \to F$ ein stetige linearer Operator zwischen den normierten Räumen E und F und es sei $B_E \coloneqq \{x \in E :\ \|x\|_E \leqslant 1\}$, dann schreibe*

$$e_n(S) = e_n(S : E \to F) \coloneqq e_n(SB_E, \|\cdot\|_F). \quad \bowtie$$

Bemerkungen (ohne Beweis). A. $e_1(S) = \|S\|$.

B. $e_{n+1}(S) \leqslant e_n(S)$ für $n \geqslant 1$.

C. $\lim_{n\to\infty} e_n(S) = 0$ genau dann, wenn S kompakt.

D. $e_{n+k-1}(S+T) \leqslant e_n(S) + e_k(T)$.

E. $e_{n+k-1}(S \circ T) \leqslant e_n(S) e_k(T)$.

F. Sei $d = \operatorname{rank} S = \dim SE$. Dann ist $d < \infty$ genau dann, wenn

$$\exists\, c_1, c_2 : c_1 2^{-\frac{n-1}{d}} \leqslant e_n(S) \leqslant c_2 2^{-\frac{n-1}{d}}.$$

G. Sei $X \subset \mathbb{R}^d$ offen und beschränkt und

$$C_b^m(X) := \left\{ f \in C^m(X \to \mathbb{R}) : \max_{|\alpha|\leqslant m} \left\|\partial^\alpha f\right\|_\infty < \infty \right\}$$

versehen mit der Norm

$$\left\|f\right\|_{C_b^m(X)} = \max_{|\alpha|\leqslant m} \left\|\partial^\alpha f\right\|_\infty.$$

Dann ist $e_n(\mathrm{id} : C_b^m(X) \to C_b(X)) \leqslant C_{m,d}(X) n^{-\frac{m}{d}}$ und die Abschätzung ist scharf, d.h. es gibt ein $\tilde{C}_{m,d}(X)$ mit $e_n(\ldots) \geqslant \tilde{C}_{m,d}(x) n^{-\frac{m}{d}}$. Weiterhin ist

$$C_{m,d}(r \cdot X) = r^m C_{m,d}(X), \qquad r \geqslant 1. \quad \multimap$$

4-2.3 **Lemma** *Sei (T,d) MR und $a, q > 0$ mit*

$$e_n(T,d) \leqslant a n^{-\frac{1}{q}}, \qquad n \geqslant 1.$$

Dann gilt

$$\log(\mathcal{N}(T,d,\varepsilon)) \leqslant \log(4) \left(\frac{a}{\varepsilon}\right)^q, \qquad \varepsilon > 0. \quad \bowtie$$

» Für $\varepsilon > a$ ist $\mathcal{N}(T,d,\varepsilon) = 1$, d.h. $\log(\mathcal{N}(T,d,\varepsilon)) = 0$ und es ist nichts zu beweisen. Sei $\delta > 0$ und $\varepsilon \in (0,a]$, dann existiert ein $n \geqslant 1$ mit

$$a(1+\delta)(n+1)^{-\frac{1}{q}} \leqslant \varepsilon \leqslant a(1+\delta)n^{-\frac{1}{q}}. \tag{*}$$

Da $e_n(T,d) < a(1+\delta)n^{-\frac{1}{q}}$, gibt es ein $a(1+\delta)n^{-\frac{1}{q}}$-Netz der Größe 2^{n-1}, d.h.

$$\mathcal{N}(T,d,a(1+\delta)n^{-\frac{1}{q}}) \leqslant 2^{n-1}.$$

Nun ist

$$a(1+\delta)n^{-\frac{1}{q}} = a(1+\delta)(n+1)^{-\frac{1}{q}} \underbrace{\left(\frac{n+1}{n}\right)^{\frac{1}{q}}}_{\leqslant 2} \leqslant 2^{\frac{1}{q}} a(1+\delta)(n+1)^{-\frac{1}{q}}$$
$$\overset{(*)}{\leqslant} 2^{\frac{1}{q}}\varepsilon.$$

Und ferner folgt mit der rechten Seite von (*),

$$n \leqslant \left(\frac{a(1+\delta)}{\varepsilon}\right)^q.$$

Somit ist

$$\log(\mathcal{N}(T,d,2^{1/q}\varepsilon)) \leqslant \log(\mathcal{N}(T,d,a(1+\delta)n^{-1/q})) \leqslant \log(2^{n-1})$$
$$\leqslant n\log 2 \leqslant (\log 2)\left(\frac{a(1+\delta)}{\varepsilon}\right)^q$$

Für $\delta \to 0$ und $2^{1/q}\varepsilon = \tilde{\varepsilon}$ folgt die Behauptung. «

4-2.4 **Orakelungleichung II** *Sei $L : X \times Y \times \mathbb{R} \to [0,\infty)$ eine lokal lipschitz-stetige Verlustfunktion und P ein W-Maß auf $(X \times Y)$. Ferner sei*

$$\mathcal{F} \subset \mathcal{L}_\infty(X) = \left\{f : X \to \mathbb{R} : f \text{ messbar und beschränkt}\right\}$$

für die $B > 0$ und $M > 0$ existieren, so dass

$$\|f\|_\infty < M, \qquad \text{für alle } f \in \mathcal{F},$$
$$L(x,y,f(x)) \leqslant B, \qquad \text{für alle } (x,y) \in (X \times Y),\ f \in \mathcal{F}.$$

Dann gilt für ERM über f,

$$P^n\left(\left\{D : \mathcal{R}_{L,P}(f_D) < \mathcal{R}^*_{L,P,\mathcal{F}} + B\sqrt{\frac{2\tau + 2\log\left(2\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)\right)}{n}} + 4\varepsilon\,|L|_{M,1}\right\}\right)$$
$$\geqslant 1 - e^{-\tau}, \qquad \textit{für alle } n \geqslant 1, \varepsilon > 0 \textit{ und } \tau > 0. \quad \rtimes$$

Beachte, dass falls $\log(\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)) \leqslant \mu(\varepsilon)$ für $\varepsilon > 0$, der Ausdruck nach ε optimiert werden kann. Somit erhält man eine Abschätzung in $n \geqslant 1$ und $\tau > 0$.

» Fixiere $\varepsilon > 0$ und ein minimales ε-Netz $\mathcal{F}_\varepsilon$ von $\mathcal{F}$ bezüglich $\|\cdot\|_\infty$, d.h.

$$\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) = \left|\mathcal{F}_\varepsilon\right|$$

und

$$\forall\ f \in \mathcal{F}\ \exists\ g \in \mathcal{F}_\varepsilon \text{ mit } \left\|f - g\right\|_\infty \leqslant \varepsilon.$$

Somit ist für solches f, g

$$\begin{aligned}
&\left|\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)\right| \\
&\leqslant \left|\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,P}(g)\right| + \left|\mathcal{R}_{L,P}(g) - \mathcal{R}_{L,D}(g)\right| + \left|\mathcal{R}_{L,D}(g) - \mathcal{R}_{L,D}(f)\right| \\
&\overset{??}{\leqslant} 2\varepsilon\,|L|_{M,1} + \left|\mathcal{R}_{L,P}(g) - \mathcal{R}_{L,D}(g)\right|.
\end{aligned}$$

D.h.

$$\sup_{f\in\mathcal{F}}\left|\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)\right| \leqslant 2\varepsilon\,|L|_{M,1} + \sup_{g\in\mathcal{F}_\varepsilon}\left|\mathcal{R}_{L,P}(g) - \mathcal{R}_{L,D}(g)\right|.$$

Der Beweis von 4-1.2 zeigte außerdem, dass

$$\begin{aligned}
\mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P,\mathcal{F}} &\leqslant 2\sup_{f\in\mathcal{F}}\left|\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)\right| \\
&\leqslant 4\varepsilon\,|L|_{M,1} + 2\sup_{g\in\mathcal{F}_\varepsilon}\left|\mathcal{R}_{L,P}(g) - \mathcal{R}_{L,D}(g)\right|.
\end{aligned}$$

Anwendung der Hoeffding Ungleichung und es union bounds liefert,

$$\begin{aligned}
&P^n\left(\left\{D : \mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P,F} \geqslant B\sqrt{\frac{2\tau}{n}} + 4\varepsilon\,|L|_{M,1}\right\}\right)\\
\leqslant &P^n\left(\left\{D : \sup_{g\in\mathcal{F}_\varepsilon}\left|\mathcal{R}_{L,P}(g) - \mathcal{R}_{L,D}(g)\right| \geqslant B\sqrt{\frac{\tau}{2n}}\right\}\right)\\
\leqslant &2\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)\, e^{-\tau} = \exp\left(\log(2\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)) - \tau\right).
\end{aligned}$$

Die Behauptung folgt nun mit einer Variablentransformation

$$-\tau' = \log(2\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)) - \tau. \quad «$$

5 Reproduzierende Kernhilberträume (RKHS)

Es gibt zwei schöne Klassen von Banachräumen

- Funktionenräume.
- Hilberträume.

RKHS sind beides, der $L^2([0,1])$ übrigens *kein* Funktionenraum!

5-1 Kerne und Beispiele

5-1.1 **Definition** *Eine Funktion $k : X \times X \to \mathbb{R}$ für die ein Hilbertraum H und ein $\Phi : X \to H$ existieren, so dass*

$$k(x,x') = \big\langle \Phi(x), \Phi(x') \big\rangle_H, \qquad \forall\, x,x' \in X,$$

heißt Kern. ⋊

Bemerkungen. A. Im Allgemeinen sind H und Φ nicht eindeutig.
Betrachte z.B. $k(x,x') = x \cdot x'$ für x,x', so ist k ein Kern, denn

$$H_1 = \mathbb{R}, \quad \Phi_1 = \mathrm{id}_{\mathbb{R}}, \quad \big\langle \Phi_1(x), \Phi_1(x') \big\rangle = x \cdot x'.$$
$$H_2 = \mathbb{R}^2, \quad \Phi_2(x) = \frac{1}{\sqrt{2}}\,(x,x)$$

B. H heißt Featurespace (Merkmalraum), Φ heißt Featuremap. ⊸

5-1.2 **Lemma** *Seien $f_n : X \to \mathbb{R}$, so dass $(f_n(x))_{n\geqslant 1} \in l_2$ für alle $x \in X$, so definiert*

$$k(x,x') = \sum_{n\geqslant 1} f_n(x) f_n(x'), \qquad x,x' \in X$$

einen Kern. ⋊

» k ist wohl definiert, denn nach Cauchy-Schwarz ist

$$\sum_{n\geqslant 1} \left|f_n(x) f_n(x')\right| \leqslant \left\|(f_n(x))\right\|_{l_2} \left\|(f_n(x'))\right\|_{l_2} < \infty.$$

Sei $H = l^2$ und $\Phi : X \to H$ gegeben durch

$$\Phi(x) = (f_n(x))_{n\geqslant 1}, \qquad x \in X.$$

Dann ist $\langle \Phi(x), \Phi(x')\rangle = k(x,x')$ und daher k ein Kern. «

5-1.3 **Lemma** *Ist k ein Kern auf X und $A : X' \to X$ eine Abbildung, dann definiert*

$$k'(x,x') = k(A(x), A(x')), \qquad x,x' \in X$$

einen Kern auf X'. ⋊

» Sei H ein Featurespace von k und $\Phi : X \to H$ die zugehörige Featuremap. Setze

$$\Psi(x) = \Phi(A(x)),$$

so gilt

$$k'(x,x') = k(A(x), A(x')) = \left\langle \Phi(A(x)), \Phi(A(x'))\right\rangle_H = \left\langle \Psi(x), \Psi(x')\right\rangle_H,$$

d.h. Ψ ist Featuremap von k'. «

5-1.4 **Lemma** *Sind k_1 und k_2 Kerne auf X und $\alpha \geqslant 0$. Dann sind auch $k_1 + k_2$ und $\alpha \cdot k_1$ Kerne auf X.* ⋊

» Der Beweis ist eine leichte Übung. «

5-1.5 **Lemma** *Sind k_1 und k_2 Kerne auf X, so ist auch $k_1 \cdot k_2$ ein Kern auf X.* ⋊

» Der Beweis erfolgt unter Verwendung des Tensorprodukts von Hilberträumen. Wir überspringen diesen, da wir die Aussage nicht benötigen. «

5-1.6 **Lemma für Polynom-Kerne** *Seien $m \geqslant 0$, $d \geqslant 1$ natürliche Zahlen und $0 \leqslant c \in \mathbb{R}$. Dann definiert*

$$k(x,x') = \left(\langle x,x'\rangle_{\mathbb{R}^d} + c\right)^m, \qquad x,x' \in \mathbb{R}^d$$

einen Kern auf $\mathbb{R}^d$. ⋊

» Nach Konstruktion existiert ein Polynom p der Ordnung m mit

$$k(x,x') = p(\langle x,x'\rangle), \qquad x,x' \in \mathbb{R}^d,$$

wobei p nur nichtnegative Koeffizienten besitzt. Es genügt daher zu zeigen, dass die Monome

$$(x,x') \mapsto \langle x,x'\rangle^m$$

einen Kern definieren. Nun ist

$$\langle x,x'\rangle^m_{\mathbb{R}^d} = \left(\sum_{i=1}^d x_i x_i'\right)^m$$

also können wir Lemma 5-1.2 anwenden mit $f_i(x) = \pi_i(x)$ der Projektion auf die i-te Koordinate und deren Potenzen. «

5-1.7 **Lemma für Taylor-Kerne** *Sei $f : (-r,r) \to \mathbb{R}$ mit $r \in (0,\infty]$,*

$$f(x) = \sum_{n=0}^\infty a_n t^n, \qquad t \in (-r,r).$$

Falls $a_n \geqslant 0$ für alle $n \geqslant 0$, definiert

$$k(x,x') = f(\langle x,x'\rangle_{\mathbb{R}^d}) = \sum_{n=0}^\infty a_n \langle x,x'\rangle^n_{\mathbb{R}^d}$$

einen Kern auf $\sqrt{r}\mathring{B}_{\mathbb{R}^d} = \left\{x \in \mathbb{R}^d : \|x\|_2 < \sqrt{r}\right\}$. ⋊

» Für $x, x' \in \sqrt{r}\mathring{B}_{\mathbb{R}^d}$ ist

$$\left|\left\langle x, x'\right\rangle_{\mathbb{R}^d}\right| \leqslant \|x\|_2 \left\|x'\right\|_2 < r$$

und daher die Konstruktion sinnvoll. Ferner seien

$$c_{j_1,\dots,j_d} = \frac{n!}{\prod_{i=1}^d j_i!}, \qquad \sum_{j=1}^d j_i = n,$$

dann gilt

$$\begin{aligned} k(x,x') &= \sum_{n=0}^{\infty} a_n \left(\sum_{i=1}^d x_i x_i'\right)^n = \sum_{n=0}^{\infty} a_n \sum_{\substack{j_1,\dots,j_d \geqslant 0 \\ j_1+\dots+j_d=n}} c_{j_1,\dots,j_d} \prod_{i=1}^d x_i^{j_i} (x_i')^{j_i} \\ &= \sum_{n=0}^{\infty} a_n \sum_{\substack{j_1,\dots,j_d \geqslant 0 \\ j_1+\dots+j_d=n}} c_{j_1,\dots,j_d} \left(\prod_{i=1}^d x_i^{j_i}\right)\left(\prod_{i=1}^d (x_i')^{j_i}\right). \end{aligned}$$

Setze $\Phi : X \to l_2(\mathbb{N}_0^d)$, wobei $l_2(I) \coloneqq \left\{(a_i)_{i\in I} : \|(a_i)\|_2 < \infty\right\}$, dann ist

$$\Phi(x) = \left(\sqrt{a_{j_1+\dots+j_d} c_{j_1,\dots,j_d}} \prod_{i=1}^d x_i^{j_i}\right)_{(j_1,\dots,j_d)\in\mathbb{N}_0^d}$$

eine Featuremap, denn

$$\begin{aligned} &\sum_{n=0}^{\infty} a_n \sum_{\substack{j_1,\dots,j_d \geqslant 0 \\ j_1+\dots+j_d=n}} c_{j_1,\dots,j_d} \left(\prod_{i=1}^d x_i^{j_i}\right)\left(\prod_{i=1}^d (x_i')^{j_i}\right) \\ &= \sum_{n=0}^{\infty} \sum_{\substack{j_1,\dots,j_d \geqslant 0 \\ j_1+\dots+j_d=n}} \sqrt{a_{j_1+\dots+j_d} c_{j_1,\dots,j_d}} \left(\prod_{i=1}^d x_i^{j_i}\right) \sqrt{a_{j_1+\dots+j_d} c_{j_1,\dots,j_d}} \left(\prod_{i=1}^d (x_i')^{j_i}\right) \\ &= \left\langle \Phi(x), \Phi(x')\right\rangle_{l^2(\mathbb{N}_0^d)}. \quad « \end{aligned}$$

5-1.8 **Bsp** *Exponential Kern.* Die Abbildung

$$k(x,x') = \exp\left(\left\langle x, x'\right\rangle_{\mathbb{R}^d}\right), \qquad x, x' \in \mathbb{R}^d$$

ist ein Kern. ■

5-1.9 **Bsp** *Gauß Kern* oder *Gauß'scher RBF-Kern.* Die Abbildung

$$k_\sigma(x,x') = \exp\left(-\sigma^2 \left\|x-x'\right\|_{\mathbb{R}^d}^2\right), \qquad x,x' \in \mathbb{R}^d$$

ist ein Kern.

» Eine einfache Rechnung zeigt

$$\begin{aligned}\left\|x-x'\right\|^2 &= \left\langle x-x', x-x'\right\rangle = \langle x,x\rangle - 2\left\langle x,x'\right\rangle + \left\langle x',x'\right\rangle \\ &= \|x\|^2 + \left\|x'\right\|^2 - 2\left\langle x,x'\right\rangle.\end{aligned}$$

Setzen wir $h(x) = \exp\left(\sigma^2 \|x\|_{\mathbb{R}^d}^2\right)$, so können wir den Gauß-Kern schreiben als

$$\exp\left(-\sigma^2 \left\|x-x'\right\|_{\mathbb{R}^d}^2\right) = \frac{\exp\left(2\sigma^2 \langle x,x'\rangle_{\mathbb{R}^d}\right)}{h(x)h(x')}.$$

Ferner ist $(x,x') \mapsto \exp(2\sigma^2 \langle x,x'\rangle)$ ein Kern nach BSP 5-1.8 und Lemma 5-1.3. Betrachten wir einen zugehörigen Featurespace H und eine Featuremap Φ und setzen

$$\Phi_\sigma = \frac{1}{h}\Phi : X \to H,$$

so gilt

$$\begin{aligned}\left\langle \Phi_\sigma(x), \Phi_\sigma(x')\right\rangle &= \left\langle \frac{\Phi(x)}{h(x)}, \frac{\Phi(x')}{h(x')}\right\rangle = \frac{1}{h(x)h(x')}\left\langle \Phi(x), \Phi(x')\right\rangle \\ &= \frac{\exp(2\sigma^2\langle x,x'\rangle)}{h(x)h(x')}. \quad \text{«} \quad \blacksquare\end{aligned}$$

5-1.10 **Definition** *Eine Funktion $k : X \times X \to \mathbb{R}$ heißt positiv definit, falls*

$$\sum_{i,j=1}^n \alpha_i\alpha_j k(x_i,x_j) \geqslant 0$$

für alle $n \geqslant 1$ und $\alpha_1,\dots,\alpha_n \in \mathbb{R}$, d.h. die Grammatritzen $(k(x_i,x_j))_{i,j=1}^n$ sind positiv definit.

k heißt strikt positiv definit, falls für alle paarweise verschiedenen $x_1,\dots,x_n$ und $\alpha_1,\dots,\alpha_n \in \mathbb{R}$ gilt

$$\sum_{i,j=1}^n \alpha_i\alpha_j k(x_i,x_j) = 0 \Rightarrow \alpha_1 = \dots = \alpha_n = 0.$$

k heißt symmetrisch falls $k(x,x') = k(x',x)$ für alle $x,x' \in X$. ⋊

5-1.11 **Satz** *$k : X \times X \to \mathbb{R}$ ist genau dann ein Kern, wenn k positiv definit und symmetrisch.* ⋊

» „⇒“: Sei $\Phi : X \to H$ eine Featuremap von k, dann gilt

$$k(x, x') = \left\langle \Phi(x), \Phi(x') \right\rangle$$

und jedes Skalarprodukt ist positiv definit und symmetrisch

$$\begin{aligned} 0 \leqslant \langle h, h \rangle &= \left\langle \sum_{i=1}^{n} \alpha_i \Phi(x_i), \sum_{j=1}^{n} \alpha_j \Phi(x_j) \right\rangle \\ &= \sum_{i,j=1}^{n} \alpha_i \alpha_j \left\langle \Phi(x_i), \Phi(x_j) \right\rangle = \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j). \end{aligned}$$

„⇐“: Wir müssen Featurespace und -map konstruieren. Dazu setzen wir

$$H_{\text{pre}} = \left\{ \sum_{i=1}^{n} \alpha_i k(x_i, \cdot) \,:\, n \geqslant 1,\ \alpha_1, \ldots, \alpha_n \in \mathbb{R},\ x_1, \ldots, x_n \in X \right\}$$

und für $f = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)$ und $g = \sum_{j=1}^{n} \beta_j k(x_j, \cdot)$ setze in H_{pre},

$$\left\langle f, g \right\rangle_{H_{\text{pre}}} = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \beta_j k(x_i, x_j).$$

Weiterhin ist

$$\begin{aligned} \left\langle f, g \right\rangle_{H_{\text{pre}}} &= \sum_{j=1}^{m} \beta_j f(x'_j), \\ \left\langle f, g \right\rangle_{H_{\text{pre}}} &= \sum_{i=1}^{m} \alpha_i g(x_i), \end{aligned}$$

also ist $\langle \cdot, \cdot \rangle_{H_{\text{pre}}}$ von der speziellen Darstellung von f oder g unabhängig und daher wohldefiniert.

Offensichtlich ist $\langle \cdot, \cdot \rangle_{H_{\text{pre}}}$ bilinear und symmetrisch, da k symmetrisch.

Zu zeigen ist also noch, dass $\left\langle f, f \right\rangle = 0 \Rightarrow f = 0$. Wir zeigen dazu, dass $\langle \cdot, \cdot \rangle_{H_{\text{pre}}}$ die Cauchy-Schwarz'sche Ungleichung erfüllt.

1. Fall. Ist $\langle f, f\rangle = 0$ und $\langle g, g\rangle = 0$, dann ist

$$0 \leqslant \langle f+g, f+g\rangle = \langle f, f\rangle + 2\langle f, g\rangle + \langle g, g\rangle = 2\langle f, g\rangle,$$
$$0 \leqslant \langle f-g, f-g\rangle = \langle f, f\rangle - 2\langle f, g\rangle + \langle g, g\rangle = -2\langle f, g\rangle.$$

Folglich ist $\langle f, g\rangle = 0$ und die Cauchy-Schwarz'sche Ungleichung folgt.

2. Fall. Ohne Einschränkung ist $\langle g, g\rangle > 0$, setzen wir also $\alpha = -\frac{\langle f,g\rangle}{\langle g,g\rangle}$. Dann folgt

$$0 \leqslant \langle f+\alpha g, f+\alpha g\rangle = \ldots = \langle f, f\rangle - \frac{\langle f, g\rangle^2}{\langle g, g\rangle}.$$

Sei nun $f \in H_{\text{pre}}$ mit $\langle f, f\rangle = 0$ und $f = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$. Dann ist

$$\begin{aligned}\big|f(x)\big|^2 &= \left|\sum_{i=1}^n \alpha_i k(x_i, \cdot)\right|^2 = \big|\langle f, k(x, \cdot)\rangle\big|^2 \\ &\leqslant \langle f, f\rangle \langle k(x, \cdot), k(x, \cdot)\rangle = 0.\end{aligned}$$

Folglich ist $\langle \cdot, \cdot\rangle_{H_{\text{pre}}}$ ein Skalarprodukt und damit H_{pre} ein Prähilbertraum.

Zu H_{pre} existiert eine Vervollständigung H mit einer isometrischen Einbettung $J : H_{\text{pre}} \to H$. Setze nun $\Phi : X \to H$, $x \mapsto JK(x, \cdot)$, dann gilt

$$\begin{aligned}\big\langle \Phi(x), \Phi(x')\big\rangle_H &= \big\langle Jk(x, \cdot), Jk(x', \cdot)\big\rangle_H = \big\langle k(x, \cdot), k(x', \cdot)\big\rangle_{H_{\text{pre}}} \\ &= k(x, x'). \quad «\end{aligned}$$

5-1.12 **Korollar** *Sind k_n, $n \geqslant 1$ Kerne auf X und gibt es eine Funktion $k : X \times X \to \mathbb{R}$ mit $k_n(x, x') \to k(x, x')$ für alle $x, x' \in X$, so ist k ein Kern.* ⋊

» Alle k_n sind symmetrisch und daher ist es die Grenzfunktion k auch. Weiterhin gilt

$$0 \leqslant \sum_{i,j=1}^d \alpha_i \alpha_j k_n(x_i, x_j) \to \sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j)$$

und daher ist k positiv definit. «

5-2 RKHS (Reproduzierende Kern-Hilbert-Räume)

Bisher waren Featurespace und -map nicht eindeutig, es gibt aber eine kanonische Wahl, den RKHS.

5-2.1 **Definition** (a) *Ein Hilbertraum H, der aus Funktionen $f : X \to \mathbb{R}$ besteht, heißt Hilbertfunktionenraum über X. (HFS)*

(b) *Ein HFS H über X heißt RKHS, falls die Dirac-Funktionale*

$$\delta_x : H \to \mathbb{R}, \quad f \mapsto f(x)$$

für alle $x \in X$ stetig sind.

(c) *Eine Funktion $k : X \times X \to \mathbb{R}$ heißt reproduzierender Kern eines HFS über X, falls für jedes $x \in X$ $k(x, \cdot) \in H$ und die reproduzierende Eigenschaft erfüllt ist,*

$$f(x) = \langle f, k(x, \cdot) \rangle, \qquad \text{für alle } f \in H,\ x \in X. \quad \bowtie$$

Bemkerungen. A. $L_2([0,1])$ ist kein RKHS (nicht mal ein HFS).

B. Norm-Konvergenz in einem RKHS H impliziert punktweise Konvergenz.

» Sei $x \in X$ und gelte $\|f_n - f\|_H \to 0$, dann gilt

$$f(x) = \delta_x(f) = \delta_x\left(\lim_{n\to\infty} f_n\right) = \lim_{n\to\infty} \delta_x(f_n) = \lim_{n\to\infty} f_n(x). \quad «$$

C. (c) wurde für den im Beweis von Satz 5-1.11 konstruierten Hilbertraum schon „fast“ gezeigt. ⊸

5-2.2 **Lemma** *Sei H ein HFS über X und $k : X \times X \to \mathbb{R}$ ein reproduzierender Kern von H. Dann gelten*

1) H ist ein RKHS.

2) k ist ein Kern, H ist ein FS von k und

$$\Phi : X \to H, \qquad x \mapsto k(x, \cdot)$$

die kanonische Featuremap von k. $\bowtie$

» 1) Eine direkte Rechung zeigt

$$\left|\delta_x(f)\right| = \left|f(x)\right| = \left|\left\langle f, k(x,\cdot)\right\rangle_H\right| \leqslant \left\|f\right\|_H \|k(x,\cdot)\|_H .$$

Somit ist δ_x stetig und $\|\delta_x\| \leqslant \|k(x,\cdot)\|$.

2) Für $x' \in X$ fest schreibe $f \coloneqq k(x',\cdot)$. Dann ist $f \in H$ per definitionem und

$$\begin{aligned}\left\langle \Phi(x), \Phi(x')\right\rangle_H = \left\langle k(x,\cdot), f\right\rangle_H = \left\langle f, k(x,\cdot)\right\rangle_H = f(x) = k(x',x)\\ = \left\langle \Phi(x'), \Phi(x)\right\rangle . \quad «\end{aligned}$$

5-2.3 **Existenz und Eindeutigkeit des reproduzierenden Kerns** *Sei H ein RKHS über X. Dann gilt*

$$k : X \times X \to \mathbb{R}, \qquad k(x,x') = \langle \delta_x, \delta_{x'}\rangle ,$$

ist der einzige reproduzierende Kern von H. Ist ferner $(e_i)_{i\in I}$ eine Orthonormalbasis (ONB), so gilt

$$k(x,x') = \sum_{i\in I} e_i(x) e_i(x'), \qquad x, x' \in X. \quad \bowtie$$

» *Existenz.* Der Darstellungssatz von Fréchet-Riesz besagt, dass

$$I : H \to H', \qquad f \mapsto \left\langle f, \cdot\right\rangle_H$$

ein isometrischer Isomorphismus ist, d.h. $H = H'$. Sei $J = I^{-1} : H' \to H$, so ist auch J ein isometrischer Isomorphismus und für $f \in H$ und $\omega \in H'$ gilt

$$\omega(f) = \left\langle J\omega, f\right\rangle_H . \tag{*}$$

Für x, x' folgt

$$k(x,x') = \langle \delta_x, \delta_{x'}\rangle_{H'} = \langle J\delta_x, J\delta_{x'}\rangle_H \overset{(*)}{=} \delta_x(J\delta_{x'}) = J\delta_{x'}(x).$$

Folglich ist $J\delta_{x'} = k(\cdot, x')$ und damit

$$f(x') - \delta_{x'}(f) = \left\langle f, J\delta_{x'}\right\rangle_H = \left\langle f, k(\cdot,x')\right\rangle_H = \left\langle f, k(x',\cdot)\right\rangle_H ,$$

d.h. k ist reproduzierender Kern von H.

Eindeutigkeit. Sei $\tilde{k}$ ein beliebiger reproduzierender Kern von H.

Für $x' \in X$ gilt $k(x', \cdot) \in H$. Für die ONB $(e_i)_{i\in\mathcal{I}}$ folgt

$$k(x', \cdot) = \sum_{i\in\mathcal{I}} \left\langle k(x', \cdot), e_i \right\rangle_H e_i = \sum_{i\in\mathcal{I}} e_i(x') e_i,$$

wobei

$$k(x', x) = \sum_{i\in\mathcal{I}} e_i(x') e_i(x)$$

punktweise existiert. Insbesondere ist $\tilde{k}$ der einzige reproduzierende Kern von H und die Darstellung mithilfe der ONB $(e_i)_{i\in\mathcal{I}}$ hängt nicht von der Wahl der ONB ab. «

5-2.4 **Satz** *Sei k ein Kern über X mit Featurespace H_0 und Featuremap $\Phi_0 : X \to H_0$. Dann gelten*

1) $H \coloneqq \left\{ f : X \to \mathbb{R} \,:\, \exists\, \omega \in H_0 : f(x) = \langle \omega, \Phi_0(x) \rangle_{H_0} \;\forall\, x \in X \right\}$ *mit*

$$\left\| f \right\|_H \coloneqq \inf \left\{ \|\omega\|_{H_0} \,:\, \langle \omega, \Phi_0(\cdot) \rangle_{H_0} = f \right\}$$

ist der einzige RKHS von dem k ein reproduzierender Kern ist.

2) *Der Operator $V : H_0 \to H$, $\omega \mapsto \langle \omega, \Phi_0(\cdot) \rangle_{H_0}$ ist eine metrische Surjektion, d.h. linear, beschränkt und $V\mathring{B}_{H_0} = \mathring{B}_H$.*

3) $H_{\text{pre}} = \left\{ \sum_{i=1}^n \alpha_i k(x_i, \cdot) \,:\, n \geqslant 1,\ \alpha_i \in \mathbb{R},\ x_i \in X \right\}$ *ist dicht in H.*

4) *Für $f \in H_{\text{pre}}$ mit $f = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$ gilt*

$$\left\| f \right\|_H^2 = \sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j). \quad \rtimes$$

Verleich. 5-1.11: Der FS ist eine Vervollständigung von H_{pre}.

5-2.4: Es existiert eine Vervollständigung, die aus Funktionen $X \to \mathbb{R}$ besteht und die RKHS ist. ⊸

» 1) Zeige H ist HFS über X. Offensichtlich besteht H aus Funktionen $X \to \mathbb{R}$ und $V : H_0 \to H$ ist linear. Für $f \in H$ gilt ferner

$$\|f\|_H = \inf\left\{\|\omega\|_{H_0} : \omega \in V^{-1}(\{f\})\right\}.$$

Zeige $\|\cdot\|_H$ ist Hilbertraumnorm auf H.

a. $\ker V = \{\omega \in H_0 : V\omega = 0\} = V^{-1}(\{0\})$ ist abgeschlossen, da V stetig.

b. Sei $\tilde{H}$ das orthogonale Komplement von $\ker V$ in H_0. Dann ist

$$V|_{\tilde{H}} : \tilde{H} \to H$$

injektiv nach Konstruktion und außerdem surjektiv, denn V ist surjektiv, d.h. zu $f \in H$ existiert ein $\omega \in H_0$ mit $V\omega = f$. Dann ist

$$\begin{aligned}&\omega = \tilde{\omega} + \omega_0, \quad \tilde{\omega} \in \tilde{H},\ \omega_0 \in H_0,\\ &f = V(\tilde{\omega} + \omega_0) = V(\tilde{\omega}) + \underbrace{V(\omega_0)}_{=0} = V|_{\tilde{H}}(\tilde{\omega}).\end{aligned}$$

c. Zeige $\|f\|_H^2 = \|(V|_{\tilde{H}})f\|_H$.

$$\begin{aligned}\|f\|_H^2 &= \inf\left\{\|\omega_0 + \tilde{\omega}\|_{H_0}^2 : \omega_0 \in \ker V, \tilde{\omega} \in H\right\}\\ &= \inf\left\{\|\omega_0\|_{H_0}^2 + \|\tilde{\omega}\|_{H_0}^2 : \omega_0 \in \ker V, \tilde{\omega} \in H\right\}\\ &= \left\|V|_{\tilde{H}}^{-1}(f)\right\|_{\tilde{H}}^2.\end{aligned}$$

Damit ist H ein Hilbertraum, da $\tilde{H}$ ein Hilbertraum ist, $V|_{\tilde{H}} : \tilde{H} \to H$ bijektiv linear und isometrisch.

Ferner folgt, dass V eine metrische Surjektion ist.

2) Zeige H ist ein RKHS und k ein reproduzierender Kern von H.

a. Für $x \in X$ gilt

$$k(x, \cdot) = \langle \Phi_0(x), \Phi_0(\cdot)\rangle_{H_0} = V\Phi_0(x)$$

und folglich ist $k(x, \cdot) \in H$.

b. $\Phi_0(x) \in \tilde{H} = (\ker V)^\perp$, denn es gilt für $\omega \in \ker V$,

$$\langle \omega, \Phi_0(x) \rangle_{H_0} = V_\omega(x) = 0, \qquad x \in X.$$

c. Zeige, dass $f(x) = \left\langle f, k(x,\cdot) \right\rangle_H$. Da $f = V(V\big|_{\tilde{H}})^{-1} f$ und $V\big|_{\tilde{H}}$ Isometrie, folgt

$$f(x) = \left\langle (V\big|_{\tilde{H}})^{-1} f, \Phi_0(x) \right\rangle_{H_0} = \left\langle f, V\big|_{\tilde{H}}, \Phi_0(x) \right\rangle_H = \left\langle f, k(x,\cdot) \right\rangle_H .$$

Somit ist k ein reproduzierender Kern und nach Lemma 5-2.2 H ein RKHS.

3) Zeige 3) und 4) gilt für jeden RKHS $\overline{H}$ für den k reproduzierender Kern ist. Wir zeigen dazu, dass

$$H_{\text{pre}} = \left\{ \sum_{i=1}^n \alpha_i k(x_i) \,:\, n \geqslant 1,\ \alpha_i \in \mathbb{R},\ x_i \in X \right\}$$

dicht ist in $\overline{H}$ und für $f = \sum_{i=1}^n \alpha_i k(x_i,\cdot)$ gilt

$$\left\| f \right\|_{\overline{H}}^2 = \sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j).$$

a. $k(x,\cdot) \in H_{\text{pre}}$ und folglich ist $H_{\text{pre}} \subset \overline{H}$ und $k(x,\cdot) \in \overline{H}$.

b. Zeige, dass $\overline{H_{\text{pre}}^{\overline{H}}} = \overline{H}$. Angenommen $\overline{H_{\text{pre}}^{\overline{H}}} \subsetneq \overline{H}$, dann ist ${H_{\text{pre}}}^\perp \neq (0)$, d.h. es existiert ein $f \in {H_{\text{pre}}}^\perp$ und ein $x \in X$ mit $f(x) \neq 0$ und $f \perp k(x,\cdot)$.

$$\Rightarrow 0 \neq f(x) = \left\langle f, k(x,\cdot) \right\rangle_{\overline{H}} = 0. \lightning$$

D.h. $\overline{H_{\text{pre}}^{\overline{H}}} = \overline{H}$, also ist H_{pre} dicht in $\overline{H}$.

Sei nun $f = \sum_{i=1}^n \alpha_i k(x_i,\cdot)$, dann gilt

$$\begin{aligned} \left\| f \right\|_{\overline{H}}^2 = \left\| \sum_{i=1}^n \alpha_i k(x_i,\cdot) \right\|_{\overline{H}}^2 &= \left\langle \sum_{i=1}^n \alpha_i k(x_i,\cdot), \sum_{j=1}^n \alpha_j k(x_j,\cdot) \right\rangle \\ &= \sum_{i,j=1}^n \alpha_i \alpha_j \left\langle k(x_i,\cdot), k(x_j,\cdot) \right\rangle_{\overline{H}} = \sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j) \end{aligned}$$

aufgrund der repräsentierenden Eigenschaft angewandt auf $g = k(x_i, x\cdot)$.

4) Zeige schließlich H ist der einzige RKHS für den k ein reproduzierender Kern ist.

Seien dazu H_1 und H_2 RKHS für die k ein reproduzierender Kern ist. Nach dem bisher Gezeigten ist H_{pre} dicht in H_1 und H_2 und die Normen von H_1 und H_2 sind auf H_{pre} identisch. Wir zeigen nun, dass $H_1 \subset H_2$ und $\|f\|_{H_1} = \|f\|_{H_2}$ für alle $f \in H_1$. Dann folgt die Gleichheit aus der Symmetrie des Problems und H ist eindeutig.

Sei also $f \in H_1$, dann existiert eine Folge (f_n) in H_{pre} mit

$$\|f_n - f\|_{H_1} \to 0, \qquad n \to \infty.$$

Folglich ist (f_n) Cauchyfolge in H_{pre} bezüglich $\|\cdot\|_{H_1}$ und damit auch bezüglich $\|\cdot\|_2$, also ist (f_n) Cauchyfolge in H_2, d.h. es existiert ein $g \in H_2$, so dass

$$\|f_n - g\|_{H_2} \to 0, \qquad n \to \infty.$$

Da die Konvergenz in $H_{1,2}$ der punktweisen Konvergenz entspricht gilt daher auch

$$f(x) = \lim_{n\to\infty} f_n(x) = g(x), \qquad x \in X.$$

Folglich ist $f = g \in H_2$ und daher $H_1 \subset H_2$. Da außerdem

$$\|f_n - f\|_{H_2} \to 0, \qquad n \to \infty$$

folgt $\|f_n\|_{H_2} \to \|f\|_{H_2}$, d.h.

$$\|f\|_{H_1} = \lim_{n\to\infty} \|f_n\|_{H_1} = \lim_{n\to\infty} \|f_n\|_{H_2} = \|f\|_{H_2}. \quad \text{«}$$

Interpretation.

- Zu jedem Kern k existiert genau ein RKHS, so dass k ein reproduzierender Kern von H ist.
- Zu jedem RKHS H existiert genau ein reproduzierender Kern von H und dieser ist auch ein Kern.

Zwischen Kern und RKHS besteht also eine 1:1 Relation.

Der RKHS ist der „kleinste Featurespace“ im Sinne von Surjektionen. H besteht aus den Funktionen, die entstehen, wenn Daten in einen Featurespace H_0 abgebildet werden und dann ein linearer Ansatz gemacht wird. ⊸

5-3 Eigenschaften von Kernen und RKHS

Ziel dieses Abschnittes ist es, eine Beziehung zwischen den Eigenschaften von Kernen und den Eigenschaften der Funktionen eines RKHS herzustellen.

5-3.1 **Lemma** *Sei k ein Kern auf X und H der RKHS von k, dann sind äquivalent*

(i) k ist beschränkt.

(ii) Alle $f \in H$ sind beschränkt.

In diesem Fall gilt zudem

$$\|\mathrm{id} : H \to \mathcal{L}_\infty(X)\| = \|k\|_\infty, \qquad \|k\|_\infty^2 = \sup_{x\in X} k(x,x) < \infty. \quad \bowtie$$

» „(i)⇒(ii)“: Für $f \in H$ und $x \in X$ gilt da $k(x,\cdot) \in H_{\mathrm{pre}}$ nach Satz 5-2.4,

$$\begin{aligned} |f(x)| &= \left|\langle f, k(x,\cdot)\rangle_H\right| \leqslant \|f\|_H \, \|k(x,\cdot)\|_H = \|f\|_H \sqrt{k(x,x)} \\ &\leqslant \|f\|_H \underbrace{\|k\|_\infty}_{<\infty}. \end{aligned}$$

Somit ist $f \in \mathcal{L}_\infty(X)$ und $\|f\|_\infty \leqslant \|k\|_\infty \|f\|_H$. Also gilt insbesondere

$$\|\mathrm{id} : H \to \mathcal{L}_\infty(X)\| \leqslant \|k\|_\infty.$$

„(ii)⇒(i)“: $\mathrm{id} : H \to \mathcal{L}_\infty(X)$ ist wohldefiniert nach Voraussetzung. Wir zeigen, dass $\mathrm{id} : H \to \mathcal{L}_\infty(X)$ stetig ist. Dazu benutzen wir den Satz vom abgeschlossenen Graphen. Sei dazu (f_n) eine Folge in H und $g \in \mathcal{L}_\infty(X)$ mit

$$\|f_n - f\|_H \to 0, \qquad \|\mathrm{id}(f_n) - g\|_{\mathcal{L}_\infty(X)} \to 0, \tag{*}$$

so ist zu zeigen, dass $f = g$.

Nach (*) gelten

$$f_n(x) \to f(x), \qquad f_n(x) \to g(x), \qquad \forall\, x \in X.$$

Somit ist $f(x) = g(x)$ für alle $x \in X$ und daher ist

$$\mathrm{id} : H \to \mathcal{L}_\infty(X)$$

stetig. Weiterhin ist

$$\begin{aligned} k(x,x) &\leqslant \|k(x,\cdot)\|_\infty \leqslant \|\mathrm{id} : H \to \mathcal{L}_\infty(x)\| \, \|k(x,\cdot)\|_H \\ &= \|\mathrm{id} : H \to \mathcal{L}_\infty(x)\| \sqrt{k(x,x)} \end{aligned}$$

und daher

$$\begin{aligned} &\sqrt{k(x,x)} \leqslant \|\mathrm{id} : H \to \mathcal{L}_\infty(x)\| , \\ \Rightarrow\ &\|k\|_\infty \leqslant \|\mathrm{id} : H \to \mathcal{L}_\infty(x)\| . \end{aligned}$$

Insbesondere ist k auf der Diagonalen beschränkt. Schließlich ist

$$\begin{aligned} \left|k(x,x')\right| &= \left|\left\langle k(x,\cdot), k(x',\cdot)\right\rangle_H\right| \leqslant \|k(x,\cdot)\|_H \left\|k(x',\cdot)\right\|_H \\ &= \sqrt{k(x,x)}\sqrt{k(x',x')} \leqslant \|k\|_\infty^2 . \quad « \end{aligned}$$

5-3.2 **Lemma** *Sei k ein Kern auf X und H der RKHS von k, dann sind äquivalent*

(i) $k(x,\cdot)$ ist messbar für alle $x \in X$.

(ii) Alle $f \in H$ sind messbar. ⋈

» „(ii)⇒(i)“: $k(x,\cdot) \in H$.
„(i)⇒(ii)“: $k(x,\cdot)$ ist messbar und folglich besteht

$$H_{\mathrm{pre}} = \{h(x,\cdot) : x \in X\}$$

aus messbaren Funktionen. Sei nun $f \in H$, dann existiert eine Folge (f_n) in H_{pre} mit $\left\|f_n - f\right\|_H \to 0$ und folglich ist $f(x) = \lim\limits_{n\to\infty} f_n(x)$ für alle $x \in X$. «

5-3.3 **Lemma** *Sei k Kern auf X mit RKHS H und $k(x,\cdot)$ messbar für alle $x \in X$ und H separabel. Dann ist k messbar und die kanonische Featuremap $\Phi : X \to H$ ist messbar.* ⋈

» Wir überspringen den Beweis. «

5-3.4 **Lemma** *Sei X ein Hausdorff-Raum, k ein Kern auf X mit RKHS H. Dann sind äquivalent*

(i) *k ist beschränkt und separat stetig, d.h. $k(x,\cdot)$ ist stetig für alle $x \in X$.*

(ii) *Alle $f \in H$ sind stetig und beschränkt. In diesem Fall gilt*

$$\|\mathrm{id} : H \to C_b(X)\| = \|k\|_\infty < \infty. \quad \bowtie$$

» „(i)⇒(ii)“: $H_{\mathrm{pre}} \subset C_b(X)$ nach Voraussetzung. Sei $f \in H$, dann existiert eine Folge (f_n) in H_{pre} mit $\|f_n - f\|_H \to 0$. Da k beschränkt folgt mit Lemma 5-3.1, dass

$$\|f_n - f\|_\infty \to 0$$

und folglich ist $f \in C_b(X)$.

„(ii)⇒(i)“: $k(x,\cdot) \in H$ ist stetig nach Voraussetzung und beschränkt nach Lemma 5-3.1. Somit ist k beschränkt und $\|k\|_\infty = \|\mathrm{id} : H \to C_b(X)\|$. «

5-3.5 **Definition** *Sei k ein Kern auf X mit RKHS H und kanonischer Featuremap $\Phi : X \to H$. Dann heißt $d_k : X \times X \to [0,\infty)$*

$$d_k(x,x') := \|\Phi(x) - \Phi(x')\|_H = \sqrt{k(x,x) + k(x',x') - 2k(x,x')}$$

die Kernmetrik von k. $\bowtie$

Man sieht leicht ein, dass d_k genau dann eine Metrik ist, wenn Φ injektiv ist. Im Allgemeinen ist d_k lediglich eine Semi-Metrik.

5-3.6 **Lemma** *Sei (X,τ) ein Hausdorff-Raum, k ein Kern auf X und H ein RKHS von k, $\Phi : X \to H$ die kanonische Featuremap. Dann sind äquivalent*

(i) *k ist stetig.*

(ii) *k ist separat stetig und $x \mapsto k(x,x)$ ist stetig.*

(iii) *$\Phi : X \to H$ ist stetig.*

(iv) *$\mathrm{id} : (X,\tau) \to (X,d_k)$ ist stetig, d.h. jede d_k-offene Kugel ist τ-offen.* $\bowtie$

» „(i)⇒(ii)“: Trivial.
„(ii)⇒(iv)“: $d_k(\cdot, x) : (X, \tau) \to \mathbb{R}$ ist stetig und daher ist die Kugel

$$B_\varepsilon(x', d_k) = \left\{ x' : d_k(x, x') < \varepsilon \right\}$$

τ-offen. Diese Kugeln erzeugen aber gerade die Topologie von d_k, d.h. die Toplogie von τ ist größer oder gleich der Topologie von d_k.
„(iv)⇒(iii)“: $\Phi : (X, d_k) \to H$ ist stetig nach Konstruktion von d_k und somit ist $\Phi : (X, \tau) \to H$ stetig nach Voraussetzung.
„(iii)⇒(i)“: Seien (x_n) und (y_n) Folgen in X mit $x, y \in X$, so dass $x_n \to x$ und $y_n \to y$, so gilt

$$\begin{aligned} \left| k(x_n, y_n) - k(x, y) \right| &= \left| \left\langle \Phi(y_n), \Phi(x_n) - \Phi(x) \right\rangle + \left\langle \Phi(x), \Phi(y_n) - \Phi(y) \right\rangle \right| \\ &\leqslant \left\| \Phi(y_n) \right\| \left\| \Phi(x_n) - \Phi(x) \right\| + \left\| \Phi(x) \right\| \left\| \Phi(y_n) - \Phi(y) \right\| \\ &\leqslant \left\| \Phi(y_n) \right\| d_k(x_n, x) + \left\| \Phi(x) \right\| d_k(y_n, y). \end{aligned}$$

Nach Voraussetzung ist Φ stetig, d.h. $\left\| \Phi(y_n) \right\|$ ist beschränkt. «

5-3.7 **Satz** *Sei X ein kompakter Hausdorffraum und k ein stetiger Kern mit RKHS H. Dann ist die Abbildung*

$$\mathrm{id} : H \to C(X)$$

wohldefiniert und kompakt. ⋊

» Da X kompakt und k stetig, ist k auch beschränkt. Foglich ist $\mathrm{id} : H \to C(X)$ wohldefiniert und stetig (Lemma 5-3.4).

Sei $C(X, d_k)$ der Raum der d_k-stetigen Funktionen $f : X \to \mathbb{R}$. Offensichtlich ist $C_b(X, d_k) \subset l_\infty(X)$. Außerdem ist X kompakt und Φ stetig, d.h. $\Phi(X) \subset H$ ist kompakt. Da Φ Isometrie bezüglich d_k, ist (X, d_k) kompakt und folglich

$$C(X, d_k) = C_b(X, d_k).$$

Für $f \in B_H$ und $x, y \in H$ gilt

$$\left| f(x) - f(y) \right| = \left| \left\langle f, \Phi(x) - \Phi(y) \right\rangle_H \right| \leqslant \left\| f \right\| \left\| \Phi(x) - \Phi(y) \right\| \leqslant d_k(x, y).$$

Also sind alle Funktionen in B_H lipschitz bezüglich d_k mit lipschitz-Konstante $\leqslant 1$. Der Satz von Arzelà-Ascoli besagt nun, dass

$$\overline{B_H} \text{ kompakt ist in } C(X, d_k).$$

Somit ist $\mathrm{id}_{H \to C(X,d_k)}(B_H)$ relativkompakt und daher $\mathrm{id} : H \to C(X, d_k)$ kompakt. Da $C(X, d_k) \subset C(X)$ mit Normgleichheit und

$$H \xrightarrow{\mathrm{id}} C(X, d_k) \overset{\text{stetig}}{\hookrightarrow} C(X)$$

ist $\mathrm{id} : H \to C(X)$ kompakt. «

Bemerkung. Ist X *nicht* kompakt aber k stetig und beschränkt, so ist

$$\mathrm{id} : H \to C_b(X)$$

wohldefiniert und stetig aber im Allgemeinen *nicht* kompakt.

Betrachte z.B. $X = \mathbb{R}$ und den Gauß-Kern $k_\sigma(x, y) = \exp(-\sigma^2 \left\|x - y\right\|_2^2)$. Seien $m, n \in \mathbb{N}$ mit $n \neq m$, so gilt

$$\begin{aligned} \|k_\sigma(m, \cdot) - k_\sigma(n, \cdot)\|_\infty &\geqslant |k_\sigma(m, m) - k_\sigma(n, m)| \\ &= 1 - \exp(-\sigma^2 |m - n|^2) \geqslant 1 - e^{-\sigma^2}, \end{aligned}$$

da $|m - n| \geqslant 1$. Es ist aber

$$\|k_\sigma(n, \cdot)\|_H = \sqrt{k_\sigma(n, n)} = 1$$

und folglich $\mathrm{id} : H \to C_b(X)$ nicht kompakt. ⊸

5-3.8 **Lemma** *Sei X ein separabler, metrischer Raum und k ein stetiger Kern mit RKHS H. Dann ist H separabel.* ⋊

» $\Phi : X \to H$ ist stetig, da k stetig und somit ist $\Phi(X)$ separabel. Folglich ist auch $H_{\mathrm{pre}} = \operatorname{span} \Phi(X)$ separabel, denn man kann span durch Linearkombinationen mit rationalen Koeffizienten approximieren, und damit ist auch $H = \overline{H_{\mathrm{pre}}}$ separabel. «

Wir wollen nun untersuchen, wie sich die Differenzierbarkeit des Kerns auf die Featuremap und den RKHS übertragen. Dazu betrachten wir einen Kern $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ und die Abbildung

$$\tilde{k} : \mathbb{R}^{2d} \to \mathbb{R}, \qquad (x, y) \mapsto \tilde{k}((x, y)) = k(x, y).$$

Falls die partiellen Ableitungen von i und $i + d$ von $\tilde{k}$ existieren, setzen wir

$$\partial_i \partial_{i+d} k \coloneqq \partial_i \partial_{i+d} \tilde{k}.$$

Analog lässt sich dies definieren, wenn $X \subset \mathbb{R}^d$ offen.

5-3-9 **Lemma** *Sei $X \subset \mathbb{R}^d$ offen, k ein Kern mit Featurespace H und Featuremap $\Phi : X \to H$, so dass $\partial_i \partial_{i+d} k$ für $1 \leqslant i \leqslant d$ existiert und stetig ist. Dann existiert auch*

$$\partial_i \Phi : X \to H$$

und ist stetig und es gilt

$$\left\langle \partial_i \Phi(x), \partial_i \Phi(y) \right\rangle_H = \partial_i \partial_{i+d} k(x, y) = \partial_{i+d} \partial_i k(x, y). \quad \bowtie$$

» Wir beweisen nur den Fall $X = \mathbb{R}^d$. Sei e_i der i-te Einheitsvektor, dann definiere für festes $h > 0$,

$$\Delta_h \Phi(x) \coloneqq \Phi(x + h e_i) - \Phi(x).$$

Wir zeigen nun, dass $(h_n^{-1} \Delta_{h_n} \Phi(x))_n$ für alle Folgen $h_n \to 0$ mit $h_n \neq 0$ konvergiert. Dann ist $\lim_{n \to \infty} h_n^{-1} \Delta_{h_n} \Phi(x) \eqqcolon \partial_i \Phi(x)$ unabhängig von der Folge (h_n). Da H vollständig ist, genügt es zu zeigen, dass $(h_n^{-1} \Delta_{h_n} \Phi(x))$ eine Cauchyfolge ist. Es ist

$$\begin{aligned}
&\left\| h_n^{-1} \Delta_{h_n} \Phi(x) - h_m^{-1} \Delta_{h_m} \Phi(x) \right\|_H^2 \\
&\quad = \left\| h_n^{-1} \Delta_{h_n} \Phi(x) \right\|_H^2 + \left\| h_m^{-1} \Delta_{h_m} \Phi(x) \right\|_H^2 \\
&\quad - 2 \left\langle h_n^{-1} \Delta_{h_n} \Phi(x), h_m^{-1} \Delta h_m \Phi(x) \right\rangle \\
&\quad \leqslant 2 \left(\varepsilon + \partial_i \partial_{i+d} k(x, x) - \left\langle h_n^{-1} \Delta_{h_n} \Phi(x), h_m^{-1} \Delta_{h_m} \Phi(x) \right\rangle \right) \qquad (*)
\end{aligned}$$

Wir definieren für festes n und x,

$$K_{x,n}(y) = k(x + h_n e_i, y) - k(x, y),$$

dann ist

$$\begin{aligned}\left\langle \Delta_{h_n}\Phi(x), \Delta_{h_m}\Phi(y)\right\rangle_H &= \left\langle \Phi(x + h_n e_i) - \Phi(x), \Phi(y + h_m e_i) - \Phi(y)\right\rangle_H \\ &= k(x + h_n e_i, y + h_m e_i) - k(x, y + h_m e_i) \\ &\quad - k(x + h_n e_i, y) + k(x, y) \\ &= K_{x,n}(y + h_m e_i) - K_{x,n}(y).\end{aligned}$$

Wenden wir den Mittelwertsatz auf $K_{x,n}$ an, erhalten wir ein $\xi_{m,n} \in [-|h_m|, |h_m|]$ mit

$$\begin{aligned}h_m^{-1}\left\langle \Delta_{h_n}\Phi(x), \Delta_{h_m}\Phi(y)\right\rangle &= h_m^{-1}\left(K_{x,n}(y + h_m e_i) - K_{x,n}(y)\right) \\ &= \partial_i K_{x,n}(y + \xi_{m,n} e_i) \\ &= \partial_{i+d}\left(k(x + h_n e_i, y + \xi_{m,n} e_i) - k(x, y + \xi_{m,n} e_i)\right)\end{aligned}$$

Wir wenden nun den Mittelwertsatz auf die erste Variable an und erhalten ein $\eta_{n,m} \in [-|h_n|, |h_n|]$, so dass

$$\left\langle h_n^{-1}\Delta_{h_n}\Phi(x), h_m^{-1}\Delta_{h_m}\Phi(y)\right\rangle_H = \partial_i\partial_{i+d}k(x + \eta_{n,m}e_i, y + \xi_{m,n}e_i)$$

Somit ist nach (*),

$$\begin{aligned}&\left\|h_n^{-1}\Delta_{h_n}\Phi(x) - h_m^{-1}\Delta_{h_m}\Phi(x)\right\|_H^2 \\ &\quad\leqslant 2\left(\varepsilon + \partial_i\partial_{i+d}k(x, x) - \partial_i\partial_{i+d}k(x + \eta_{n,m}e_i, x + \xi_{m,n}e_i)\right) \\ &\quad< 4\varepsilon, \qquad n, m \geqslant n_0,\end{aligned}$$

denn $\partial_i\partial_{i+d}k$ ist stetig. Also ist $(h_n^{-1}\Delta_{h_n}\Phi(x))$ Cauchyfolge und daher existiert $\partial_i\Phi(x)$ und ist stetig.

Weiterhin gilt

$$\left\langle h_n^{-1}\Delta_{h_n}\Phi(x), h_m^{-1}\Delta_{h_m}\Phi(y)\right\rangle_H \to \partial_i\partial_{i+d}k(x, y)$$

und da $h_n^{-1}\Delta_{h_n}\Phi(x) \to \partial_i\Phi(x)$, folgt die Formel. «

5-3.10 **Definition** *Ein Kern k auf $X \subset \mathbb{R}^d$ offen heißt m-fach stetig differenzierbar, wenn*

$$\partial^{\alpha,\alpha} k : X \times X \to \mathbb{R}$$

existiert und stetig ist für alle $\alpha \in \mathbb{N}_0^d$ mit $|\alpha| \leqslant m$. ⋊

5-3.11 **Korollar** *Sei k ein m-fach stetig differenzierbarer Kern mit RKHS H. Dann sind alle $f \in H$ m-fach stetig differenzierbar und es gilt*

$$\left|\partial^\alpha f(x)\right| \leqslant \left\|f\right\| \sqrt{\partial^{\alpha,\alpha} k(x,x)},$$

für alle $x \in X$ und $\alpha \in \mathbb{N}_0^d$ mit $|\alpha| \leqslant m$. ⋊

» Iteration von Lemma 5-3.9 impliziert, dass $\partial^\alpha \Phi : X \to H$ Featuremap von $\partial^{\alpha,\alpha} k$. Da $\langle f, \cdot \rangle$ stetig, folgt

$$\left\langle f, \partial^\alpha \Phi(x) \right\rangle_H = \partial^\alpha \left\langle f, \Phi(x) \right\rangle = \partial^\alpha f(x)$$

und nach Cauchy-Schwarz

$$\left|\partial^\alpha f(x)\right| = \left|\left\langle f, \partial^\alpha \Phi(x) \right\rangle\right| \leqslant \left\|f\right\|_H \left\|\partial^\alpha \Phi(x)\right\|_H = \left\|f\right\|_H \sqrt{\partial^{\alpha,\alpha} k(x,x)}. \quad «$$

5-4 Große RKHS

Wann kann ein RKHS „viele“ Funktionen approximieren?

5-4.1 **Definition** *Sei (X,d) ein kompakter metrischer Raum, dann heißt ein stetiger Kern k auf X universell, wenn der RKHS H von k dicht in $C(X)$ ist, d.h.*

$$\forall\, g \in C(X), \varepsilon > 0\ \exists\, f \in H : \left\|f - g\right\|_\infty \leqslant \varepsilon.$$ ⋊

5-4.2 **Lemma** *Sei (X,d) ein kompakter metrischer Raum und k ein universeller Kern auf X. Dann gelten*

1) *Jede Featuremap von k ist injektiv.*

2) *$k(x,x) > 0$ für alle $x \in X$.*

3) *Der Kern* $k^* : X \times X \to \mathbb{R}$ *gegeben durch*

$$k^*(x,y) = \frac{k(x,y)}{\sqrt{k(x,x)k(y,y)}}$$

ist universell.

4) *Ist* $M \subset X$ *abgeschlossen, dann ist* $k\big|_{M\times M}$ *universell.* ⋊

» 1) Seien $x_1 \neq x_2 \in X$ und $g : X \to \mathbb{R}$ mit $x \in X$,

$$g(x) \coloneqq \frac{d(x_1,x)}{d(x_1,x)+d(x_2,x)} - \frac{d(x_2,x)}{d(x_1,x)+d(x_2,x)},$$

$g(x_1) = -1$ und $g(x_2) = 1$ und g stetig auf X, also $g \in C(X)$. Sei nun H_0 ein Featurespace von k und $\Phi_0 : X \to H_0$ eine Featuremap und ferner H der RKHS von k. Da k universell, existiert ein $f \in H$ mit

$$\|f-g\|_\infty \leqslant \frac{1}{2} \Rightarrow f(x_1) \leqslant -\frac{1}{2} \text{ und } f(x_2) \geqslant \frac{1}{2}.$$

Nach Satz 5-2.4 existiert ein $\omega \in H_0$ mit $f = \langle \omega, \Phi_0(\cdot)\rangle$ und daher

$$\left.\begin{aligned} \langle \omega, \Phi_0(x_1)\rangle_H &\leqslant -\tfrac{1}{2}, \\ \langle \omega, \Phi_0(x_2)\rangle_H &\geqslant \tfrac{1}{2}, \end{aligned}\right\} \Rightarrow \Phi_0(x_1) \neq \Phi_0(x_2).$$

2) Zeige $k(x,x) > 0$ für alle $x \in X$. In 1) haben wir gesehen, dass es für $x \in X$ ein $g \in C(X)$ mit $g(x) = 1$ gibt. Dazu gibt es ein $f \in H$ mit $\|f-g\|_\infty \leqslant \frac{1}{2}$ und folglich

$$\frac{1}{2} \leqslant f(x) = \big\langle f, \Phi(x)\big\rangle_H \Rightarrow \Phi(x) \neq 0.$$

Somit ist $k(x,x) = \langle \Phi(x), \Phi(x)\rangle_H > 0$.

3) Zeige, dass der normalisierte Kern universell ist. Schreibe dazu

$$\alpha(x) \coloneqq (k(x,x))^{-1/2},$$

dann ist $\alpha\Phi : X \to H$ eine Featuremap von k^*. Sei nun $g \in C(X)$ und $\varepsilon > 0$. Sei weiterhin

$$c \coloneqq \|\alpha\|_\infty < \infty.$$

Dann existiert ein $f \in H$ mit $\left\| f - \frac{g}{\alpha} \right\|_\infty \leqslant \frac{\varepsilon}{c}$ und damit

$$\left\| \left\langle f, \alpha \Phi(\cdot) \right\rangle_H - g \right\|_\infty \leqslant \underbrace{\|\alpha\|_\infty}_{\leqslant c} \left\| f - \frac{g}{\alpha} \right\|_\infty \leqslant \varepsilon.$$

4) Sei $M \subset X$ abgeschlossen, also kompakt, so ist $k\big|_{M\times M}$ universell. Tietzes Fortsetzungssatz besagt

$$\forall\, g \in C(M)\ \exists\, \hat{g} \in C(X) : \hat{g}\big|_M = g.$$

Zu jedem $g \in C(X)$ und $\varepsilon > 0$ existiert eine Fortsetzung $\hat{g} \in C(X)$ und ein $f \in H$ mit $\left\| \hat{g} - f \right\|_\infty \leqslant \varepsilon$. Dann ist auch

$$\left\| \hat{g}\big|_M - f\big|_M \right\|_\infty \leqslant \varepsilon,$$

wobei $f\big|_M$ im RKHS von $k\big|_{M\times M}$. «

5-4-3 **Satz** *Sei (X,d) ein kompakter metrischer Raum, k stetiger Kern mit $k(x,x) > 0$ für alle $x \in X$. Ferner sei $\Phi : X \to l_2$ eine injektive Featuremap von k. Schreibe $\Phi_n : X \to \mathbb{R}$ für die n-te Komponente von Φ, d.h. $\Phi(x) = (\Phi_n(x))_{n\geqslant 1}$ für alle $x \in X$.*

Schreibe $\mathcal{A} := \operatorname{span}\{\Phi_n : n \geqslant 1\}$. Ist $\mathcal{A}$ eine Algebra, dann ist k universell. ⋊

Zum Beweis des Satzes benötigen wir den

Satz von Stone-Weierstraß *Sei (X,d) ein kompakter metrischer Raum, $\mathcal{A} \subset C(X)$ Algebra mit*

1) Zu jedem $x \in X$ existiert ein $f \in \mathcal{A}$ mit $f(x) \neq 0$. „$\mathcal{A}$ verschwindet nicht“

2) Zu $x \neq y \in X$ existiert ein $f \in \mathcal{A}$ mit $f(x) \neq f(y)$. „$\mathcal{A}$ separiert“.

Dann gilt $\mathcal{A} \subset C(X)$ ist dicht. ⋊

» 1) $\|(\Phi_n(x))_{n\geqslant 1}\|_{l^2}^2 = \langle \Phi(x), \Phi(x)\rangle_{l^2} = k(x,x) > 0$ und daher existiert ein $n \in \mathbb{N}$, so dass $\Phi_n(x) > 0$.

2) Sei nun $x \neq y$, dann ist $\Phi(x) \neq \Phi(y)$ nach Voraussetzung, d.h. es gibt ein $n \in \mathbb{N}$, so dass $\Phi_n(x) \neq \Phi_n(y)$.

3) Zu zeigen ist noch, dass $\mathcal{A} \subset C(X)$. Da k stetig, ist $\Phi : X \to l^2$ stetig und daher ist auch Φ_n stetig für alle $n \geqslant 1$, d.h. $\mathcal{A} \subset C(X)$. Damit zeigt Stone-Weierstraß, dass $\overline{A}^{\|\cdot\|_\infty} = C(X)$. Für $g \in C(X)$ und $\varepsilon > 0$ existiert somit ein $f \in \mathcal{A}$ mit $\|f - g\|_\infty \leqslant \varepsilon$ und

$$f = \sum_{j=1}^{m} \alpha_j \Phi_{n_j}.$$

Setze

$$\omega_n := \begin{cases} \alpha_j, & \text{falls } n_j = n, \\ 0, & \text{sonst,} \end{cases}$$

und $\omega := (\omega_n)_{n\geqslant 1}$. Somit ist $\omega \in l^2$ und $f = \langle \omega, \Phi(\cdot)\rangle_{l^2}$ nach Konstruktion. «

5-4.4 **Korollar** *Für $r \in (0,\infty]$ sei $h : (-r,r) \to \mathbb{R}$ mit*

$$h(t) = \sum_{n\geqslant 0} a_n t^n, \qquad t \in (-r,r).$$

Falls $a_n > 0$ für alle $n \geqslant 0$, ist der Taylorkern universell auf allen kompakten Teilmengen $X \subset \sqrt{r}\mathring{B}_{\mathbb{R}^d}$. ⋊

» Wir haben schon gesehen, dass $\Phi : X \to l^2(\mathbb{N}_0^d)$,

$$\Phi(x) = \left(\underbrace{\sqrt{a_{j_1} \dots a_{j_d} c_{j_1,\dots,j_d}}}_{>0} \prod_{i=1}^{d} x_i^{j_i} \right)_{j_1,\dots,j_d \geqslant 0}$$

eine Featuremap von k ist.

1) k ist stetig, da h und $\langle \cdot,\cdot\rangle_{\mathbb{R}^d}$ stetig.

2) $a_0 > 0$ und daher ist $k(x,x) = \sum_{n\geqslant 0} a_n (\langle x,x\rangle)^n \geqslant a_0 > 0$.

3) Φ ist injektiv. Sei $x \neq y$ dann existiert ein $i \in \{1, \dots, d\}$ mit $x_i \neq y_i$. Nimm $j_i = 1$, alle anderen $j_l = 0$. Folglich ist $\Phi_{j_1,\dots,j_d}(x) \neq \Phi_{j_1,\dots,j_d}(y)$ und daher $\Phi(x) \neq \Phi(y)$.

   Zu zeigen ist, dass $\mathcal{A} = \operatorname{span}\left\{\Phi_{j_1,\dots,j_d} : j_1, \dots, j_d \geqslant 0\right\}$ ist Algebra. Da $a_n \geqslant 0$ ist $\mathcal{A}$ Algebra, denn $\mathcal{A}$ enhält alle Monome und Linearkombinationen davon. Also ist Satz 5-4.3 anwendbar. «

Bsp Der *Exponentialkern*

$$h(x, x') \coloneqq \exp\left(\left\langle x, x' \right\rangle_{\mathbb{R}^d}\right)$$

ist universell auf allen Kompakta, da $a_n = \frac{1}{n!} > 0$. ■

Bsp Für $\sigma > 0$ ist der *Gaußkern*

$$k_\sigma(x, x') = \exp(-\sigma^2 \left|x - x'\right|^2)$$

universell auf allen Kompakta. Da

$$h_\sigma(x, x') = \frac{\exp(2\sigma^2 \langle x, x\rangle)}{\exp(\sigma^2 |x|^2) \exp(\sigma^2 |x'|^2)},$$

$\exp(2\sigma^2 \langle x, x\rangle)$ universell und h_σ eine normalisierte Fassung davon. ■

Bemerkungen. A. Dahmen und Michelli zeigten 1987, dass ein Kern k genau dann universell ist, wenn

$$a_0 > 0 \text{ und } \sum_{a_{2n}>0} \frac{1}{2n} = \sum_{a_{2n+1}>0} \frac{1}{2n+1} = \infty.$$

B. Eine leichte Übung zeigt, dass jeder universelle Kern strikt positiv ist.

C. Pirkus zeigte 2004, dass Taylorkerne genau dann strikt positiv sind, wenn

$$a_0 > 0 \text{ und } |\{n : a_{2n} > 0\}| = |\{n : a_{2n+1} > 0\}| = \infty.$$

D. Man kann auch untersuchen, wann $H \subset L_p(\mu)$ dicht ist.

   Für Gaußkerne ist $H_\sigma \subset L_p(\mu)$ dicht für alle $p \in [1, \infty)$ und alle W-Maße μ auf $\mathbb{R}^d$ und $\mu = \lambda^d$ auf $\mathbb{R}^d$.

E. Sei (X, d) kompakter metrischer Raum, dann existiert ein universeller Kern auf X.

F. Sei (X, τ) kompakter Hausdorffraum, dann existiert genau dann ein universeller Kern auf X, wenn (X, τ) metrisierbar ist.

G. In der Regel gilt für universelle RKHS H, dass $\dim H = \infty$, insbesondere gilt dies für $X \subset \mathbb{R}^d$ mit $\overset{\circ}{X} \neq \emptyset$. ⊸

6 Support Vector Machines (SVMs)

6-1 Definition, einfache Eigenschaften und Beispiele

6-1.1 **Definition** *Sei H ein RKHS und $L : X \times Y \times \mathbb{R} \to [0, \infty)$ eine konvexe Verlustfunktion. Ferner sei $\lambda > 0$ „Regularisierungsparameter". Dann heißt eine Lernmethode Support Vector Machine (SVM), wenn*

(a) *Für jedes $D \in (X \times Y)^m$ die Entscheidungsfunktion $f_{D,\lambda} \in H$.*

(b) $\lambda \left\| f_{D,\lambda} \right\|_H^2 + \mathcal{R}_{L,D}(f_{D,\lambda}) = \inf_{f \in H} \lambda \left\| f \right\|_H^2 + \mathcal{R}_{L,D}(f)$.

D.h. $f_{D,\lambda}$ minimiert das regularisierte empirische Risiko,

$$\lambda \|\cdot\|_{H^2} + \mathcal{R}_{L,D}(\cdot), \qquad \text{über } H. \quad \rtimes$$

6-1.2 **Eindeutigkeit** *Für alle $D \in (X \times Y)^n$, $n \geqslant 1$, $\lambda > 0$ existiert höchstens ein $f_{D,\lambda} \in H$.* $\rtimes$

» Angenommen es gibt $f_1 \neq f_2 \in H$ und beide minimieren $\lambda \|\cdot\|_{H^2} + \mathcal{R}_{L,D}(\cdot)$ über H. Sei $f = \frac{1}{2}(f_1 + f_2)$. L konvex, dann ist $\mathcal{R}_{L,D}(\cdot)$ konvex

$$\mathcal{R}_{L,D}(f) \leqslant \frac{1}{2}\mathcal{R}_{L,D}(f_1) + \frac{1}{2}\mathcal{R}_{L,D}(f_2)$$

Parallelogrammgleichung in H:

$$\left\| f_1 + f_2 \right\|_H^2 + \left\| f_1 - f_2 \right\|_H^2 = 2 \left\| f_1 \right\|_H^2 + 2 \left\| f_2 \right\|_H^2.$$

Da $f_1 \neq f_2$ ist

$$\left\| \frac{1}{2}(f_1 + f_2) \right\|_H^2 < \frac{1}{2} \left\| f_1 \right\|_H^2 + \frac{1}{2} \left\| f_2 \right\|_H^2$$

und damit

$$\begin{aligned}\lambda \left\|f\right\|_H^2 + \mathcal{R}_{L,D}(f) &< \frac{1}{2}\left(\left\|f_1\right\|_H^2 + \mathcal{R}_{L,D}(f_1)\right) + \frac{1}{2}\left(\left\|f_2\right\|_H^2 + \mathcal{R}_{L,D}(f_2)\right) \\ &= \inf_{f\in H} \lambda \left\|f\right\|_H^2 + \mathcal{R}_{L,D}(f). \lightning \quad «\end{aligned}$$

6-1.3 **Satz** *Ist* $\dim H < \infty$*, dann existiert ein* $f_{D,\lambda} \in H$. ⋊

» L ist konvex, d.h. $L(x,y,\cdot) : \mathbb{R} \to [0,\infty)$ konvex und lokal lipschitz für alle x, y also ist L insbesondere stetig. Betrachte für $D \in (X \times Y)^n$ das Risikofunktional

$$\mathcal{R}_{L,D} : H \to [0,\infty), \qquad f \mapsto \mathcal{R}_{L,D}(f).$$

Dieses ist konvex und stetig, da

$$f \mapsto \mathcal{R}_{L,D}(f) = \frac{1}{n}\sum_{i=1}^{n} L(x_i, y_i, f(x_i))$$

und die Konvergenz in H die punktweise Konvergenz impliziert. Ferner ist die Abbildung

$$H \to [0,\infty), \qquad f \mapsto \lambda \left\|f\right\|_H^2$$

stetig und konvex (aufgrund der Parallelogrammungleichung sogar strikt konvex) und damit ist auch die Abbildung

$$H \to [0,\infty), \qquad f \mapsto \lambda \left\|f\right\|_H^2 + \mathcal{R}_{L,D}(f) \tag{*}$$

stetig und konvex. Setze $m \coloneqq \mathcal{R}_{L,D}(0) < \infty$ und

$$A = \left\{f \in H : \lambda \left\|f\right\|_H^2 + \mathcal{R}_{L,D}(f) \leqslant m\right\}.$$

Dann gelten

- $0 \in A$.
- A ist abgeschlossen, da (*) stetig.

- A ist beschränkt, da für $f \in A$ gilt

$$\lambda \left\|f\right\|_H^2 \leqslant \lambda \left\|f\right\|_H^2 + \mathcal{R}_{L,D}(f) \leqslant m,$$

ist $\left\|f\right\|_H \leqslant \sqrt{m/\lambda}$.

Somit ist A nichtleer und kompakt, d.h. (*) besitzt ein Minimum $f^* \in A$ mit

$$\lambda \left\|f^*\right\|_H^2 + \mathcal{R}_{L,D}(f^*) \leqslant m$$

und für $f \notin A$ gilt

$$\lambda \left\|f\right\|_H^2 + \mathcal{R}_{L,D}(f) > m \geqslant \lambda \left\|f^*\right\|_H^2 + \mathcal{R}_{L,D}(f^*),$$

also ist f^* sogar ein globales Minimum. «

6-1.4 **Lemma** *Falls $f_{D,\lambda} \in H$ existiert gelten*

1) $\left\|f_{D,\lambda}\right\|_H \leqslant \left(\lambda^{-1}\mathcal{R}_{L,D}(0)\right)^{1/2}$.

Inbesondere ist $L(x,y,0) \leqslant 1$ für alle $(x,y) \in X \times Y$ und folglich

$$\left\|f_{D,\lambda}\right\|_H \leqslant \lambda^{-1/2}.$$

2) Falls ein $f^* \in H$ existiert mit $\mathcal{R}_{L,D}(f^*) \leqslant \mathcal{R}_{L,D}(0)$, ist $f_{D,\lambda} \neq 0$. ⋊

» 1) Man rechnet direkt nach, dass

$$\begin{aligned}\lambda \left\|f_{D,\lambda}\right\|^2 &\leqslant \lambda \left\|f_{D,\lambda}\right\|^2 + \mathcal{R}_{L,D}(f_{D,\lambda}) = \min_{f\in H} \lambda \left\|f\right\|_H^2 + \mathcal{R}_{L,D}(f) \\ &\leqslant \lambda \left\|0\right\|_H^2 + \mathcal{R}_{L,D}(0) = \mathcal{R}_{L,D}(0).\end{aligned}$$

2) Für jedes $\alpha \in [0,1]$ sei

$$h(\alpha) := 2\lambda\alpha \left\|f^*\right\|_H^2 + \mathcal{R}_{L,D}(f^*) - \mathcal{R}_{L,D}(0)$$

Da $\mathcal{R}_{L,D}(f^*) \leqslant \mathcal{R}_{L,D}(0)$ und $\alpha \mapsto h(\alpha)$ quadratisch, besitzt

$$h : [0,1] \to [0,\infty)$$

ein globales Minimum $\alpha^* \in (0,1]$, denn

$$\begin{aligned} & 2\lambda\alpha \left\|f^*\right\|_H^2 + \mathcal{R}_{L,D}(f^*) - \mathcal{R}_{L,D}(0) = 0 \\ \Leftrightarrow\ & \alpha^* = \frac{\mathcal{R}_{L,D}(0) - \mathcal{R}_{L,D}(f^*)}{2\lambda \left\|f^*\right\|_H^2}. \end{aligned}$$

Somit ist

$$\begin{aligned} \lambda \left\|\alpha^* f^*\right\|_H^2 + \mathcal{R}_{L,D}(\alpha^* f^*) &\leqslant h(\alpha^*) < h(0) = \mathcal{R}_{L,D}(0) \\ &= \lambda \left\|0\right\|_H^2 + \mathcal{R}_{L,D}(0). \end{aligned}$$

Also ist $f_{D,\lambda} \neq 0$. «

6-1.5 **Representer Theorem** *Sei H ein RKHS mit Kern k, L eine konvexe Verlustfunktion und $\lambda > 0$, $D \in (X \times Y)^n$. Schreibe $D = ((x_1, y_1), \ldots, (x_n, y_n))$, dann existieren ein $f_{D,\lambda}$ und $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ mit*

$$f_{D,\lambda} = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot). \quad \bowtie$$

» Sei $X' := \{x_1, \ldots, x_n\}$ und

$$H\big|_{X'} := \left\{ \sum_{i=1}^{n} \alpha_i k(x_i, \cdot) \,:\, \alpha_i \in \mathbb{R} \right\} = \operatorname{span}\{k(x_i, \cdot) \,:\, i = 1, \ldots, n\}.$$

Wir zeigen, dass $H\big|_{X'}$ der RKHS von $k\big|_{X' \times X'}$ ist.

Satz 5-2.4 sagte nämlich $\operatorname{span}\{k(x_i, \cdot) \,:\, i = 1, \ldots, n\}$ ist dicht im RKHS von $k\big|_{X' \times X'}$. Da $\dim \operatorname{span}\{k(x_i, \cdot) \,:\, i = 1, \ldots, n\} < \infty$, ist $\operatorname{span}\{k(x_i, \cdot) \,:\, i = 1, \ldots, n\}$ abgeschlossen im RKHS von $k\big|_{X' \times X'}$, d.h.

$$\operatorname{span}\{k(x_i, \cdot)\} = \overline{\operatorname{span}\{k(x_i, \cdot)\}} = \text{ RKHS von } k\big|_{X' \times X'}.$$

Betrachte das SVM Problem bezüglich $k\big|_{X' \times X'}$,

$$\inf_{f \in H\big|_{X'}} \lambda \left\|f\right\|_H^2 + \mathcal{R}_{L,D}(f). \tag{*}$$

Nach Lemma 6-1.3 existiert ein $f_{D,\lambda,H|_{X'}} \in H|_{X'}$. Dieses ist Lösung von (*), denn $\dim H|_{X'} < \infty$. Da $H|_{X'} \subset H$ abgeschlossen, existiert das orthogonale Komplement $H|_{X'}^{\perp}$. Für $f \in H|_{X'}^{\perp}$ gilt dann

$$f(x,\cdot) = \bigl\langle f, k(x_i,\cdot)\bigr\rangle_H = 0.$$

Schreibe $P_{X'} : H \to H$ und $P_{X'}^{\perp} : H \to H$ für die orthogonale Projektion auf $H|_{X'}$ bzw. $H|_{X'}^{\perp}$. Für $f \in H$ gilt somit $f = P_{X'}f + P_{X'}^{\perp}f$ und nach (*) ist

$$P_{X'}^{\perp} f(x_i) = 0, \qquad i = 1,\ldots,n.$$

Somit ist auch

$$\mathcal{R}_{L,D}(f) = \mathcal{R}_{L,D}(P_{X'}f + P_{X'}^{\perp}f) = \mathcal{R}_{L,D}(P_{X'}f)$$

und da ferner $\bigl\|P_{X'}f\bigr\|_H^2 \leqslant \bigl\|f\bigr\|_H^2$, gilt

$$\begin{aligned}
\inf_{f\in H} \lambda \bigl\|f\bigr\|_H^2 + \mathcal{R}_{L,D}(f) &\leqslant \inf_{f\in H|_{X'}} \lambda \bigl\|f\bigr\|_H^2 + \mathcal{R}_{L,D}(f) \\
&= \inf_{f\in H} \lambda \bigl\|P_{X'}f\bigr\|_H^2 + \mathcal{R}_{L,D}(P_{X'}f) \\
&\leqslant \inf_{f\in H} \lambda \bigl\|f\bigr\|_H^2 + \mathcal{R}_{L,D}(f).
\end{aligned}$$

Insbesondere ist daher

$$\begin{aligned}
\inf_{f\in H} \lambda \bigl\|f\bigr\|_H^2 + \mathcal{R}_{L,D}(f) &= \inf_{f\in H|_{X'}} \lambda \bigl\|f\bigr\|_{H|_{X'}}^2 + \mathcal{R}_{L,D}(f) \\
&= \lambda \Bigl\| f_{D,\lambda,H|_{X'}} \Bigr\|_{H|_{X'}}^2 + \mathcal{R}_{L,D}(f_{D,\lambda,H|_{X'}}).
\end{aligned}$$

Da $f_{D,\lambda,H|_{X'}} \in H$ löst $f_{D,\lambda,H|_{X'}}$ auch das original SVM-Problem und da $f_{D,\lambda}$ eindeutig und $f_{D,\lambda} \in H|_{X'}$ folgt die Darstellung. «

Als Konsequenz ist für $f = \sum_{i=1}^n \alpha_i k(x_i,\cdot)$,

$$\begin{aligned}
\min_{f\in H} \lambda \bigl\|f\bigr\|_H^2 + \mathcal{R}_{L,D}(f) &= \min_{\alpha_1,\ldots,\alpha_n\in\mathbb{R}} \lambda \bigl\|f\bigr\|_H^2 + \mathcal{R}_{L,D}(f) \\
&= \min_{\alpha_1,\ldots,\alpha_n\in\mathbb{R}} \lambda \sum_{i,j=1}^n \alpha_i\alpha_j k(x_i,x_j) + \mathcal{R}_{L,D}\left(\sum_{i=1}^n \alpha_i k(x_i,\cdot)\right)
\end{aligned}$$

Dies ist ein endlichdimensionales und *konvexes* Optimierungsproblem.

BSP Verlustfunktion der kleinsten Quadrate $L(x, y, t) = (y - t)^2$. Es ist

$$\frac{\delta}{\delta\alpha_{i_0}}\left(\lambda\sum_{i,j=1}^{n}\alpha_i\alpha_j k(x_i,x_j) + \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{n}\alpha_j k(x_i,x_j)\right)^2\right)$$
$$= 2\lambda\sum_{i=1}^{n}\alpha_i k(x_i,x_{i_0}) + \frac{\delta}{\delta\alpha_{i_0}}\left(\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{n}\alpha_j k(x_i,x_j)\right)^2\right), \quad (*)$$

wobei

$$\frac{\delta}{\delta\alpha_{i_0}}\left(\left(y_i - \sum_{j=1}^{n}\alpha_j k(x_i,x_j)\right)^2\right)$$
$$= -2\left(y_i - \sum_{j=1}^{n}\alpha_j k(x_i,x_j)\right)\underbrace{\frac{\delta}{\delta_{a_{i_0}}}\left(\sum_{j=1}^{n}\alpha_j k(x_i,x_j)\right)}_{k(x_i,x_{i_0})}.$$

Somit gilt

$$(*) = 2\lambda\sum_{i=1}^{n}\alpha_i k(x_i,x_{i_0}) - \frac{2}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{n}\alpha_j k(x_i,x_j)\right)k(x_i,x_{i_0})$$
$$\stackrel{!}{=} 0.$$

Schreibe k als $n \times n$-Matrix $k = \left(k(x_i,x_j)\right)_{i,j=1}^{n}$. Dann ist $k = k^\top$ und die Optimierung ist gleichbedeutend mit

$$2\lambda k\alpha - \frac{2}{n}ky + \frac{2}{n}kk\alpha \stackrel{!}{=} 0 \Leftrightarrow \lambda nk\alpha + kk\alpha = ky$$
$$\Leftrightarrow k(\lambda nE_n\alpha + k\alpha) = ky \quad (**)$$

wobei E_n die n-te Einheitsmatrix bezeichne. Falls $\lambda_n E_n\alpha + k\alpha = y$, dann ist (**) erfüllt.

Da k ein Kern ist die Matrix k positiv semi-definit, d.h. k hat n nichtnegative Eigenwerte,

$$\lambda n \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix} + \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}.$$

$\lambda n E_n + k$ hat folglich strikt positive Eigenwerte und ist daher invertierbar.

$$\alpha = (\lambda n E_n + k)^{-1} y$$

ist die eindeutige Lösung und

$$f_{D,\lambda} = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot). \quad \blacksquare$$

6-1-1 Ausflug in das Reich der konvexen Analysis

Sei E ein endlichdimensionaler Vektorraum, $P : E \to \mathbb{R}$ konvex und stetig differenzierbar sowie $h_i : E \to \mathbb{R}$ affin linear für $i = 1, \dots, m$.

Primales Optimierungsproblem *Wir suchen*

$$P^* = \inf_{\omega \in E} P(\omega)$$

unter den Nebenbedingungen $h_i(\omega) \leqslant 0$ *für alle* $i = 1, \dots, m$. ⋊

Wir nehmen für alles Weitere an, dass ein ω^* existiert mit $h_i(\omega^*) \leqslant 0$ für $i = 1, \dots, m$ und $P(\omega^*) = P^*$. Unser Ziel ist es das primale Optimierungsproblem auf ein duales Optimierungsproblem zurückzuführen.

Definition *Die Funktion*

$$L(\omega, \beta) := P(\omega) + \sum_{i=1}^{m} \beta_i h_i(\omega), \qquad \omega \in E,\ \beta \in \mathbb{R}^n$$

heißt Lagrangefunktion.

$$D(\beta) := \inf_{\omega \in E} L(\omega, \beta)$$

heißt die zur Lagrangefunktion gehörige duale Funktion. ⋊

Da P konvex und die h_i affin linear, ist auch L konvex.

Duales Optimierungsproblem

$$D^* = \sup_{\beta \geqslant 0} D(\beta). \quad \rtimes$$

Wie hängen das duale Problem mit seinen Lösungen mit dem primalen Problem zusammen?

Lemma *Für $\beta \geqslant 0$ und $\omega \in E$ mit $h_i(\omega) \leqslant 0$ für $i = 1, \ldots, m$ gilt*

$$D(\beta) \leqslant P(\omega)$$

und damit insbesondere $D^* \leqslant P^*$. $\rtimes$

» Man rechnet direkt nach, dass

$$D(\beta) = \inf_{\omega' \in E} L(\omega', \beta) \leqslant L(\omega, \beta) = P(\omega) + \sum_{i=1}^{m} \beta_i h_i(\omega) \leqslant P(\omega). \quad «$$

Satz $D^* = P^*$. $\rtimes$

» Wir überspringen den Beweis, da dieser zu weit in die konvexe Analysis hineinführt. «

Korollar *Sei $\beta^* \geqslant 0$ mit $D(\beta^*) = D^*$, dann gilt*

$$D^* = \max_{\beta \geqslant 0} \inf_{\omega} L(\omega, \beta) = \min_{\omega} \sup_{\beta \geqslant 0} L(\omega, \beta) = P^*. \quad \rtimes$$

» „$\leqslant$“: gilt immer.
„$\geqslant$“: Sei $\omega^* \in E$ eine primale Lösung, d.h.

$$h_i(\omega^*) \leqslant 0, \quad i = 1, \ldots, m, \qquad P(\omega^*) = P^*.$$

Dann ist

$$D^* = D(\beta^*) = \max_{\beta \geqslant 0} \inf_{\omega \in E} L(\omega, \beta),$$

$$P^* = P(\omega^*) = \min_{\omega, h_i(\omega) = 0} P(\omega) = \min_{\omega \in E} \sup_{\beta \geqslant 0} P(\omega) + \sum_{i=1}^{m} \beta_i h_i(\omega) \qquad (*)$$

denn

$$\sup_{\beta \geqslant 0} P(\omega) + \sum_{i=1}^{m} \beta_i h_i(\omega) = \begin{cases} P(\omega), & h_i(\omega) \leqslant 0,\ i = 1, \ldots, m, \\ \infty, & \text{sonst.} \end{cases}$$

und folglich ist

$$(*) = \min_{\omega \in E} \sup_{\beta \geqslant 0} L(\omega, \beta).$$

Die Behauptung folgt, dass $P^* = D^*$. «

Satz *Sei β^* die duale Lösung, d.h. $\beta^* \geqslant 0$ und $D(\beta^*) = D^*$ und ω^* die primale Lösung, d.h. $h_i(\omega^*) \leqslant 0$ für $i = 1, \ldots, m$ mit $P(\omega^*) = P^*$. Dann folgt*

$$\max_{\beta \geqslant 0} L(\omega^*, \beta) = L(\omega^*, \beta^*) = \min_{\omega \in E} L(\omega, \beta^*)$$

und $L(\omega^*, \beta^*) = D^* = P^*$. ⋊

(ω^*, β^*) ist Sattelpunkt von L.

» Es gilt

$$D^* = \max_{\beta \geqslant 0} \inf_{\omega \in E} L(\omega, \beta) = \min_{\omega \in E} \sup_{\beta \geqslant 0} L(\omega, \beta) = P(\omega^*) = \sup_{\beta \geqslant 0} L(\omega^*, \beta).$$

und

$$P^* = \min_{\omega \in E} \sup_{\beta \geqslant 0} L(\omega, \beta) = \max_{\beta \geqslant 0} \inf_{\omega \in E} L(\omega, \beta) = D^* = D(\beta^*) = \inf_{\omega \in E} L(\omega, \beta^*)$$

sowie

$$L(\omega^*, \beta^*) \geqslant \inf_{\omega \in E} L(\omega, \beta^*) = P^*,$$
$$L(\omega^*, \beta^*) \leqslant \sup_{\beta \geqslant 0} L(\omega^*, \beta) = D^*,$$

und folglich ist

$$L(\omega^*, \beta^*) = \sup_{\beta \geqslant 0} L(\omega^*, \beta) = \inf_{\omega \in E} L(\omega, \beta^*) = P^* = D^*$$

und sup ist max und inf ist min. «

Korollar *Seien β^* und ω^* wie oben. Dann ist*

$$\beta_i^* h_i(\omega^*) = 0 \text{ für alle } i = 1, \dots, m. \quad \bowtie$$

» $P(\omega^*) = P^* = L(\omega^*, \beta^*) = P(\omega^*) + \sum_{i=1}^m \beta_i h_i(\omega^*)$. «

Korollar *Sei $E = E_1 \times E_2$, $\omega = (\omega_1, \omega_2) \in E_1 \times E_2$ und L derart, dass:*

Für alle $\beta^* \geqslant 0$ mit $D(\beta^*) = D^*$ existiert genau ein $\omega_1 \in E_1$ und ein (oder mehrere) $\omega_2 \in E_2$ mit $h_i(\omega_1, \omega_2) \leqslant 0$ für $i = 1, \dots, m$ und

$$L((\omega_1, \omega_2), \beta^*) = D(\beta^*) = \inf_{\omega \in E} L(\omega, \beta^*).$$

Dann gilt für jede primale Lösung $\omega^* = (\omega_1^*, \omega_2^*)$,

$$\omega_1 = \omega_1^*.$$

Insbesondere: Falls wir wissen, dass es eine primale Lösung ω^* gibt, so finden wir die erste Komponente ω_1^* indem wir

$$\inf_{\omega \in E} L(\omega, \beta^*)$$

lösen und für die Lösung $\omega = (\omega_1, \omega_2)$ gilt $\omega_1^* = \omega_1$. $\bowtie$

» Sei ω^* eine Lösung, dann ist

$$h_i(\omega^*) = h_i(\omega_1^*, \omega_2^*) \leqslant 0, \qquad i = 1, \dots, m$$

und der vorige Satz zeigte $L(\omega^*, \beta^*) = D^* = D(\beta^*)$. Mit der vorausgesetzten Eindeutigkeit in der ersten Komponente folgt $\omega_1^* = \omega_1$. «

Als Anwendung betrachten wir SVMs mit Hinge-loss.

Bsp SVMs mit Hinge-loss. Das Optimierungsproblem

$$\min_{f \in H} \lambda \left\| f \right\|_H^2 + \frac{1}{n} \sum_{i=1}^n \max \left\{ 0, 1 - y_i f(x) \right\}$$

ist ein SVM Optimierungsproblem, da $L_{\text{hinge}}(y,t) = \max\{0, 1 - yt\}$. Dieses ist äquivalent zu dem Optimierungsproblem

$$\min_{\substack{f\in H\\ \xi\in\mathbb{R}^n}} \frac{1}{2}\langle f, f\rangle + c\sum_{i=1}^{n}\xi_i, \qquad c = \frac{1}{2\lambda n}$$

mit den Nebenbedingungen

$$\xi_i \geqslant 0, \quad \xi_i \geqslant 1 - y_i f(x_i) \Leftrightarrow \xi_i \geqslant \max 0, 1 - y_i f(x_i).$$

Die Lagrangefunktion des Problems ist

$$L(f,\xi,\beta,\gamma) = \frac{1}{2}\langle f, f\rangle + c\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\beta_i\xi_i + \sum_{i=1}^{n}\gamma_i(1 - y_i f(x_i) - \xi_i).$$

L ist differenzierbar und

$$\frac{\delta}{\delta f}L(f,\xi,\beta,\gamma) = f - \sum_{i=1}^{n}\gamma_i y_i k(x_i,\cdot) \stackrel{!}{=} 0$$

führt auf die eindeutige Lösung

$$f = \sum_{i=1}^{n}\gamma_i y_i k(x_i,\cdot).$$

Weiterhin führt

$$\frac{\delta}{\delta\xi_i}L(f,\xi,\beta,\gamma) = c - \beta_{i_0} - \gamma_{i_0} \stackrel{!}{=} 0$$

auf $\beta_{i_0} + \gamma_{i_0} = c$.

Ist $\beta + \gamma = c$, so minimiert jedes ξ. Ist andererseits $\beta + \gamma \neq c$, so gibt es keine Lösung.

Setzen wir dies in unser Optimierungsproblem ein, so erhalten wir

$$\begin{aligned} D(\beta,\gamma) &= \frac{1}{2}\sum_{i=1}^{n}\gamma_i y_j \gamma_i y_j k(x_i,x_j) + \sum_{i=1}^{n}\gamma_i - \sum_{i=1}^{n}\gamma_i y_j \gamma_i y_j k(x_i,x_j) \\ &= \sum_{i=1}^{n}\gamma_i - \frac{1}{2}\sum_{i=1}^{n}\gamma_i y_j \gamma_i y_j k(x_i,x_j) \end{aligned}$$

mit $\beta \geqslant 0$, $\gamma \geqslant 0$ und $\beta + \gamma = c$, d.h. $\gamma \in [0,c]^n$ und $\beta \coloneqq c - \gamma$.

Damit ist das duale Problem

$$\max_{\gamma \in [0,c]^n} \sum_{i=1}^{n} \gamma_i - \frac{1}{2} \sum_{i,j=1}^{n} \gamma_i \gamma_j \gamma_i \gamma_j k(x_i, x_j). \tag{*}$$

- mit der Definition $\mathbf{k} = (\gamma_i \gamma_j k(x_i, x_j))_{i,j=1}^n$ -

$$\max_{\gamma \in [0,c]^n} \langle \gamma, 1 \rangle - \frac{1}{2} \langle \gamma, \mathbf{k}\gamma \rangle .$$

Ist γ^* Lösung des dualen Problems, so ist

$$f_{D,x} = \sum_{i=1}^{n} \gamma_i \gamma_i^* k(x_i, \cdot)$$

die Lösung der ersten Komponente des primalen Problems.

$$\xi_i^* = \max\left\{0, 1 - \gamma_i f_{D,x}(x_i)\right\}$$

ist die zweite Komponente.

Bemerkung. Sei $\gamma \in [0,c]^n$ und $f_\gamma = \sum_{i=1}^n \gamma_i \gamma_i k(x_i, \cdot)$. Gilt nun

$$P(f_\gamma, \xi_\gamma) - D(\gamma) \leqslant \varepsilon,$$

dann ist auch

$$P(f_\gamma) - P^* \leqslant \varepsilon,$$

da $D(\gamma) \leqslant P^*$.

$$P(f_\gamma, \xi_\gamma) = \sum_{i,j=1}^{n} \gamma_i \gamma_j \gamma_i \gamma_j k(x_i, x_j) + c \sum_{i=1}^{n} \max\left\{0, 1 - \gamma_i f_\gamma(x_i)\right\} - \sum_{i=1}^{n} \gamma_i. \quad \multimap$$

Dann ist

$$\lambda \left\| f_\gamma \right\|_H^2 + \mathcal{R}_{L,D}(f_\gamma) - \min_{f \in H} \left(\lambda \left\| f \right\|_H^2 + \mathcal{R}_{L,D}(f) \right) = 2\lambda \left(P(f_\gamma, \xi_\gamma) - P^* \right).$$

Gilt nun

$$P(f_\gamma, \xi_\gamma) - P^* \leqslant \frac{\varepsilon}{2\lambda} \tag{+}$$

so ist

$$\lambda \left\| f_\gamma \right\|_H^2 + \mathcal{R}_{L,D}(f_\gamma) \leqslant \inf_{f \in H} \left(\lambda \left\| f \right\|_H^2 + \mathcal{R}_{L,D}(f) \right) + \varepsilon. \quad \blacksquare$$

Wie löst man das duale Problem?

- Standardsoftware.
- Gradient assent. Zu γ berechne den Gradienten von (*) und mache einen Step in Richtung des steilsten Anstiegs. Wiederhole den Vorgang, bis (+) erreicht ist.

  Ein Problem bei diesem Verfahren ist, dass die Kernmatrix ein enormer Platzfresser ist. Es gibt jedoch zahlreiche Verfahren, die nicht die gesamte Kernmatrix in den RAM laden muss.
- Man betrachtet die Abbildung

  $$\gamma_i \mapsto D(\gamma + \gamma_i e_i).$$

  Diese ist eine eindimensionale konkave und quadratische Funktion und ihr Optimierungsproblem ist explizit lösbar.

  (a) Suche Richtung i^* mit maximalem Gewinn in D bei Optimierung in Richtung i^*.

  (b) Optimiere in Richtung i^*.

  (c) Gehe zu (a), falls (*) nicht erfüllt ist.

  Die meisten verfügbaren Programme machen dies so.

6-2 Orakelungleichungen für SVMs

6-2.1 **Satz** *Sei (X, d) ein kompakter, metrischer Raum, L eine konvexe und lokal lipschitzstetige Verlustfunktion mit $L(x, y, 0) \leqslant 1$ für alle $(x, y) \in X \times Y$. Ferner sei H*

ein RKHS über X mit stetigem Kern k und $\|k\|_\infty \leqslant 1$ und P sei ein Wahrscheinlichkeitsmaß auf $X \times Y$.

Für alle $\lambda > 0$, $n \geqslant 1$, $\varepsilon > 0$ und $\tau > 0$ gilt mit einer Wahrscheinlichkeit P^n nicht kleiner als $1 - \varepsilon^{-\tau}$:

$$\begin{aligned}
&\lambda \left\|f_{D,\lambda}\right\|_H^2 + \mathcal{R}_{L,P}(f_{D,\lambda}) - \mathcal{R}_{L,P}^* < \inf_{f \in H} \left(\lambda \left\|f\right\|_H^2 + \mathcal{R}_{L,P}(f) - \mathcal{R}_{L,P}^*\right) \\
&+ 4\varepsilon \left|L\right|_{2\lambda^{-1/2},1} + \left(2 \left|L\right|_{2\lambda^{-1/2},1} \lambda^{-1/2} + 1\right) \\
&\cdot \sqrt{\frac{2\tau + 2\log 2\mathcal{N}(B_H, \|\cdot\|_\infty, \frac{1}{2}\lambda^{1/2}\varepsilon)}{n}}.
\end{aligned}$$

⋊

» Satz 5-3.7 zeigte, dass die Abbildung

$$\mathrm{id} : H \to C(X)$$

kompakt ist. Daraus folgt

$$\log \mathcal{N}(B_H, \|\cdot\|_\infty, \varepsilon) < \varepsilon, \qquad \varepsilon > 0.$$

Lemma 6-1.4 und 5-3.1 zeigten

$$\begin{aligned}
\left\|f_{D,\lambda}\right\|_H &\leqslant \lambda^{-1/2}, \qquad \lambda > 0, \\
\left\|f_{D,\lambda}\right\|_\infty &\leqslant \lambda^{-1/2}, \qquad \text{da } \|\mathrm{id} : H \to C(X)\| = \|k\|_\infty \leqslant 1.
\end{aligned}$$

Sei nun $\delta > 0$, dann existiert ein $f_\delta \in H$ mit

$$\lambda \left\|f_\delta\right\|_H^2 + \mathcal{R}_{L,P}(f_\delta) \leqslant \inf_{f \in H} \left(\lambda \left\|f\right\|_H^2 + \mathcal{R}_{L,P}(f)\right) + \delta,$$

und damit

$$\begin{aligned}
&\lambda \left\|f_{D,\lambda}\right\|_H^2 + \mathcal{R}_{L,P}(f_{D,\lambda}) - \inf_{f \in H} \left(\lambda \left\|f\right\|_H^2 + \mathcal{R}_{L,P}(f)\right) \\
&\leqslant \lambda \left\|f_{D,\lambda}\right\|_H^2 + \mathcal{R}_{L,P}(f_{D,\lambda}) - \lambda \left\|f_\delta\right\|_H^2 - \mathcal{R}_{L,P}(f_\delta) + \delta \\
&\quad - \mathcal{R}_{L,D}(f_{D,\lambda}) + \mathcal{R}_{L,D}(f_{D,\lambda}) \qquad (*)
\end{aligned}$$

Weiterhin ist

$$\lambda \left\|f_{D,\lambda}\right\|_H^2 + \mathcal{R}_{L,D}(f_D) \leqslant \lambda \left\|f_\delta\right\|_H^2 + \mathcal{R}_{L,D}(f_\delta),$$

da $f_{D,\lambda}$ diese Funktion minimiert. Es folgt

$$\begin{aligned}(*) &\leqslant \lambda \left\|f_\delta\right\|_H^2 + \mathcal{R}_{L,P}(f_{D,\lambda}) - \lambda \left\|f_\delta\right\|_H^2 - \mathcal{R}_{L,P}(f_\delta) + \delta \\ &\quad - \mathcal{R}_{L,D}(f_{D,\lambda}) + \mathcal{R}_{L,D}(f_\delta) \\ &= \mathcal{R}_{L,P}(f_{D,\lambda}) - \mathcal{R}_{L,D}(f_{D,\lambda}) - \left(\mathcal{R}_{L,P}(f_\delta) - \mathcal{R}_{L,D}(f_\delta)\right) + \delta \\ &\leqslant \left|\mathcal{R}_{L,P}(f_{D,\lambda}) - \mathcal{R}_{L,D}(f_{D,\lambda})\right| + \left|\left(\mathcal{R}_{L,P}(f_\delta) - \mathcal{R}_{L,D}(f_\delta)\right)\right| + \delta. \qquad (**)\end{aligned}$$

Dann gilt ferner

$$\begin{aligned}\lambda \left\|f_\delta\right\|_H^2 &\leqslant \lambda \left\|f_\delta\right\|_H^2 + \mathcal{R}_{L,P}(f_\delta) \leqslant \inf_{f\in H}\left(\lambda \left\|f\right\|_H^2 + \mathcal{R}_{L,P}(f)\right) + \varepsilon \\ &\leqslant \mathcal{R}_{L,P}(0) + \delta \leqslant 1 + \delta \leqslant 4,\end{aligned}$$

falls $\delta \leqslant 3$. Damit ist

$$\begin{aligned}f_\delta &\in 2\lambda^{-1/2} B_H, \\ f_{D,\lambda} &\in \lambda^{-1/2} B_H \subset 2\lambda^{-1/2} B_H.\end{aligned}$$

Folglich ist

$$(**) \leqslant 2 \sup_{f\in 2\lambda^{-1/2}B_H} \left|\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)\right| + \delta$$

für jedes $\delta > 0$ also gilt auch

$$(**) \leqslant 2 \sup_{f\in 2\lambda^{-1/2}B_H} \left|\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,D}(f)\right|.$$

Definiere nun

$$B := 2\lambda^{-1/2} \left|L\right|_{2\lambda^{-1/2}} + 1.$$

Für $f \in 2\lambda^{-1/2} B_H$ gilt dann

$$\begin{aligned}L(x,y,f(x)) &\leqslant \left|L(x,y,f(x)) - L(x,y,0)\right| + \underbrace{L(x,y,0)}_{\leqslant 1} \\ &\leqslant \left|L\right|_{2\lambda^{-1/2},1} \left|f(x) - 0\right| + 1 \leqslant \left|L\right|_{2\lambda^{-1/2},1} 2\lambda^{-1/2} + 1 \leqslant B.\end{aligned}$$

Sein nun $\mathcal{F}_\varepsilon$ ein ε-Netz von $2\lambda^{-1/2}B_H$ und

$$\left|\mathcal{F}_\varepsilon\right| = \mathcal{N}(2\lambda^{-1/2}B_H, \|\cdot\|_\infty, \varepsilon) = \mathcal{N}(B_H, \|\cdot\|_\infty, \frac{1}{2}\lambda^{1/2}\varepsilon).$$

Für $f \in 2\lambda^{-1/2}B_H$ gibt es $g \in \mathcal{F}_\varepsilon$ mit $\left\|f - g\right\|_\infty < \varepsilon$ und

$$\begin{aligned}\left|\mathcal{R}_{L,D}(f) - \mathcal{R}_{L,P}(f)\right| &\leqslant \left|\mathcal{R}_{L,D}(f) - \mathcal{R}_{L,D}(g)\right| + \left|\mathcal{R}_{L,D}(g) - \mathcal{R}_{L,P}(g)\right| \\ &\quad + \left|\mathcal{R}_{L,P}(g) - \mathcal{R}_{L,P}(f)\right| \\ &\leqslant 2\varepsilon\,|L|_{2\lambda^{-1/2},1} + \left|\mathcal{R}_{L,D}(g) - \mathcal{R}_{L,P}(g)\right|.\end{aligned}$$

Somit ist

$$\begin{aligned}&\lambda\left\|f_{D,\lambda}\right\|_H^2 + \mathcal{R}_{L,P}(f_{D,\lambda}) - \inf_{f\in H}\left(\lambda\left\|f\right\|_H^2 + \mathcal{R}_{L,P}(f)\right) \\ &\quad\leqslant 4\varepsilon\,|L|_{2\lambda^{-1/2},1} + 2\sup_{g\in\mathcal{F}_\varepsilon}\left|\mathcal{R}_{L,P}(g) - \mathcal{R}_{L,D}(g)\right|.\end{aligned}$$

Wir schätzen weiterhin ab,

$$\begin{aligned}&P^n\left(\left\{D : \lambda\left\|f_{D,\lambda}\right\|_H^2 + \mathcal{R}_{L,P}(f_{D,\lambda}) - \inf_{f\in H}\left(\lambda\left\|f\right\|_H^2 + \mathcal{R}_{L,P}(f)\right)\right.\right. \\ &\qquad\left.\left.\geqslant B\sqrt{\frac{2\tau}{m}} + 4\varepsilon\,|L|_{2\lambda^{-1/2}}\right\}\right) \\ &\leqslant P^n\left(\left\{D : 2\sup_{g\in\mathcal{F}_\varepsilon}\left|\mathcal{R}_{L,P}(g) - \mathcal{R}_{L,D}(g)\right| \geqslant B\sqrt{\frac{2\tau}{m}}\right\}\right) \\ &\overset{!}{\leqslant} 2\left|\mathcal{F}_\varepsilon\right| e^{-\tau} = \exp(-\tau + \log 2\mathcal{N}(B_H, \|\cdot\|_\infty, \frac{1}{2}\lambda^{1/2}\varepsilon)),\end{aligned}$$

wobei der union bound und die Hoeffdings Ungleichung verwendet wurden. Mit einer Variablentransformation für τ folgt schließlich die Behauptung. «

Bemerkung. Mit etwas mehr Theorie über Hilberträume kann man zeigen, dass das inf tatsächlich angenommen wird. ⊸

6-2.2 **Korollar** *Es gelten die Voraussetzungen des Satzes 6-2.1 und außerdem sei L lipschitz stetig mit $|L|_1 \leqslant 1$. Ferner existieren Konstanten $a \geqslant 1$, $p > 0$ mit*

$$\log \mathcal{N}(B_H, \|\cdot\|_\infty, \varepsilon) \leqslant a\varepsilon^{-2p}, \qquad \varepsilon > 0,$$

dann folgt, dass für alle $\lambda \in (0,1]$, $\tau > 0$ *und* $n \geqslant 1$ *mit einer Wahrscheinlichkeit* P^n *nicht kleiner als* $1 - e^{-\tau}$ *gilt,*

$$\lambda \left\| f_{D,\lambda} \right\|_H^2 + \mathcal{R}_{L,P}(f_{D,\lambda}) - \mathcal{R}_{L,P}^* \leqslant \inf_{f \in H} \lambda \left\| f \right\|_H^2 + \mathcal{R}_{L,P}(f) - \mathcal{R}_{L,P}^* + 2\lambda^{-1/2} \left(\frac{a}{n} \right)^{(2+2p)^{-1}} + 4\lambda^{-1/2} \sqrt{\frac{2\tau + 2}{n}}. \quad \bowtie$$

» Eine einfache Rechnung zeigt …

$$\sqrt{2\tau + 2\log(2\mathcal{N}(B_H, \|\cdot\|_\infty, \tfrac{1}{2}\lambda^{1/2}\varepsilon))} \leqslant \sqrt{2\tau + 2\log 2 + a\left(\tfrac{1}{2}\lambda^{1/2}\varepsilon\right)^{-2p}}$$

und da $2\log 2 \leqslant 2$, $\sqrt{s+r} \leqslant \sqrt{s} + \sqrt{r}$ für $s, r > 0$,

$$\leqslant \sqrt{2(\tau+1)} + a^{1/2} 2^p \lambda^{-p/2} \varepsilon^{-p}.$$

Da $4\varepsilon \, |L|_{2\lambda^{-1/2},1} \leqslant 4\varepsilon$ und analog

$$2 \, |L|_{2\lambda^{-1/2},1} \, \lambda^{-1/2} + 1 \leqslant 2\lambda^{-1/2} + 1 \leqslant 4\lambda^{-1/2},$$

erhalten wir

$$4\varepsilon \, |L|_{2\lambda^{-1/2},1} + \left(2 \, |L|_{2\lambda^{-1/2},1} \, \lambda^{-1/2} + 1\right) \cdot \sqrt{\frac{2\tau + 2\log 2\mathcal{N}(B_H, \|\cdot\|_\infty, \frac{1}{2}\lambda^{1/2}\varepsilon)}{n}}$$
$$\leqslant 4\varepsilon + 4\lambda^{-1/2} \left(\sqrt{\frac{2(\tau+1)}{n}} + \frac{a^{1/2} 2^p \lambda^{-p/2} \varepsilon^{-p}}{\sqrt{n}} \right).$$

Betrachten wir nun

$$h(\varepsilon) := \varepsilon + \lambda^{-1/2} 2^p \lambda^{-p/2} \varepsilon^{-p} \left(\frac{a}{n} \right)^{1/2},$$

so führt das Lösen von $h'(\varepsilon) = 0$ auf die Minimalstelle

$$\varepsilon^* = p^{(1+p)^{-1}} 2^{(1+p)^{-1}} \lambda^{-1/2} \left(\frac{a}{n} \right)^{(2+2p)^{-1}}.$$

Setzen wir diese ein, ergibt sich,

$$h(\varepsilon^*) = \underbrace{2^{\frac{p}{1+p}}}_{\leqslant 2} p^{\frac{1-p}{1+p}} \lambda^{-1/2} \left(\frac{a}{n} \right)^{\frac{1}{2+2p}}$$

Weiterhin besitzt

$$q(p) = p^{\frac{1-p}{1+p}},$$
$$q'(p) = -\frac{p^{-\frac{2p}{1+p}}\left(2\ln p^p - 1 + p^2\right)}{(1+p)^2}$$

ein Maximum in $p = 1$. Folglich ist

$$h(\varepsilon^*) \leqslant 2\lambda^{-1/2}\left(\frac{a}{n}\right)^{\frac{1}{2+2p}}.$$

Setzen wir dies in die Orakelungleichung 6-2.1 ein, folgt die Behauptung. «

Bsp Sei $X \subset \mathbb{R}^n$ kompakt, $O = X^\circ$ nichtleer.

$$C_b^m(\overline{O}) \coloneqq \left\{f \in C(X) \cap C^m(O) \,:\, f^{(k)} \text{ stetig fortsetzbar auf } X,\ 1 \leqslant k \leqslant m\right\}.$$

Weiterhin sei $k \in C_b^m(\overline{O})$ ein Kern mit RKHS H. Mit Korollar 5-3.11 folgt, dass

$$\mathrm{id} : H \to C_b^m(\overline{O})$$

eine stetige Abbildung ist. Aus den Bemerkungen in Kapitel 4-2 folgt für die Entropiezahl e_n,

$$e_n\left(\mathrm{id} : C_b^m(\overline{O}) \to l_\infty(\overline{O})\right) \leqslant cn^{-\frac{m}{d}}, \qquad n \geqslant 1$$

mit einer Konstanten c.

Aus der Multiplikativität von Entropiezahlen und $e_i(\cdot) = \|\cdot\|$ folgt, weiter

$$e_n(\mathrm{id} : H \to l_\infty(\overline{O})) \leqslant cn^{-\frac{m}{d}}\left\|\mathrm{id} : H \to C_b^m(\overline{O})\right\|.$$

Wir können nun Lemma 4-2.3 anwenden und erhalten

$$\log \mathcal{N}(B_H, \|\cdot\|_\infty, \varepsilon) \leqslant \log 4\left(\frac{c}{\varepsilon}\right)^{d/m}.$$

Mit $a = \ln 4c^{d/m}$ und $2p = \frac{d}{m}$ ist nun Korollar 6-2.2 anwendbar.

Betrachten wir nun den Gauß-Kern

$$\Omega_\sigma(x, x') = \exp(-\sigma^2\left\|x - x'\right\|^2)$$

mit RKHS H_Ω, dann können wir m beliebig groß wählen. Es gilt dann

$$e_n(\mathrm{id} : H_\sigma \to l_\infty(X)) \leqslant c\sigma^m n^{-m/d}, \qquad \sigma \geqslant 1, \quad m \geqslant 1, \quad n \geqslant 1$$

und c einer von n, m, σ unabhängigen Konstanten. Folglich ist

$$\log \mathcal{N}(B_\sigma, \|\cdot\|_\infty, \varepsilon) \leqslant \log 4 \left(\frac{c\sigma^m}{\varepsilon}\right)^{\frac{d}{m}} = \underbrace{c^{d/m} \log(4) \sigma^d}_{=:a} \varepsilon^{-\frac{d}{m}}.$$

Die rechte Seite der Orakelungleichung ist somit bis auf Konstanten, die von n, τ, m, λ unabhängig sind

$$\lambda^{-1/2} \sigma^{d \cdot \frac{1}{2+2\frac{d}{2m}}} n^{-\frac{1}{2+2\frac{d}{2m}}} + 4\lambda^{-1/2} \sqrt{\frac{2\tau+2}{n}} = \lambda^{-1/2} \sigma^{\frac{dm}{2m+d}} n^{-\frac{m}{2m+d}} + 4\lambda^{-1/2} \sqrt{\frac{2\tau+2}{n}}$$

Für $m >> 1$ entspricht dies ungefähr

$$\lambda^{-1/2} \sigma^{d/2} n^{-1/2} + 4\lambda^{-1/2} \sqrt{\frac{2\tau+2}{n}}. \quad \blacksquare$$

6-3 Die Funktion $A(\lambda)$

Wir definieren die Funktion A durch

$$A(\lambda) := \inf_{f \in H} \left(\lambda \|f\|_H^2 + \mathcal{R}_{L,P}(f)\right) - \mathcal{R}^*_{L,P}, \qquad \lambda > 0.$$

Im Folgenden untersuchen wir, wann $A(\lambda) \approx 0$.

6-3.1 **Satz** *Sei P ein W-Maß auf $X \times Y$ und L eine P-integrierbare NVF. Dann gilt für jedes $p \in (0, \infty]$*

$$\mathcal{R}^*_{L,P} = \inf_{f \in \mathcal{L}_p(P_X)} \mathcal{R}_{L,P}(f) = \mathcal{R}^*_{L,P,\mathcal{L}_p(P_X)}. \quad \bowtie$$

Es genügt also beschränkte, messbare Funktionen zur Berechnung des Bayes-Risikos zu betrachten.

» „$p = \infty$“: Sei $f : Y \to \mathbb{R}$ messbar mit $\mathcal{R}_{L,P}(f) < \infty$. (Falls es keine solche Funktion gibt, sind wird fertig.) Setze nun

$$f_n := f \cdot \chi_{\{|f| \leqslant n\}},$$

dann ist $f_n \in \mathcal{L}_\infty$ für alle $n \geqslant 1$. Weiterhin gilt

$$\begin{aligned}\left|\mathcal{R}_{L,P}(f_n) - \mathcal{R}_{L,P}(f)\right| &\leqslant \int_{X\times Y} \left|L(x,y,f_n(x)) - L(x,y,f(x))\right| \,\mathrm{d}P(x,y) \\ &= \int_{\{|f|>n\}} \left|L(x,y,0) - L(x,y,f(x))\right| \,\mathrm{d}P(x,y) \\ &\leqslant \int_{\{|f|>n\}} L(x,y,0) + L(x,y,f(x)) \,\mathrm{d}P(x,y)\end{aligned}$$

und da $L(x,y,t) \leqslant b(x,y) + h(|t|)$ folgt

$$\leqslant \int_{\{|f|>n\}} \underbrace{b(x,y) + h(0)}_{(*)} + \underbrace{L(x,y,f(x))}_{(**)} \,\mathrm{d}P(x,y).$$

$(*) \in L^1(P)$, denn L eine P-integrierbare NVF, $(**) \in L^1(P)$, da $\mathcal{R}_{L,P}(f) < \infty$ und folglich

$$\leqslant \int_{X\times Y} \chi_{\{|t|>n\}} g \,\mathrm{d}P \to 0, \qquad n \to \infty,$$

denn $g_n \leqslant g$ und $g_n \to 0$. Für $\varepsilon \downarrow 0$ gibt es daher eine beschränkte Funktion f_ε mit

$$\left|\mathcal{R}_{L,P}(f_\varepsilon) - \mathcal{R}_{L,P}(f)\right| \leqslant \varepsilon.$$

Dann ist

$$\mathcal{R}_{L,P}(f) - \varepsilon \leqslant \mathcal{R}_{L,P}(f_\varepsilon) \leqslant \mathcal{R}_{L,P}(f) + \varepsilon$$

und die Behauptung folgt.

„$p < \infty$“: $\mathcal{L}_\infty(P_Y) \subset \mathcal{L}_p(P_X)$. Damit sind wir fertig. «

6-3.2 **Satz** *Sei P ein W-Maß auf $X \times Y$ und L eine P-integrierbare NVF der Ordnung $p \in [1, \infty)$ und ist H ein universeller RKHS (also insbesondere ist X kompakt). Dann gilt*

$$\mathcal{R}^*_{L,P} = \mathcal{R}^*_{L,P,H} \coloneqq \inf_{f \in H} \mathcal{R}_{L,P}(f). \quad \rtimes$$

» Satz 2-1.6 besagt, dass die Abbildung

$$\mathcal{R}_{L,P} : \mathcal{L}_p(P_X) \to \mathbb{R}$$

stetig ist; weiterhin sind die Abbildungen

$$\mathrm{id} : H \to C(X), \qquad \mathrm{id} : C(X) \to \mathcal{L}_p(P_X)$$

stetig und haben dichtes Bild. Folglich existiert zu $g \in \mathcal{L}_p(P_X)$ eine Folge (f_n) in H mit $\|f_n - g\|_{\mathcal{L}_p(P_X)} \to 0$. Aufgrund der Stetigkeit von $\mathcal{R}_{L,P}$ ist

$$\lim_{n \to \infty} \mathcal{R}_{L,P}(f_n) = \mathcal{R}_{L,P}(g)$$

und mit Satz 6-3.1 folgt die Behauptung

$$\mathcal{R}^*_{L,P,H} = \mathcal{R}^*_{L,P,\mathcal{L}_p(X)} = \mathcal{R}^*_{L,P}. \quad «$$

Korollar *Falls die Voraussetzungen des Satzes 6-3.2 erfüllt sind, ist*

$$A(\lambda) = \inf_{f \in H} \left(\lambda \|f\|^2_H + \mathcal{R}_{L,P}(f) \right) - \mathcal{R}^*_{L,P,H}, \qquad \lambda \geqslant 0. \quad \rtimes$$

Bemerkung. Satz 6-3.2 gilt auch für nicht universelle Kerne, falls der RKHS H dicht in $\mathcal{L}_p(P_X)$ liegt. ⊸

6-3.3 **Lemma** *Sei L eine Verlustfunktion, H ein RKHS über X und P ein W-Maß auf $X \times Y$, so dass $\mathcal{R}^*_{L,P,H} = \mathcal{R}^*_{L,P}$. Dann erfüllt die Funktion $A : [0, \infty) \to [0, \infty]$ die folgenden Eigenschaften:*

(i) *A ist monoton steigend, stetig und konkav.*

(ii) *$A(0) = 0$.*

(iii) $\mu^{-1}A(\mu) \leqslant \lambda^{-1}A(\lambda)$ *für* $0 < \lambda < \mu$.

(iv) $A(\lambda) \leqslant \mathcal{R}_{L,P}(0) - \mathcal{R}^*_{L,P,H}$ *für alle* $\lambda \geqslant 0$.

(v) A *ist subadditiv, d.h.* $A(\lambda + \mu) \leqslant A(\lambda) + A(\mu)$ *für* $\lambda, \mu \geqslant 0$.

(vi) Falls eine Funktion $h : [0,1] \to [0,\infty)$ *mit* $\lim_{\lambda \to 0} h(\lambda) = 0$ *und* $A(\lambda) \leqslant \lambda h(\lambda)$ *für* $\lambda \in [0,1]$ *existiert, dann ist*

$$A(\lambda) = 0, \qquad \lambda \geqslant 0. \quad \bowtie$$

» „(ii)“: Man rechnet direkt nach, dass

$$A(0) = \inf_{f \in H} \left(0 \cdot \|f\|_H^2 + \mathcal{R}_{L,P}(f)\right) - \mathcal{R}^*_{L,P,H} = \inf_{f \in H} \mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P,H} = 0.$$

„(i)“: *A ist konkav.* Zu $\lambda \in [0,\infty)$ und $f \in H$ definiere

$$h_f(\lambda) := \lambda \|f\|_H^2 + \mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P,H}.$$

Dann ist h_f affin linear und

$$A(\lambda) = \inf_{f \in H} h_f(\lambda).$$

Seien $\varepsilon > 0$, $\lambda, \mu \geqslant 0$ und $\alpha \in [0,1]$. Wir zeigen nun, dass

$$\alpha(A\lambda) + (1-\alpha)A(\mu) \leqslant A(\alpha\lambda + (1-\alpha)\mu).$$

Da A das Infimum über h_f ist, existieren $f_1, f_2, f_3 \in H$, so dass

$$\begin{aligned} A(\lambda) &\leqslant h_{f_1}(\lambda) \leqslant A(\lambda) + \varepsilon, \\ A(\mu) &\leqslant h_{f_2}(\mu) \leqslant A(\mu) + \varepsilon, \\ h_{f_3}(\alpha\lambda + (1-\alpha)\mu) &\leqslant A(\alpha\lambda + (1-\alpha)\mu) + \varepsilon \end{aligned}$$

Somit gilt

$$\begin{aligned} \alpha(A\lambda) + (1-\alpha)A(\mu) &\leqslant \alpha h_{f_1}(\lambda) + (1-\alpha)h_{f_2}(\mu) \\ &\leqslant \alpha A(\lambda) + (1-\alpha)A(\mu) + \varepsilon \\ &\leqslant \alpha h_{f_3}(\lambda) + (1-\alpha)h_{f_3}(\mu) + \varepsilon \end{aligned}$$

und da h_f affin linear also konkav folgt

$$\ldots \leqslant h_{f_3}(\alpha\lambda + (1-\alpha)\mu) + \varepsilon \leqslant A(\alpha\lambda + (1-\alpha)\mu) + 2\varepsilon.$$

Grenzübergang $\varepsilon \to 0$ liefert, dass A konkav ist.

A ist stetig in 0. Sei $\varepsilon > 0$, dann existiert ein $f_\varepsilon \in H$ mit

$$\mathcal{R}_{L,P}(f_\varepsilon) - \mathcal{R}^*_{L,P,H} \leqslant \varepsilon.$$

Ohne Einschränkung ist $f_\varepsilon \neq 0$, denn sonst ist $A \equiv 0$. Für $\lambda \leqslant \left\|f_\varepsilon\right\|_H^{-2} \varepsilon$ folgt daher,

$$0 \leqslant A(\lambda) = \inf_{f\in H}\left(\lambda\left\|f\right\|_H^2 + \mathcal{R}_{L,P}(f)\right) - \mathcal{R}^*_{L,P,H} \leqslant \underbrace{\lambda\left\|f_\varepsilon\right\|^2}_{\leqslant\varepsilon} + \underbrace{\mathcal{R}_{L,P}(f_\varepsilon) - \mathcal{R}^*_{L,P,H}}_{\leqslant\varepsilon} \leqslant 2\varepsilon.$$

Somit folgt $A(\lambda) \to 0 = A(0)$ für $\lambda \to 0$.

A ist monoton steigend. Seien $0 \leqslant \lambda \leqslant \mu$, dann ist

$$h_f(\lambda) \leqslant h_f(\mu), \qquad f \in H$$

und folglich

$$A(\lambda) = \inf_{f'\in H} h_{f'}(\lambda) \leqslant h_f(\mu)$$

Dies gilt für jedes f, also auch für das Minimum

$$A(\lambda) \leqslant A(\mu).$$

„(iii)“: Sei $\lambda \leqslant \mu$, dann ist

$$A(\lambda) = A\left(\frac{\lambda}{\mu}\mu + \left(1-\frac{\lambda}{\mu}\right)\cdot 0\right) \geqslant \frac{\lambda}{\mu}A(\mu) + \left(1-\frac{\lambda}{\mu}\right)A(0) = \frac{\lambda}{\mu}A(\mu)$$
$$\Rightarrow \frac{1}{\mu}A(\mu) \leqslant \frac{1}{\lambda}A(\lambda).$$

„(v)“: Ohne Einschränkung sei $\lambda \leqslant \mu$ und folglich,

$$A(\lambda+\mu) \leqslant \frac{\lambda+\mu}{\mu}A(\mu) = \frac{\lambda}{\mu}A(\mu) + A(\mu) \leqslant A(\lambda) + A(\mu)$$

A ist stetig. Sei $\lambda > 0$, dann gilt für $\mu \geqslant 0$,

$$A(\mu) \leqslant A(\lambda + \mu) \leqslant A(\lambda) + A(\mu) \to A(\mu), \qquad \lambda \to 0.$$

„(iv)“:

$$\begin{aligned} A(\lambda) &= \inf_{f \in H} h_f(\lambda) = \inf_{f \in H} \left(\lambda \left\| f \right\|_H^2 + \mathcal{R}_{L,P}(f) \right) + \mathcal{R}^*_{L,P,H} \\ &\leqslant \mathcal{R}_{L,P}(0) - \mathcal{R}^*_{L,P,H}. \end{aligned}$$

„(vi)“: Für $\lambda \in (0,1]$ gilt nach (iii)

$$A(1) \leqslant \lambda^{-1} A(\lambda) \leqslant h(\lambda).$$

Konvergiert $h(\lambda) \to 0$ für $\lambda \to 0$, so ist $A(1) = 0$. Da A konkav und nichtnegativ folgt $A \equiv 0$. «

6-3.4 **Satz** *Existiert ein $f^* \in H$ mit*

$$\mathcal{R}_{L,P}(f^*) = \mathcal{R}^*_{L,P} = \mathcal{R}^*_{L,P,H}.$$

Dann ist

$$A(\lambda) \leqslant \lambda \left\| f^* \right\|_H^2, \qquad \lambda \geqslant 0. \quad \bowtie$$

» Sei $\lambda \geqslant 0$, dann ist

$$A(\lambda) = \inf_{f \in H} \left(\lambda \left\| f \right\|_H^2 + \mathcal{R}_{L,P}(f) \right) - \mathcal{R}^*_{L,P} \leqslant \lambda \left\| f^* \right\|_H^2. \quad «$$

Bemerkungen. A. Sei L konvex und k beschränkt, dann gilt

$$\exists\, c \geqslant 0 \;\forall\, \lambda \geqslant 0 : A(\lambda) \leqslant c\lambda \;\Leftrightarrow\; \exists\, f^* \in H : \mathcal{R}_{L,P}(f^*) = \mathcal{R}^*_{L,P}. \qquad (*)$$

B. Ein schnelleres Konvergenzverhalten als in (*) ist nur in trivialen Situationen möglich. Lemma 6-3.3 zeigte, dass dann $A(\lambda) = 0$ für $\lambda \geqslant 0$, d.h.

$$\mathcal{R}_{L,P}(0) = \mathcal{R}^*_{L,P}.$$

c. In vielen Situationen ist $A(\lambda) \leqslant c\lambda^{\beta}$ für $\lambda \geqslant 0$, mit $c \geqslant 0$ und $\beta \in (0,1]$ (z.B. gilt dies für die Verlustfunktion der kleinsten Quadrate), wobei

$$\exists\, c \geqslant 0 \;\forall\; \lambda \geqslant 0 : A(\lambda) \leqslant c\lambda^{\beta}, \quad \beta \in (0,1] \Leftrightarrow \exists\, f^*_{L,P} \in [L_2(P_X), H]_{\beta,\infty}\,. \qquad \multimap$$

6-3.5 **Lemma** *Sei $\mathcal{R}^*_{L,P} = \mathcal{R}^*_{L,P,H}$ und $I \subset (0,\infty)$ ein beschränktes Intervall und Λ sei ein endliches ε-Netz von I.*

Für Konstanten $\alpha, c \in (0,\infty)$ gilt dann

$$\min_{\lambda\in\Lambda} \left(A(\lambda) + c\lambda^{-\alpha}\right) \leqslant A(2\varepsilon) + \inf_{\lambda\in I} \left(A(\lambda) + c\lambda^{-\alpha}\right). \qquad \bowtie$$

» Ohne Einschränkgung ist $\Lambda := \{\lambda_1, \ldots, \lambda_m\}$ mit $\lambda_{i-1} \leqslant \lambda_i$ und $\lambda_0 := \inf I$.

Es gilt $0 < \lambda_i - \lambda_{i-1} \leqslant 2\varepsilon$ für alle $i = 1, \ldots, m$, denn $B_\varepsilon(\lambda_i) \cap B_\varepsilon(\lambda_{i+1}) \neq \emptyset$, sonst wäre Λ kein ε-Netz.

Sei nun $\delta > 0$, dann existiert ein $\lambda^* \in I$ mit

$$A(\lambda^*) + c(\lambda^*)^{-\alpha} \leqslant \inf_{\lambda\in I} \left(A(\lambda) + c\lambda^{-\alpha}\right) + \delta$$

und folglich existiert ein i mit $\lambda_{i-1} \leqslant \lambda^* \leqslant \lambda_i$, also $\lambda_i \leqslant \lambda^* + 2\varepsilon$. Dann gilt

$$\begin{aligned}\min_{\lambda\in\Lambda} \left(A(\lambda) + c\lambda^{-\alpha}\right) &\leqslant A(\lambda_i) + c(\lambda_i)^{-\alpha} \leqslant A(\lambda^* + 2\varepsilon) + c(\lambda^*)^{-\alpha} \\ &\leqslant A(2\varepsilon) + \inf_{\lambda\in I} \left(A(\lambda) + c\lambda^{-\alpha}\right) + \delta.\end{aligned}$$

Im Limes für $\delta \to 0$ folgt die Behauptung. «

6-4 Konsistenz und Lernraten für SVMs

6-4.1 **Satz** *Sei (X,d) ein kompakter, metrischer Raum, L eine konvexe, lipschitz-stetige Verlustfunktion mit $|L|_1 \leqslant 1$ und $L(x,y,0) \leqslant 1$ für alle $(x,y) \in X \times Y$. Ferner sei H ein universeller RKHS mit Kern k, $\|k\|_\infty \leqslant 1$ und*

$$\log \mathcal{N}(B_H, \|\cdot\|_\infty, \varepsilon) \leqslant a\varepsilon^{-2p}, \qquad \varepsilon > 0,$$

wobei $a \geqslant 1$ und $p > 0$ Konstanten. Dann ist für alle Folgen (λ_n) in $[0,1]$ mit

$$\lambda_n \to 0 \quad \textit{und} \quad n\lambda_n^{1+p} \to \infty$$

die Lernmethode

$$(X, Y)^n \ni D \mapsto f_{D,\lambda_n}$$

universell konsistent. ⋈

SVMs mit geeigneter Regularisierung sind also universell konsistent.

» L ist konvex, lipschitz-stetig und erfüllt $L(x, y, 0) \leqslant 1$. Somit ist L eine P-integrierbare NVF der Ordnung 1 und es folgt

$$\mathcal{R}^*_{L,P,H} = \mathcal{R}^*_{L,P}.$$

Korollar 6-2.2 sichert weiterhin, dass

$$\mathcal{R}_{L,P}(f_{D,\lambda_n}) - \mathcal{R}^*_{L,P} \leqslant A(\lambda_n) + 2\lambda_n^{-1/2}\left(\frac{a}{n}\right)^{\frac{1}{2+2p}} + 4\lambda_n^{-1/2}\sqrt{\frac{2\tau+2}{n}}$$

mit einer Wahrscheinlichkeit nicht kleiner als $1 - e^{-\tau}$.

Da $A(\lambda_n) \to 0$ und $n\lambda^{1+p} \to \infty$ folgt, dass

$$\lambda_n^{-1/2}\left(\frac{1}{n}\right)^{\frac{1}{2+2p}} = \left(\frac{1}{\lambda_n^{1+p} n}\right)^{\frac{1}{2+2p}} \to 0,$$

sowie

$$4\lambda_n^{-1/2}\sqrt{\frac{2\tau+2}{n}} = 4\sqrt{2\tau+2}\frac{1}{\sqrt{\lambda_n n}} \leqslant 4\sqrt{2\tau+2}\frac{1}{\sqrt{\lambda_n^{1+p} n}} \to 0. \quad «$$

Bsp Betrachte die Verlustfunktion $L = L_{\text{hinge}}$, dann besagt Zhang's Ungleichung 2-2.8

$$\mathcal{R}_{L_{\text{class}},P}(f) - \mathcal{R}^*_{L_{\text{class}},P} \leqslant \mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}.$$

Damit ist die SVM, die L benutzt, unter den Voraussetzungen des Satzes 6-4.1 nicht nur universell konsistent bezüglich L, sondern auch universell klassifikationskonsistent. ⊸

6-4.2 **Satz** *Es gelten die Voraussetzungen von Satz 6-4.1. Zudem seien $c > 0$ und $\beta \in (0,1]$ Konstanten mit*

$$A(\lambda) \leqslant c\lambda^{\beta}.$$

Für $n \geqslant 1$ definiere dann

$$\lambda_n := n^{-\frac{1}{(1+p)(1+2\beta)}}.$$

Dann lernt die Lernmethode

$$(X \times Y)^n \ni D \mapsto f_{D,\lambda_n}$$

in Verteilung mit der Rate $n^{-\frac{1}{(1+p)(1+2\beta)}}$. ⋊

» Mit Korollar 6-2.2 folgt,

$$\lambda \left\| f_{D,\lambda_n} \right\|_H^2 + \mathcal{R}_{L,P}(f_{D,\lambda_n}) - \mathcal{R}_{L,P}^* \leqslant c\lambda_n^{\beta} + 2\lambda_n^{-1/2} \left(\frac{a}{n}\right)^{(2+2p)^{-1}} + 4\lambda_n^{-1/2} \sqrt{\frac{2\tau+2}{n}}$$

mit einer Wahrscheinlichkeit nicht kleiner als $1 - \varepsilon^{-\tau}$. Einsetzen ergibt die Behauptung. «

Bemerkungen. A. Die Definition von λ_n ist asymptotisch (bezgl. n) für die Orakelungleichung aus Korollar 6-2.2 die beste Wahl.

B. Um λ_n so definieren zu können, benötigt man Wissen über β. In der Regel verfügt man darüber jedoch nicht! ⊸

6-4.3 **Definition** *Sei L eine Verlustfunktion, die bei 1 abgeschnitten werden kann, H ein RKHS über X und $\Lambda = (\Lambda_n)$ eine Familie von endlichen Teilmengen von $(0,1]$. Für $n \geqslant 3$ und $D = ((x_1,y_1),\ldots,(x_n,y_n)) \in (X \times Y)^n$ definiere*

$$m := \left\lfloor \frac{n}{2} \right\rfloor,$$
$$D_1 = ((x_1,y_1),\ldots,(x_m,y_m)),$$
$$D_2 = ((x_{m+1},y_{m+1}),\ldots,(x_n,y_n)),$$

und betrachte

$$f_{D_1,\lambda} = \operatorname*{argmin}_{f\in H}\left(\lambda \left\|f\right\|_H^2 + \mathcal{R}_{L,D_1}(f)\right),$$
$$\lambda_{D_2} \in \operatorname*{argmin}_{\lambda\in\Lambda_n} \mathcal{R}_{L,D_2}(\widehat{f}_{D_1,\lambda}).$$

Dann heißt $D \mapsto f_{D,\lambda_{D_2}}$ *TV-SVM (Training/Validation-Support vector machine).* ⋊

6-4.4 **Satz** *Es gelten die Voraussetzungen von Satz 6-4.1 und L sei bei 1 abschneidbar. Ferner sei Λ_n ein ε_n-Netz von $(0,1]$, wobei $\varepsilon_n > 0$. Für $\tau \geqslant 1$ sei ferner*

$$\tau_n = 2\tau + 4\log|\Lambda_n| + 1.$$

Dann gilt mit einer Wahrscheinlichkeit P^n nicht kleiner als $1 - e^{-\tau}$,

$$\mathcal{R}_{L,P}(\widehat{f}_{D_1,\lambda_2}) - \mathcal{R}^*_{L,P} \leqslant \inf_{\lambda\in(0,1]}\left(A(\lambda) + 3\lambda^{-1/2}\left(\frac{a}{n}\right)^{(2+2p)^{-1}} + 12\lambda^{-1/2}\sqrt{\frac{\tau_n}{n}}\right) + A(2\varepsilon_n).$$

Insbesondere ist die TV-SVM universell konsistent, falls $\varepsilon_n \to 0$ und $n^{-1}\log|\Lambda_n| \to 0$. Ist außerdem $A(\lambda) \leqslant c\lambda^\beta$ für $\lambda \geqslant 0$ und geeignetes $c \geqslant 0$ und $\beta \in (0,1]$, so lernt die TV-SVM in Verteilung mit Rate $n^{-\frac{\beta}{(1+p)(1+2\beta)}}$, sofern $\varepsilon \leqslant \min\left\{n^{-1}, |\Lambda_n|\right\}$, verhält sich die Lernrate polynomial in n. ⋊

» Da $m = \lfloor \frac{n}{2} \rfloor$ ist $m > \frac{n}{2}$. Mit Korollar 6-2.2 folgt somit

$$\begin{aligned}\mathcal{R}_{L,P}(f_{D_1,\lambda}) - \mathcal{R}^*_{L,P} &\leqslant A(\lambda) + 2\lambda^{-1/2}\left(\frac{a}{n}\right)^{(2+2p)^{-1}} + 4\lambda^{-1/2}\sqrt{\frac{2\tau+2}{m}} \\ &\leqslant A(\lambda) + 3\lambda^{-1/2}\left(\frac{a}{n}\right)^{(2+2p)^{-1}} + 8\lambda^{-1/2}\sqrt{\frac{\tau+1}{n}}\end{aligned}$$

mit einer Wahrscheinlichkeit nicht kleiner als $1 - e^{-\tau}$. Mit dem union bound folgt, dass diese Abschätzung für alle $\lambda \in \Lambda_n$ simultan mit einer Wahrscheinlichkeit P^n nicht kleiner als $1 - |\Lambda_n|\, e^{-\tau}$ gilt. Ferner gilt

$$L(x,y,\widehat{t}) \leqslant \underbrace{|L|_1}_{\leqslant 1} + \underbrace{L(x,y,0)}_{\leqslant 1} \leqslant 2 =: B.$$

Satz 4-1.2 impliziert mit $n - m \geqslant n/2 - 1 \geqslant n/4$, dass

$$\mathcal{R}_{L,P}(f_{D_1}, \lambda_{D_2}) \leqslant \inf_{\lambda \in \Lambda_n} \mathcal{R}_{L,P}(\widehat{f}_{D,\lambda}) + 4\sqrt{\frac{2\tau + 2\log 2\,|\Lambda_n|}{n}}$$

mit Wahrscheinlichkeit P^{n-m} nicht kleiner als $1 - e^{-\tau}$. Man sieht mit Lemma 6-3.5 leicht ein, dass somit auch mit einer Wahrscheinlichkeit P^n nicht kleiner als $1 - (1 + |\Lambda_n|)e^{-\tau}$ gilt

$$\begin{aligned}
\mathcal{R}_{L,P}(\widehat{f}_{D_1,\lambda_{D_2}}) - \mathcal{R}^*_{L,P} &\leqslant \inf_{\lambda \in \Lambda_n} \left(\mathcal{R}_{L,P}(\widehat{f}_{D,\lambda}) - \mathcal{R}^*_{L,P}\right) + 4\sqrt{\frac{2\tau + 2\log 2\,|\Lambda_n|}{n}} \\
&\leqslant \inf_{\lambda \in \Lambda_n} \left(A(\lambda) + 3\lambda^{-1/2} \left(\frac{a}{n}\right)^{(2+2p)^{-1}} + 8\lambda^{-1/2}\sqrt{\frac{\tau+1}{n}} \right) \\
&\quad + 4\sqrt{\frac{2\tau + 2\log 2\,|\Lambda_n|}{n}} \\
&\leqslant \inf_{\lambda \in (0,1]} \left(A(\lambda) + 3\lambda^{-1/2} \left(\frac{a}{n}\right)^{(2+2p)^{-1}} + 8\lambda^{-1/2}\sqrt{\frac{\tau+1}{n}} \right) \\
&\quad + A(2\varepsilon) + 4\sqrt{\frac{2\tau + 2\log 2\,|\Lambda_n|}{n}}.
\end{aligned}$$

Weiterhin ist

$$2\tau + 2\log 2\,|\Lambda_n| \leqslant 2\tau + 2 + 2\log|\Lambda_n| \leqslant \tau_n$$

sowie $\tau + 1 \leqslant \tau_n$ und $\lambda^{-1/2} \geqslant 1$ für $\lambda \in (0, 1]$. Somit folgt die Behauptung. «

Bsp Wähle Λ_n als $\frac{1}{n}$-Netz mit Kardinalität $\leqslant 2 - n$. Dann ist die Konstruktion von β unabhängig von D und die TV-SVM lernt mit Rate $n^{-\frac{\beta}{(1+p)(1+2\beta)}}$. ■

Bemkerungen. A. Die TV-SVM kann bei geeigneter Modifikation auch weitere Parameter adaptiv bestimmen, wie z.B. den Kernparameter σ des Gaußkerns.

B. In der Praxis werden kleinere Netze Λ_n benutzt, z.B. $10 \leqslant |\Lambda_n| \leqslant 20$ mit geometrischer Verteilung.

c. Die benutzte Technik mit $\|\cdot\|_\infty$-Überdeckungszahlen war Stand der Forschung bis ca. 2002. Dies kann noch deutlich verbessert werden, denn in Satz 6-2.1 haben wir ausgenutzt, dass $f_{D,\lambda} \in \lambda^{-1/2} B_H$. Dies führte zu $b \approx \lambda^{-1/2}$ in Hoeffdings Ungleichung. Man kann dies aus mehreren Gründen noch deutlich verbesern

1.) Sei $f_{D,\lambda} \in \operatorname{argmin}_{f\in H} \left(\lambda \left\|f\right\|_H^2 + \mathcal{R}_{L,P}(f)\right)$, dann gilt

$$\lambda \left\|f_{D,\lambda}\right\|_H^2 \leqslant \lambda \left\|f_{D,\lambda}\right\|_H^2 + \mathcal{R}_{L,P}(f_{D,\lambda}) - \mathcal{R}^*_{L,P} = A(\lambda).$$

Für $A(\lambda) \leqslant c\lambda^\beta$ folgt somit

$$\left\|f_{D,\lambda}\right\|_H \leqslant \sqrt{c}\lambda^{\frac{\beta-1}{2}}, \qquad \lambda \geqslant 0.$$

Wir sind bisher also stets von $\beta = 0$ und $c = 1$ ausgegangen...

2.) Sei λ_n eine Nullfolge mit $\lambda_n^{-1/2} n^{-(2+2p)^{-1}} \to 0$ polynomial. Dann besagt Korollar 6-2.2, dass mit hoher Wahrscheinlichkeit gilt

$$\begin{aligned}\lambda_n \left\|f_{D,\lambda_n}\right\|^2 &\leqslant \lambda_n \left\|f_{D,\lambda_n}\right\|_H^2 + \mathcal{R}_{L,P}(f_{D,\lambda_n}) - \mathcal{R}^*_{L,P} \\ &\leqslant A(\lambda_n) + \lambda_n^{-1/2} n^{-(2+2p)},\end{aligned}$$

was polynomial gegen Null konvergiert, falls $A(\lambda) \leqslant c\lambda^\beta$. Sei α der entsprechende Exponent, dann folgt

$$\left\|f_{D,\lambda_n}\right\|_H^2 \leqslant \kappa \lambda^{-1/2} n^{-\alpha}, \tag{*}$$

mit hoher Wahrscheinlichkeit. Betrachte nun im Beweis von Satz 6-2.1 nur noch die Datensätze für die (*) gilt. Dies ergibt eine bessere Orakelungleichung, welche wiederum auf eine Verbesserung von (*) führt. Iteration dieses Arguments liefert eine deutlich verbesserte Orakelungleichung.

3.) Kann L abgeschnitten werden, z.B. bei $M = 1$, so folgt

$$\left\|\widehat{f}_{D,\lambda}\right\|_\infty \leqslant 1,$$

wobei wir bisher verwendet haben, dass

$$\left\|f_{D,\lambda}\right\|_\infty \leqslant \lambda^{-1/2}.$$

Man sollte daher eigentlich

$$\mathcal{R}_{L,P}(\widehat{f}_{D,\lambda}) - \mathcal{R}^*_{L,P}$$

abschätzen.

4.) In vielen Fällen liegt ein sogenanntes variance bound vor,

$$\mathbf{E}(L \circ f - L \circ f^*_{L,P})^2 \leqslant V(\mathbf{E}(L \circ f - L \circ f^*_{L,P}))^{\vartheta},$$

wobei $L \circ f(x,y) = L(x,y,f(x))$ und $V > 0$, $\vartheta \in [0,1]$ Konstanten.

Vergleichen wir nun die Fehlerterme von Bernsteins-

$$\sqrt{\frac{\sigma^2\tau}{n}} + \frac{B}{n}$$

und Hoeffdings-Ungleichung

$$B\sqrt{\frac{\tau}{n}},$$

so können wir σ^2 durch $\mathbf{E}(L \circ f - L \circ f^*_{L,P})^2$ „ersetzen“. Dies führt zu einer neuen Orakelungleichung, die zeigt, dass

$$\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P} \qquad (**)$$

mit hoher Wahrscheinlichkeit klein ist (z.B. $O(n^{-\alpha\vartheta})$). Das variance bound zeigt dann, dass $\sigma^2 \hat{=} \mathbf{E}(L \circ f - L \circ f^*_{L,P})^2$ ebenfalls mit hoher Wahrscheinlichkeit klein ist und folglich ist (**) noch kleiner (z.B. $O(n^{-\alpha\vartheta-1/2})$). Iteration dieses Arguments ergibt eine deutliche Verbesserung.

5.) Dies zusammengenommen kann Lernraten bis zu $O(n^{-1})$ ergeben. Vergleiche dies mit unserem Ergebnis $O(n^{-\frac{\beta}{(1+p)(1+2\beta)}})$, welches nie kleiner ist als $O(n^{-1/3})$.

6.) Bisher haben wir nur Überdeckungszahlen bezüglich der Supremumsnorm verwendet. Dies erfordert, dass der Eingaberaum kompakt ist. Oftmals sind die Daten jedoch auf ganz $\mathbb{R}^n$ verteilt (z.B. Gauß-verteilt). Man kann anstatt $\mathcal{N}(B_H, \|\cdot\|_\infty, \varepsilon)$ bessere Überdeckungszahlen wie $\mathcal{N}(B_H, \|\cdot\|_{L^2(P_X)}, \varepsilon)$ betrachten. Dies geht jedoch nicht mehr mit elementaren Methoden. Dafür erhält man z.B. für die Verlustfunktion der kleinsten Quadrate optimale Lernraten. Resultate in dieser Richtung sind relativ neu (2007-2009). ⊸

7 Verlustfunktionen II

Bisher haben wir die Verlustfunktionen studiert, die das Lernziel gut beschrieben (L_{class}, L_{hinge}, L_{LS}, …). Wir haben aber auch schon gesehen, dass L_{hinge} in einem günstigen Verhältnis zu L_{class} steht (Zhang's Ungleichung)

$$\mathcal{R}_{L_{\text{class}},P}(f) - \mathcal{R}^*_{\text{class},P} \leqslant \mathcal{R}_{L_{\text{hinge}},P}(f) - \mathcal{R}^*_{L_{\text{hinge}},P}.$$

Während die Minimierung des Überschussrisikos von L_{class} ein kombinatorisches Optimierungsproblem darstellt und somit äußerst diffiziel werden kann, beschreibt die Minimierung des Überschussrisikos von L_{hinge} ein konvexes Optimierungsproblem, zu dessen Lösung wir bereits Methoden kennengelernt haben (siehe Kapitel 6-1). Dies motiviert den Einsatz von L_{hinge} z.B. in SMVs. Weiterhin haben wir in den Übungen gesehen, dass

$$\mathcal{R}_{L_{\text{class}},P}(f) - \mathcal{R}^*_{L_{\text{class}},P} \leqslant \sqrt{\mathcal{R}_{L_{\text{LS}},P}(f) - \mathcal{R}^*_{L_{\text{LS}},P}}.$$

Ziel dieses Abschnitts ist es nun zu untersuchen, wie man eine Zielverlustfunktion L, welche das Lernziel beschreibt, durch eine Ersatzverlustfunktion ersetzen kann, die einer „solchen" Ungleichung genügt.

7-1 Marginbasierte Verlustfunktionen

7-1.1 **Satz von Bartlett, Jordan, Mc Auliffe (2007)** *Sei L eine konvexe, marginbasierte Verlustfunktion, die durch*

$$\phi : \mathbb{R} \to [0, \infty)$$

dargestellt wird ($L(y,t) = \phi(y-t)$). Dann sind äquivalent

(i) ϕ ist in 0 differenzierbar mit $\phi'\big|_0 = 0$.

(ii) Es existiert eine Funktion $Y : [0,1] \to [0,\infty)$ streng monoton und

$$Y(\mathcal{R}_{L_{\text{class}},P}(f) - \mathcal{R}^*_{L_{\text{class}},P}) \leqslant \mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P} \qquad (*)$$

für alle W-Maße P auf $X \times \{-1,1\}$ und $f : X \to \mathbb{R}$.

In diesem Fall gilt

$$Y(t) = \phi(0) - \inf_{s\in\mathbb{R}} \left(\frac{t+1}{2}\phi(s) + \left(1 - \frac{t+1}{2}\right)\phi(-s) \right)$$
$$\overset{\eta=\frac{t+1}{2}}{=} \phi(0) - \inf_{s\in\mathbb{R}} \left(\eta\phi(s) + (1-\eta)\phi(s)\right). \quad \bowtie$$

Bsp Die Äquivalenz gilt für L_{hinge} und L_{LS} mit

L_{hinge},	$Y(t) = t$,	für alle $t \in [0,1]$,
L_{LS},	$Y(t) = t^2$,	für alle $t \in [0,1]$. ■

Satz *Existieren $q \in [0,\infty]$, $c_p > 0$ mit*

$$P_X(\{x \in X : |2\eta(x) - 1| < t\}) \leqslant (c_p t)^q, \qquad t \in [0,\infty), \qquad (**)$$

und gilt für das in (*) definiert Y, $Y(t) \geqslant ct^p$ für ein $p > 1$. Dann gilt

$$\mathcal{R}_{L_{\text{class}},P}(f) - \mathcal{R}^*_{L_{\text{class}},P} \leqslant 2c^{-\frac{q+1}{q+p}} c_p^{\frac{q(p-1)}{q+p}} \left(\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}\right)^{\frac{q+1}{q+p}}. \quad \bowtie$$

(**) heißt Tsybakov's Noise Assumption (2004). Es ist

$$\eta = \frac{1}{2} \Leftrightarrow |2\eta - 1| = 0, \qquad \eta \approx \frac{1}{2} \Leftrightarrow |2\eta - 1| \approx 0.$$

Die linke Seite in (**) misst die Masse der x, bei denen $\eta(x)$ nahe bei $\frac{1}{2}$ ist. Die rechte Seite beschränkt diese Masse. Für $t \to 0$ sagt die Noise Assumption

$$P_X(\{x \in X : |2\eta(x) - 1| < t\}) = O(t^q).$$

Wir definieren

$$L_{\alpha-\text{class}}(y,t) := \begin{cases} 1-\alpha, & y = 1, \quad t < 0, \\ \alpha, & y = -1, \quad t \geqslant 0, \\ 0, & \text{sonst.} \end{cases}$$

7-1.2 **Satz** *Sei L eine konvexe Verlustfunktion, die durch ϕ dargestellt werden kann, mit*

$$L_{\alpha\text{-class}}(y,t) = \begin{cases} 1-\alpha\phi(t), & y = 1, \\ \alpha\phi(t), & y = -1. \end{cases}$$

Dann sind äquivalent

(i) *ϕ ist in 0 differenzierbar mit $\phi'(0) = 0$.*

(ii) *Es gibt eine streng monoton steigende Funktion $Y : [0,1] \to [0,\infty)$ mit*

$$Y\left(\mathcal{R}_{\alpha\text{-class},P}(f) - \mathcal{R}^*_{\alpha\text{-class},P}\right) \leqslant \mathcal{R}_{L_\alpha,P}(f) - \mathcal{R}^*_{L_\alpha,P}. \quad \bowtie$$

Bsp Für $0 < \alpha \leqslant \frac{1}{2}$ sind

$$L_{\text{hinge}}, \qquad Y(t) = t,$$

$$L_{\text{LS}}, \qquad Y(t) = \frac{t^2}{2\alpha(1-\alpha) + (1-2\alpha)}. \quad \blacksquare$$

7-2 Distanzbasierte Verlustfunktionen

Für die Verlustfunktion der kleinsten Quadrate L_{LS} ist

$$\left\|f - f^*_{L,P}\right\|^2_{L^2(P_X)} = \mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}, \qquad f^*_{L,P} = \mathbf{E}(Y \mid x).$$

Wie wollen nun den bedingten Median schätzen. Dazu nehmen wir an, dass

$$\text{Median}(Y \mid x) = \left\{f^*(x)\right\}, \qquad P_X - \text{ f.s.}$$

7-2.1 **Satz** *Sei $L(y,t) = \left|y - t\right|$ für $y \in \mathbb{R}$ und $t \in \mathbb{R}$. Dann sind für $f : X \to \mathbb{R}$ äquivalent*

(i) *$\mathcal{R}_{L,P}(f) = \mathcal{R}^*_{L,P}$,*

(ii) *$f = f^*$ P_X-f.s.*

D.h. der Minimierer des Überschussrisikos entspricht dem Median. $\bowtie$

Wenn wir bereits wissen, das $\mathcal{R}_{L,P}(f) \approx \mathcal{R}_{L,P}(f^*)$, wie nahe ist dann f an f^*?

Definition *Eine Verteilung Q auf $[-1,1]$ hat den Mediantyp $q \in (1,\infty)$, wenn $\alpha_Q \in [0,2]$ und ein $b_Q > 0$ existieren, so dass für $t^* = \operatorname{Median}(Q)$ und jedes $s \in [0,\alpha_Q]$ gelten*

$$Q((t^*-s,t^*)) \geqslant b_Q s^{q-1}, \qquad Q((t^*,t^*+s)) \geqslant b_Q s^{q-1}. \quad \bowtie$$

Definition *Sei $p \in (0,\infty]$, $q \in (1,\infty)$ und P ein W-Maß auf $X \times [-1,1]$. Dann hat P einen p-mittelbaren Mediantyp q, wenn folgende Eigenschaften erfüllt sind*

(a) *$P(Y \mid x)$ hat Mediantyp q P_X- f.s..*

(b) *Die Abbildung $\gamma : X \to \mathbb{R}$, $x \mapsto b_{P(Y|x)} \alpha_{P(Y|x)}^{q-1}$ erfüllt $\gamma^{-1} \in L_p(P_X)$.* $\bowtie$

Bsp Alle Dichten von $P(Y \mid x)$ erfüllen für $p = \infty$ und $q = 2$,

$$\min\left\{1 + f^*(x), 1 - f^*(x)\right\} \geqslant \alpha^*.$$

Unter diesen Voraussetzungen gilt

$$r := \frac{p_q}{p+1},$$
$$\left\|f - f^*\right\|_{L_r(P_X)} \leqslant c\left(\mathcal{R}_{L,P}(f) - \mathcal{R}_{L,P}^*\right)^{1/q}$$

für alle $f : X \to [-1,1]$ und $r = q = 2$. ■

Index

Literaturverzeichnis

[1] F. Cucker and D.X. Zhou, Learning Theory: An Approximation Theory Viewpoint, Cambridge University Press (2007)

[2] L. Devroye, L. Györfi und G. Lugosi: A Probabilistic Theory of Pattern Recognition, Springer (1996)

[3] L. Györfi, M. Kohler, A. Krzyzak und Harro Walk, A Distribution-Free Theory of Nonparametric Regression, Springer (2002)

[4] T. Hastie, R. Tibshirani und J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, Springer (2009)

[5] I. Steinwart und A. Christmann, Support Vector Machines, Springer (2008)

[6] Bauer, H. Wahrscheinlichkeitstheorie; de Gruyter, Berlin (1990), 4. Aufl.

[7] Bauer, H. Maß- und Integrationstheorie; de Gruyter, Berlin (1998), 2. Aufl.

[8] Werner, Dirk: Funktionalanalysis. 6. Auflage, Springer 2007