UNIVERSITÄT
BAYREUTH

Herleitung oberer Schranken für die Approximation mit radialen Basisfunktionen unter Verwendung von Beweistechniken der statistischen Lerntheorie

Masterarbeit im Fachbereich Mathematik
von Dominik Köhler

FAKULTÄT FÜR MATHEMATIK, PHYSIK UND INFORMATIK
Lehrstuhl für Angewandte und Numerische Analysis

Datum: 12. Dezember 2022
Betreuung: Prof. Dr. H. Wendland

Inhaltsverzeichnis

Einleitung

Das Ziel dieser Arbeit ist es, gleichmäßige obere Schranken für die Approximation mit radialen Basisfunktionen zu finden, aufbauend auf den Ideen in [Gir95] und [GS08]. Im Speziellen betrachten wir eine Schranke der folgenden Form:

$$\sup_{f\in\mathcal{F}}\left|\int_{\mathbb{R}^k} f(z)\lambda(z)dz - \|\lambda\|_1\frac{1}{n}\sum_{i=1}^{n} f(z_i)\right| \leq \alpha, \quad \alpha \overset{n\to\infty}{\longrightarrow} 0.$$

Hierbei ist $\mathcal{F}$ eine Menge an Funktionen aus $\mathbb{R}^k \to \mathbb{R}$. Das Besondere an dieser Grenze ist, dass diese gleichmäßig für alle Funktionen aus $\mathcal{F}$ auf den festen Punkten $z_1, \ldots, z_n$ gilt. Dies lässt sich nur erreichen, wenn wir in der Menge $\mathcal{F}$ nicht alle messbaren Funktionen aus $\mathbb{R}^k \to \mathbb{R}$ betrachten, sondern hierfür eine geeignete Methode finden, diese einzuschränken. Analog zu Beweistechniken in der statistischen Lerntheorie betrachten wir hierzu sogenannte Komplexitätsmaße für Mengen an Funktionen.

In der statistischen Lerntheorie beschäftigt man sich mit dem Problem, wie weit der empirische Erwartungswert und der Erwartungswert voneinander entfernt sind. Man sucht also möglichst kleine obere Schranken für die Abschätzung

$$P\left(z_1, \ldots, z_n \in \mathbb{R}^k : \sup_{f\in\mathcal{F}}\left|E\left[f(z)\right] - \frac{1}{n}\sum_{i=1}^{n} f(z_i)\right| \leq \alpha'(\mathcal{F}, n, \delta)\right) \geq 1-\delta, \tag{0.1}$$

ebenfalls gleichmäßig für die Menge $\mathcal{F}$. Hiermit will man angeben, ob man für eine Funktion f nur durch die Funktionsauswertungen in den Punkten $z_1, \ldots, z_n$ etwas über den Erwartungswert von f aussagen kann, also ob das Wissen, welches aus den Punkten $z_1, \ldots, z_n$ gewonnen wird, verallgemeinerbar ist. Im Kontext der statistischen Lernentheorie heißen solche Schranken deswegen auch *generalization bounds*.

Könnte man für jedes Modell eigene n Punkte wählen, so könnte man schärfere Schranken herleiten, zum Beispiel mit der *Hoeffding*-Ungleichung, welche wir später in Satz 2.1 betrachten werden. Die Besonderheit bei unserer Vorgehensweise ist, dass wir alle Modelle nur auf den fest gewählten Punkten $z_1, \ldots, z_n$ beurteilen.

Vor allem durch die Arbeit von V. Vapnik und A. Chervonenkis in [VC71] wurde es möglich, die Hoeffding-Ungleichung für abzählbar viele Funktionen

gleichzeitig zu verwenden, um eine Schranke in der Form von (0.1) zu erhalten. Hierzu beschränken wir die Menge der betrachteten Funktionen $\mathcal{F}$ durch ein Komplexitätsmaß, die Vapnik-Chervonenkis-Dimension, kurz VC-Dimension oder VCD. Ursprünglich wurde dies für $\{0,1\}$-wertige Funktionen bewiesen, später allerdings auf Funktionen verallgemeinert, welche auf ein Intervall $[A,B] \subset \mathbb{R}$ abbilden. Den Beweis hierzu wollen wir hier vollständig wiedergeben.

Dabei beziehen wir uns allerdings wieder auf $\{0,1\}$-wertige Funktionen. Dieser Schritt lässt sich umgehen, wenn man statt der VC-Dimension das Komplexitätsmaß *Rademacher-Komplexität* verwendet. Hiermit können wir wieder eine Schranke in der Form von (0.1) herleiten. Weiterhin lässt sich die Rademacher-Komplexität von unten gegen die VC-Dimension abschätzen. Damit erhalten wir einen direkten Vergleich zwischen den beiden Abschätzungen und sehen, dass die Abschätzungen mit der Rademacher-Komplexität schärfer sind.

Um die Resultate auf eine spezielle Menge $\mathcal{F}$ an Funktionen anzuwenden, müssen wir also die VC-Dimension von $\mathcal{F}$ berechnen. In dieser Arbeit wollen wir einige Grundtechniken hierzu angeben, aber vor allem die VC-Dimension von radialen Funktionen

$$\phi(\|\cdot - t\|_2) : \mathbb{R}^k \to \mathbb{R}, \qquad t \in \mathbb{R}^k \text{ fest}, \tag{0.2}$$

bestimmen. Ist die Funktion $\phi : \mathbb{R}_0^+ \to \mathbb{R}$ dabei die Identität, so wurde für die VC-Dimension der resultierende Menge

$$\mathcal{F} = \{\|\cdot - t\|_2 : \mathbb{R}^k \to \mathbb{R}_0^+, t \in \mathbb{R}^k\}$$

bereits in [Dud79] hergeleitet. In dieser Arbeit wollen wir die Berechnung auf allgemeine radiale Funktionen erweitern. Als technischen Trick verwenden wir, dass das Anwenden von monotonen Funktionen auf eine Menge $\mathcal{F}$ die VC-Dimension nicht vergrößert, also für jede monotone Funktion $f_m : \mathbb{R} \to \mathbb{R}$ gilt:

$$\mathrm{VCD}(f_m \circ \mathcal{F}) = \mathrm{VCD}\left(\{f_m \circ f \mid f \in \mathcal{F}\}\right) = \mathrm{VCD}(\mathcal{F}).$$

Resultierend aus dieser Idee zerlegen wir die Funktion $\phi : \mathbb{R}_0^+ \to \mathbb{R}$ aus (0.2) in Intervalle, auf welchen ϕ jeweils monoton ist:

$$\phi(z) = \sum_{i=1}^{p} \phi|_{\mathcal{I}_i}(z), \qquad \mathcal{I}_1 \cup \ldots \cup \mathcal{I}_p = \mathbb{R}_0^+, \quad \mathcal{I}_i \cap \mathcal{I}_j = \emptyset,$$
$$\phi \text{ auf } \mathcal{I}_i \text{ monoton}, \quad 1 \le i, j \le p. \tag{0.3}$$

In diesem Fall sagen wir, dass ϕ die *Monotoniebedingung für p Intervalle* erfüllt und bezeichnen dies mit

$$\phi \in \mathcal{M}_p.$$

Als Ergebnis erhalten wir für die Menge

$$\mathcal{F} = \left\{\phi(\|\cdot - t\|_2), \qquad \phi \in \mathcal{M}_p, t \in \mathbb{R}^k\right\}$$

eine obere Schranke für die VC-Dimension mit einer Konstanten c_p:

$$\mathrm{VCD}(\mathcal{F}) \leq c_p(2k+3).$$

Diese Methode lässt sich erweitern, sodass wir insgesamt die VC-Dimension für Mengen an verketteten reellwertige Funktionen berechnen können. Dabei ist k die Raumdimension und c_p eine Konstante abhängig von p. Insgesamt erhalten wir mit einer Wahrscheinlichkeit von mindestens $1-\delta$ unabhängig und identisch verteilte Punkte $z_1, \ldots, z_n$, sodass gilt:

$$\begin{aligned}&\sup_{\Phi(\cdot,t)\in\mathcal{F}} \left| \int_Z \Phi(z,t)\lambda(z)dz - \frac{1}{n}\sum_{i=1}^{n} \Phi(z_i,t)\|\lambda\|_1 \right| \\ \leq &(B-A)Cc_p\|\lambda\|_1 \frac{1}{\sqrt{n}} \max\left\{\sqrt{2k+3}, -\frac{\ln(\delta)}{\sqrt{n}}, \sqrt{-\delta}\right\}. \qquad (0.4)\end{aligned}$$

Die Konstante C ist dabei unabhängig von λ und $\mathcal{F}$ und die Konstante c_p ist nur abhängig von p.

In Kapitel 1 geben wir eine kurze Einführung in die statistische Lerntheorie. In den Kapiteln 2 bis 4 leiten wir die VC-Schranke her, eine Abschätzung in der Form von (0.1). Dabei gehen wir in Kapitel 3 näher auf die VC-Dimension ein. In Kapitel 5 übertragen wir die gewonnene Abschätzung auf die Approximationstheorie und erhalten eine Schranke in der Form von (0.4). In Kapitel 6 betrachten wir dann die *Rademacher-Komplexität*. Hierfür leiten wir Abschätzungen analog zur VC-Schranke her und erhalten wieder Abschätzungen der Form (0.1) und (0.4). Die Rademacher-Komplexität schätzen wir in Kapitel 6.5 von unten gegen die VC-Dimension ab. In Kapitel 7 verwenden wir die Ergebnisse für radiale Funktionen und betrachten einige radiale Kerne genauer.

Wir wollen noch kurz auf die verwendete Notation in dieser Arbeit eingehen. Im Folgenden werden wir immer als Grundraum $\mathbb{R}^k$ verwenden. Dabei

bezeichnen wir Teilmengen aus $\mathbb{R}^k$ meist mit Z und betrachten dann Punkte $z, z_1, \ldots, z_n \in Z \subset \mathbb{R}^k$. Weiterhin bezeichnen wir mit $\mathcal{F} \subset \mathbb{R}^{\mathbb{R}^k}$ eine Menge an messbaren Funktionen von $\mathbb{R}^k \to \mathbb{R}$ und Variablen aus $\mathbb{R}$ mit $x, x_1, \ldots, x_n \in \mathbb{R}$. Die Variable $d \in \mathbb{R}$ verwenden wir für die VC-Dimension oder Rademacher-Komplexität einer Menge $\mathcal{F}$.

Ebenfalls wird es nötig sein, nur die Vorzeichen der Funktionen zu betrachten. Dafür verwenden wir die Heaviside-Funktion:

$$\mathcal{H}(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0. \end{cases} \tag{0.5}$$

Wenden wir die Heaviside-Funktion auf jedes Element einer Menge $\mathcal{F}$ an und verschieben dies um einen beliebigen Wert $\beta \in \mathbb{R}$, so bezeichnen wir die resultierende Menge mit

$$\mathcal{F}_{\mathcal{H}} := \{\mathcal{H}(f(\cdot) - \beta) : Z \to \{0, 1\} : f \in \mathbb{R}^k, \beta \in \mathbb{R}\}.$$

1 Kurzübersicht statistische Lerntheorie

Die Einführung folgt [Vap98], Kapitel 1. In der statistischen Lerntheorie betrachten wir Daten als endliche Teilmenge S_X einer größeren Menge $X \subset \mathbb{R}^k$ mit zugehörigen Werten aus einer Menge $Y \subset \mathbb{R}$. Insgesamt beobachten wir also eine Menge an Paaren $(x, y) \in S_X \times Y$. Dabei bezeichnen wir mit $S \subset X \times Y$ die (endliche) Menge aller beobachteten Paare. Dabei nehmen wir an, dass die beobachteten Werte aus Y von den Werten aus X abhängen, und zwar, dass die Werte in Y durch die Verteilungsfunktion $F(y|x)$ beschrieben werden können. Zusätzlich nehmen wir an, dass die beobachtete Menge S_X zufällig aus den Werten aus X gewählt ist, also einer Verteilung $F(x)$ folgt. Insgesamt beobachten wir also die gemeinsame Wahrscheinlichkeitsdichte auf $X \times Y$, gegeben durch $F(x, y) = F(x)F(y|x)$. Diesen Zusammenhang wollen wir durch die beobachteten Tupel nun möglichst gut bestimmen.

Allgemein betrachten wir eine Menge an möglichen Funktionen oder Hypothesen, welche wir als $\mathcal{F}$ bezeichnen, und suchen diejenige Hypothese aus dieser Menge, welche unser Problem am besten beschreibt. In der Realität sind die Daten auf $X \times Y$ durch ein Wahrscheinlichkeitsmaß P gegeben, welches allerdings für uns unbekannt ist. Betrachten wir nun eine Hypothese $f \in \mathcal{F}$ und werten diese auf der Menge S_X aus, dann könnte es $(x, y) \in S$ geben, sodass $y \neq f(x)$ gilt. Wir wollen im Folgenden aus der Menge $\mathcal{F}$ diejenige Funktion $f \in \mathcal{F}$ finden, von welcher wir erwarten können, dass diese Werte aus Y wählt, welche die eigentlichen Werten aus $X \times Y$ möglichst gut approximieren, das heißt $(x, f(x))$ soll nah an (x, y) sein. Um zu beurteilen, was es heißt, dass $(x, f(x))$ nah an (x, y) ist, benötigen wir eine Funktion, die den Abstand zwischen diesen Punkten misst. Den Unterschied nennen wir dann auch Fehler und diesen wollen wir durch eine *Verlustfunktion* messen, wie in [SC08], Kapitel 2.1:

Definition 1.1 (Verlustfunktion)**.** *Sei $(X, \mathcal{A}_X)$ messbar und $Y \subset \mathbb{R}$ abgeschlossen. Dann heißt eine Funktion*

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

Verlustfunktion, falls diese messbar ist.

Dabei bezeichnen wir $L(x, y, f(x))$ als den Verlust, welcher gemessen durch L entsteht, wenn man nicht y, sondern $f(x)$ vorhersagt. Oft nimmt man für $y = f(x)$ an, dass $L(x, y, f(x)) = 0$ gilt. Damit können wir den gemittelten Fehler gegeben L auf S berechnen, was wir als *empirisches L-Risiko*

bezeichnen:

$$\widehat{R}_L(f) = \frac{1}{\#S} \sum_{(x,y)\in S} L(x, y, f(x)).$$

Das Ziel ist es allerdings zu betrachten, wie gut die gewählte Funktion auf allen Daten aus X die zugehörigen für uns unbekannten Werte aus Y annähert. Den erwarteten Fehler auf ganz $X \times Y$, gegeben einer Verteilung der Werte $(x, y) \in X \times Y$,

$$R_L(f) := \int_{X\times Y} L(x, y, f(x))dP(x, y)$$

bezeichnen wir als *L-Risiko* oder kurz *Risiko*.

Bemerkung 1.1. *Oft schreiben wir auch R statt R_L, genauso $\widehat{R}$ statt $\widehat{R}_L$. Um hervorzuheben, dass das empirische Risiko von n Stützpunkten abhängt, schreiben wir auch $\widehat{R}_n$.*

Für eine gegebene Verlustfunktion L und einem unbekannten Wahrscheinlichkeitsmaß P auf $X \times Y$ heißt die messbare Funktion mit geringstem L-Risiko *Bayes-Entscheidungsfunktion* und wird mit f^* bezeichnet. Das zugehörige Risiko heißt *Bayes-Risiko* und wird mit R_L^* bezeichnet, also:

$$R_L^* := \inf\{R_L(f)|\ f : X \to \mathbb{R} \text{ messbar}\}.$$

Für die Einführung hier folgen wir [Vap98], Kapitel 1.13.

1.1 Zerlegung des Risikos

In diesem Kapitel folgen wir [MRT12], Kapitel 2.4, sowie [BB08]. Ziel ist es, eine Funktion $f \in \mathcal{F} \subset Y^X$ zu finden, für die die Paare $(x, f(x))$ möglichst nah, im Sinne der gewählten Verlustfunktion, an den Paaren (x, y) sind, für alle $x \in X$. Dazu betrachten wir eine beliebige Funktion $f \in \mathcal{F}$ und berechnen für eine Verlustfunktion $L(x, y, f(x))$ das L-Risiko dieser Funktion. Betrachten wir dies für alle Funktionen der Menge $\mathcal{F}$ auf diese Art und Weise, können wir diese mithilfe des L-Risikos vergleichen. Das wollen wir im Folgenden genauer betrachten. Wir bezeichnen hierzu mit

- $\hat{f}$ die Funktion aus $\mathcal{F}$, welche das empirische Risiko auf S minimiert. Wir nehmen hier an, dass diese existiert, also dass ein $\hat{f} \in \mathcal{F}$ existiert, mit $\widehat{R}(\hat{f}) \leq \widehat{R}(f)\, \forall f \in \mathcal{F}$.

- $R_{\mathcal{F}}$ das Infimum der Menge $\{R(f)|f \in \mathcal{F}\}$,
- $f^* := \arg\inf\{R(f)|f \in \mathcal{F}\}$,
- R^* das Infimum der Menge $\{R(f)|f : X \to Y \text{ messbar}\}$, auch Bayes-Risiko genannt.

Wir sind daran interessiert, wie nah wir mit einer Funktion f aus der Menge an Funktionen $\mathcal{F}$ an das Bayes-Risiko kommen, gegeben der gewählten Verlustfunktion. Dazu zerlegen wir dieses Problem:

$$R(f) - R^* = (R(f) - R(\hat{f})) + (R(\hat{f}) - R_{\mathcal{F}}) + (R_{\mathcal{F}} - R^*). \tag{1.1}$$

Der erste Term gibt für eine Funktion f an, wie viel besser oder schlechter diese auf allen Daten ist, gemessen der Verlustfunktion L, im Vergleich zu der Funktion mit dem geringsten empirischen Risiko auf S. Man nennt diesen Fehler auch Optimierungsfehler, engl. *optimization error*.

Der zweite Term gibt den Fehler an, welcher entsteht, falls man das empirische Risiko statt dem Risiko über $\mathcal{F}$ minimiert. Dabei kann es passieren, dass $R(\hat{f}) > \widehat{R}(\hat{f})$ gilt. In diesem Fall sagt man auch, dass $\hat{f}$ *overfittet*. Diesen Fehler bezeichnen wir als Bewertungsfehler, engl. *estimation error*.

Der dritte Term gibt an, wie nah das geringst mögliche Risiko für Funktionen aus $\mathcal{F}$, $R_{\mathcal{F}}$, an dem Bayes-Risiko R^* ist. Dies wird auch Approximationsfehler genannt, engl. *approximation error*.

Hilfreich für das Abschätzen des Bewertungsfehlers ist die folgende Abschätzung, vgl. Formelzeile (2.26) in [MRT12]:

$$\begin{aligned} R(\hat{f}) - R_{\mathcal{F}} &= R(\hat{f}) - \widehat{R}(\hat{f}) + \widehat{R}(\hat{f}) - \hat{R}(f^*) + \widehat{R}(f^*) - R(f^*) \\ &\leq 2 \sup_{f \in \mathcal{F}} |\widehat{R}(f) - R(f)|. \end{aligned} \tag{1.2}$$

In der ersten Zeile ist der Ausdruck $\widehat{R}(\hat{f}) - \widehat{R}(f^*)$ nach Definition von $\hat{f}$ höchstens 0.

Insgesamt werden wir im Folgenden den Fehler in der obigen Formelzeile (1.2) abschätzen. Wir sind also interessiert, wie viele Beobachtungen n man benötigt, damit der Term $|\widehat{R}(f) - R(f)|$ kleiner ist als ein frei wählbares $\varepsilon > 0$, falls es überhaupt möglich ist. Da allerdings der Berechnung des Risikos, sowie die Verteilung der beobachteten Daten aus S dem unbekannten

Wahrscheinlichkeitsmaß P auf $X \times Y$ zu Grunde liegt, können wir die Abschätzung nur mit einer gewissen Wahrscheinlichkeit $\delta \in (0, 1]$ bestimmen. Somit suchen wir Grenzen der Form:

$$P(|R(f) - \widehat{R}(f)| \geq \varepsilon) \leq \delta. \tag{1.3}$$

1.2 Verallgemeinerung der Verlustfunktion

In der statistischen Lerntheorie nutzt man die Verlustfunktion, um den Abstand der Vorhersage der Funktion $f(x)$ zum eigentlichen (wahren) Wert y zu messen. Dies nutzen wir allerdings nicht weiter aus bei den Beweisen aus dieser Arbeit. Statt einer Verlustfunktion $L(x, y, f(x))$ betrachten wir im Folgenden deswegen allgemein eine messbare Funktion $f(z) : \mathbb{R}^k \to \mathbb{R}$. Hierfür definieren wir für das Risiko und das empirische Risiko für ein gegebenes Wahrscheinlichkeitsmaß P auf $Z \subset \mathbb{R}^k$:

$$R(f) = \int_Z f(z) dP(z),$$
$$\widehat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} f(z_i).$$

In dieser Arbeit nutzen wir immer diese Darstellung. Es lässt sich allerdings für $Z = X \times Y$ und $z = (x, y)$ die Darstellung $f(z) = L(x, y, f(x))$ nutzen. Damit sind die folgenden Beweise auch in diesem Kontext gültig.

2 Herleitung einer Schranke für gleichmäßige Konvergenz

In diesem Kapitel folgen wir [vLS11] und [BBL03], sowie [Vap98] für das Unterkapitel 2.2.

In diesem Kapitel wollen wir Bedingungen für eine Ungleichung der folgenden Form herleiten, für $\delta \in (0, 1]$:

$$P\left(\forall f \in \mathcal{F} : |R(f) - \widehat{R}(f)| \leq \alpha_{\mathcal{F},n,\delta}\right) \geq 1 - \delta. \tag{2.1}$$

Das empirische Risiko wird dabei auf n Punkten berechnet, welche zufällig und unabhängig gemäß P verteilt sind. Somit meinen wir mit der Zeile (2.1):

$$\begin{aligned} &P\left(\forall f \in \mathcal{F} : |R(f) - \widehat{R}(f)| \leq \alpha_{\mathcal{F},n,\delta}\right) \\ =&P\left(z_1, \ldots, z_n \in Z : \forall f \in \mathcal{F} : \left|\int_Z z dP(z) - \frac{1}{n}\sum_{i=1}^{n} f(z_i)\right| \leq \alpha_{\mathcal{F},n,\delta}\right) \end{aligned} \tag{2.2}$$

Die Werte $z_1, \ldots z_n$ sind formal die Auswertungen von Zufallsvariablen $Z_1, \ldots, Z_n : Z \to \mathbb{R}$, also:

$$z_1 = Z_1(z), \ldots, z_n = Z_n(z).$$

Dabei ist α nur abhängig von der gewählten Menge an Funktionen $\mathcal{F} \subset \mathbb{R}^Z$, der Anzahl der beobachteten Punkten n und δ, nicht von dem (unbekannten) Wahrscheinlichkeitsmaß P. Das Ziel hierbei ist es, die *Hoeffding*-Ungleichung und die *Boolesche* Ungleichung zu kombinieren, wie in den ersten beiden Schritten erklärt. Allerdings ist dies nur möglich, falls wir nur endlich viele Funktionen betrachten. In den Punkten 3 und 4 wird erklärt, wieso wir statt unendlich vielen Funktionen auch nur endlich viele betrachten können. Damit ist es uns möglich, eine Schranke in Form von (2.1) für unendlich viele Funktionen in $\mathcal{F}$ zu betrachten.

1. Für eine feste Funktion gilt eine Ungleichung der Form (2.1) nach der *Hoeffding*-Ungleichung. Dies zeigen wir in Satz 2.1. Eine Anforderung hierfür ist, dass die betrachtete Funktion nur Werte in $[A, B] \subset \mathbb{R}$ annehmen darf.

2. Unter Verwendung der *Booleschen* Ungleichung

$$P(\cup_{i=1}^{N} A_i) \leq \sum_{i=1}^{N} P(A_i) \tag{2.3}$$

können wir die Ungleichung aus Satz 2.1 auf endlich viele Funktionen erweitern. Damit erhalten wir in Satz 2.2 eine Grenze der Form (2.1) für $\#\mathcal{F} < \infty$. Die Boolesche Ungleichung ist allerdings immer erreicht, falls man $N = \infty$ betrachtet und falls ein $\varepsilon > 0$ existiert, sodass: $\forall i \in \mathbb{N} : P(A_i) > \varepsilon$.

3. Diese Vorüberlegung wollen wir nun nutzen, um eine Grenze für unendlich viele Funktionen zu erhalten. Dazu zeigen wir in zwei Schritten, dass bereits endlich viele Funktionen für die Berechnung ausreichend sind.

   - Zuerst zeigen wir, dass es genügt, nur die Vorzeichen der verwendeten Funktionen zu betrachten. Somit können wir uns nach Satz 2.3 auf Funktionen mit Bild in $\{0, 1\}$ beschränken. Damit müssen wir auf den n Datenpunkten höchstens 2^n verschiedene Funktionen betrachten.
   - Um die Anzahl der Funktionen in der Abschätzung

     $$\sup_{f \in \mathcal{F}} |R(f) - \widehat{R}(f)|$$

     beschränken zu können, müssen wir noch das Integral durch eine endliche Summe ersetzen. In Satz 2.4 sehen wir, dass wir eine nur um Konstanten schlechtere Abschätzung erhalten, falls wir das Integral durch das empirische Risiko auf n weiteren Datenpunkten ersetzen. Damit können wir auf diesen n Punkte die Boolesche Ungleichung (2.3) für $N = 2^{2n}$ Mengen anwenden und erhalten die Abschätzung:

     $$\sup_{f \in \mathcal{F}} |R(f) - \widehat{R}(f)| \leq (B - A) \sqrt{\frac{\ln\left(2^{2n} \frac{2}{\delta}\right)}{2n}}. \tag{2.4}$$

4. Wir sehen in der Ungleichung (2.4), dass die rechte Seite für $n \to \infty$ nur gegen 0 geht, falls wir den Wert $N = 2^n$ mit einem Wert ersetzen

können, der nicht exponentiell von n abhängt. In Kapitel 3 betrachten wir eine Möglichkeit, Anforderungen an $\mathcal{F}$ zu stellen, damit dies nur polynomial in n steigt.

In Kapitel 4 fügen wir die Ergebnisse zusammen und leiten die sog. *VC-Schranke* her, welche mit einer Wahrscheinlichkeit von mindestens $(1-\delta)$ gilt:

$$\sup_{f\in\mathcal{F}}|R(f)-\widehat{R}(f)| \leq (B-A)^2\sqrt{2\frac{d\ln\left(\frac{2en}{d}\right)+\ln\left(\frac{4}{\delta}\right)}{n}}. \tag{2.5}$$

Dabei ist die Konstanten d nur abhängig von der gewählten Menge $\mathcal{F}$. Weiterhin fordern wir, dass alle Funktionen aus $\mathcal{F}$ nur Werte in dem Intervall $[A,B]$ annehmen.

2.1 $\#\mathcal{F}$ endlich

Als Ausgangspunkt hierfür nehmen wir Hoeffdings Ungleichung, siehe, auch für den Beweis, [Hoe63], Satz 2 und [Mas07], Proposition 2.7.

Satz 2.1. *Seien $Z_1,\ldots,Z_n$ unabhängige, reellwertige Zufallsvariablen auf $Z\subset\mathbb{R}^k$, welche nach P verteilt sind, mit $A\leq Z_i\leq B$, $1\leq i\leq n$ und $[A,B]\subset\mathbb{R}$. Dann gilt für alle $\alpha>0$:*

$$P\left(\left|\int_Z Z_i(z)dP(z)-\frac{1}{n}\sum_{i=1}^{n}Z_i(z)\right|\geq\alpha\right)\leq 2\exp\left(-2n\alpha^2(B-A)^{-2}\right). \tag{2.6}$$

Setzt man in den obigen Satz 2.1 für die Zufallsvariablen X_i die messbare Funktion $f: Z\to\mathbb{R}$ angewandt auf Zufallsvariablen $Z_i: Z\to Z$, $1\leq i\leq n$ ein, so erhält man:

$$P(|\widehat{R}(f)-R(f)|\geq\alpha)\leq 2\exp\left(-2n\alpha^2(B-A)^{-2}\right). \tag{2.7}$$

Bezeichnen wir die rechte Seite der obigen Ungleichung (2.7) mit δ und lösen

dies nach α auf, so erhalten wir:

$$\delta = 2\exp\left(-2n\alpha^2(B-A)^{-2}\right) \tag{2.8}$$

$$(B-A)^2\ln\left(\frac{2}{\delta}\right) = -2n\alpha^2 \tag{2.9}$$

$$(B-A)\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}} = \alpha. \tag{2.10}$$

Damit können wir die Grenze (2.7) umformen und erhalten für alle $\delta \in (0,1]$ ein α wie in Formelzeile (2.10), sodass eine Schranke der Form (2.1) gilt:

$$P\left(|\widehat{R}(f) - R(f)| < (B-A)\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}}\right) \geq 1-\delta. \tag{2.11}$$

Dies gilt nur für die fest gewählte Funktion f. Allerdings können wir mit der Booleschen Ungleichung (engl. *union bound*), vgl. auch [Hau95a], Satz 10, eine Schranke herleiten, welche auch für endliche viele durch $-\infty < A < B < \infty$ beschränkte Funktionen $f_1, \ldots, f_N$ gleichzeitig gilt. Die Boolesche Ungleichung besagt für ein Wahrscheinlichkeitsmaß P und Ereignisse $A_1, \ldots, A_N$ aus dem zugehörigen Wahrscheinlichkeitsraum:

$$P(\cup_{i=1}^{N} A_i) \leq \sum_{i=1}^{N} P(A_i). \tag{2.12}$$

Diese Idee führt zu dem nächsten Satz.

Satz 2.2. *Es gilt für $\mathcal{F} \subset [A,B]^Z$, mit $[A,B] \subset \mathbb{R}$, $\#\mathcal{F} = N$ und alle $\delta \in (0,1]$ mit einer Wahrscheinlichkeit von mindestens $(1-\delta)$:*

$$\max_{f\in\mathcal{F}} |R(f) - \widehat{R}(f)| < (B-A)\sqrt{\frac{\ln\left(N\frac{2}{\delta}\right)}{2n}}. \tag{2.13}$$

Beweis. Wenden wir die Boolesche Ungeichung (2.12) N-mal auf die Hoeffding-Ungleichung (2.6) an, so erhalten wir:

$$P\left(\exists f \in \mathcal{F} : |R(f) - \widehat{R}(f)| \geq \varepsilon\right) \leq 2N\exp\left(-2n\varepsilon^2(B-A)^{-2}\right). \tag{2.14}$$

Formen wir diese Grenze analog zu den Formelzeilen (2.8) bis (2.10) um, erhalten wir:

$$(B-A)\sqrt{\frac{\ln\left(N\frac{2}{\delta}\right)}{2n}} = \varepsilon.$$

Analog zur Herleitung von Ungleichung (2.11) gilt für jedes $\delta \in (0,1]$ mit einer Wahrscheinlichkeit von mindestens $(1-\delta)$:

$$\forall f \in \mathcal{F} : |R(f) - \widehat{R}(f)| < (B-A)\sqrt{\frac{\ln\left(N\frac{2}{\delta}\right)}{2n}}, \tag{2.15}$$

und somit die Behauptung. □

2.2 Übergang von $[A,B]$ zu $\{0,1\}$

Für diesen Abschnitt folgen wir [Vap98], Kapitel 5.2.

Satz 2.3. *Sei $[A,B] \subset \mathbb{R}$ und $\mathcal{F}$ eine Teilmenge aus $[A,B]^Z$. Seien wieder $z_1, \ldots, z_n$ die beobachteten Punkte aus Z und*

$$\mathcal{F}_{\mathcal{H}} = \{\mathcal{H}(f(z) - \beta) : f \in \mathcal{F}, \beta \in \mathbb{R}\}.$$

Dann gilt:

$$\begin{aligned} &P\left(\left\{\sup_{f\in\mathcal{F}}|R(f) - \widehat{R}(f)| > \varepsilon\right\}\right) \\ \leq &P\left(\left\{\sup_{f_{\mathcal{H}}\in\mathcal{F}_{\mathcal{H}}}\left|\int_Z f_{\mathcal{H}}(z)dP(z) - \frac{1}{n}\sum_{i=1}^{n} f_{\mathcal{H}}(z_i)\right| > \frac{\varepsilon}{B-A}\right\}\right). \end{aligned}$$

Beweis. Für den Beweis betrachten wir das Risiko und das empirische Risiko als Integrale. Dafür wiederholen wir kurz unsere Notation hierzu:

$$\begin{aligned} R(f) &= \int_Z f(z)dP(z), \\ \widehat{R}(f) &= \frac{1}{n}\sum_{i=1}^{n} f(z_i). \end{aligned}$$

Somit ist $R(f)$ ein definiertes Integral gegen die Wahrscheinlichkeitsverteilung P, da die Funktionen aus $\mathcal{F}$ nach unten und oben beschränkt sind. Im Folgenden betrachten wir für eine feste Funktion $f \in \mathcal{F}$ die um A verschobene Funktion $f_A(z) := f(z) - A$ mit Werten in $[0, B - A]$. Zur Vereinfachung schreiben wir $\widetilde{B} := B - A$.

Wie in der Definition des Integrals wollen wir nun die Funktion f_A durch Treppenfunktionen $f_{A,l}$ herleiten.

$$\begin{aligned} f_{A,\ell}(z) :=& \sum_{i=0}^{\ell-1} \frac{\widetilde{B}}{\ell} I\left\{ f_A(z) > \frac{i\widetilde{B}}{\ell} \right\} \\ =& \sum_{i=1}^{l} \frac{(i-1)\widetilde{B}}{\ell} I\left\{ f_A^{-1}\left(\frac{(i-1)\widetilde{B}}{\ell}, \frac{i\widetilde{B}}{\ell} \right) \right\}. \end{aligned}$$

Dabei beschreibt $I\{\cdot\}$ die Indikatorfunktion. Die Funktionenfolge $(f_{A,l})_{l\in\mathbb{N}}$ konvergiert dann punktweise von unten gegen die Funktion f_A. Damit kann man den Satz der monotonen Konvergenz, vgl. [Kön06], Kapitel 7, Satz 4, anwenden. Man erhält:

$$\lim_{\ell\to\infty} \int_Z f_{A,\ell}(z) dP(z) = \int_Z f_A(z) dP(z).$$

Dabei gilt:

$$\int_Z f_{A,\ell}(z) dP(z) = \sum_{i=0}^{\ell-1} \frac{\widetilde{B}}{\ell} P\left(\left\{ f_A(z) > \frac{i\widetilde{B}}{\ell} \right\} \right).$$

Somit können wir das Risiko nun als Grenzwert einer Summe betrachten:

$$\int_Z f_A(z) dP(z) = \lim_{\ell\to\infty} \sum_{i=0}^{\ell-1} \frac{\widetilde{B}}{\ell} P\left(\left\{ f_A(z) > \frac{i\widetilde{B}}{\ell} \right\} \right). \tag{2.16}$$

Ebenso wollen wir das empirische Risiko $\frac{1}{n}\sum_{i=1}^n f_A(z_i)$ ähnlich darstellen, allerdings unter der Verwendung eines anderen Wahrscheinlichkeitsmaßes. Dazu bemerken wir, dass sich das empirische Risiko ebenfalls von unten annähern lässt:

$$\frac{1}{n}\sum_{i=1}^n f_A(z_i) = \lim_{\ell\to\infty} \frac{1}{n} \sum_{i=1}^n \max_{j\in\{0,\ldots,\ell-1\}} \left\{ \frac{j\widetilde{B}}{\ell} : \frac{j\widetilde{B}}{\ell} < f_A(z_i) \right\}.$$

Diese Summe können wir für festes ℓ wie folgt umschreiben:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\max_{j\in\{0,\dots,\ell-1\}}\left\{\frac{j\widetilde{B}}{\ell}:\frac{j\widetilde{B}}{\ell}<f_A(z_i)\right\}\\
=&\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{\ell-1}\frac{\widetilde{B}}{\ell}I\left\{\frac{j\widetilde{B}}{\ell}<f_A(z_i)\right\}\\
=&\sum_{j=0}^{\ell-1}\frac{\widetilde{B}}{\ell}\frac{1}{n}\sum_{i=1}^{n}I\left\{\frac{j\widetilde{B}}{\ell}<f_A(z_i)\right\}.
\end{aligned}$$

Die innere Summe können wir umschreiben:

$$\sum_{i=1}^{n}I\{\frac{j\widetilde{B}}{\ell}<f_A(z_i)\}=\#\left\{i\in\{1,\dots,n\}:\frac{j\widetilde{B}}{\ell}<f_A(z_i)\right\}.$$

Im Folgenden beschreibe S die Menge der Datenpunkte z_i, also:

$$S:=\{z_1,\dots,z_n\}.$$

Nutzen wir dies, um ein Wahrscheinlichkeitsmaß ν zu definieren:

$$\forall\mathcal{A}\subset Z:\nu(\mathcal{A}):=\frac{\#\{z\in S\cap\mathcal{A}\}}{n}.$$

Dies ist ein Wahrscheinlichkeitsmaß, da $\#S=n$. Insgesamt erhalten wir also:

$$\frac{1}{n}\sum_{i=1}^{n}f_A(z_i)=\lim_{\ell\to\infty}\sum_{j=0}^{\ell-1}\frac{\widetilde{B}}{\ell}\nu\left(\left\{\frac{j\widetilde{B}}{\ell}<f_A(z)\right\}\right). \tag{2.17}$$

Mithilfe dieser Darstellungen (2.16) und (2.17) erhalten wir nun die folgende Abschätzung:

$$\begin{aligned}
&\left|\int_Z f_A(z)dP(z)-\frac{1}{n}\sum_{i=1}^{n}f_A(z_i)\right|\\
=&\left|\lim_{\ell\to\infty}\sum_{i=0}^{\ell-1}\frac{\widetilde{B}}{\ell}P\left(\left\{f_A(z)>\frac{i\widetilde{B}}{\ell}\right\}\right)-\lim_{\ell\to\infty}\sum_{i=0}^{\ell-1}\frac{\widetilde{B}}{\ell}\nu\left(\left\{f_A(z)>\frac{i\widetilde{B}}{\ell}\right\}\right)\right|\\
\leq&\lim_{\ell\to\infty}\sum_{i=0}^{\ell-1}\frac{\widetilde{B}}{\ell}\left|P\left(\left\{f_A(z)>\frac{i\widetilde{B}}{\ell}\right\}\right)-\nu\left(\left\{f_A(z)>\frac{i\widetilde{B}}{\ell}\right\}\right)\right|
\end{aligned}$$

Die innere Abschätzungen erfolgen immer gegen denselben Wert $\frac{i\widetilde{B}}{\ell}$. Deswegen lässt sich dieser Wert auch abschätzen gegen das Supremum über alle möglichen Werte hierfür, also alle $\beta \in [0, \widetilde{B})$. Dadurch ist der Wert der Summe $\sum_{i=0}^{\ell-1} \frac{1}{\ell} = 1$.

$$
\begin{aligned}
&\leq \lim_{\ell\to\infty} \sum_{i=0}^{\ell-1} \frac{\widetilde{B}}{\ell} \left(\sup_{\beta\in[0,\widetilde{B})} |P(\{f_A(z) > \beta\}) - \nu(\{f_A(z) > \beta\})| \right) \\
&= \widetilde{B} \cdot \sup_{\beta\in[0,\widetilde{B})} |P(\{f_A(z) > \beta\}) - \nu(\{f_A(z) > \beta\})| \\
&= \widetilde{B} \cdot \sup_{\beta\in[A,B)} |P(\{f(z) > \beta\}) - \nu(\{f(z) > \beta\})| \\
&= \widetilde{B} \cdot \sup_{\beta\in[A,B)} \left| \int_Z \mathcal{H}(f(z) - \beta) dP(z) - \frac{1}{n} \sum_{i=1}^{n} \mathcal{H}(f(z) - \beta) \right|.
\end{aligned}
$$

Für die letzte Umformung haben wir wieder die Definition der Wahrscheinlichkeitsmaße P und ν genutzt.

Damit haben wir die Grenze für eine jede feste Funktion $f \in \mathcal{F}$ einzeln hergeleitet, insbesondere auch für

$$\tilde{f} := \arg\sup_{f\in\mathcal{F}} |R(f) - \widehat{R}(f)|.$$

Dafür gilt dann:

$$
\begin{aligned}
&P\left(\left\{ \sup_{f\in\mathcal{F}} |R(f) - \widehat{R}(f)| > \varepsilon \right\} \right) \\
=&P\left(\left\{ |R(\tilde{f}) - \widehat{R}(\tilde{f})| > \varepsilon \right\} \right) \\
\leq&P\left(\left\{ \left| \int_Z \mathcal{H} \circ \tilde{f}(z) dP(z) - \frac{1}{n} \sum_{i=1}^{n} \mathcal{H} \circ \tilde{f}(z_i) \right| > \frac{\varepsilon}{B - A} \right\} \right) \\
\leq&P\left(\left\{ \sup_{f_{\mathcal{H}}\in\mathcal{F}_{\mathcal{H}}} \left| \int_Z f_{\mathcal{H}}(z) dP(z) - \frac{1}{n} \sum_{i=1}^{n} f_{\mathcal{H}}(z_i) \right| > \frac{\varepsilon}{B - A} \right\} \right).
\end{aligned}
$$

□

2.3 Abschätzung des Bewertungsfehlers

Um das Risiko gegen das empirische Risiko abzuschätzen, haben wir im Satz 2.3 gesehen, dass es genügt, eine Klasse an Funktionen $\mathcal{F}_{\mathcal{H}}$ zu betrachten,

welche nur Werte in $\{0,1\}$ annehmen. Im nächsten Satz sehen wir, dass wir die Differenz des Risikos und des empirischen Risikos weiterhin abschätzen können, indem wir, statt das Risiko auf Z zu berechnen, ein empirisches Risiko auf n weiteren Punkten betrachten.

Dafür folgen wir [BBL03], Kapitel 4.4.

Satz 2.4. *Sei $\mathcal{F}_{\mathcal{H}}$ eine Menge an Funktionen aus $\{0,1\}^Z$. Seien weiterhin $z_1, \dots, z_{2n}$ Datenpunkte aus Z, unabhängig und identisch verteilt. Das empirische Risiko auf den Punkten $z_1, \dots, z_n$, bzw. auf den Punkten $z_{n+1}, \dots, z_{2n}$ bezeichnen wir für $f \in \mathcal{F}_{\mathcal{H},L}$:*

$$\begin{aligned}\widehat{R}(f) :=& \frac{1}{n} \sum_{i=1}^{n} f(z_i),\\ \widehat{R}'(f) :=& \frac{1}{n} \sum_{i=n+1}^{2n} f(z_i).\end{aligned}$$

Gelte weiterhin für $\varepsilon > 0$ und $n \in \mathbb{N}$: $n\varepsilon^2 \geq 2$. So gilt:

$$P\left(\sup_{f \in \mathcal{F}_{\mathcal{H}}} |R(f) - \widehat{R}(f)| > \varepsilon\right) \leq 2P\left(\sup_{f \in \mathcal{F}_{\mathcal{H}}} |\widehat{R}(f) - \widehat{R}'(f)| > \frac{\varepsilon}{2}\right). \tag{2.18}$$

Beweis. Sei im Folgenden $f^* \in \mathcal{F}_{\mathcal{H}}$ beliebig, aber fest gewählt. Zuerst bemerken wir, dass auf Grund der Dreiecksungleichung gilt:

$$\begin{aligned}P\left(\sup_{f \in \mathcal{F}_{\mathcal{H}}} |\widehat{R}(f) - \widehat{R}'(f)| > \frac{\varepsilon}{2}\right) &\geq P\left(|\widehat{R}(f^*) - \widehat{R}'(f^*)| > \frac{\varepsilon}{2}\right)\\ &\geq P\left(|R(f^*) - \widehat{R}(f^*)| \geq \varepsilon, |R(f^*) - \widehat{R}'(f^*)| \leq \frac{\varepsilon}{2}\right).\end{aligned}$$

Da die Punkte $z_1, \dots, z_n$ unabhängig von den Punkten $z_{n+1}, \dots, z_{2n}$ gewählt sind, kann man die rechte Seite auch als Produkt schreiben:

$$\begin{aligned}&P\left(|R(f^*) - \widehat{R}(f^*)| \geq \varepsilon, |R(f^*) - \widehat{R}'(f^*)| \leq \frac{\varepsilon}{2}\right)\\ =&P\left(|R(f^*) - \widehat{R}(f^*)| \geq \varepsilon\right) \cdot P\left(|R(f^*) - \widehat{R}'(f^*)| \leq \frac{\varepsilon}{2}\right).\end{aligned}$$

Nun wollen wir den rechten Faktor abschätzen gegen $\frac{1}{2}$. Dafür schreiben wir zuerst:

$$P\left(|R(f^*) - \widehat{R}'(f^*)| \leq \frac{\varepsilon}{2}\right) = 1 - P\left(|R(f^*) - \widehat{R}'(f^*)| > \frac{\varepsilon}{2}\right).$$

Darauf können wir nun die Tschebyscheff-Ungleichung anwenden, vgl. [Kle13], Satz 5.11:

$$P\left(|R(f^*) - \widehat{R}'(f^*)| > \frac{\varepsilon}{2}\right) \leq \left(\frac{\varepsilon}{2}\right)^{-2} \frac{1}{n^2}(n \operatorname{Var}(f^*)) = \frac{4 \operatorname{Var}(f^*)}{n\varepsilon^2}.$$

Die Ungleichung gilt dabei, da die Werte $z_1, \ldots, z_{2n}$ unabhängig und identisch verteilt sind. Da die Funktion f^* nur Werte in $\{0, 1\}$ annimmt, folgt mit Hilfe der Popoviciu-Ungleichung für Varianzen, siehe zum Beispiel Korollar 1 in [BD00], dass die Varianz von f^* höchstens $\frac{1}{4}$ sein kann. Man sieht dies mit $f^{*^2}(z) \leq f^*(z)$ auch anhand einer kleinen Rechnung für $E[f^*] = p \in [0, 1]$:

$$E[(f^* - p)^2] = E[f^{*^2}] - 2pE[f^*] + p^2 \leq p - p^2 = p(1 - p) \leq \frac{1}{4}. \tag{2.19}$$

Da wir gefordert haben, dass $n\varepsilon^2 \geq 2$ gilt, folgt:

$$\frac{4 \operatorname{Var}(f^*)}{n\varepsilon^2} \leq \frac{1}{n\varepsilon^2} \leq \frac{1}{2}.$$

Damit gilt insgesamt:

$$P\left(|R(f^*) - \widehat{R}'(f^*)| \leq \frac{\varepsilon}{2}\right) \geq \frac{1}{2},$$

also:

$$\begin{aligned}
&P\left(\left|\widehat{R}(f^*) - \widehat{R}'(f^*)\right| > \frac{\varepsilon}{2}\right) \\
\geq &P\left(|R(f^*) - \widehat{R}(f^*)| \geq \varepsilon\right) \cdot P\left(|R(f^*) - \widehat{R}'(f^*)| \leq \frac{\varepsilon}{2}\right) \\
\geq &\frac{1}{2} P\left(|R(f^*) - \widehat{R}(f^*)| \geq \varepsilon\right).
\end{aligned}$$

Insgesamt haben wir dies somit auch bewiesen für die Funktion, bei welcher das Supremum der linken Seite angenommen wird, also für

$$f^* := \arg \sup_{f \in \mathcal{F}_\mathcal{H}} |R(f) - \widehat{R}(f)|.$$

Damit ist f^* insbesondere abhängig von den Daten $z_1, \ldots, z_n$. Aus der Definition des Supremums folgt für alle Funktionen $f \in \mathcal{F}_\mathcal{H}$:

$$\sup_{f \in \mathcal{F}_\mathcal{H}} |\widehat{R}(f) - \widehat{R}'(f)| \geq |\widehat{R}(f^*) - \widehat{R}'(f^*)|$$

und somit die Behauptung.

□

Solche Beweistechniken nennt man in der Literatur auch *Symmetrisierung*. Die dabei eingeführten Variablen $z_{n+1}, \ldots, z_{2n}$ bezeichnet man dann als Geistvariablen, engl. *ghost variables*. Diese Aussage mit Beweis befindet sich ebenfalls in [Vap98], Kapitel 4.5.1 und 4.5.2. Dort wird allerdings der Term $\frac{\varepsilon}{2}$ in der rechten Seite der Ungleichung (2.18) durch den Term $\varepsilon - \frac{1}{n}$ ersetzt.

3 Die VC-Dimension

In diesem Kapitel lernen wir die Technik der Vapnik-Chervonenkis-Dimension, kurz VC-Dimension oder VCD, genauer kennen, um die Größe der Funktionenmenge $\mathcal{F}$ beschränken zu können. Zuerst definieren wir dabei die sog. *Wachstumsfunktion* und die VC-Dimension. Anschließend schätzen wir die Wachstumsfunktion gegen einen von der VC-Dimension abhängigen Term ab. Im Kapitel 3.3 beschreiben wir Möglichkeiten, die VC-Dimension zu berechnen oder nach oben abzuschätzen.

3.1 Definition der VC-Dimension

3.1.1 Die VCD für Funktionen mit Bild in $\{0,1\}$

Wir wollen nun die Zahl der verschiedenen Funktionen aus $\mathcal{F}_\mathcal{H} \subset \{0,1\}^Z$, eingeschränkt auf die n Punkte aus $S \subset Z$, abschätzen. Dazu betrachten wir die folgenden Definitionen, analog zu [vLS11], Kapitel 5, [Vap98], Kapitel 4.9 und [Sch01].

Definition 3.1. *Sei $S^+ \subset S$ fest. Wir sagen, dass ein festes $f \in \mathcal{F}_\mathcal{H}$ die Teilmenge $S^+ \subset S$ induziert (engl. to induce), falls gilt:*

$$z \in S \implies \qquad z \in S^+ \Leftrightarrow f(z) = 1.$$

Der Begriff ist in der Literatur nicht eindeutig, manchmal wird auch *to carve out* benutzt, vgl. [Des14] oder *to cut*, vergleiche den Titel von [Dud79]. In [SC08], Definition 4.53, sagt man auch, dass die Menge S^+ von der Menge $S \backslash S^+$ *getrennt* wird (engl. *to seperate*).

Jede Funktion $f \in \mathcal{F}_\mathcal{H}$ induziert dabei die folgende Teilmenge S_f von S:

$$S_f = S \cap f^{-1}(1).$$

Die maximale Anzahl der induzierten Teilmengen einer Menge mit n Elementen bezeichnen wir mit der *Wachstumsfunktion* $\mathscr{G}_{\mathcal{F}_\mathcal{H}}(n)$, engl. *growth function*.

Definition 3.2. *Sei wieder $\mathcal{F}_\mathcal{H} \subset \{0,1\}^Z$, dann ist für jedes $S \subset Z$:*

$$\mathcal{F}_\mathcal{H}|_S := \{f|_S : f \in \mathcal{F}_\mathcal{H}\}.$$

Die Wachstumsfunktion $\mathscr{G}_{\mathcal{F}_\mathcal{H}}$ von $\mathcal{F}_\mathcal{H}$ definieren wir für alle $n \in \mathbb{N}$ als:

$$\mathscr{G}_{\mathcal{F}_\mathcal{H}}(n) := \max_{S \subset Z, \#S = n} \#\mathcal{F}_\mathcal{H}|_S.$$

Dabei setzen wir $\mathscr{G}_{\mathcal{F}_\mathcal{H}}(0) := 1$.

Wenn die verwendete Funktionenmenge aus dem Kontext klar ist, schreiben wir oft auch nur $\mathscr{G}(n)$. Die Wachstumsfunktion $\mathscr{G}(n)$ wird in der Literatur auch oft mit $\Gamma(n)$ oder $\Pi(n)$ bezeichnet, vgl. [Son98] und [BEHW89].

Offensichtlich gilt immer $\mathscr{G}_{\mathcal{F}_\mathcal{H}}(n) \leq 2^n$, da der Bildraum nur aus zwei Elementen besteht. In diesem Fall werden also alle Teilmengen einer Menge $S \subset Z$, $\#S = n$ durch $\mathcal{F}_\mathcal{H}$ induziert. In diesem Fall sagen wir auch, dass S durch $\mathcal{F}$ *gesplittert* werden, engl. *to shatter*.

Definition 3.3. *Die Menge $\mathcal{F}_\mathcal{H}$ splittert die Menge $S \subset Z$, mit $\#S = n \in \mathbb{N}$, falls die Mächtigkeit der Menge $\{f_{|S} | f \in \mathcal{F}_\mathcal{H}\}$ genau 2^n entspricht.*

Mit der Vapnik-Chervonenkis-Dimension, kurz VC-Dimension oder VCD, halten wir die größte Kardinalität einer Menge S fest, welche durch $\mathcal{F}_\mathcal{H}$ gesplittert wird.

Definition 3.4. *Sei wieder $\mathcal{F} \subset \{0,1\}^X$, dann definieren wir:*

$$\mathrm{VCD}(\mathcal{F}) := \begin{cases} \max\{n \in \mathbb{N}_0 | \mathscr{G}_\mathcal{F}(n) = 2^n\}, & \textit{falls existent,} \\ \infty, & \textit{sonst.} \end{cases}$$

Aus der Definition der VCD sieht man, dass die VCD endlich ist, falls die Wachstumsfunktion in n nur polynomial wächst. Betrachten wir als Beispiel

$$\mathcal{F}_1 := \{f_i(j) : \mathbb{N} \to \{0,1\}, f_i(j) = \delta_{i,j}, \quad i \in \mathbb{N}\},$$

mit dem Kronecker-Delta $\delta_{i,j}$. Hier können wir auf einer beliebigen Teilmenge $S \subset \mathbb{N}$, $\#S = n$ alle Teilmengen mit genau einem Element induzieren, sowie die leere Menge. Dies ergibt $n+1$ Teilmengen. Das größte $n \in \mathbb{N}$, welches $n + 1 = 2^n$ erfüllt, ist 1, somit gilt $\mathrm{VCD}(\mathcal{F}_1) = 1$.

Setzen wir in die Abschätzung (2.13) statt N die Wachstumsfunktion $\mathscr{G}(n)$ ein, so erhalten wir:

$$\sup_{f \in \mathcal{F}} |R(f) - \widehat{R}(f)| \leq \sqrt{\frac{\ln(\mathscr{G}(n)) + \ln\left(\frac{2}{\delta}\right)}{2n}}. \tag{3.1}$$

Dabei erkennt man, dass für alle festen $\delta \in (0,1]$ die rechte Seite der obigen Ungleichung (3.1) nur gegen 0 geht, falls gilt:

$$\lim_{n \to \infty} \frac{\ln(\mathscr{G}(n))}{n} \to 0. \tag{3.2}$$

Dies ist erreicht, falls $\mathscr{G}(n)$ nur polynomiell in n wächst. In Satz 3.7 sehen wir, dass dies äquivalent dazu ist, dass die VC-Dimension der Menge $\mathcal{F}$ endlich ist.

3.1.2 Die VCD für Teilmengen von $\mathcal{P}(Z)$

Wir können die VCD auch für Mengen von Teilmengen von X betrachten. Sei hierfür $\mathcal{C} \subset \mathcal{P}(Z)$, wobei $\mathcal{P}(Z)$ die Mengen aller Teilmengen von Z beschreibt. Wir nennen $\mathcal{C}$ auch Konzeptklasse, engl. *concept class*, nach [Son98]. Mit der folgenden Identifikation lassen sich die obigen Definitionen 3.1 bis 3.4 auch auf $\mathcal{C}$ anwenden.

Definition 3.5. *Sei $\mathcal{C} \subset \mathcal{P}(Z)$ fest. Wir definieren für alle $C \in \mathcal{C}$ eine Funktion $f_C : Z \to \{0, 1\}$, mit:*

$$f_C(z) := \begin{cases} 1, & x \in C, \\ 0, & \text{sonst.} \end{cases}$$

Sei weiterhin:

$$\mathcal{F}_\mathcal{C} := \{f_c : c \in \mathcal{C}\}.$$

Dann definieren wir für Teilmengen $S^+ \subset S \subset Z$:

- *$S^+ \subset S$ wird genau dann durch $C \in \mathcal{C}$ induziert, wenn S^+ durch f_C induziert wird.*
- *$\mathscr{G}_\mathcal{C}(n) := \mathscr{G}_{\mathcal{F}_\mathcal{C}}(n)$ beschreibt die Wachstumsfunktion von $\mathcal{C}$.*
- *S wird genau dann durch $\mathcal{C}$ gesplittert, wenn S durch $\mathcal{F}_\mathcal{C}$ gesplittert wird.*
- $\mathrm{VCD}(\mathcal{C}) := \mathrm{VCD}(\mathcal{F})$.

3.1.3 Die VC-Dimension für Funktionen mit Bild in $\mathbb{R}$

Hierbei folgen wir [Vap98], Kapitel 5.2.3. Als Motivation betrachten wir Satz 2.3. Mit dessen Hilfe konnten wir die Schranke aus Satz 2.2 für reellwertige Funktionen auch für Funktionen anwenden, welche auf $\{0, 1\}$ abbilden. Dabei haben wir statt einer Funktion $f(z)$ die Funktion $\mathcal{H}(f(z) - \beta)$ verwendet, mit $\beta \in \mathbb{R}$ und $\mathcal{H}$ aus Formelzeile (0.5).

Dazu führen wir die folgende Definition analog zu Definition 3.5 ein.

Definition 3.6. *Sei $A, B \in \mathbb{R}$ und $\mathcal{F} \subset \mathbb{R}^Z$. Dann definieren wir für jedes $f \in \mathcal{F}$ und jedes $\beta \in \mathbb{R}$:*

$$f_{\mathcal{H}} : Z \to \{0,1\}, \qquad z \mapsto \mathcal{H}(f(z) - \beta).$$

Diese Funktionen fassen wir zusammen in der Menge

$$\mathcal{F}_{\mathcal{H}} := \{\mathcal{H}(f(\cdot) - \beta) : f \in \mathcal{F}, \beta \in \mathbb{R}\}.$$

Dann definieren wir für Teilmengen $S^+ \subset S \subset X$:

- *$S^+ \subset S$ wird durch genau dann durch $(f, \beta) \in \mathcal{F} \times \mathbb{R}$ induziert, wenn S^+ durch f_β induziert wird.*
- *$\mathscr{G}_{\mathcal{F}}(n) := \mathscr{G}_{\mathcal{F}_\beta}(n)$ beschreibt die Wachstumsfunktion von $\mathcal{F}$.*
- *S wird genau dann durch $\mathcal{F}$ gesplittert, wenn S durch $\mathcal{F}_\beta$ gesplittert wird.*
- $\mathrm{VCD}(\mathcal{F}) := \mathrm{VCD}(\mathcal{F}_\beta)$.

Wir sagen, dass f die Menge S^+ induziert, falls ein $\beta \in \mathbb{R}$ existiert, sodass (f, β) die Menge S^+ induziert.

Zur Veranschaulichung betrachten wir zwei Beispiele, wie wir die VC-Dimension für reellwertige Funktionen berechnen können. Im ersten Beispiel berechnen wir die Wachstumsfunktion, im zweiten Beispiel finden wir eine Menge mit drei Elementen, welche nicht mehr induziert werden kann. Somit ist die VC-Dimension in diesem Beispiel höchstens zwei.

Als erstes Beispiel betrachten wir dabei die Menge der Geraden in $\mathbb{R}$,

$$\mathcal{F}_2 := \{f : \mathbb{R} \to \mathbb{R}, x \mapsto ax, \quad a \in \mathbb{R}\}.$$

Für die Bestimmung der VC-Dimension von $\mathcal{F}_2$ müssen wir also die VC-Dimension der Menge

$$\mathcal{F}_{2,\mathcal{H}} := \{\mathcal{H}(f(\cdot) - \beta) : \mathbb{R} \to \{0,1\}, \quad f \in \mathcal{F}\}$$

bestimmen. Diese entspricht der Menge der Indikatorfunktionen aller halboffenen Intervalle in $\mathbb{R}$, also

$$\{I(-\infty, a), a \in \mathbb{R}\} \cup \{I(a, \infty), a \in \mathbb{R}\}.$$

Dementsprechend induziert die Menge $\mathcal{F}_2$ auf beliebigen n Punkten aus $\mathbb{R}$ genau $2n$ Mengen. Dies sieht man daran, dass für eine Menge an Punkten

$$x_1 < \ldots < x_n$$

für jedes $1 \leq k \leq n$ die beiden Mengen

$$\{x_1, \ldots, x_k\}, \quad \{x_{k+1}, \ldots, x_n\}$$

induziert werden. Dabei ist die Menge $\{x_{n+1}, \ldots, x_n\}$ leer. Das maximale $n \in \mathbb{N}$, welches $2n = 2^n$ erfüllt, ist 2, somit gilt $\mathrm{VCD}(\mathcal{F}_2) = 2$.

Als zweites Beispiel betrachten wir die Menge aller nach unten geöffneten Parabeln

$$\mathcal{F}_3 := \{f : \mathbb{R} \to \mathbb{R}, x \mapsto -(x-a)(x-b), a, b \in \mathbb{R}\}.$$

Wir wollen nun zeigen, dass zwei Punkte aus $\mathbb{R}$ durch $\mathcal{F}_3$ gesplittert werden, allerdings keine drei. Damit ergibt sich $\mathrm{VCD}(\mathcal{F}_3) = 2$. Dafür wählen wir $x_1 < x_2 \in \mathbb{R}$. Eine Funktion $-(x-a)(x-b) \in \mathcal{F}_3$ ist genau dann größer als 0, wenn x in (a, b) liegt. Damit können wir die 4 Teilmengen von $\{x_1, x_2\}$ induzieren. Dabei sei $0 < \varepsilon < |x_2 - x_1|$.

- $\emptyset$ wird zum Beispiel durch die Wahl $(a, b) = (x_2 + \varepsilon, x_2 + 2\varepsilon)$ und $\beta = 0$ induziert.
- $\{x_1\}$ wird zum Beispiel durch die Wahl $(a, b) = (x_1 - \varepsilon, x_1 + \varepsilon)$ und $\beta = 0$ induziert.
- $\{x_2\}$ wird zum Beispiel durch die Wahl $(a, b) = (x_2 - \varepsilon, x_2 + \varepsilon)$ und $\beta = 0$ induziert.
- $\{x_1, x_2\}$ wird zum Beispiel durch die Wahl $(a, b) = (x_1 - \varepsilon, x_2 + \varepsilon)$ und $\beta = 0$ induziert.

Allerdings lassen sich keine drei Punkte $x_1 < x_2 < x_3 \in \mathbb{R}$ splittern, hier lässt sich die Menge $\{x_1 x_3\}$ nicht induzieren. Dies liegt daran, dass jedes Intervall $(a, b) \subset \mathbb{R}$, welches x_1 und x_3 enthält, auch den Punkt x_2 enthalten muss. Somit ist jede nach unten geöffnete Parabel aus $\mathcal{F}_3$, welche in den Punkten x_1, x_3 größer als 0 ist, auch im Punkt x_2 größer als 0.

3.2 Abschätzung der Wachstumsfunktion

Im nächsten Satz sehen wir, dass sich für VCD($\mathcal{F}$) die Wachstumsfunktion $\mathscr{G}_{\mathcal{F}}(n)$ gegen ein Polynom vom Grad VCD($\mathcal{F}$) abschätzen lässt. Dafür verwenden wir die Anschauung aus Definition 3.5. Der Satz geht zurück auf Sauer [Sau72] und Shelah [She72]. Für den Beweis folgen wir [Son98], Satz 3.

Satz 3.7. *Sei $\mathcal{F} \subset \mathbb{R}^X$ mit endlicher VC-Dimension d. Dann lässt sich die Wachstumsfunktion für alle $n \in \mathbb{N}$ abschätzen durch:*

$$\mathscr{G}(n) \begin{cases} = 2^n, & n \leq d \\ \leq \left(\frac{en}{d}\right)^d, & n > d. \end{cases} \tag{3.3}$$

Dabei beschreibt e die Euler'sche Zahl $\exp(1)$*. Weiterhin zeigen wir:*

$$\mathscr{G}(n) \leq \sum_{i=0}^{d} \binom{n}{i}.$$

Beweis. Wir betrachten wieder, analog zu Definition 3.6, die Menge der Funktionen, welche auf $\{0, 1\}$ abbilden:

$$\mathcal{F}_{\mathcal{H}} := \{\mathcal{H}(f - \beta), f \in \mathcal{F}, \beta \in \mathbb{R}\}.$$

Für den Fall $n \leq d$ folgt die Abschätzung direkt aus den Definitionen für die Wachstumsfunktion und die VC-Dimension. Für den Fall $n > d$ definieren wir zuerst zwei Funktionen zur Vereinfachung der Notation. Sei dabei $z_1, \ldots, z_n \in Z$ und $\binom{n}{i}$ der Binomialkoeffizient:

$$\begin{aligned} \gamma(z_1, \ldots, z_n) :=& \#\{(f(z_1), \ldots, f(z_n))|\ f \in \mathcal{F}_{\mathcal{H}}\}, \\ \Phi(n, d) :=& \sum_{i=0}^{d} \binom{n}{i}. \end{aligned}$$

Damit wollen wir nun die folgende Ungleichheitskette beweisen:

$$\gamma(z_1, \ldots, z_n) \leq \Phi(n, d) \leq \left(\frac{en}{d}\right)^d. \tag{3.4}$$

Die Funktion γ können wir statt der Wachstumsfunktion $\mathscr{G}$ verwenden, da gilt:

$$\mathscr{G}(n) = k \in \mathbb{N} \Rightarrow \exists\{z_1, \ldots, z_n\} \subset Z : \gamma(z_1, \ldots, z_n) = k.$$

Sei nun $S = \{z_1, \ldots, z_n\} \subset Z$ eine beliebige Teilmenge von Z.

Beweisen wir nun den ersten Teil der Ungleichung (3.4). Für $n = d$ ist die Aussage klar, da $\Phi(d, d) = 2^d$, die Anzahl aller Teilmengen einer Menge mit d Elementen. Sei im Folgenden also $n > d$. Wir beweisen die Ungleichung durch Widerspruch. Angenommen:

$$\gamma(z_1, \ldots, z_n) = s > \Phi(n, d).$$

Dann können wir alle verschiedenen Funktionen auf $\{z_1, \ldots, z_n\}$ als Matrix schreiben:

$$M = \begin{pmatrix} f_1(z_1) & \ldots & f_s(z_1) \\ f_1(z_2) & \ldots & f_s(z_2) \\ \vdots & \vdots & \vdots \\ f_1(z_n) & \ldots & f_s(z_n) \end{pmatrix}.$$

Diese Matrix hat paarweise verschiedene Spalten, da alle betrachteten Funktionen auf S paarweise verschieden sind. Die i-te Spalte induziert hierbei $f_i^{-1}(1) \cap S$. Wir wollen nun durch Induktion über n zeigen, dass wir eine $(d+1) \times 2^{d+1}$-Untermatrix von M mit paarweise verschiedenen Spalten finden können. Damit hätten wir eine Teilmenge $S^+ \subset S$ mit $d+1$ Punkten gefunden, die durch $\mathcal{F}$ gesplittert wird, also wäre dann die VC-Dimension von $(\mathcal{F})$ mindestens $d+1$.

Induktionsanfang: Sei $n = d + 1$. Dann gilt:

$$\Phi(n, d) = \Phi(d+1, d) = \sum_{i=0}^{d} \binom{d+1}{i} = 2^{d+1} - 1,$$

also gilt $s \geq 2^{d+1}$. Damit ist M bereits eine $(d+1) \times 2^{d+1}$-Matrix mit paarweise verschiedenen Spalten.

Induktionsschritt $n \mapsto n+1$: Sei die Behauptung für ein allgemeines $n > d$ bewiesen. Betrachten wir nun eine Matrix M_1 mit $n+1$ Zeilen und

$$s_1 = \Phi(n+1, d)$$

Spalten. Betrachten wir hiervon die Untermatrix, die entsteht, wenn man die erste Zeile weglässt. Sind in dieser Untermatrix zwei Spalten gleich, so müssen diese also unterschiedliche Werte in der ersten Zeile annehmen. Da

alle Werte aus $\{0,1\}$ stammen, können höchstens zwei Spalten in der $n \times s_1$-Untermatrix von M_1 gleich sein. Folglich können wir die Spalten der Matrix M_1 wie folgt umordnen:

$$\begin{pmatrix} 0\cdots0 & 1\cdots1 & *\cdots* \\ A & A & B \end{pmatrix} \tag{3.5}$$

Sei r_1 die Anzahl der Spalten in A. Dann lässt sich die Behauptung durch Fallunterscheidung zeigen:

- $r_1 > \Phi(n, d-1)$: Damit erhält man direkt die Behauptung aus der Induktionsvoraussetzung für A, da A somit eine $d \times 2^d$-Untermatrix mit paarweise verschiedenen Spalten besitzt. Folglich besitzt M_1 eine $(d+1) \times 2^{d+1}$ Untermatrix mit paarweise verschiedenen Spalten, nach der Konstruktion in (3.5).

- $r_1 \leq \Phi(n, d-1)$: Dann hat die Matrix (AB) allerdings mehr als $s_1 - r_1$ verschiedene Spalten. Es gilt somit:

  $$s_1 - r_1 > \Phi(n+1, d) - \Phi(n, d-1) \overset{\text{Pascalsche Identität}}{=} \Phi(n, d).$$

  Dabei ist mit der *Pascal'schen Identität* folgende Gleichheit gemeint, vgl. [Bru12], Kapitel 5.1:

  $$\binom{n}{d} + \binom{n}{d+1} = \binom{n+1}{d+1}. \tag{3.6}$$

  Somit haben wir hier durch die Induktionsvoraussetzung bereits eine $(d+1) \times 2^{d+1}$ Untermatrix von (AB) gefunden, also insbesondere von M_1.

Nun wollen wir die zweite Ungleichung aus der Formelzeile (3.4) beweisen. Diese folgt aus folgender äquivalenten Aussage:

$$\begin{aligned} \left(\frac{d}{n}\right)^d \sum_{i=0}^{d} \binom{n}{i} &\overset{\frac{d}{n}\leq 1}{\leq} \sum_{i=0}^{d} \binom{n}{i} \left(\frac{d}{n}\right)^i \\ &\leq \sum_{i=0}^{n} \binom{n}{i} \left(\frac{d}{n}\right)^i \overset{\text{Binomischer Lehrsatz}}{=} \left(1 + \frac{d}{n}\right)^n \\ &\leq e^d. \end{aligned}$$

□

In dem Rest des Kapitel geben wir einige für die Berechnung der VCD hilfreiche Lemmas an.

3.3 Hilfreiche Lemmas zur Berechnung

Zunächst fassen wir einige hilfreiche Resultate zusammen, welche bei der Bestimmung der VCD für verschiedene Mengen an Funktionen hilfreich sind. Dabei folgen wir [Son98] und [SSBD14], Kapitel 6. Wir betrachten vier Lemmas:

- In Lemma 3.8 betrachten wir einige elementare Eigenschaften der VC-Dimension.
- In Lemma 3.9 bemerken wir, dass sich die VCD einer Menge an Funktionen nicht ändert, falls man diese mit einer Menge an streng monotonen Funktionen, welche alle fallen oder alle steigen, verknüpft. Dies verallgemeinern wir in Korollar 3.11 auf alle monotonen Funktionen.
- In Lemma 3.10 betrachten wir, wie sich die VCD für die Vereinigung von Mengen ändert.
- In Lemma 3.12 berechnen wir die VCD eines Vektorraum an Funktionen.

In Unterkapitel 3.4 betrachten wir zusätzlich die VCD für Funktionen $\mathbb{R} \to \mathbb{R}$. Hierbei können wir die Bestimmung der VC-Dimension vereinfachen auf das Bestimmen des Monotonieverhaltens der betrachteten Funktionen.

Im folgenden Lemma findet sich der erste Punkt zum Beispiel in [SSBD14], Aufgabe 6.1 und der zweite Punkt in [Son98], Kapitel 3. Der dritte Punkt ist ein Spezialfall von Lemma 3.9.

Lemma 3.8. *Seien im Folgenden* $\{f\}, \mathcal{F}, \mathcal{F}_1, \mathcal{F}_2 \subset \mathbb{R}^Z$. *Sei weiterhin* $S \subset Z$ *mit* $\#S = n$ *und*

$$-\mathcal{F} := \{-f : f \in \mathcal{F}\}. \tag{3.7}$$

Dann gilt:

1. $\mathcal{F}_1 \subset \mathcal{F}_2 \implies VCD(\mathcal{F}_1) \leq VCD(\mathcal{F}_2)$.
2. *Für jede Menge* $S^+ \subset S$, *welche durch* (f, β) *induziert wird, existiert ein* $\tilde{\beta} \in \mathbb{R}$, *sodass* S^+ *auch durch* $(f, \tilde{\beta})$ *induziert, mit* $f(x) - \tilde{\beta} \neq 0$ $\forall x \in S$.
3. $\mathrm{VCD}(-\mathcal{F}) = \mathrm{VCD}(\mathcal{F})$

4. $\mathrm{VCD}(\{f\}) = 1$

Beweis. Die Aussage gilt, da alle Teilmengen einer Menge $S \subset \mathbb{R}^k$, welche durch eine Funktion $f \in \mathcal{F}_1$ induziert werden, auch durch dieselbe Funktion in $\mathcal{F}_2$ induziert werden.

Für den zweiten Punkt bemerken wir, dass S nur $n < \infty$ Elemente enthält. Für die Menge $S^+ \subset S$, welche durch (f, β) induziert wird, existiert also ein $\varepsilon > 0$, sodass gilt:

$$z \in S \backslash S^+ \Rightarrow f(z) - (\beta - \varepsilon) < 0.$$

Gleichzeitig gilt

$$z \in S^+ \rightarrow f(z) - \beta \geq 0 \Rightarrow f(z) - (\beta - \varepsilon) > 0$$

und damit ist die Behauptung mit $\tilde{\beta} := \beta - \varepsilon$ bewiesen.

Beweisen wir nun den dritten Punkt: Sei $S \subset Z$ eine Menge, welche durch $\mathcal{F}$ gesplittert wird. Somit existieren für jede Teilmenge $S^+ \subset S$ Funktionen $f^+, f^- \in \mathcal{F}$, welche die Mengen $S^+, S^- := S \backslash S^+$ induzieren. Nach dem zweiten Punkt existieren somit auch Werte $\beta^+, \beta^- \in \mathbb{R}$, sodass (f^+, β^+) die Menge S^+ und (f^-, β^-) die Menge S^- induzieren, mit:

$$f^+(x) - \beta^+ \neq 0, f^-(x) - \beta^- \neq 0, \qquad \forall x \in S.$$

Somit induziert $(-f^+, -\beta^+)$ die Menge S^- und $(-f^-, -\beta^-)$ die Menge S^+, mit $-f^+, -f^- \in -\mathcal{F}$.

Für den vierten Punkt betrachten wir $z_1 < z_2 \in Z$. Im Folgenden wollen wir die Menge $S = \{z_2\}$ durch $\{f\}$ splittern. Gilt $f(z_1) = f(z_2)$, so induziert mit $\beta = f(z_2)$ $(f, f(z_2))$ die Menge $S^+ = \{z_2\} \subset S$ und $(f, f(z_2) + \varepsilon)$ die leere Menge. Damit gilt $\mathrm{VCD}(\{f\}) = 1$. Ansonsten nehmen wir ohne Beschränkung der Allgemeinheit $f(z_1) < f(z_2)$ an. Damit lässt sich kein $\beta \in \mathbb{R}$ finden, sodass (f, β) die Menge $\{x_1\} \subset S := \{x_1, x_2\}$ induziert und es folgt $\mathrm{VCD}(\{f\}) \leq 1$. Analog zum Fall $f(x_1) = f(x_2)$ folgt hier auch $\mathrm{VCD}(\{f\}) = 1$. □

Das nächste Lemma wird durch [Gir95], Beweis zu Präposition 3.2, impliziert.

Lemma 3.9. *Sei $\mathcal{G} \subset \mathbb{R}^Z$. Weiterhin sei $\mathcal{F}$ eine Menge der streng monoton Funktionen aus $\mathbb{R}^{\mathbb{R}}$, welche alle dieselbe Monotonierichtung besitzen, also alle*

Funktionen aus $\mathcal{F}$ sind entweder streng monoton steigend oder streng monoton fallend. Wir bezeichnen:

$$\mathcal{F} \circ \mathcal{G} := \{f \circ g : f \in \mathcal{F}, g \in \mathcal{G}\}. \tag{3.8}$$

Dann gilt:

$$\mathrm{VCD}(\mathcal{F} \circ \mathcal{G}) = \mathrm{VCD}(\mathcal{G}).$$

Insbesondere folgt also für alle $a, b \in \mathbb{R}$ mit $a \neq 0$:

$$\mathrm{VCD}(a\mathcal{G} + b) = \mathrm{VCD}(\mathcal{G}), \qquad a\mathcal{G} + b := \{ag + b : g \in \mathcal{G}\}.$$

Sind die Funktionen aus $\mathcal{F}$ nur monoton statt streng monoton, so folgt:

$$\mathrm{VCD}(\mathcal{F} \circ \mathcal{G}) \leq \mathrm{VCD}(\mathcal{G}).$$

Beweis. Wir sehen, dass mit der Wahl $f \in \mathcal{F}, x \mapsto x$ die Einbettung $\mathcal{G} \subset \mathcal{F} \circ \mathcal{G}$ gilt und mit Lemma 3.8 folgt die Richtung

$$\mathrm{VCD}(\mathcal{G}) \leq \mathrm{VCD}(\mathcal{F} \circ \mathcal{G})$$

Für die Richtung $\mathrm{VCD}(\mathcal{F} \circ \mathcal{G}) \leq \mathrm{VCD}(\mathcal{G})$ nehmen wir also eine Menge $S \subset Z$ an, welche durch $\mathcal{F} \circ \mathcal{G}$ gesplittert wird. Zu jeder festen Teilmenge $S^+ \subset S$ existieren also Funktionen $f \in \mathcal{F}, g \in \mathcal{G}$ und $\beta \in \mathbb{R}$, sodass gilt:

$$f \circ g(z) \begin{cases} \geq \beta, & z \in S^+, \\ < \beta, & z \in S \backslash S^+. \end{cases}$$

Da f auf $\mathbb{R}$ streng monoton ist, existiert die Umkehrfunktion f^{-1} auf dem Bild $f(\mathbb{R})$, mit $f^{-1}(f(x)) = x$, $\forall x \in \mathbb{R}$. Ist $f(x)$ streng monoton steigend, so auch $f^{-1}(x)$ und es gilt:

$$f^{-1} \circ f \circ g(z) \begin{cases} \geq f^{-1}(\beta), & z \in S^+, \\ < f^{-1}(\beta), & z \in S \backslash S^+. \end{cases}$$

In diesem Fall wird S^+ also bereits durch g induziert. Ist umgekehrt f streng monoton fallend, so ist auch f^{-1} streng monoton fallend. Dann können wir den zweiten Punkt aus Lemma 3.8 anwenden um nur noch strikte Ungleichungen zu betrachten. Damit induziert g die Menge $S \backslash S^+$. Insgesamt wird dadurch in beiden Fällen S auch durch $\mathcal{G}$ gesplittert.

Der zweite Punkt folgt direkt aus dem ersten Punkt und der dritte Punkt folgt direkt aus Lemma 3.8. □

Im nächsten Punkt nutzen wir das Lemma von Sauer, Satz 3.7, um die VCD der Vereinigung von Mengen abzuschätzen. Dies findet sich auch in [SSBD14], Aufgabe 6.11.

Lemma 3.10. *Seien $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_r$ Mengen aus $\mathbb{R}^{\mathbb{R}^k}$, wofür die VCD kleiner gleich $d \in \mathbb{N}$ sei. Dann gilt:*

1. *VCD$(\mathcal{F}_1 \cup \mathcal{F}_2) \leq 2d + 1$.*
2. *VCD$(\cup_{i=1}^{r} \mathcal{F}_i) \leq \max\{4d\log(2d), 2\log(r)\}$, für $d \geq 3$.*

Beweis. Für den ersten Punkt betrachten wir Umformungen aus dem Sauer-Shelah-Lemma, vgl. Satz 3.7, mit Hilfe der Wachstumsfunktion aus Definition 3.2. Dabei bemerken wir, dass $\mathscr{G}_{\mathcal{F}_1 \cup \mathcal{F}_2}(n) \leq \mathscr{G}_{\mathcal{F}_1}(n) + \mathscr{G}_{\mathcal{F}_2}(n)$ gilt, da jede Funktion aus $\mathcal{F}_1 \cup \mathcal{F}_2$, welche eine Teilmenge einer Menge S mit n Punkten aus $\mathbb{R}^k$ induziert wird, in $\mathcal{F}_1$ oder $\mathcal{F}_2$ liegen muss. Somit wird diese Teilmenge von S auch durch eine Funktion aus $\mathcal{F}_1$ oder $\mathcal{F}_2$ induziert.

Sei also $\mathscr{G}_{\mathcal{F}_1}(2d+1)$ die Wachstumsfunktion von $\mathcal{F}_1$ bezüglich $2d+1$ Punkten aus $\mathbb{R}^k$. Wir zeigen unter der Verwendung von Satz 3.7 und der Pascal'sche Identität (3.6), dass für $l \geq 2$ keine $2d+l$ Punkte durch $\mathcal{F}_1 \cup \mathcal{F}_2$ gesplittert werden. Damit folgt dann $\mathrm{VCD}(\mathcal{F}_1 \cup \mathcal{F}_2) \leq 2d+1$.

$$\begin{aligned}
\mathscr{G}_{\mathcal{F}_1}(2d+l) &\leq \sum_{i=0}^{d} \binom{2d+l}{i} \\
&= 1 + \sum_{i=1}^{d} \binom{2d+l-1}{i} + \binom{2d+l-1}{i-1} \\
&= \sum_{i=0}^{d} \binom{2d+l-1}{i} + \sum_{i=d+l}^{2d+l-1} \binom{2d+l-1}{i} \\
&< \sum_{i=0}^{2d+l-1} \binom{2d+l-1}{i} \\
&= 2^{2d+l-1}.
\end{aligned}$$

Damit folgt

$$\mathscr{G}_{\mathcal{F}_1 \cup \mathcal{F}_2}(2d+l) \leq \mathscr{G}_{\mathcal{F}_1}(2d+l) + \mathscr{G}_{\mathcal{F}_2}(2d+l) < 2 \cdot 2^{2d+l-1} = 2^{2d+l}$$

und damit die Behauptung, dass keine Menge mit $2d+2$ Elementen durch $\mathcal{F}_1 \cup \mathcal{F}_2$ gesplittert wird.

Für den zweiten Punkt definieren wir für jedes $\mathcal{F}_i, 1 \leq i \leq r$, die Mengen:

$$\mathcal{C}_{\mathcal{F}_i} := \{\mathcal{H}(f(\cdot) - \beta) : \mathbb{R}^k \to \{0,1\} |\ f \in \mathcal{F}_i, \beta \in [A,B]\}.$$

Mit dem Lemma von Sauer, Satz 3.7, und der Definition der Wachstumsfunktion für reellwertige Funktionen aus Definition 3.2 erhalten wir für $n > d$:

$$\mathscr{G}_{\mathcal{F}_i}(n) \leq \left(\frac{en}{d}\right)^d.$$

Insgesamt gilt also für die Wachstumsfunktion $\mathscr{G}(n)$ von $\cup_{i=1}^r \mathcal{F}_i$:

$$\mathscr{G}_{\cup_{i=1}^r \mathcal{F}_i}(n) \leq r \cdot \left(\frac{en}{d}\right)^d.$$

Falls m Punkte durch $\cup_{i=1}^r \mathcal{F}_i$ gesplittert werden, so muss gelten:

$$2^m \leq r \left(\frac{em}{d}\right)^d.$$

Wir suchen also das größtmögliche m, welches die obige Ungleichung erfüllt. Dabei gilt für $d \geq 3$:

$$m \leq \log(r) + d \log(m).$$

Somit gilt einer der folgenden beiden Fälle:

- $m \leq 2d \log(m)$,
- $m \leq 2 \log(r)$.

Für den ersten Fall können wir die folgende Ungleichung benutzen:

$$m \leq 2d \log(m) \Rightarrow m \leq 4d \log(2d).$$

Dies leitet man her, indem man die Funktion

$$f(m) = 2d \log(m) - m$$

betrachtet. Die erste Ableitung hiervon ist $\frac{2d}{m} - 1$ und kleiner als 0, falls $m > 2d$ gilt. Um eine möglichst gute Grenze für m zu erhalten, suchen wir also eine bestmögliche obere Schranke für die Nullstelle von $f(m)$, welche größer als $2d$ ist. Diese existiert, da $f(m) \to -\infty$ divergiert für $m \to \infty$.

Wir behaupten, dass $m = 4d\log(2d)$ eine solche obere Schranke ist. Nachrechnen ergibt:

$$\begin{aligned} f(4d\log(2d)) &= 2d\log(4d\log(2d)) - 4d\log(2d) \\ &= 2d(\log(4d\log(2d)) - 2\log(2d)) \\ &= 2d\left(\log\left(\frac{4d\log(2d)}{(2d)^2}\right)\right) < 0. \end{aligned}$$

Damit ist die Behauptung bewiesen. □

Damit können wir das folgende Korollar herleiten, welches das Lemma 3.9 auf alle Monotonen Funktionen erweitert.

Korollar 3.11. *Sei $\mathcal{G} \subset \mathbb{R}^Z$ und $\mathcal{F}$ die Menge der streng monotonen Funktionen aus $\mathbb{R}^{\mathbb{R}}$. Mit der Notation aus (3.8) folgt:*

$$\mathrm{VCD}(\mathcal{F} \circ \mathcal{G}) = 2\,\mathrm{VCD}(\mathcal{G}) + 1.$$

Sind die Funktionen aus $\mathcal{F}$ nur monoton statt streng monoton, so folgt:

$$\mathrm{VCD}(\mathcal{F} \circ \mathcal{G}) \leq 2\,\mathrm{VCD}(\mathcal{G}) + 1.$$

Beweis. Sei $\mathcal{F}_s$ die Menge der (streng) monoton steigenden Funktionen auf $\mathbb{R}^{\mathbb{R}}$ und $\mathcal{F}_f$ die Menge der streng monoton fallenden Funktionen auf $\mathbb{R}^{\mathbb{R}}$, so folgt:

$$\mathcal{F} \circ \mathcal{G} = \mathcal{F}_s \circ \mathcal{G} \cup \mathcal{F}_f \circ \mathcal{G}.$$

Da zusätzlich aus Lemma 3.9 folgt, dass

$$\mathrm{VCD}(\mathcal{F}_s \circ \mathcal{G}) \leq \mathrm{VCD}(\mathcal{G}), \quad \mathrm{VCD}(\mathcal{F}_f \circ \mathcal{G}) \leq \mathrm{VCD}(\mathcal{G}),$$

so folgt mit Lemma 3.10 die Behauptung. □

Im nächsten Lemma berechnen wir die VCD für den Span aus endlich vielen Funktionen. Dabei folgen wir [Son98], Satz 1.

Lemma 3.12. *Sei die Menge $\mathcal{F} \subset \mathbb{R}^Z$ ein endlich-dimensionaler Vektorraum und $\mathcal{F}_\beta = \mathcal{F} + \mathrm{Span}(1)$. Dann gilt:*

$$\mathrm{VCD}(\mathcal{F}) = \dim(\mathcal{F}_\beta).$$

Beweis. Wir nehmen an, dass $\mathrm{VCD}(\mathcal{F}) = n$ gilt. Sei $S = \{z_1, \ldots, z_n\} \subset Z$ eine Menge, welche durch $\mathcal{F}$ gesplittert wird. Für die Richtung $\mathrm{VCD}(\mathcal{F}) \leq \dim(\mathcal{F}_\beta)$ konstruieren wir die $2^n \times n$ Matrix

$$F = (f_j(z_i) - \beta_j)_{1 \leq i \leq n, 1 \leq j \leq 2^n}$$

mit paarweise verschiedenen Zeilen. Diese existiert, da $\mathrm{VCD}(\mathcal{F}) = n$. Nach Lemma 3.8 existiert eine solche Matrix sogar ohne einen Nulleintrag, insbesondere also ohne einer Nullspalte. Wir wollen zeigen, dass die Matrix F mindestens Rang n besitzt. Sei dafür $0 \neq \nu \in \mathbb{R}^n$ beliebig. Dann existiert eine Zeile aus F, sodass:

$$\operatorname{sign}(F)_{i,j} = \operatorname{sign}(\nu_i), \qquad \forall 1 \leq i \leq n,$$

mit $\operatorname{sign}(0) = 0$. Somit gilt insbesondere $F\nu \neq 0$ und es folgt:

$$F\nu = 0 \implies \nu = 0.$$

Somit ist auch der Rang von F mindestens n und damit auch der Rang von dem Vektorraum $\mathcal{F}_\beta$.

Für die andere Richtung betrachten wir den Vektorraum $\mathcal{F}_\beta$ und $S^+ \subset S = \{z_1, \ldots, z_n\} \subset X$. Wir wählen Funktionen $f_1 - \beta_1, \ldots, f_n - \beta_n \in \mathcal{F}_\beta$, sodass die Vektoren $(f_i(z_1) - \beta_i, \ldots, f_i(z_n) - \beta_i)^T$ für $1 \leq i \leq n$ linear unabhängig sind. Diese bilden die invertierbare Matrix

$$F = (f_i(z_j) - \beta_i)_{1 \leq i,j \leq n},$$

Sei ε der Vektor aus $\{-1, 1\}^n$, welcher in der i-ten Komponente genau dann 1 ist, falls $z_i \in S^+$ gilt, für $1 \leq i \leq n$. Dann findet sich ein $\nu = F^{-1}\varepsilon \in \mathbb{R}^n$, sodass $F\nu = \varepsilon$. Damit gilt für alle $z \in S$, für alle $1 \leq j \leq n$:

$$\mathcal{H}\left(\sum_{i=1}^{n} \nu_i (f_i(z) - \beta_i)\right) = 1 \Leftrightarrow z \in S^+$$

und die Behauptung ist bewiesen. □

3.4 Die VCD für Funktionen aus $\mathbb{R}^\mathbb{R}$

Für Funktionen von $\mathbb{R}$ nach $\mathbb{R}$ vereinfacht sich die Berechnung der VCD etwas. Sei dafür $S = \{x_1 < x_2 < \ldots < x_n\} \subset \mathbb{R}$ und $S^+ = \{x_2, x_4, \ldots\} \subset S$.

Damit eine stetige Funktion $f - \beta$, $f \in \mathcal{F}, \beta \in \mathbb{R}$ die Menge S^+ induziert, müsste $f - \beta$ zwischen jeweils zwei nebeneinanderliegenden Punkten eine Nullstelle besitzen. Um die Menge S^+ zu splittern, bräuchte $f - \beta$ demnach mindestens $n - 1$ Nullstellen. Dementsprechend betrachten wir im nächsten Lemma 3.13 eine Menge an Funktionen $\mathcal{F}$, in welcher jede Funktion $f \in \mathcal{F}$ nur endlich viele Vorzeichenwechsel besitzt.

Wir sagen, dass $f : \mathbb{R} \to \mathbb{R}$ einen Vorzeichenwechsel an der Stelle x_0 besitzt, falls gilt:

$$\begin{aligned}\exists \tilde{\varepsilon} > 0 \, \forall \varepsilon \in (0, \tilde{\varepsilon}) : &\operatorname{sign}(f(x_0 + \varepsilon)) \neq \operatorname{sign}(f(x_0 - \varepsilon)) \\ &\vee \operatorname{sign}(f(x_0 + \varepsilon)) \neq \operatorname{sign}(f(x_0)).\end{aligned}$$

Falls für ein $x_0 \in \mathbb{R}$ kein $\tilde{\varepsilon} > 0$ existert, sodass das Vorzeichen für alle Werte in $(x_0, x_0 + \varepsilon)$ gleich ist, so sagen wir, dass die Funktion f unendlich viele Stellen mit Vorzeichenwechsel besitzt.

Lemma 3.13. *Sei $\mathcal{F}$ die Menge der Funktionen aus $\mathbb{R}^{\mathbb{R}}$, für welche die Funktionen $f - \beta$ für alle $f \in \mathcal{F}, \beta \in \mathbb{R}$ an höchstens p Stellen Vorzeichenwechsel (*VZW*) besitzen, also*

$$\mathcal{F} = \{f : \mathbb{R} \to \mathbb{R}, f(x) - \beta \text{ besitzt höchstens } p \, \text{VZW} \, \forall \beta \in \mathbb{R}\}.$$

Dann gilt:

$$\operatorname{VCD}(\mathcal{F}) = p + 1.$$

Beweis. Wir zeigen zuerst eine Konstruktion, aus welcher $\operatorname{VCD}(\mathcal{F}) \geq p + 1$ folgt. Danach zeigen wir, dass $\mathcal{F}$ keine Menge mit $p + 2$ Punkten splittert, woraus $\operatorname{VCD}(\mathcal{F}) = p + 1$ folgt. Hierzu bezeichnen wir mit S eine beliebige Menge mit Punkten, welche wir mit $x_1 < \ldots < x_{p+1}$ bezeichnen. Dann existiert ein $0 < \varepsilon$, sodass für alle $2 \leq i \leq p+1$ gilt: $\varepsilon < x_i - x_{i-1}$. Dann wird eine beliebige Teilmenge $S^+ \subset S$ durch die folgende Funktion f induziert:

$$f(x) := \sum_{x_i \in S \setminus S^+} -I|_{[x_i, x_i + \varepsilon)}(x) + 1.$$

Diese Summe induziert offensichtlich die Menge S^+. Betrachten wir nun die Menge $\widetilde{S} = S \cup \{x_{p+2}\}$, mit $x_{p+1} < x_{p+2}$. Sei $S^+ = \{x_2, x_4, x_6, \ldots\} \subset \widetilde{S}$ die Menge der Elemente mit geradem Index aus $\widetilde{S}$. Diese Menge wird von keiner Funktion f aus $\mathcal{F}$ induziert, da hierzu zwischen allen benachbarten der $p+2$

Indizes die Funktion $f - \beta$, $\beta \in \mathbb{R}$ einen Vorzeichenwechsel besitzen muss, also insgesamt $p+1$ Vorzeichenwechsel. Wir hatten aber vorausgesetzt, dass die Funktion $f - \beta$ höchstens p Vorzeichenwechsel haben darf. □

Korollar 3.14. *Sei $\mathcal{F} \subset \mathbb{R}^Z$ eine Menge an Funktionen. Wir fordern, dass für jede Funktion $f \in \mathcal{F}$ höchstens p Intervalle existieren, auf denen die Funktion f jeweils monoton ist, also analog zur Formelzeile (0.3):*

$$f(x) = \sum_{i=1}^{p} \phi|_{\mathcal{I}_i}(x), \qquad \mathcal{I}_1 \cup \ldots \cup \mathcal{I}_p = \mathbb{R}, \quad \mathcal{I}_i \cap \mathcal{I}_j = \emptyset,$$
$$f \text{ auf } \mathcal{I}_i \text{ monoton}, \quad 1 \leq i, j \leq p.$$

Dann gilt:

$$\mathrm{VCD}(\mathcal{F}) \leq p + 1.$$

Beweis. Da jede Funktion in jedem Intervall, auf welchem sie monoton ist, höchstens einen Vorzeichenwechsel besitzen kann, folgt die Behauptung direkt mit Lemma 3.13. □

Sei im Folgenden

$$\pi_p = \mathrm{Span}\,\{1, \ldots, x^p\} \tag{3.9}$$

der Raum der Polynome mit Grad höchstens p.

Lemma 3.15. *Sei $\mathcal{F}_{p,q}$ die Menge der rationalen Funktionen auf $\mathbb{R}$, also:*

$$\mathcal{F}_{p,q} := \left\{ \frac{f}{g} \mid f \in \pi_p, g \in \pi_q \right\}.$$

Dann gilt:

$$\mathrm{VCD}(\mathcal{F}_{p,q}) \leq p + q + 1.$$

Beweis. Wir wollen Lemma 3.13 anwenden. Damit eine Funktion $\frac{f}{g}$ aus $\mathcal{F}_{p,q}$ einen Vorzeichenwechsel hat, müsste f oder g eine Nullstelle haben. Insgesamt haben f und g höchstens $p + q$ Nullstellen, also gilt die Behauptung mit Lemma 3.13. □

Für den Raum π_p folgt mit Lemma 3.15:

$$\mathrm{VCD}(\pi_p) = p + 1.$$

Diese Aussage befindet sich auch in [Son98] und wurde dort mit Hilfe von Lemma 3.12 aus dieser Arbeit bewiesen. In [Son98] wurde allerdings eine andere Definition der VC-Dimension für reellwertige Funktionen genutzt, dort wurden die affinen Verschiebungen um β, also $f - \beta$, nicht betrachtet, sondern nur die Mengen $\mathcal{H}(f(x))$. In dem Fall der Polynomräume π_p fallen die Definitionen allerdings zusammen.

4 Die VC-Schranke

Insgesamt können wir mit den bisherigen Resultaten den folgenden Satz beweisen. Dabei orientieren wir uns an [vLS11], Kapitel 5 und [BBL03], Satz 2.

Satz 4.1. *Sei wieder $\mathcal{F} \subset [A,B]^Z$, $\mathrm{VCD}(\mathcal{F}) = d$ und seien $z_1, \ldots, z_n$ Punkte aus Z, unabhängig und identisch verteilt gemäß eines Wahrscheinlichkeitsmaßes P auf Z. Beschreibe hier wieder $\widehat{R}(f)$ das empirische Risiko $\frac{1}{n}\sum_{i=1}^n f(z_i)$ und $R(f)$ das Risiko $\int_Z f(z)dP(z)$. Für ein $\delta \in (0,1]$ definieren wir:*

$$\alpha := (B-A)\sqrt{8\frac{d\ln\left(\frac{2en}{d}\right) - \ln\left(\frac{\delta}{4}\right)}{n}}.$$

Es soll dabei $n\alpha^2(B-A)^{-2} \geq 2$ gelten, also:

$$8\left(d\ln\left(\frac{2en}{d}\right) - \ln\left(\frac{4}{\delta}\right)\right) \geq 2. \tag{4.1}$$

Dann gilt mit einer Wahrscheinlichkeit von mindestens $(1-\delta)$:

$$\sup_{f\in\mathcal{F}} |R(f) - \widehat{R}(f)| \leq \alpha. \tag{4.2}$$

Beweis. Für den Beweis betrachten wir die Schranke (4.2) wieder in der Form der Schranke (2.7). Wir wollen dann die Sätze 2.3, 2.4 und 3.7 anwenden. Wir verwenden wieder die Notation

$$\mathcal{F}_{\mathcal{H}} := \{\mathcal{H} \circ (f(\cdot) - \beta) : f \in \mathcal{F}, \beta \in \mathbb{R}\} \subset \{0,1\}^{X\times Y}.$$

und die verwendete Notation aus Formelzeile (2.2). Als ersten Beweisschritt wenden wir Satz 2.3 an und erhalten

$$\begin{aligned}
&P\left(\sup_{f\in\mathcal{F}}\left|R(f) - \widehat{R}(f)\right| \geq \alpha\right)\\
=&P\left(z_1,\ldots,z_n \in Z : \sup_{f\in\mathcal{F}}\left|\int_Z f(z)dP(z) - \frac{1}{n}\sum_{i=1}^n f(z_i)\right| \geq \alpha\right)\\
\leq&P\left(\sup_{f\in\mathcal{F}_{\mathcal{H}}}\left|\int_Z f(z)dP(z) - \frac{1}{n}\sum_{i=1}^n f(z_i)\right| \geq \frac{\alpha}{B-A}\right).
\end{aligned}$$

Seien $z_{n+1}, \ldots, z_{2n}$ weitere n Punkte, sodass $z_1, \ldots, z_{2n}$ unabhängig und identisch gemäß dem Wahrscheinlichkeitsmaß P verteilt sind. Für Satz 2.4 benötigen wir die Voraussetzung $n\left(\frac{\alpha}{B-A}\right)^2 \geq 2$, gegeben durch (4.1).

$$\begin{aligned} & P\left(\sup_{f\in\mathcal{F}_{\mathcal{H}}}\left|\int_Z f(z)dP(z) - \frac{1}{n}\sum_{i=1}^{n} f(z_i)\right| \geq \frac{\alpha}{B-A}\right) \\ \leq & 2P\left(\sup_{f\in\mathcal{F}_{\mathcal{H}}}\left|\frac{1}{n}\sum_{i=1}^{n} f(z_{i+n}) - \frac{1}{n}\sum_{i=1}^{n} f(z_i)\right| \geq \frac{\alpha}{2(B-A)}\right). \end{aligned}$$

Für die Berechnung hierfür sind nur endlich viele Funktionen aus $\mathcal{F}_{\mathcal{H}}$ relevant, wobei wir eine obere Grenze hierfür durch die Wachstumsfunktion $\mathscr{G}_{\mathcal{F}_{\mathcal{H}}}(2n)$ angeben können. Die Funktionen $f(z_{i+n}) - f(z_i)$ nehmen dabei Werte in $[-1, 1]$ an. Somit können wir jetzt mit Hilfe der Hoeffding-Ungleichung aus Satz 2.1 und der Booleschen Ungleichung (2.12) folgern:

$$\leq 2 \cdot 2\mathscr{G}_{\mathcal{F}_{\mathcal{H}}}(2n) \exp\left(-2n\left(\frac{\alpha}{2(B-A)}\right)^2 (1-(-1))^{-2}\right).$$

Durch die Beschränkung der Wachstumsfunktion durch ein Polynom mit Grad der VC-Dimension d von $\mathcal{F}_{\mathcal{H}}$ aus Satz 3.7 erhält man:

$$\leq 4\left(\frac{2en}{d}\right)^d \exp\left(-n\frac{\alpha^2}{8}(B-A)^{-2}\right).$$

Wir erhalten die Behauptung analog zu den Umformungen aus den Zeilen (2.8)-(2.10):

$$\begin{aligned} \delta &= 4\left(\frac{2en}{d}\right)^d \exp\left(-n\frac{\alpha^2}{8}(B-A)^{-2}\right) \\ \ln\left(\frac{\delta}{4}\right) &= d\ln\left(\frac{2en}{d}\right) - (B-A)^{-2}\frac{n\alpha^2}{8} \\ \Longrightarrow \qquad \alpha &= (B-A)\sqrt{8\frac{d\ln\left(\frac{2en}{d}\right) - \ln\left(\frac{\delta}{4}\right)}{n}}. \end{aligned}$$

Somit finden wir, analog zur Formelzeile (2.11), für alle $\delta \in (0, 1]$ ein $\alpha \in \mathbb{R}$, sodass die folgende Ungleichung gilt:

$$P\left(z_1, \ldots z_n \in Z : \sup_{f\in\mathcal{F}}\left|R(f) - \widehat{R}(f)\right| \leq \alpha\right) \geq 1 - \delta.$$

□

5 Die VC-Dimension in der Approximationstheorie

Für dieses Kapitel folgen wir [Gir95] und [Vap98], Kapitel 6.5.4.

Die Ungleichung (4.2) gibt für jede Funktion f einer Menge $\mathcal{F}$ die Konvergenz des empirischen Risikos zum Risiko an. Diese gleichmäßige Konvergenz wollen wir im Kontext der Approximationstheorie betrachten. Dabei berechnen wir nicht das Integral über eine Verlustfunktion, sondern allgemein ein Integral über eine Funktion gegen eine Wahrscheinlichkeitsdichte.

Satz 5.1. *Seien im Folgenden $Z, T \subset \mathbb{R}^k$ und $\mathcal{K} \subset \mathbb{R}^{Z \times T}$ eine Menge an messbaren Funktionen. Sei weiterhin*

$$\mathcal{F} := \{K(\cdot, t) : Z \to \mathbb{R}, t \in T, K \in \mathcal{K}\},$$

wobei gelten soll:

$$A \leq f(z) \leq B, \quad A, B \in \mathbb{R}, \forall z \in Z, f \in \mathcal{F}, \quad d = \mathrm{VCD}(\mathcal{F}).$$

Sei weiterhin

$$\alpha_\delta := \|\lambda\|_1 (B - A) \sqrt{8 \frac{d \ln\left(\frac{2en}{d}\right) + \ln\left(\frac{4}{\delta}\right)}{n}}.$$

Dann existieren für jedes $\delta \in (0, 1]$, mit

$$n \alpha_\delta^2 (B - A)^{-2} \geq 2,$$

analog zu Formelzeile (4.1) und jedes nicht-negative $\lambda \in \mathcal{L}_1(\mathbb{R}^k), \|\lambda\|_1 \neq 0$, Punkte $\tilde{z}_1, \ldots, \tilde{z}_n \in Z$, sodass gilt:

$$\sup_{K(\cdot,t) \in \mathcal{F}} \left| \int_Z K(z, t) \lambda(z) dz - \frac{1}{n} \sum_{i=1}^{n} \|\lambda\|_1 K(\tilde{z}_i, t) \right| \leq \alpha_\delta.$$

Mit $\delta \to 1$ erhalten wir damit dann Punkte $\tilde{z}_1, \ldots, \tilde{z}_n$ auch für die Schranke mit α_1.

Beweis. Wir wollen hierfür Satz 4.1 anwenden. Seien $z_1, \ldots, z_n$ die beobachteten Punkte aus Satz 4.1. Dann gilt nach Ungleichung (4.2) mit einer Wahrscheinlichkeit von mindestens $(1 - \delta)$:

$$\sup_{t \in T, K \in \mathcal{K}} \left| \int_Z K(z, t) P(z) dz - \frac{1}{n} \sum_{i=1}^{n} K(z_i, t) \right| \leq \alpha_\delta, \tag{5.1}$$

mit:

$$\alpha_\delta := (B-A)\sqrt{8\frac{d\ln\left(\frac{2en}{d}\right)+\ln\left(\frac{4}{\delta}\right)}{n}}.$$

Betrachten wir als nächstes das Wahrscheinlichkeitsmaß P auf Z. Für Satz 4.1 war dies beliebig, aber fest. Im Folgenden betrachten wir Wahrscheinlichkeitsmaße auf Z, welche über Funktionen aus $\mathcal{L}_1(Z)$ definiert werden:

Falls $\lambda \in \mathcal{L}_1(Z)$ mit $\|\lambda\|_1 \neq 0$ und $\lambda(z) > 0$ außer auf einer Lebesgue-Nullmenge, so kann man über λ eine Wahrscheinlichkeitsdichtefunktion und somit ein Wahrscheinlichkeitsmaß auf Z beschreiben:

$$P(z) := \frac{\lambda(z)}{\|\lambda\|_1}.$$

Multiplizieren wir alle Funktionen aus $\mathcal{K}$ mit der Konstanten $\|\lambda\|_1$, erhalten wir somit:

$$\begin{aligned}
&\int_Z \|\lambda\|_1 K(z,t)P(z)dz \\
=&\int_Z \|\lambda\|_1 K(z,t)\frac{\lambda(z)}{\|\lambda\|_1}dz \\
=&\int_Z K(z,t)\lambda(z)dz.
\end{aligned}$$

Folglich benötigen für die Schranke aus Satz 4.1 die VC-Dimension der Menge

$$\{\|\lambda\|_1 K(\cdot,t) : Z \to \mathbb{R} | t \in T, K \in \mathcal{K}\}.$$

Da $\|\lambda\|_1$ allerdings konstant und ungleich 0 ist, können wir mit Lemma 3.9 auch die VC-Dimension der Menge $\mathcal{F} = \{K(\cdot,t) : Z \to \mathbb{R} | t \in T\}$ betrachten. Diese ist nach Voraussetzung d. Die Funktionen $\|\lambda\|_1 K(\cdot,t)$ bilden dabei nicht auf $[A,B]$ ab, sondern auf $[\|\lambda\|_1 A, \|\lambda\|_1 B]$. Dies erklärt den Faktor $\|\lambda\|_1$ in der Definition von α_δ.

□

6 Die Rademacher-Komplexität

Ziel dieses Kapitels ist es, eine schärfere Schranke als die VC-Schranke aus den Kapiteln 4 und 5 herzuleiten. Dafür wollen wir die Verwendung der Booleschen Ungleichung (2.12) umgehen und zusätzlich das Wahrscheinlichkeitsmaß P einbeziehen. Hierfür führen wir ein neues Komplexitätsmaß ein, die *Rademacher-Komplexität*. Wie wir im Unterkapitel 6.5 sehen werden, lässt sich die Rademacher-Komplexität von unten gegen die VC-Dimension abschätzen. Dabei ist die resultierende Schranke schärfer als die VC-Schranke aus Kapitel 4. Um dies zu zeigen und einzuführen folgen wir [Men03]. Weitere Einführungen zur Rademacher-Komplexität finden sich in [BBL03], [BM02], [MRT12] oder [SSBD14]. Zuerst wurde die Rademacher-Komplexität für die Lerntheorie in [Kol01] benutzt.

Wir wollen kurz die verwendete Notation wiederholen. Wir betrachten eine Menge $Z \subset \mathbb{R}^k$, eine endliche Teilmenge $S \subset Z$ und $\mathcal{F} \subset \{f : Z \to \mathbb{R}\}$. Mit der Heaviside-Funktion $\mathcal{H} : \mathbb{R} \to \{0, 1\}$ aus Formelzeile (0.5) schreiben wir

$$\mathcal{F}_{\mathcal{H}} := \{\mathcal{H}(f(\cdot) - \beta), \beta \in \mathbb{R}, f \in \mathcal{F}\}.$$

Elemente aus $\mathcal{F}_{\mathcal{H}}$ bezeichnen wir meist mit $f_{\mathcal{H}}$. Für eine gegebene Wahrscheinlichkeit P auf Z und Punkten $z_1 \ldots, z_n \in Z$ sowie einer gegebenen Funktion $f \in \mathcal{F}$ schreiben wir für das Risiko $R(f)$ und das empirische Risiko $\widehat{R}(f)$:

$$R(f) = \int_Z f(z) dP(z), \qquad \widehat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} f(z_i).$$

6.1 Der bedingte Erwartungswert

Da wir auch oft mit dem bedingten Erwartungswert rechnen, wiederholen wir auch kurz die Definition des bedingten Erwartungswertes für reelle Zufallsvariablen und wichtige Rechenregeln hierfür. Dabei folgen wir [SC08], Theorem A.4.5 und Lemma A.4.6. Weitere Referenzen sind [Bil95], Kapitel 34 und [Kle13], Kapitel 8.2.

Definition 6.1. *Sei $(Z, \mathcal{B}, P)$ ein Wahrscheinlichkeitsraum, $\xi : Z \to \mathbb{R}$ eine Zufallsvariable, welche P-integrierbar oder nicht-negativ ist und $\mathcal{A} \subset \mathcal{B}$ eine σ-Unteralgebra. Dann existiert eine P-fast sicher eindeutige, $\mathcal{A}$-messbare*

Funktion $\eta : Z \to \mathbb{R}$, sodass gilt:

$$\int_A \eta dP = \int_A \xi dP, \qquad \forall A \in \mathcal{A}.$$

Dann nennen wir η den bedingten Erwartungswert von ξ und schreiben

$$\eta = E[\xi|\mathcal{A}] = E_P[\xi|\mathcal{A}].$$

Lemma 6.2. *Gegeben sei eine Zufallsvariable ξ aus einem Wahrscheinlichkeitsraum $(Z, \mathcal{B}, P)$, deren Erwartungswert existiert, und eine σ-Unteralgebra $\mathcal{A} \subset \mathcal{B}$.*

1. *Ist $\sigma(\xi)$ unabhängig von $\mathcal{A}$, so gilt: $E[\xi|\mathcal{A}] = E[\xi]$.*

   Dabei meinen wir, dass die σ-Algebren $\sigma(\xi) = \{\xi^{-1}(B), B \in \mathcal{B}_\mathbb{R}\}$ und $\mathcal{A}$ unabhängig sind. Dabei ist $\mathcal{B}_\mathbb{R}$ die Borel-σ-Algebra von $\mathbb{R}$.

2. *Ist ξ $\mathcal{A}$-messbar, so gilt: $E[\xi|\mathcal{A}] = \xi$.*

3. *Es gilt die Turmeigenschaft, also für $\mathcal{A}_1 \subset \mathcal{A}_2 \subset \mathcal{A}$ haben wir:*

   $$E[E[\xi|\mathcal{A}_2]|\mathcal{A}_1] = E[\xi|\mathcal{A}_1] = E[E[\xi|\mathcal{A}_1]|\mathcal{A}_2].$$

4. *Es gilt die Jensensche Ungleichung, also für $f : \mathbb{R} \to \mathbb{R}$ konvex erhalten wir:*

   $$f(E[\xi|\mathcal{A}]) \leq E[f(\xi)|\mathcal{A}]. \tag{6.1}$$

Der letzte Punkt gilt insbesondere auch für den Erwartungswert, also $\mathcal{A} = \mathcal{B}$. Für den Beweis verweisen wir dabei auf Kapitel 8.2 in [Kle13].

6.2 Einführung und grundlegende Umformungen

Hier folgen wir [GS08] und [Men03], Kapitel 2.3. Zuerst geben wir die Definition der Rademacher-Komplexität an, nach [Men03], Definition 2.22. Dann wollen wir Konvergenzgrenzen ähnlich zu der VC-Schranke in Satz 4.1 mithilfe der Rademacher-Komplexität angeben.

Mit der Rademacher-Komplexität wollen wir, analog zur VC-Dimension, festhalten, wie gut für $Y \subset \mathbb{R}$ Funktionen einer Menge $\mathcal{F} \subset Y^Z$ auf festen

n Punkten $S = \{z_1, \ldots, z_n\} \subset Z$ verschiedene Werte aus Y^n darstellen können. Hierfür wählen wir mit $Y = \{-1, 1\}$ zufällige Werte $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \in \{-1, 1\}^n$ und zählen, wie viele der resultierenden Paare (z_i, ε_i) durch eine Funktion f als Paar $(z_i, f(z_i))$ dargestellt werden können, wie folgt:

$$\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \varepsilon_i f(z_i).$$

Diese Summe nimmt dabei Werte in $[-n, n]$ an. Wir nehmen an, dass die Werte aus ε zufällig gewählt und auf $\{-1, 1\}$ gleichverteilt sind. Um dies etwas besser für unsere Schranken $\left|\int_Z f(z)dP(z) - \frac{1}{n}\sum_{i=1}^{n} f(z_i)\right|$ verwenden zu können, betrachten wir auch den Betrag der Summe, also:

$$\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \varepsilon_i f(z_i) \right| = \sup_{f \in \mathcal{F} \cup -\mathcal{F}} \sum_{i=1}^{n} \varepsilon_i f(z_i). \tag{6.2}$$

Mit dem Erwartungswert über alle möglichen Werte $\varepsilon \in \{-1, 1\}^n$ erhalten wir eine genauere Vorstellung über die Komplexität von $\mathcal{F}$:

$$\mathcal{R}_{\mathcal{F}} := E_\varepsilon \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \varepsilon_i f(z_i) \right| = \frac{1}{2^n} \sum_{\varepsilon \in \{-1,1\}^n} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \varepsilon_i f(z_i) \right|. \tag{6.3}$$

Dabei ist der Wert $\mathcal{R}_{\mathcal{F}}$ in $[0, n]$. Seien hierbei $\tilde{\mathcal{F}} \subset \mathcal{F}$ zwei Mengen an Funktionen aus $\{-1, 1\}^Z$, so gilt auf Grund des Supremums $\mathcal{R}_{\tilde{\mathcal{F}}} \leq \mathcal{R}_{\mathcal{F}}$. Sind in der Menge $\mathcal{F}|_S$ alle 2^n möglichen Funktionen, so ist die Summe $\mathcal{R}_{\mathcal{F}} = n$. Somit misst die Rademacher-Komplexität, analog zur VC-Dimension, wie viele zufällige Werte aus $\{-1, 1\}$ auf endlich vielen Punkten $S \subset Z$ durch $\mathcal{F}$ dargestellt werden können. Dies führen wir jetzt formal ein, auch für reellwertige Funktionen.

Definition 6.3. *Seien $Z \subset \mathbb{R}^k$, $\mathcal{F} \subset \mathbb{R}^Z$ Mengen und $S = \{z_1, \ldots, z_n\} \subset Z$. Seien $\varepsilon_1, \ldots, \varepsilon_n$ unabhängig und identisch gleichverteilt auf {-1,1} und weiterhin $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$. Diese werden auch als Rademacher-Zufallsvariablen bezeichnet. Damit definieren wir zuerst die empirische Rademacher-Komplexität, in Bezug auf die n Beobachtungen aus S:*

$$\widehat{\mathcal{R}}_S(\mathcal{F}) := E_\varepsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{\sqrt{n}} \left| \sum_{i=1}^{n} \varepsilon_i f(z_i) \right| \right].$$

Sei P ein Wahrscheinlichkeitsmaß auf Z und $z_1, \ldots, z_n$ unabhängig und identisch gemäß P verteilt. Dann definiert sich die Rademacher-Komplexität für das Wahrscheinlichkeitsmaß P auf Z:

$$\mathcal{R}_n(\mathcal{F}) := E_P\left[\widehat{\mathcal{R}}_S(\mathcal{F})\right].$$

Oftmals schreiben wir auch $\widehat{\mathcal{R}}(\mathcal{F}) := \widehat{\mathcal{R}}_S(\mathcal{F})$. Wir normieren die Summe nur mit den Faktor $\frac{1}{\sqrt{n}}$, statt dem Faktor $\frac{1}{n}$, damit wir später in den Sätzen 6.13 und 6.16 die Konvergenzrate $\frac{1}{\sqrt{n}}$ in den Ungleichungen klar erkennen können. Ebenfalls werden in der Literatur für die Definition der Rademacher-Komplexität oft die Betragsstriche weggelassen, zum Beispiel auch in [SSBD14], [BBL03], [MRT12] oder [Lia20]. Durch Formelzeile (6.2) können die beiden Definitionen aber ineinander übergeführt werden.

Betrachtet man eine Funktionen $f \in \mathcal{F}$, ausgewertet an den Punkten $z_1, \ldots, z_n$, als Vektor $(f(z_1), \ldots, f(z_n))$, so erhält man analog eine geometrische Definition:

Definition 6.4. *Sei $A \subset \mathbb{R}^k$, so definiert sich die Rademacher-Komplexität von A analog zur Definition 6.3, für Rademacher-Zufallsvariablen ε:*

$$\widehat{\mathcal{R}}(A) := E_\varepsilon\left[\sup_{a=(a_1,\ldots,a_k)\in A} \frac{1}{\sqrt{k}} \left|\sum_{i=1}^{k} \varepsilon_i a_i\right|\right].$$

Im Allgemeinen lässt sich die Rademacher-Komplexität einer Menge $\mathcal{F}$, verglichen mit der VC-Dimension, nicht sehr leicht berechnen. Im folgenden Satz, sowie im Kapitel 6.3, wollen wir dennoch einige Ergebnisse festhalten. In Kapitel 6.5 schätzen wir die Rademacher-Komplexität von unten gegen die VC-Dimension ab, somit können wir danach auch die Mittel der VC-Dimension nutzen. Für die Rechenregeln im nächsten Satz folgen wir [Men03], Satz 2.25, [BM02], Satz 12 und [Lia20], Kapitel 3. In [Lia20] werden in der Definition der Rademacher-Komplexität die Betragsstriche weggelassen.

Satz 6.5. *Seien im Folgenden $\mathcal{F}$ und $\mathcal{G}$ Mengen an reellwertigen Funktionen, $z_1, \ldots, z_n$ Punkte aus Z, $c, c' \in \mathbb{R}$ und $A, B \in \mathbb{R}^k$. Dann gilt:*

1. *Für $\mathcal{F} \subset \mathcal{G}$ gilt: $\mathcal{R}_n(\mathcal{F}) \leq \mathcal{R}_n(\mathcal{G})$.*
2. *$\mathcal{R}_n(c\mathcal{F}) = |c|\mathcal{R}_n(\mathcal{F})$.*
3. *$\mathcal{R}_n(\mathcal{F}+\mathcal{G}) \leq \mathcal{R}_n(\mathcal{F})+\mathcal{R}_n(\mathcal{G})$. Dabei ist $\mathcal{F}+\mathcal{G} := \{f+g, f \in \mathcal{F}, g \in \mathcal{G}\}$.*

4. $\mathcal{R}_n(\mathcal{F}) = \mathcal{R}_n(\text{conv}(\mathcal{F})) = \mathcal{R}_n(\text{absconv}(\mathcal{F}))$. *Dabei gilt:*

- $\text{conv}(\mathcal{F})$ *ist die konvexe Hülle von* $\mathcal{F}$*:*

$$\text{conv}(\mathcal{F}) = \left\{ \sum_{i=1}^{m} \alpha_i f_i | \sum_{i=1}^{m} \alpha_i = 1, \alpha_i \geq 0, f_i \in \mathcal{F}, \forall 1 \leq i \leq m \right\} \tag{6.4}$$

- $\text{absconv}(\mathcal{F})$ *beschreibt die symmetrische konvexe Hülle von* $\mathcal{F}$*, also* $\text{absconv}(\mathcal{F}) = \text{conv}(\mathcal{F} \cup -\mathcal{F})$*, mit* $-\mathcal{F} := \{-f : f \in \mathcal{F}\}$.

Beweis. Der erste Punkt folgt direkt aus der Definition der empirischen Rademacher-Komplexität.

Beim zweiten Punkt können wir die Konstante c für $c > 0$ aus dem Supremum und dem Erwartungswert ziehen:

$$E_P E_\varepsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{\sqrt{n}} \left| \sum_{i=1}^{n} c\varepsilon_i f(z_i) \right| \right] = cE_P E_\varepsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{\sqrt{n}} \left| \sum_{i=1}^{n} \varepsilon_i f(z_i) \right| \right].$$

Für $c = -|c| < 0$ benutzen wir, dass die Rademacher-Zufallsvariablen ε der gleichen Verteilung folgen wie die Variablen $-\varepsilon$. Insgesamt können wir also $-|c|(-1) = |c|$ aus dem Supremum und dem Erwartungswert ziehen.

Der dritte Punkt folgt direkt aus der Linearität des Erwartungswertes und der Darstellung von $\mathcal{F} + \mathcal{G}$ im Supremum sowie der Dreiecksungleichung.

Für den vierten Punkt gilt die Ungleichung

$$\mathcal{R}_n(\mathcal{F}) \leq \mathcal{R}_n(\text{conv}(\mathcal{F})) \leq \mathcal{R}_n(\text{absconv}(\mathcal{F}))$$

nach dem ersten Punkt. Wir müssen noch die umgekehrte Richtung zeigen. Die erste Ungleichung gilt, da $\text{absconv}(\mathcal{F})$ symmetrisch ist. Somit können wir für jede Realisierung $z_1, \ldots, z_n$ und alle $\varepsilon_1, \ldots, \varepsilon_n \in \{-1, 1\}$ die Funktion $\tilde{f} \in \text{absconv}(\mathcal{F})$, bei der das Supremum erreicht wird, durch $-\tilde{f} \in \text{absconv}(\mathcal{F})$ ersetzen. Damit ist die Summe sicher nicht-negativ, analog zur Formelzeile (6.2):

$$\sup_{\tilde{f} \in \text{absconv}(\mathcal{F})} \left| \sum_{i=1}^{n} \varepsilon_i \tilde{f}(z_i) \right| = \sup_{\tilde{f} \in \text{absconv}(\mathcal{F})} \sum_{i=1}^{n} \varepsilon_i \tilde{f}(z_i).$$

Da $\tilde{f}$ die Darstellung $\tilde{f} = \sum_{j=1}^{m} \lambda_j f_j$ mit $\sum_{j=1}^{m} |\lambda_j| = 1$ besitzt, erhalten wir:

$$\begin{aligned}
&\sup_{\tilde{f} \in \text{absconv}(\mathcal{F})} \sum_{i=1}^{n} \varepsilon_i \tilde{f}(z_i) \\
&= \sup_{\sum_{j=1}^{m} |\lambda_j| = 1, f_1, \ldots, f_m \in \mathcal{F}} \sum_{j=1}^{m} \lambda_j \sum_{i=1}^{n} \varepsilon_i f_j(z_i) \\
&\leq \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \varepsilon_i f(z_i) \leq \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \varepsilon_i f(z_i) \right| ..
\end{aligned}$$

□

6.3 Rademacher-Komplexität von Mengen an Funktionen

Mit dem nächsten Lemma, welches oftmals auch als Massarts Lemma bezeichnet wird, erhalten wir eine weitere Möglichkeit, die Rademacher-Komplexität nach oben abzuschätzen, vgl. [Mas00b], Lemma 5.2 und [MRT12], Satz 3.3.

Lemma 6.6. *Sei $\mathcal{F} \subset \mathbb{R}^Z$ endlich, $Z \subset \mathbb{R}^k$ und $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ eine Zufallsvariable, wobei ε_i unabhängig und identisch verteilte Rademacher-Zufallsvariablen sind und $z_1, \ldots, z_n \in Z$ beliebig. Dann gilt für die empirische Rademacher-Komplexität:*

$$E_\varepsilon \left[\max_{f \in \mathcal{F}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i f(z_i) \right] \leq \max_{f \in \mathcal{F}} \left(\sum_{i=1}^{n} f(z_i)^2 \right)^{1/2} \sqrt{\frac{2 \ln(\#\mathcal{F})}{n}}.$$

Mit der Darstellung aus Formelzeile (6.2) erhalten wir somit:

$$\widehat{R}_n(\mathcal{F}) = E_\varepsilon \left[\max_{f \in \mathcal{F} \cup -\mathcal{F}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i f(z_i) \right] \leq \max_{f \in \mathcal{F}} \left(\sum_{i=1}^{n} f(z_i)^2 \right)^{1/2} \sqrt{\frac{2 \ln(2\#\mathcal{F})}{n}}.$$

Der Beweis baut auf der Jensen'schen Ungleichung (6.1) und der Hoeffding-Ungleichung aus Satz 2.1 auf und findet sich zum Beispiel in [Mas00b].

Bemerkung 6.1. *Die Anzahl an $\{-1,1\}$-wertigen Funktionen einer Menge $\mathcal{F}_\pm$ auf n Punkten lässt sich mit der Wachstumsfunktion und der VC-Dimension abschätzen. Mit Hilfe von Massarts Lemma erhält man hiermit eine Abschätzung der Rademacher-Komplexität gegen die VC-Dimension. Dieses Vorgehen wird auch in [MRT12], Korollar 3.1, näher erläutert. Für reellwertige Funktionen betrachten wir eine solche Abschätzung in Satz 6.14.*

Für lineare Funktionen

$$\mathcal{H}_p := \{h : \mathbb{R}^k \to \mathbb{R}, x \mapsto w^T x : w \in \mathbb{R}^k \|w\|_p \leq \tau\}, \qquad \tau > 0. \tag{6.5}$$

lässt sich die Rademacher-Komplexität durch Ausnutzung der Definition der dualen Norm bestimmen. Sei $p \in [1,\infty)$ und q der zu p konjugierte Hölder-Exponent, also $\frac{1}{p} + \frac{1}{q} = 1$. Dann ist die duale Norm der ℓ_p-Norm die ℓ_q-Norm, vgl. [Wer05], Korollar II.2.2:

$$\sup_{\|t\|_{\ell_p} \leq 1} \{t^T z\} = \|t\|_{\ell_q}.$$

Ergebnisse für allgemeine $p \in [1,\infty]$ finden sich zum Beispiel in [AFM20]. Dort wird die Chintschin Ungleichung benutzt, siehe hierfür [Gar07], Satz 12.3.1 und [Haa81]. Wir wollen hier kurz ein Ergebnis für $p = 1$ und $p = 2$ mit $\tau = 1$ nennen, analog zu [SSBD14], Lemma 26.10 und Lemma 26.11.

Satz 6.7. *Sei $S = \{z_1, \ldots, z_n\} \subset Z \subset \mathbb{R}^k$ eine Menge an Vektoren. Wir definieren hierfür*

$$\mathcal{H}_p \circ S := \{(w^T x_1, \ldots, w^T x_n) : \|w\|_p \leq 1\}.$$

Dann gilt:

$$\mathcal{R}(\mathcal{H}_P \circ S) \leq \begin{cases} \max\limits_{z \in S} \|z\|_\infty \sqrt{\dfrac{2\ln(2k)}{n}}, & p = 1, \\ \dfrac{\max_{z \in S} \|z\|_2}{\sqrt{n}}, & p = 2. \end{cases}$$

6.4 Eine Schranke für den Approximationsfehler mit der Rademacher-Komplexität

Das Vorgehen für dieses Kapitel ist, soweit möglich, analog zu dem in den Kapiteln 2 und 4, wir orientieren uns hierbei an [Men03]. Wir haben wieder ein unbekanntes Wahrscheinlichkeitsmaß P auf $Z \subset \mathbb{R}$, sowie Punkte

$z_1, \ldots, z_n \in Z$ gegeben und leiten eine Ungleichung her, um für Funktionen $f \in \mathcal{F} \subset [A, B]^Z$ den Abstand zwischen dem Risiko und dem empirischen Risiko

$$D := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} (f(z_i) - E_P[f(z)]) \right|. \tag{6.6}$$

abzuschätzen. Mithilfe der Rademacher-Komplexität können allerdings schärfere Schranken bewiesen werden, da das zugrunde liegende Wahrscheinlichkeitsmaß mit einbezogen wird. Insbesondere können wir somit auf die Varianz innerhalb der Datenpunkte eingehen. Als erste Ungleichung hierzu betrachten wir eine Konzentrationsungleichung, Bernsteins Ungleichung, vgl. [Men03], Satz 2.20, analog zur Hoeffding-Ungleichung aus Satz 2.1. Danach betrachten wir im Satz 6.10 ein Symmetrisierungsargument, analog zu Satz 2.4. Damit erhalten wir die Ergebnisse:

1. Satz 6.11 mit Korollar 6.12 gibt, analog zu Satz 4.1, eine Schranke in Wahrscheinlichkeit für D an. Hierbei schätzen wir D aus (6.6) nach oben gegen ein α ab, wobei in α das Wahrscheinlichkeitsmaß P nur noch in der Rademacher-Komplexität vorkommt.

2. Satz 6.13 gibt, analog zu Satz 5.1, eine Schranke für die Approximation an.

Durch die Abschätzung der Rademacher-Komplexität nach oben durch die VC-Dimension im Unterkapitel 6.5 können wir die Ergebnisse direkt vergleichen und bemerken, dass wir mit Hilfe der Rademacher-Komplexität schärfere Grenzen herleiten können, insbesondere für δ nahe 1.

Satz 6.8. *Seien P ein Wahrscheinlichkeitsmaß auf $Z \subset \mathbb{R}^k$ und $z_1, \ldots, z_n$ unabhängige Zufallsvariablen auf Z, gemäß P verteilt. Weiterhin bezeichnen wir für eine messbare Funktion $f : Z \to \mathbb{R}$: $\nu := E\left[\sum_{i=1}^{n} f(z_i)^2\right]$. Dann gilt für alle $\varepsilon > 0$:*

$$P\left(\left| \int_Z f(z) dP(z) - \sum_{i=1}^{n} f(z_i) \right| \geq \varepsilon \right) \leq 2 \exp\left(-\frac{3\varepsilon^2}{2\left(\nu + \|f\|_\infty \varepsilon\right)} \right).$$

Mit Hilfe dieser Ungleichung lässt sich Talagrands Konzentrationsungleichung beweisen, zu finden in [Tal96], Satz 1.4. Später wurde diese durch Massart modifiziert, vergleiche hierzu [Men03], Satz 2.21, sowie [Mas00a], Satz 4.

Satz 6.9. *Seien P ein Wahrscheinlichkeitsmaß auf Z und $z_1, \ldots, z_n$ unabhängig identisch nach P verteilt. Sei $\mathcal{F}$ eine Menge an messbaren Funktionen aus $[0,1]^Z$ und bezeichne analog zur Formelzeile (6.6):*

$$D = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} (f(z_i) - E_P[f(z)]) \right|.$$

Dann existiert eine Konstante $C \in \mathbb{R}$ unabhängig von $\mathcal{F}$ und P, sodass für alle $\delta \in (0,1)$ mit einer Wahrscheinlichkeit von mindestens $(1-\delta)$ gilt:

$$nD \leq 2E_P[nD] + C\left(\sqrt{-n\ln(\delta)} + \ln(\delta)\right)$$

Beweis. Der Beweis befindet sich in [Mas00a], Abschnitt 4.1. In [Men03], Satz 2.21 wird hier statt $\sqrt{n}$ der Ausdruck $\sqrt{\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \operatorname{Var}(f(z_i))}$ benutzt. Da allerdings die Varianz von Zufallsvariablen mit Werten in $[0,1]$ analog zur Formelzeile (2.19) immer $\leq \frac{1}{4}$ ist, können wir diesen Ausdruck mit $\sqrt{n}$ ersetzen. □

Im nächsten Satz schätzen wir den Erwartungswert über dem unbekannten Wahrscheinlichkeitsmaß P gegen die Rademacher-Komplexität ab, wie in [Men03], Satz 2.23. Dies können wir dann in Satz 6.9 einsetzen und erhalten eine obere Schranke für den Abstand zwischen dem empirischen Risiko und dem Risiko, ausgedrückt mit Hilfe der Rademacher-Komplexität. Der Beweis nutzt wieder, analog zu Satz 2.4, ein Symmetrisierungsargument. Dabei folgen wir [Men03].

Satz 6.10. *Seien wieder P ein Wahrscheinlichkeitsmaß auf $Z \subset \mathbb{R}^k$ und $\mathcal{F} \subset \mathbb{R}^Z$ eine Menge an messbaren Funktionen. Seien $z_1, \ldots, z_n \in Z$ unabhängig und identisch gemäß P verteilt. Sei wieder wie in Formelzeile (6.6):*

$$D = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(z_i) - E_P f \right|,$$

dann gilt:

$$E_P[D] \leq 2\frac{\mathcal{R}_n(\mathcal{F})}{\sqrt{n}}.$$

Mithilfe von Satz 6.9 und Satz 6.10 lässt sich folgender Satz herleiten, vergleiche zusätzlich [GS08], Satz 4.1 und Korollar 2.24 in [Men03].

Satz 6.11. *Sei $\mathcal{F}$ eine Menge an Funktionen aus $[0,1]^Z$ und P ein Wahrscheinlichkeitsmaß auf Z. Weiterhin seien die Zufallsvariablen $z_1, \ldots, z_n$ unabhängig und identisch gemäß P verteilt. Dann existiert eine Konstante C, sodass für jedes $\delta \in (0,1)$ mit einer Wahrscheinlichkeit von mindestens $(1-\delta)$ gilt:*

$$\sup_{f \in \mathcal{F}} |R(f) - \widehat{R}(f)| \leq C \frac{1}{\sqrt{n}} \max \left\{ \mathcal{R}_n(\mathcal{F}), -\frac{\ln(\delta)}{\sqrt{n}}, \sqrt{-\delta} \right\}. \tag{6.7}$$

Dabei gilt für $\delta \geq \frac{1}{e}$, dass $-\ln(\delta) \geq \sqrt{-\ln(\delta)}$.

Beweis. Wir schreiben wieder analog zur Formelzeile (6.6):

$$D := \sup_{f \in \mathcal{F}} |R(f) - \widehat{R}(f)|.$$

Nach Satz 6.9 existiert eine Konstante $\tilde{C} \geq 1$, sodass für alle $\delta \in (0,1)$ gilt:

$$D \leq 2E_P D + \frac{\tilde{C}}{n} \left(\sqrt{-n \ln(\delta)} - \ln(\delta) \right).$$

Setzen wir die Ungleichung aus Satz 6.10 ein mit $\|L(\cdot, f)\|_\infty \leq 1$, erhalten wir mit einer Wahrscheinlichkeit von mindestens $\delta \in (0,1)$:

$$D \leq \frac{4\mathcal{R}_n(\mathcal{F})}{\sqrt{n}} + \frac{\tilde{C}\sqrt{-\ln(\delta)}}{\sqrt{n}} - \frac{\tilde{C}}{n} \ln(\delta).$$

Dabei können wir die drei Terme zusammenfassen:

$$D \leq \frac{4 + \tilde{C}}{\sqrt{n}} \max \left\{ \mathcal{R}_n(\mathcal{F}), \sqrt{-\ln(\delta)}, \frac{-\ln(\delta)}{\sqrt{n}} \right\}.$$

und mit $C := 4 + \tilde{C}$ folgt die Behauptung.

□

Dieses Resultat lässt sich ausweiten auf Funktionen, welche auf $[A, B] \subset \mathbb{R}$ abbilden, vgl. [GS08], Satz 4.2.

Korollar 6.12. *Sei P ein Wahrscheinlichkeitsmaß auf $Z \subset \mathbb{R}^k$ und $\mathcal{F} \subset [A,B]^{\mathbb{R}^k}$ eine Menge an Funktionen mit Rademacher-Komplexität $\mathcal{R}_n(\mathcal{F})$. Dann existiert eine Konstante $C \geq 1$, sodass für alle $\delta \in (0,1)$ mit einer Wahrscheinlichkeit von mindestens $(1-\delta)$ die folgende Schranke gilt:*

$$\sup_{f\in\mathcal{F}} |R(f) - \widehat{R}(f)| \leq (B-A)C\frac{1}{\sqrt{n}} \max\left\{ \frac{\mathcal{R}_n(\mathcal{F}-A)}{B-A}, -\frac{\ln(\delta)}{\sqrt{n}}, \sqrt{-\ln(\delta)} \right\}.$$

Beweis. Für jede Funktion $f \in \mathcal{F}$ sei $f_A := \frac{f-A}{B-A}$ eine Funktion mit Werten in $[0,1]$. Dafür bezeichnen wir:

$$\mathcal{F}_A := \left\{ \frac{f-A}{B-A} : f \in \mathcal{F} \right\}.$$

Setzen wir also $f_A \in \mathcal{F}_A$ in die Gleichung (6.7) ein, erhalten wir für die linke Seite mit $z_1, \ldots, z_n \in Z$:

$$\begin{aligned}
&\sup_{f_A\in\mathcal{F}_A} \left| \int_Z f_A(z) dP(z) - \frac{1}{n}\sum_{i=1}^n f_A(z_i) \right| \\
=&\sup_{f\in\mathcal{F}} \frac{1}{B-A} \left| \int_Z (f(z)-A) dP(z) - \frac{1}{n}\sum_{i=1}^n (f(z_i)-A) \right| \\
=&\frac{1}{B-A} \sup_{f\in\mathcal{F}} \left| \int_Z f(z) dP(z) - \frac{1}{n}\sum_{i=1}^n f(z_i) \right|.
\end{aligned}$$

Mit Hilfe von Satz 6.5 erhalten wir:

$$\mathcal{R}_n(\mathcal{F}_A) = \mathcal{R}_n\left(\frac{\mathcal{F}-A}{B-A}\right) = \frac{1}{B-A}\mathcal{R}_n(\mathcal{F}-A)$$

und damit nach Satz 6.11 die Behauptung. □

Damit können wir mit der Rademacher-Komplexität ein analoges Resultat zu Satz 4.1 herleiten, vgl. Satz 4.3 in [GS08]:

Satz 6.13. *Seien die Voraussetzungen wie in Satz 5.1, insbesondere $Z \subset \mathbb{R}^k$, $\mathcal{K} \subset [A,B]^{\mathbb{Z}\times Z}$ eine Menge an Kernen auf Z und*

$$\mathcal{F} = \{K(\cdot,t) : \mathbb{R}^k \to [A,B], K \in \mathcal{K}, t \in \mathbb{R}^k\}.$$

Dann existiert eine von $\mathcal{F}$ unabhängige Konstante C, sodass für jedes feste, nicht-negative $\lambda \in \mathcal{L}_1(Z), \|\lambda\|_1 \neq 0$ gilt:

- *Es existieren Punkte* $\tilde{z}_1, \ldots, \tilde{z}_n \in Z$

$$\sup_{f \in \mathcal{F}} \left| \int_Z f(z)\lambda(z)dz - \frac{1}{n}\sum_{i=1}^{n} f(\tilde{z}_i)\|\lambda\|_1 \right| \leq C\|\lambda\|_1 \mathcal{R}_n(\mathcal{F} - A)\sqrt{\frac{1}{n}}.$$

- *Für alle* $\delta \in (0,1)$ *finden wir mit einer Wahrscheinlichkeit von mindestens* $1-\delta$ *Punkte* $z_1, \ldots, z_n$*, sodass gilt:*

$$\begin{aligned} &\sup_{f \in \mathcal{F}} \left| \int_Z f(z)\lambda(z)dz - \frac{1}{n}\sum_{i=1}^{n} f(z_i)\|\lambda\|_1 \right| \\ \leq &C(B-A)\frac{1}{\sqrt{n}} \max\left\{ \frac{\|\lambda\|_1 \mathcal{R}_n(\mathcal{F}-A)}{B-A}, -\frac{\ln(\delta)}{\sqrt{n}}, \sqrt{-\ln(\delta)} \right\}. \end{aligned}$$

Beweis. Die Schritte sind analog zu dem Vorgehen in Satz 5.1. Für ein festes $\lambda \in \mathcal{L}_1(Z)$ betrachten wir hier die Rademacher-Komplexität von

$$\|\lambda\|_1 \mathcal{F} := \{\|\lambda\|_1 f : f \in \mathcal{F}\}$$

und erhalten mit der Hilfe von Korollar 6.12 die folgende Abschätzung für alle $\delta \in (0,1)$:

$$\begin{aligned} &\sup_{f \in \mathcal{F}} \left| \int_Z f(z)\lambda(z)dz - \frac{1}{n}\sum_{i=1}^{n} K_\lambda(\tilde{z}_i, t) \right| \\ \leq &\|\lambda\|_1 (B-A) C\frac{1}{\sqrt{n}} \max\left\{ \frac{\mathcal{R}_n(\|\lambda\|_1(\mathcal{F}-A))}{B-A}, -\frac{\ln(\delta)}{\sqrt{n}}, \sqrt{-\ln(\delta)} \right\}. \end{aligned}$$

Für die Rademacher-Komplexität der Menge $\|\lambda\|_1(\mathcal{F}-A)$ gilt mit Satz 6.5:

$$\mathcal{R}_n(\|\lambda\|_1(\mathcal{F}-A)) = \|\lambda\|_1 \mathcal{R}_n(\mathcal{F}-A).$$

Insgesamt erhalten wir mit der Betrachtung des Grenzwertes $\delta \to 1$, analog zum Vorgehen im Beweis zu Satz 5.1, die Behauptung. □

6.5 Abschätzung der Rademacher-Komplexität gegen die VC-Dimension

Im Folgenden wollen wir die Rademacher-Komplexität gegen die VC-Dimension abschätzen und zeigen, dass sich schärfere Grenzen durch die Rademacher-Komplexität ergeben. Insgesamt gilt es den folgenden Satz zu beweisen, vergleiche [GS08], Lemma 4.4.

Satz 6.14. *Sei $\mathcal{F} \subset [0,1]^{\mathbb{R}^k}$ eine Menge an Funktionen, dann existiert eine absolute Konstante C, insbesondere unabhängig von n und $\mathcal{F}$, sodass für alle $n \in \mathbb{N}$ gilt:*

$$\mathcal{R}_n(\mathcal{F}) \leq C\sqrt{\mathrm{VCD}(\mathcal{F})}.$$

Daraus folgt das folgende Korollar:

Korollar 6.15. *Sei $\mathcal{F} \subset [A,B]^{\mathbb{R}^k}$, so gilt unter der Voraussetzung von Satz 6.14:*

$$\mathcal{R}_n(\mathcal{F}) \leq C(B-A)\sqrt{\mathrm{VCD}(\mathcal{F})}.$$

Beweis. Wir schreiben für Konstanten $c, d \in \mathbb{R}$:

$$c\mathcal{F} + d := \{cf + d | f \in \mathcal{F}\}.$$

Es gilt die Ungleichungskette:

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}) &= \mathcal{R}_n\left(\frac{B-A}{B-A}(\mathcal{F}-A)\right) \\
&= (B-A)\mathcal{R}_n\left(\frac{1}{B-A}(\mathcal{F}-A)\right) \\
&\leq C(B-A)\sqrt{\mathrm{VCD}\left(\frac{1}{B-A}(\mathcal{F}-A)\right)} \\
&= C(B-A)\sqrt{\mathrm{VCD}(\mathcal{F})}.
\end{aligned}$$

Die letzte Gleichheit gilt dabei mit Hilfe von Lemma 3.9. □

Damit folgt, analog zu Korollar 6.12 und Satz 6.13 eine Schranke mit der VC-Dimension:

Satz 6.16. *Seien die Voraussetzungen wie in Satz 5.1, insbesondere sei $\mathcal{K} \subset [A,B]^{Z \times Z}$ eine Menge an Kernen,*

$$\mathcal{F} = \left\{K(\cdot, t) : Z \to [A,B], t \in \mathbb{R}^k, K \in \mathcal{K}\right\}$$

die Menge der betrachteten Funktionen und $d := \mathrm{VCD}(\mathcal{F})$. Dann existiert eine von $\mathcal{F}$ unabhängige Konstante C, sodass für jedes feste, nicht-negative $\lambda \in \mathcal{L}_1(Z), \|\lambda\|_1 \neq 0$, gilt:

- *Es existieren Punkte $\tilde{z}_1, \ldots, \tilde{z}_n \in Z$, sodass gilt:*

$$\sup_{f \in \mathcal{F}} \left| \int_Z f(z)\lambda(z)dz - \frac{1}{n}\sum_{i=1}^{n} f(\tilde{z}_i)\|\lambda\|_1 \right| \leq C(B-A)\|\lambda\|_1 \sqrt{\frac{d}{n}}.$$

- *Für alle $\delta \in (0,1)$ gilt mit einer Wahrscheinlichkeit von mindestens $1-\delta$:*

$$\begin{aligned} &\sup_{f \in \mathcal{F}} \left| \int_Z f(z)\lambda(z)dz - \frac{1}{n}\sum_{i=1}^{n} f(\tilde{z}_i)\|\lambda\|_1 \right| \\ \leq &C(B-A)\frac{1}{\sqrt{n}} \max\left\{ \sqrt{d}, -\frac{\ln(\delta)}{\sqrt{n}}, \sqrt{-\ln(\delta)} \right\}. \end{aligned}$$

Beweis. Wir setzen in die Ungleichung aus Satz 6.13 die Abschätzung der Rademacher-Komplexität gegen die VC-Dimension aus Korollar 6.15 ein. Dabei erhalten wir:

$$\mathcal{R}_n(\mathcal{F} - A) \leq C(B - A - 0)\sqrt{\mathrm{VCD}(\mathcal{F} - A)} = C(B-A)\sqrt{d},$$

da $d = \mathrm{VCD}(\mathcal{F})$. □

6.5.1 Beweisidee von Satz 6.14

Für den Beweis folgen wir den Beweis von Lemma 4.4 in [GS08].

1. In Satz 6.19, dem sog. *Dudleys Integral*, schätzen wir die empirische Rademacher-Komplexität nach oben gegen die sog. *ε-Überdeckungszahlen* ab.

2. Für die ε-Überdeckungszahlen $\mathcal{N}(\mathcal{F}, \|\cdot\|_p, \varepsilon)$ existieren zwei Abschätzungen nach oben:

   (a) In [Hau95b] wird bewiesen:

   $$\mathcal{N}(\mathcal{F}, \|\cdot\|_p, \varepsilon) \leq C\,\mathrm{VCD}(\mathcal{F}) \left(\frac{4e}{\varepsilon^2}\right)^{\mathrm{VCD}(\mathcal{F})}.$$

   Dies findet sich für Funktionen mit Werten in $\{0,1\}$ auch in [Men03], Satz 2.15 und wird dort in Korollar 2.32 auch dafür verwendet, die Rademacher Komplexität von unten gegen die VC-Dimension abzuschätzen.

(b) In [MV03], Satz 1, werden die ε-Überdeckungszahlen für reellwertige Funktionen gegen die fatD_ε-Dimension abgeschätzt. Diese wiederum lässt sich gegen die VC-Dimension abschätzen.

Wir wollen dabei den zweiten beschriebenen Weg gehen. Dafür leiten wir in Satz 6.20 eine Schranke der ε-Überdeckungszahlen nach oben gegen die sog. fatD_ε-Dimension her.

3. Mit Hilfe von Lemma 6.21 können wir die fatD_ε-Dimension von unten gegen die VC-Dimension abschätzen.

4. Da die VC-Dimension unabhängig von dem betrachteten Wahrscheinlichkeitsmaßes P ist, können wir auf beiden Seiten den Erwartungswert gegeben P betrachten und erhalten damit nicht nur eine Schranke für die empirische Rademacher-Komplexität, sondern auch für die Rademacher-Komplexität.

6.5.2 Definitionen der benötigten Komplexitätsmaßen

Die ε-Überdeckungszahlen, engl. *ε-covering numbers* definieren wir analog zu [AB09], Kapitel 10.2.

Definition 6.17. *Sei $\mathcal{F} \subset \mathbb{R}^Z, Z \subset \mathbb{R}^k$ eine Menge an Funktionen, $\|\cdot\|$ eine Norm auf Z und $\varepsilon > 0$. Dann ist die ε-Überdeckungszahl $\mathcal{N}(\mathcal{F}, \|\cdot\|, \varepsilon)$ definiert als die kleinste Zahl m, für welche Funktionen $f_1, \dots, f_m \in \mathcal{F}$ existieren, sodass für alle Funktionen $f \in \mathcal{F}$ ein $1 \leq i \leq m$ existiert, sodass gilt: $\|f - f_i\| \leq \varepsilon$:*

$$\mathcal{N}(\mathcal{F}, \|\cdot\|, \varepsilon) := \min_{\tilde{\mathcal{F}} \subset \mathcal{F}} \{\#\tilde{\mathcal{F}} : \forall f \in \mathcal{F} \exists \tilde{f} \in \tilde{\mathcal{F}} : \|f - \tilde{f}\| \leq \varepsilon\}.$$

Für die nächste Definition folgen wir [ABDCBH97], Definition 2.2.

Definition 6.18. *Sei $Z \subset \mathbb{R}^k$ und $\varepsilon > 0$ fest. Dann sagen wir, dass eine Menge $\mathcal{F} \subset \mathbb{R}^Z$ die Teilmenge $S \subset Z$ fat_ε-splittert, falls eine feste nur von $\mathcal{F}$ abhängige Funktion $s \in \mathbb{R}^Z$ existiert, sodass für jede Teilmenge $S^+ \subset S$ eine Funktion $f_{S^+} \in \mathcal{F}$ gefunden werden kann, sodass gilt:*

$$z \in S \implies \begin{cases} f_{S^+}(z) \leq s(z) - \varepsilon, & z \notin S^+ \\ f_{S^+}(z) \geq s(z) + \varepsilon, & z \in S^+. \end{cases}$$

Die fat_ε*-Dimension der Menge* $\mathcal{F}$ *bezeichnen wir als* $\mathrm{fatD}_\varepsilon(\mathcal{F})$ *und entspricht der Kardinalität der größten Teilmenge aus* Z*, welche* $\mathcal{F}$ fat_ε*-splittert:*

$$\mathrm{fatD}_\varepsilon(\mathcal{F}) := \max_{S \subset Z} \{\#S : \mathcal{F} \text{ } \mathrm{fat}_\varepsilon \text{ -splittert } S\}.$$

6.5.3 Beweise

Ebenfalls lässt sich die Rademacher-Komplexität gegen die ε-Überdeckungszahl abschätzen, vergleiche, auch für den Beweis, [Ver18], Satz 8.1.3, sowie [SSBD14], Lemma 27.4 und [Sri12], Lemma 107. Das Resultat wird meist als Dudley's Integral bezeichnet.

Satz 6.19. *Seien* $\mathcal{F} \subset \mathbb{R}^Z$*,* $Z \subset \mathbb{R}^k$ *und gebe* μ_n *den Abstand zweier Funktionen aus* $\mathcal{F}$ *auf den Punkten* $z_1, \ldots, z_n \in Z$ *wie folgt an:*

$$\mu_n(f,g) := \left(\frac{1}{n} \sum_{i=1}^{n} (f(z_i) - g(z_i))^2 \right)^{1/2}.$$

Sei $\mathcal{N}(\mathcal{F}, \mathcal{L}_2(\mu_n), \varepsilon)$ *die* ε*-Überdeckungszahl von* $\mathcal{F}$*. Dann existiert eine absolute Konstante* $\widehat{C}$*, sodass für alle* $n \in \mathbb{N}$ *gilt:*

$$\widehat{\mathcal{R}}(\mathcal{F}) \leq \widehat{C} \int_0^\infty (\ln \mathcal{N}(\mathcal{F}, \mathcal{L}_2(\mu_n), \varepsilon))^{\frac{1}{2}} \, d\varepsilon.$$

Dabei gilt: $\widehat{C} \leq 12$*.*

Die ε-Überdeckungszahlen lassen sich auch gegen die fat_ε-Dimension abschätzen. Dafür folgen wir [MV03], Satz 1. Dort findet sich auch der Beweis.

Satz 6.20. *Für eine Menge* $\mathcal{F} \subset [0,1]^X$*, mit* $X \subset \mathbb{R}^k, k \in \mathbb{N}$ *und* μ *ein Wahrscheinlichkeitsmaß auf* X *sei* $\mathcal{N}(\varepsilon, \mathcal{F}, \mathcal{L}_2(\mu))$ *die* ε*-Überdeckungszahl und* $\mathrm{fatD}_\varepsilon(\mathcal{F})$ *die* fat_ε*-Dimension von* $\mathcal{F}$*. Dann existieren Konstanten* $C_1, C_2 > 0$*, sodass für alle* $\varepsilon \in (0,1)$ *gilt:*

$$\mathcal{N}(\mathcal{F}, \mathcal{L}_2(\mu), \varepsilon) \leq \left(\frac{2}{\varepsilon} \right)^{C_1 \, \mathrm{fatD}_{\varepsilon C_2}(F)}.$$

Dabei sind die Konstanten C_1, C_2 *absolut, also insbesondere unabhängig von* n, ε *und* $\mathcal{F}$*.*

Betrachtet man allerdings statt dem $\mathcal{L}_2$-Maß ein $\mathcal{L}_p$-Maß, für $1 \leq p \leq \infty$, so ändern sich auch die Konstanten C_1, C_2, siehe hierzu [Men04], Satz 3.7.

Nun fehlt es noch, die fat_ε-Dimension gegen die VC-Dimension abzuschätzen. Dabei folgen wir [ABDCBH97], Lemma 2.3 und 2.4, dessen für uns relevanten Aussagen wir im nächsten Lemma zusammenfassen.

Lemma 6.21. *Für jede Menge an Funktionen $\mathcal{F} \subset [0,1]^Z$ und für alle $\epsilon > 0$ gilt:*

$$\text{fatD}_\varepsilon(\mathcal{F}) \leq \left(\frac{2}{\varepsilon} - 1\right) \text{VCD}(\mathcal{F})$$

Für den Beweis verweisen wir ebenfalls auf [ABDCBH97].

Mithilfe dieser Resultate kann man nun den Satz 6.14 beweisen, analog zu [GS08].

Beweis. Sei $\varepsilon \geq 1$. Dann gilt, dass $\mathcal{N}(\mathcal{F}, \mathcal{L}_2(\mu_n), \varepsilon) = 1$, da $\mathcal{F} \subset [0,1]^Z$. Da $\ln(1) = 0$, gilt somit:

$$\int_0^\infty \left(\ln(\mathcal{N}(\mathcal{F}, \mathcal{L}_2(\mu_n), \varepsilon))\right)^{1/2} d\varepsilon = \int_0^1 \left(\ln(\mathcal{N}(\mathcal{F}, \mathcal{L}_2(\mu_n), \varepsilon))\right)^{1/2} d\varepsilon.$$

Sei $\varepsilon \in (0,1)$. Dann gilt mit Lemma 6.21 folgende Ungleichung:

$$\text{fatD}_\varepsilon(\mathcal{F}) \leq \left(\frac{2}{\varepsilon} - 1\right) \text{VCD}(\mathcal{F}).$$

Insgesamt folgt mit Satz 6.20 die Existenz der Konstanten $C_1, C_2 > 0$:

$$\begin{aligned}
\int_0^1 \left(\ln(\mathcal{N}(\mathcal{F}, \mathcal{L}_2(\mu_n), \varepsilon))\right)^{1/2} d\varepsilon &\leq \int_0^1 \left(\ln\left(\frac{2}{\varepsilon}\right)^{C_1 \,\text{fatD}_{\varepsilon C_2}(\mathcal{F})}\right)^{1/2} d\varepsilon \\
&\leq \int_0^1 \left(\ln\left(\frac{2}{\varepsilon}\right)^{C_1\left(\frac{2}{\varepsilon}-1\right) \text{VCD}(\mathcal{F})}\right)^{1/2} d\varepsilon \\
&= \int_0^1 \left(C_1\left(\frac{2}{\varepsilon} - 1\right) \text{VCD}(\mathcal{F})\right)^{1/2} \ln\left(\frac{2}{\varepsilon}\right)^{\frac{1}{2}} d\varepsilon
\end{aligned}$$

Aus $\text{fatD}_{\varepsilon C_2} = 0$ folgt, dass die zweite Ungleichung auch für $\varepsilon C_2 > 1$ gilt. Somit können wir den Faktor C_2 auch weglassen. Da das Integral

$$\int_0^1 \left(C_1\left(\frac{2}{\varepsilon} - 1\right)\right)^{1/2} \ln\left(\frac{2}{\varepsilon}\right)^{\frac{1}{2}} d\varepsilon \leq 5C_1 =: \widetilde{C}$$

konvergiert, erhält man insgesamt:

$$\int_0^1 \left(C_1 \left(\frac{2}{\varepsilon} - 1 \right) \mathrm{VCD}(\mathcal{F}) \right)^{1/2} \ln \left(\frac{2}{\varepsilon} \right)^{\frac{1}{2}} d\varepsilon \leq \widetilde{C} (\mathrm{VCD}(\mathcal{F}))^{\frac{1}{2}}.$$

Mit Satz 6.19 erhalten wir:

$$\widehat{\mathcal{R}}(\mathcal{F}) \leq \widehat{C} \int_0^\infty \left(\ln(\mathcal{N}(\varepsilon, \mathcal{F}, \mathcal{L}_2(\mu_n))) \right)^{1/2} d\varepsilon \leq (\widehat{C}\widetilde{C})(\mathrm{VCD}(\mathcal{F}))^{\frac{1}{2}}.$$

Da $\widehat{C} \leq 12$ aus Satz 6.19 gilt, folgt:

$$C = \widehat{C}\tilde{C} \leq 60 C_1.$$

Da der Erwartungswert einer Konstanten c bezüglich eines Wahrscheinlichkeitsmaßes immer gleich der Konstanten c ist, erhalten wir das gewünschte Ergebnis auch für die nicht-empirische Rademacher-Komplexität. □

Um die Konstante C im Satz 6.14 zu bestimmen, fehlt also, die Konstante C_1 aus Satz 6.20 zu bestimmen.

7 Anwendung auf radiale Funktionen

Im Folgenden wollen wir Satz 5.1 auf radiale Funktionen, kurz RF, anwenden. RF sind, wie in [Wen05], Kapitel 6.3 beschrieben, Funktionen, welche nur vom Abstand zur 0, gemäß einer Norm, abhängen, analog zu Definition 6.15 in [Wen05].

Definition 7.1. *Eine Funktion $\Phi : \mathbb{R}^k \to \mathbb{R}^k$ heißt radial, falls eine Funktion $\phi : [0, \infty) \to \mathbb{R}$ existiert, sodass $\Phi(z) = \phi(\|z\|_2)$ gilt.*

Im Folgenden betrachten wir als Argument für die radiale Funktion Φ den euklidischen Abstand zweier Punkte $z, t \in \mathbb{R}^k$. Dies entspricht dem Kern

$$K(z,t) := \Phi(z-t) = \phi(\|z-t\|_2) : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}, \tag{7.1}$$

welcher stetig ist, falls die zugehörige Funktion $\phi : \mathbb{R}_0^+ \to \mathbb{R}$ stetig ist.

Mit Hinblick auf Satz 6.16 wollen wir nun die VC-Dimension von radialen Funktionen berechnen. Wir haben im Korollar 3.11 gesehen, dass die Verknüpfung einer Funktionenmenge $\mathcal{G} \subset \mathbb{R}^Z$ mit einer Menge an monotonen Funktionen $\mathcal{F}$ die VC-Dimension nur um den Faktor 2 vergrößert, also

$$\mathrm{VCD}(\mathcal{F} \circ \mathcal{G}) = \mathrm{VCD}(\{f \circ g : f \in \mathcal{F}, g \in \mathcal{G}\}) \leq 2\,\mathrm{VCD}(\mathcal{G}) + 1, \tag{7.2}$$

mit:

$$\mathcal{F} = \{f : \mathbb{R} \to \mathbb{R}, f \text{ monoton auf } \mathbb{R}\}.$$

In Satz 7.6 sehen wir, dass die VC-Dimension der Menge

$$\{f : \mathbb{R}^k \to \mathbb{R},\ z \mapsto \Phi(\|z-t\|_2), t \in \mathbb{R}^k\}$$

abhängig ist von der VC-Dimension von Bällen in der betrachteten Norm, sowie der Anzahl der Intervalle auf $\mathbb{R}_0^+$, auf denen $\phi(r)$ monoton ist. Diesbezüglich wollen wir die folgenden beiden Definitionen einführen. Wir erinnern dabei an die Darstellung aus Formelzeile(0.3).

Definition 7.2. *Sei $Y \subset \mathbb{R}$ ein Intervall. Wir sagen, eine Funktion $f \in Y^{\mathbb{R}}$ erfüllt die Monotoniebedingung für p Intervalle, falls für jedes $f \in \mathcal{F}$ höchstens p paarweise disjunkte Intervalle $\mathcal{I}_1 \cup \cdots \cup \mathcal{I}_p = Y$ existieren, sodass sich f schreiben lässt als*

$$f(x) = \sum_{i=1}^{p} f|_{\mathcal{I}_i}(x), \quad f|_{\mathcal{I}_i} \text{ monoton}, 1 \leq i \leq p. \tag{7.3}$$

Wir fordern dabei, dass p minimal gewählt ist. Weiterhin wählen wir die Indizes der Intervalle so, dass Intervalle mit kleinerem Index auch kleinere Werte enthält as Intervalle mit größeren Index.

Mengen, welche die Bedingung (7.3) erfüllen, fassen wir in der Menge $\mathcal{M}_p$ zusammen:

$$\mathcal{M}_p := \{f : \mathbb{R} \to Y, f \text{ erfüllt (7.3) für } p \text{ Intervalle}\}.$$

Definition 7.3. *Sei $\mathcal{F} \subset Y^{\mathbb{R}}$ wie in Definition 7.2. Für jede feste Funktion $f \in \mathcal{F}$ sei $I_{f,1} \cup \cdots \cup I_{f,p_f}$ die Zerlegung aus Formelzeile (7.3) mit minimalen p_f. Sei $p_{2,f} \leq \lfloor \frac{p_1}{2} \rfloor$ die maximale Zahl der Intervalle $I_{f,i}$, auf denen f monoton steigend ist und auf dem darauf folgenden Intervall $I_{f,i+1}$ wieder monoton fällt. Dann sagen wir, dass $\mathcal{F}$*

$$\tilde{p} := \max_{f \in \mathcal{F}} (p_f - p_{2,f}),$$

effektive Intervalle besitzt.

Die Ergebnisse wenden wir in Satz 7.11 auf Satz 6.16 an. Dabei erhalten wir eine Konstante $c_{\tilde{p},2}$, sodass für die Menge

$$\mathcal{F} = \left\{\phi(\|\cdot - t\|_2 : \mathbb{R}^k \to \mathbb{R}, t \in \mathbb{R}^k, \phi \in \mathcal{M}_p)\right\} \tag{7.4}$$

und nicht-negative $\lambda \in \mathcal{L}_1(\mathbb{R}^k)$, $\|\lambda\|_1 \neq 0$ gilt:

$$\begin{aligned}&\sup_{\Phi(\cdot,t) \in \mathcal{F}} \left| \int_{\mathbb{R}^k} \Phi(z,t)\lambda(z)dz - \frac{1}{n} \sum_{i=1}^{n} \Phi(\tilde{z}_i, t)\|\lambda\|_1 \right| \\ \leq &(B - A)c_{\tilde{p},2} C \|\lambda\|_1 \sqrt{\frac{(2k+3)}{n}}.\end{aligned}$$

Im nächsten Unterkapitel 7.1 zeigen wir, wie wir die VC-Dimension von $\mathcal{F}$ aus (7.4) bestimmen können. Dafür berechnen wir im Kapitel 7.2 die VC-Dimension der Bälle aus $\mathbb{R}^k$ in der $\|\cdot\|_2$-Norm. In den Kapiteln 7.3 und 7.4 wenden wir dies auf verschiedene radiale Kerne an. Im Kapitel 7.5 betrachten wir als Anwendung noch den Span aus endlich vielen Funktionen.

7.1 Vereinfachung von Verknüpfung auf Vereinigung

Für $\mathcal{F} \subset Y^{\mathbb{R}^k}$, mit $Y \subset \mathbb{R}$ ein Intervall und $\mathcal{G} \subset \mathbb{R}^{\mathbb{R}}$ wollen wir im Folgenden die VCD der Menge

$$\mathcal{F} \circ \mathcal{G} := \{f \circ g : f \in \mathcal{F}, g \in \mathcal{G}\}$$

abschätzen. Aufbauend auf der Idee aus Formelzeile (7.2) wollen wir dann die VC-Dimension von $\mathcal{G} \circ \mathcal{F}$ durch die Anzahl der Intervalle, auf denen die Funktionen aus $\mathcal{G}$ monoton sind, abschätzen. Dazu betrachten wir das Lemma 7.5, welches durch [BEHW89], Lemma 3.2.3, impliziert wird. Dort berechnen wir die VC-Dimension für die Vereinigung verschiedener Mengen. Anders als in Lemma 3.10 sind diesmal alle betrachteten Mengen Teilmengen einer Menge mit bekannter, endlicher VC-Dimension. Dazu führen wir die folgende Notation ein.

Definition 7.4. *Sei die Konzeptklasse $\mathcal{C}$ eine Menge an Teilmengen von $\mathbb{R}^k$, dann bezeichnen wir mit $\mathcal{C}^{\cup p}$ die Vereinigung von p nicht notwendigerweise verschiedenen Mengen aus $\mathcal{C}$:*

$$\mathcal{C}^{\cup p} := \{C_1 \cup \ldots \cup C_p : C_1, \ldots, C_p \in \mathcal{C}\}.$$

Analog definieren wir für den Schnitt:

$$\mathcal{C}^{\cap p} := \{C_1 \cap \ldots \cap C_p : C_1, \ldots, C_p \in \mathcal{C}\}.$$

Betrachten wir statt der Konzeptmenge $\mathcal{C}$ eine Menge an Funktionen $\mathcal{G}$ von $Z \subset \mathbb{R}^k$ nach $\mathbb{R}$, so betrachten wir für die Funktionen aus $\mathcal{G}$ die Konzeptmengen

$$\mathcal{C}_{\mathcal{G}} := \{S \subset Z : \exists g \in \mathcal{G}, \beta \in \mathbb{R} : z \in S \Leftrightarrow g(z) - \beta \geq 0\},$$

analog zu den Definitionen 3.5 und 3.6. Damit erhalten wir, analog zu oben:

$$\mathcal{G}^{\cup p} := \mathcal{C}_{\mathcal{G}}^{\cup p} = \{C_1 \cup \ldots \cup C_p : C_1, \ldots, C_p \in \mathcal{C}_{\mathcal{G}}\},$$
$$\mathcal{G}^{\cap p} := \mathcal{C}_{\mathcal{G}}^{\cap p} = \{C_1 \cap \ldots \cap C_p : C_1, \ldots, C_p \in \mathcal{C}_{\mathcal{G}}\}.$$

Lemma 7.5. *Sei d die VC-Dimension einer Konzeptklasse $\mathcal{C}$, nach Definition 3.5. Dann existiert für alle $p \in \mathbb{N}$ eine Konstante c_p, sodass gilt:*

$$\mathrm{VCD}(\mathcal{C}^{\cup p}) \leq c_p d,$$
$$\mathrm{VCD}(\mathcal{C}^{\cap p}) \leq c_p d.$$

In [BEHW89] wird noch impliziert, dass die Konstante c_p höchstens mit Rate $p \log(p)$ in p steigt. In [CKM18] wird gezeigt, dass auch die untere Grenze für $\mathrm{VCD}(\mathcal{C}^{\cup p})$ und $\mathrm{VCD}(\mathcal{C}^{\cap p})$ für die Abschätzungen in $\Omega(dp \log(p))$ liegt.

Beweis. Wir beweisen dies nur für $\mathcal{C}^{\cup p}$, der Beweis zu $\mathcal{C}^{\cap p}$ geht dann analog. Wir erinnern uns daran, dass die VC-Dimension einer Menge $\mathcal{C}$ die größtmögliche natürliche Zahl d ist, sodass

$$\mathscr{G}_{\mathcal{C}}(d) = 2^d$$

gilt.

Für den Beweis nutzen wir Sauers Lemma, Satz 3.7, um die Wachstumsfunktion von $\mathcal{C}^{\cup p}$ für $n \geq d$ wie folgt abzuschätzen:

$$\mathscr{G}_{\mathcal{C}^{\cup p}}(n) \leq \mathscr{G}_{\mathcal{C}}(n)^p \leq \left(\frac{en}{d}\right)^{dp}.$$

Dabei gilt die erste Abschätzung, weil man p Mengen, welche von $\mathcal{C}$ auf n Punkten induziert werden, vereinigt. Dies führt zu höchstens $\mathscr{G}_{\mathcal{C}}(n)^p$ verschiedenen Mengen. Analog ergeben sich aus den p-fachen Schnitt aus $\mathscr{G}_{\mathcal{C}}(n)$ verschiedenen Mengen höchstens $\mathscr{G}_{\mathcal{C}}(n)^p$ neue Mengen.

Wir suchen nun also nach Definition 3.4 das größtmögliche $m \in \mathbb{N}$, welches erfüllt:

$$\left(\frac{em}{d}\right)^{dp} \geq 2^m.$$

Da für hinreichend große $n \in \mathbb{N}$ die Ungleichung $\left(\frac{en}{d}\right)^{dp} < 2^n$ gilt, muss m kleiner sein als jedes n mit dieser Eigenschaft. Durch Umformungen erhalten wir:

$$\begin{aligned} \left(\frac{en}{d}\right)^{dp} &< 2^n \\ dp \log_2\left(\frac{en}{d}\right) &< n \\ 1 &< \frac{n}{dp \log_2\left(\frac{en}{d}\right)}. \end{aligned}$$

Setzen wir nun $n = \tilde{c}_p dp$ ein, erhalten wir

$$1 < \frac{\tilde{c}_p}{\log_2\left(e\tilde{c}_p p\right)}.$$

und sehen, dass die Grenze für hinreichend großes $\tilde{c}_p$ erfüllt ist. Ebenfalls sehen wir, dass $\tilde{c}_p$ nur von p abhängt. Wir setzen dann: $c_p := p\tilde{c}_p$. □

Wie wir in Formelzeile (7.2) wiederholt haben, vergrößert das Anwenden einer monotonen Funktion auf alle Funktionen aus $\mathcal{G}$ die VC-Dimension von $\mathcal{G}$ nicht. Um dies zu verallgemeinern, betrachten wir im Folgenden Funktionen, welche die *Monotoniebedingung für p Intervalle* erfüllen.

Satz 7.6. *Sei $\mathcal{G} \subset Y^Z$ eine Menge an Funktionen mit*

$$\mathrm{VCD}(\mathcal{G}) = d < \infty$$

und $Y \subset \mathbb{R}$ ein Intervall. Wir definieren hierzu:

$$\overline{\mathcal{G}} := \mathcal{G} \cup -\mathcal{G}.$$

Sei $\mathcal{F} \subset \mathbb{R}^Y$ eine Menge an Funktionen, welche die Monotoniebedingung aus Definition 7.2 für $p < \infty$ Intervalle erfüllt. Weiterhin besitzte $\mathcal{F}$ insgesamt $\tilde{p}$ effektive Intervalle aus Definition 7.3. Dann gilt:

$$\mathrm{VCD}(\mathcal{F} \circ \mathcal{G}) \leq \mathrm{VCD}((\overline{\mathcal{G}}^{\cap 2})^{\cup \tilde{p}}) \leq c_{\tilde{p},2}(2d+1). \tag{7.5}$$

Für $p = 1$ erhalten wir dabei:

$$\mathrm{VCD}(\mathcal{F} \circ \mathcal{G}) \leq 2\,\mathrm{VCD}(\mathcal{G}) + 1 = 2d + 1. \tag{7.6}$$

Sind alle Funktionen aus $\mathcal{F}$ für $p = 1$ monoton steigend, oder alle Funktionen aus $\mathcal{F}$ monoton fallend, so folgt:

$$\mathrm{VCD}(\mathcal{F} \circ \mathcal{G}) \leq \mathrm{VCD}(\mathcal{G}) = d. \tag{7.7}$$

Beweis. Für den Beweis der ersten Ungleichung betrachten wir ein beliebiges, aber festes $f \in \mathcal{F}$ und bezeichnen die zugehörigen Intervalle aus der Zerlegung (7.3) mit $\mathcal{I}_1 \cup \cdots \cup \mathcal{I}_{p_f} = Y$. Auf diesen ist f jeweils monoton. Weiterhin sei $S \subset Z$ eine beliebige, aber feste, endliche Menge. Wir zeigen, dass dann jede Teilmenge S^+ von S, welche durch $\mathcal{F} \circ \mathcal{G}$ induziert wird, auch durch $(\overline{\mathcal{G}}^{\cap 2})^{\cup p}$ induziert wird.

Sei im Folgenden die Menge S^+ durch $(f \circ g, \beta)$ induziert, für $g \in \mathcal{G}$. Sei weiterhin P die Menge der Indizes $1 \leq i \leq p_f$, für welche f auf den Intervallen $\mathcal{I}_i$ monoton steigt oder konstant ist und N die Menge der Indizes zu den Intervallen, auf denen f monoton fällt.

Dann können wir die Menge S^+ zerlegen, je nachdem, in welches Intervall $g(z)$ abbildet:

$$S_i^+ := \{z \in S^+ : g(z) \in \mathcal{I}_i\}, \qquad 1 \leq i \leq p.$$

Damit existieren für $S^+ = S_1^+ \cup \ldots \cup S_{p_f}^+$ Werte $\beta_1, \ldots, \beta_{p_f} \in \mathbb{R}$, sodass:

$$g(z) \begin{cases} \geq \beta_i, i \in P, & z \in S_i^+, \\ \leq \beta_i, i \in N, & z \in S_i^+. \end{cases}$$

Für die Werte $\beta_1, \ldots, \beta_n$ gilt die Einschränkung:

$$\forall 1 \leq i, j \leq p_f : f(\beta_i) = f(\beta_j) = \beta.$$

Von der Funktion $f \circ g$ werden diejenigen Punkte aus S induziert, für welche $f \circ g(z) \geq \beta$ gilt. Dies ist in jedem Intervall $\mathcal{I}_1, \ldots, \mathcal{I}_{p_f}$ eine Teilmenge von S_i sein. Genauso gilt für die Punkte aus $S \backslash S^+$:

$$g(z) \begin{cases} < \beta_i, i \in P, z \in S \backslash S_i^+, \\ > \beta_i, i \in N, z \in S \backslash S_i^+. \end{cases}$$

Im Folgenden bezeichnen wir

$$b_i := \overline{\mathcal{I}}_i \cap \overline{\mathcal{I}}_{i+1}.$$

Betrachten wir nun $i \in P$, dann lässt sich S_i^+ durch g und $-g$ wie folgt darstellen:

$$\begin{aligned} S_i^+ &= \{z \in S : \beta_i \leq g(z) \leq b_i\} \\ &= \{z \in S : \beta_i \leq g(z)\} \cap \{z \in S : -b_i \leq -g(z)\}. \end{aligned}$$

Genauso lassen sich die Mengen S_j^+ für $j \in N$ darstellen:

$$S_j^+ = \{z \in S : -\beta_j \leq -g(z)\} \cap \{z \in S : b_{j-1} \leq g(z)\}.$$

Ist f auf dem auf $\mathcal{I}_i$ folgenden Intervall $\mathcal{I}_{i+1}$ wieder monoton fallend, so können wir dies vereinfachen:

$$S_i^+ \cup S_{i+1}^+ = \{z \in S : \beta_i \leq g(z)\} \cap \{z \in S : -\beta_{i+1} \leq -g(z)\}.$$

Dabei sind sowohl g als auch $-g$ in der Menge $\overline{\mathcal{G}}$ und somit lässt sich S_i^+ durch $\overline{\mathcal{G}}^{\cap 2}$ induzieren.

Weiterhin existieren $p_{2,f}$ Indizes aus P, bei denen der darauf folgende Index in N liegt. Damit benötigen wir maximal

$$p_f - p_{2,f} \leq \tilde{p} = \sup_{f \in \mathcal{F}} (p_f - p_{2,f})$$

Mengen aus $\bar{\mathcal{G}}^{\cap 2}$, um S^+ zu induzieren. Dabei war $\tilde{p}$ die Zahl der effektiven Intervalle von $\mathcal{F}$.

Insgesamt können wir somit jedes beliebige $S^+ \subset S$, welches durch eine Funktion aus $\mathcal{F} \circ \mathcal{G}$ induziert wird, auch durch eine Menge in $(\overline{\mathcal{G}}^{\cap 2})^{\cup \tilde{p}}$ induzieren. Somit können wir auch die VC-Dimensionen gegeneinander abschätzen.

Betrachten wir nun die zweite Ungleichung der Formelzeile (7.5). Nach Lemma 3.10 gilt:

$$\mathrm{VCD}(\mathcal{G} \cup -\mathcal{G}) = \mathrm{VCD}(\bar{\mathcal{G}}) \leq 2d + 1.$$

Weiterhin existiert nach Lemma 7.5 eine Konstante c_2:

$$\mathrm{VCD}(\bar{\mathcal{G}}^{\cap 2}) \leq 2c_2(2d + 1). \tag{7.8}$$

Damit existiert nach Lemma 7.5 eine Konstante $c_{\tilde{p},2}$, sodass gilt:

$$\mathrm{VCD}((\bar{\mathcal{G}}^{\cap 2})^{\cup \tilde{p}}) \leq c_{\tilde{p},2}(2d + 1).$$

Die beiden Aussagen (7.6) und (7.7) zum Fall $p = 1$ folgen direkt aus Lemma 3.9 bzw. Korollar 3.11. □

Bemerkung und Beispiel 7.1. *Wir wollen hier kurz erklären, dass die hergeleitete Abschätzung mindestens linear in p und in* $\mathrm{VCD}(\mathcal{G})$ *sind. Gleichzeitig bietet diese Bemerkung Platz für Beispiele, um den Beweis von Satz 7.6 besser zu verstehen.*

Dabei wollen wir zuerst zeigen, dass die Grenze in Bezug auf die VC-Dimension von $\mathcal{G}$ *scharf ist. Wir betrachten hierzu* $\mathcal{F}_{id} = \{id_{\mathbb{R}}\}$*. Die Identität* $id_{\mathbb{R}} : \mathbb{R} \to \mathbb{R}, x \mapsto x$ *ist monoton steigend auf ganz* $\mathbb{R}$*. Gleichzeitig gilt:*

$$\mathcal{F}_{id} \circ \mathcal{G} = \{id_{\mathbb{R}}\} \circ \mathcal{G} = \mathcal{G} \Rightarrow \mathrm{VCD}(\mathcal{F}_{id} \circ \mathcal{G}) = d.$$

Weiterhin lässt sich die Identität zu allen Mengen an Funktionen $\mathcal{F}$ *hinzufügen. Erfüllt eine Menge* $\mathcal{F}$ *die Monotoniebedingung für p Intervalle, so erfüllt* $\mathcal{F} \cup \{id\}$ *dies ebenfalls. Ebenso erhöht sich die Zahl der effektiven Intervalle einer Menge nicht, falls man die Identität hinzufügt. Durch Lemma 3.8 erhalten wir:*

$$d = \mathrm{VCD}(\mathcal{G}) = \mathrm{VCD}\left(\{id_{\mathbb{R}}\} \circ \mathcal{G}\right) \leq \mathrm{VCD}\left((\mathcal{F} \cup \{id\}) \circ \mathcal{G}\right).$$

Da die hergeleiteten Grenzen in Satz 7.6 alle linear in d sind, sind diese auch scharf, bis auf Konstanten in Abhängigkeit von p und $\tilde{p}$*.*

Als zweites wollen wir die Konstante p betrachten. Dafür betrachten wir eine Menge an Funktionen $\mathcal{F}_+$, welche wir wie folgt konstruieren:

$$\mathcal{F}_{+,i} := \left\{ f : [0,p) \to \mathbb{R}, x \mapsto I_{[i-1,i)}(x) \cdot (ax+b), a, b \in \mathbb{R}, a \geq 0 \right\}$$
$$\mathcal{F}_+ := \left\{ \sum_{i=1}^{p} f_i, f_i \in \mathcal{F}_{+,i} \right\}.$$

Hierbei bezeichnen wir mit der Funktion $I : \mathbb{R} \to \{0,1\}$ die Identität.

Man sieht schnell, dass $\mathcal{F}_+$ die Monotoniebedinung für p Intervalle erfüllt. Sei $S = \{x_1, \ldots, x_p\} \subset \mathbb{R}$ eine Menge, sodass $x_i \in [i-1, i), 1 \leq i \leq p$ gilt und sei $S^+ \subset S$ eine beliebige Teilmenge. Wir wollen zeigen, dass S^+ durch $\mathcal{F}_+$ induziert wird. Hierfür konstruieren wir eine Funktion $f = \sum_{i=1}^{p} f_i \in \mathcal{F}_+$, mit $f_i \in \mathcal{F}_{+,i}$ für $1 \leq i \leq p$. Dafür wählen wir die Funktionen $f_1, \ldots, f_p$ wie folgt:

$$f_i := \begin{cases} I_{[i-1,i)}(x) \cdot (x+1), & x_i \in S^+, \\ I_{[i-1,i)}(x) \cdot (x-1), & x_i \notin S^+. \end{cases}$$

Mit $\beta = 0$ erhalten wir insgesamt:

$$x \in S \implies f(x) = \sum_{i=1}^{p} f_i(x) \geq 0 \Leftrightarrow x \in S^+,$$

also wird S^+ durch $\mathcal{F}_+$ induziert. Da $S^+ \subset S$ beliebig war, gilt:

$$p \leq \mathrm{VCD}(\mathcal{F}_+).$$

Somit erhalten wir für $\mathcal{G} = \{id_{\mathbb{R}}\}$:

$$p \cdot \mathrm{VCD}(\mathcal{G}) = p \leq VCD(\mathcal{F}_+) = \mathrm{VCD}(\mathcal{F} \circ \mathcal{G}).$$

Angenommen, die VC-Dimension von $\mathcal{G}$ wäre $d > 1$. Dann könnte es sein, dass für alle $1 \leq i \leq p$ Mengen S_i aus jeweils d Punkten existieren, sodass $\mathcal{G}(S_i) \in [i-1, i)$ gilt mit $\mathrm{VCD}(\mathcal{G}|_{S_i}) = d$. Für jede Teilmenge $S_i^+ \subset S_i$ existiert also eine Funktion $g_i \in \mathcal{G}$ und ein $\beta_i \in \mathbb{R}$, sodass gilt:

$$x \in S_i \implies g_i(x) - \beta_i \geq 0 \Leftrightarrow x \in S_i^+.$$

Verknüpft man die Menge $\mathcal{G}$ nun mit der Menge $\mathcal{F}_+$, so lassen sich die Punkte aus den Teilmengen S_i^+ miteinander vereinen und jede dieser vereinten Mengen lässt sich nun durch $\mathcal{F}_+ \circ \mathcal{G}$ induzieren, indem wir aus $\mathcal{F}_+$ die Funktion

$$f := \sum_{i=1}^{p} 1_{[i-1,i)}(x) \cdot (x - \beta_i)$$

betrachten.

Insgesamt können wir also S durch $\mathcal{F}_+ \circ \mathcal{G}$ splittern und erhalten:

$$p \cdot d = p \cdot \mathrm{VCD}(\mathcal{F}) \leq \mathrm{VCD}(\mathcal{F}_+ \circ \mathcal{G}).$$

Der Faktor c_2 aus Formelzeile (7.8) lässt sich nach Lemma 3.10 mit 2 abschätzen.

7.2 Die VC-Dimension für Bälle in der $\|\cdot\|_2$-Norm

In diesem Kapitel leiten wir die VCD von Bällen in der $\|\cdot\|_2$-Norm analog zu [Dud79] her. Zuerst wollen wir betrachten, dass die VCD von Bällen in der $\|\cdot\|_2$-Norm, also

$$B_\beta(t) := \{z \in \mathbb{R}^k : \|z - t\|_2 \leq \beta\}, \tag{7.9}$$

der VCD der Funktionen

$$\{f : \mathbb{R}^k \to \mathbb{R},\ x \mapsto \|z - t\|_2 : t \in \mathbb{R}^k\}$$

entspricht.

Lemma 7.7. *Sei*

$$\mathcal{C}_B := \{B_\beta(t), t \in \mathbb{R}^k, \beta \in \mathbb{R}_0^+\} \tag{7.10}$$

eine Konzeptmenge und

$$\mathcal{F}_B := \{f : \mathbb{R}^k \to \mathbb{R}, z \mapsto \|z - t\|_2 : t \in \mathbb{R}^k\},$$

dann gilt:

$$\mathrm{VCD}(\mathcal{C}_B) = \mathrm{VCD}(\mathcal{F}_B).$$

Beweis. Die VCD von $\mathcal{F}_B$ entspricht nach Definition 3.6 der VC-Dimension der Mengen

$$\mathcal{C}_{-B} := \left\{ \{z \in \mathbb{R}^k : \|z - t\|_2 - \beta \geq 0\} : t \in \mathbb{R}^k, \beta \in \mathbb{R} \right\}.$$

Analog zu Lemma 3.8 und Definition 3.5 für die VC-Dimension für Konzeptmengen gilt allerdings

$$\mathrm{VCD}(\mathcal{C}_B) = \mathrm{VCD}(-\mathcal{C}_B)$$

auch für die Konzeptmengen $\mathcal{C}_B, -\mathcal{C}_B = \mathcal{C}_{-B}$. □

Weiterhin benötigen wir den folgenden Satz, welcher auch als Radons Lemma bezeichnet wird. Dafür und für den Beweis beziehen wir uns auf [Mat13], Satz 1.3.1.

Satz 7.8. *Sei A eine Menge von $n+2$ Punkten in $\mathbb{R}^k$. Dann existieren zwei disjunkte Teilmengen P, N von A, sodass für deren konvexe Hüllen $conv(P), conv(N)$ gilt:*

$$conv(P) \cap conv(N) \neq \emptyset.$$

Bälle, welche Punkte aus einer Menge P oder N enthalten, enthalten auch deren konvexe Hüllen. Gleichzeitig schneiden sich zwei $\|\cdot\|_2$-Bälle aus $\mathbb{R}^k$ immer in einer Hyperebene, wie das nächste Lemma zeigt.

Lemma 7.9. *Seien $B, \widetilde{B}$ zwei verschiedene Bälle in der $\|\cdot\|_2$-Norm im Raum $\mathbb{R}^k$, mit Zentren $z = (z_1, \ldots, z_k), \tilde{z} = (\tilde{z}_1, \ldots, \tilde{z}_k)$ und Radien $r, \tilde{r}$. Dann liegt der Schnitt beider Mengen in einer Hyperebene aus $\mathbb{R}^k$.*

Beweis. Die Aussage ist klar für $B \cap \widetilde{B} = \emptyset$. Ansonsten besteht der Schnitt $B \cap \widetilde{B}$ aus den Punkten $x = (x_1, \ldots, x_k) \in \mathbb{R}^k$, welche die folgenden beiden Gleichungen erfüllen:

$$\begin{aligned} r =& (x_1 - z_1)^2 + \cdots + (x_k - z_k)^2, \\ \tilde{r} =& (x_1 - \tilde{z}_1)^2 + \cdots + (x_k - \tilde{z}_k)^2. \end{aligned}$$

Somit liegt der Schnitt auf der Hyperebene

$$\begin{aligned} \tilde{r} - r =& (2x_1 z_1 - z_1^2) + \cdots + (2x_k z_k - z_k^2) \\ &+ (-2x_1 \tilde{z}_1 + \tilde{z}_1^2) + \cdots + (-2x_k \tilde{z}_k + \tilde{z}_k^2). \end{aligned}$$

□

Satz 7.10. *Die VCD der Menge aller $\|\cdot\|_2$-Bälle in $\mathbb{R}^k$, definiert wie in den Formelzeilen (7.9) und (7.10), ist $k+1$.*

Beweis. Wir bemerken zuerst, dass nicht alle Mengen aus $k+1$ Punkte gesplittert werden können. Dies ist zum Beispiel der Fall, wenn die Punkte alle auf einer Hyperebene liegen.

Es existieren jedoch $k+1$ Punkte, welche gesplittert werden können. Sei e_i der i-te Einheitsvektor in $\mathbb{R}^k$, so bezeichne:

$$S := \{e_i, 1 \leq i \leq k\} \cup \{0\}.$$

Jede Teilmenge $S^+, \#S^+ \geq 2$ von S wird dabei durch einen Ball mit Zentrum

$$z_{S^+} = \sum_{a \in S^+} a$$

induziert. Dabei gilt für den Radius r, mit $n^+ = \#S^+$:

- Falls $0 \notin S^+$, so wähle $r_{S^+} := \sqrt{n^+ - 1}$.
- Falls $0 \in S^+$, wähle $r_{S^+} := \sqrt{n^+}$.

Da für alle Einheitsvektoren, welche nicht in S^+ liegen, der $\|\cdot\|_2$-Abstand zu z_S^+ genau $\sqrt{d+1}$ beträgt, induziert der Ball $\{x \in \mathbb{R}^k, \|x - z_{S^+}\|_2 \leq r_{S^+}\}$ die Menge S^+. Für den Fal $\#S^+ = 0$ betrachten wir als Beispiel den Ball mit Radius $\varepsilon < \frac{1}{2}$ und Zentrum $\sum_{i=1}^k e_i$. Für $\#S^+ = 1$ wählen einen Ball ebenfalls mit Radius ε um den einzigen Punkt der Menge S^+.

Nun zeigen wir, dass die VCD von solchen Bällen nicht $k+2$ betragen kann. Dazu wählen wir eine beliebige Menge $S := \{z_1, \ldots, z_{k+2}\}$ mit $k+2$ paarweise verschiedenen Elementen, welche gesplittert wird, und führen dies zum Widerspruch. Nach Satz 7.8 können wir die Menge S in zwei disjunkte Mengen P, N einteilen, deren konvexe Hüllen sich schneiden. Seien B_P und B_N die beiden Mengen, welche P bzw. N induzieren. Da $\text{conv}(P) \subset B_P$, sowie $\text{conv}(N) \subset B_N$ gilt, haben auch B_N und B_P eine gemeinsame Schnittmenge $\neq \emptyset$, welche nicht nur aus einem Punkt besteht und damit nach Lemma 7.9 eine Hyperebene beschreibt. Angenommen, diese Hyperebene trennt die Punkte aus P von den Punkten aus N, dann existieren Bälle mit hinreichend großem Radius, welche die Menge P und N induzieren und sich nicht schneiden. Dies widerspricht allerdings Satz 7.8, da die konvexen Hüllen der Punkte aus P und N Teilmengen dieser Bälle sind. Somit kann die VCD der $\|\cdot\|_2$-Bälle höchstens $k+1$ betragen. □

Der Beweis liefert auch eine untere Grenze $k+1$ für alle p-Normen mit $1 \leq p < \infty$. Dann wählen wir für für die besprochenen Mengen S, S^+ einen Ball mit Zentrum z_{S^+} und Radius $r_{S^+,p} := r_{S^+}^{2/p}$.

7.3 Anwendung auf radiale Kerne

Unsere Ergebnisse aus den Sätzen 7.6 und 7.10 wollen wir nun nutzen, um eine Abschätzung für radiale Kerne in der Form von Satz 6.16 zu erhalten.

Satz 7.11. *Sei im Folgenden $Z \subset \mathbb{R}^k$ und $\lambda \in \mathcal{L}_1(Z)$ mit $\lambda \geq 0$ fast überall und $\|\lambda\|_1 \neq 0$. Sei $\mathcal{M}_p$ eine Teilmenge aus $[0,\infty)^Z$, welche die Monotoniebedingung für p Intervalle aus Definition 7.2 erfüllt. Weiterhin habe die Menge $\mathcal{F}$ insgesamt $\tilde{p}$ effektive Intervalle nach Definition 7.3. Dann sei $\mathcal{F}$ die folgende Menge an radialen Kernen:*

$$\mathcal{F} := \{\phi(\|\cdot - t\|_2) : Z \to \mathbb{R}, \qquad \phi \in \mathcal{M}_p, t \in Z\}.$$

Dann existiert eine Konstante $c_{p,2}$ abhängig von p und eine Konstante C unabhängig von $\mathcal{F}$ und λ, sodass gilt:

1. *Es existieren Punkte $\tilde{z}_1, \ldots, \tilde{z}_n \in Z$, sodass gilt:*

$$\begin{aligned}&\sup_{\Phi(\cdot,t)\in\mathcal{F}} \left| \int_Z \Phi(z,t)\lambda(z)dz - \frac{1}{n}\sum_{i=1}^{n} \Phi(\tilde{z}_i,t)\|\lambda\|_1 \right| \\ \leq &(B-A)C\|\lambda\|_1 \sqrt{\frac{\mathrm{VCD}(\mathcal{F})}{n}}.\end{aligned}$$

2. *Mit einer Wahrscheinlichkeit von mindestens $1-\delta$ findet man Punkte $z_1, \ldots z_n \in Z$, sodass gilt:*

$$\begin{aligned}&\sup_{\Phi(\cdot,t)\in\mathcal{F}} \left| \int_Z \Phi(z,t)\lambda(z)dz - \frac{1}{n}\sum_{i=1}^{n} \Phi(z_i,t)\|\lambda\|_1 \right| \\ \leq &(B-A)C\|\lambda\|_1 \frac{1}{\sqrt{n}} \max\left\{ \sqrt{VCD(\mathcal{F})}, -\frac{\ln(\delta)}{\sqrt{n}}, \sqrt{-\ln(\delta)} \right\}.\end{aligned}$$

Dabei gilt:

$$\mathrm{VCD}(\mathcal{F}) \leq \begin{cases} 2k+3, & p = 1, \\ c_{p,2}(2k+3), & p > 1. \end{cases}$$

Sind für $p = 1$ alle betrachteten Funktionen aus $\mathcal{M}_1$ monoton steigend, oder sind alle Funktionen aus $\mathcal{M}_1$ monoton fallend, so folgt:

$$\mathrm{VCD}(\mathcal{F}) \leq k + 1. \tag{7.11}$$

Die Konstante $c_{p,2}$ steigt dabei nach Lemma 7.5 höchstens mit Rate $p \log(p)$ in p.

Beweis. Wir wollen Satz 6.16 anwenden. Hierfür sind die Voraussetzungen erfüllt, wir müssen nur die VC-Dimension der Menge $\mathcal{F}$ bestimmen. Hierfür schreiben wir analog zur Formelzeile (7.2) $\mathcal{F} = \mathcal{M}_p \circ \mathcal{F}_2$, mit:

$$\mathcal{F}_2 := \{f : \mathbb{R}^k \to \mathbb{R}_0^+, t \mapsto \|z - t\|_2 : z \in \mathbb{R}^k\}.$$

Wir erfüllen hiermit die Voraussetzungen für Satz 7.6 Es gilt nach Satz 7.10:

$$\mathrm{VCD}(\mathcal{F}_2) = k + 1,$$

somit erhalten wir:

$$\mathrm{VCD}(\mathcal{F}) \leq \begin{cases} 2k + 3, & p = 1, \\ c_{p,2}(2(k+1)+1), & p > 1. \end{cases}$$

Der Spezialfall für $p = 1$ aus Formelzeile (7.11) folgt dabei aus der Formelzeile (7.7). □

Somit haben wir eine Abschätzung für alle Funktionen radialen Kernen aus der Darstellung (7.1), bei denen ϕ auf höchstens p Intervallen $\mathcal{I}_1 \cup \ldots \cup \mathcal{I}_p = \mathbb{R}_+^0$ jeweils monoton ist. Allerdings haben wir keine Abschätzung für radiale Funktionen, bei denen ϕ unendlich oft steigt und fällt, wie zum Beispiel die radiale Poisson-Funktion für $k = 3$, entnommen aus [Fas07], Kapitel 4.3:

$$\Phi_3(z - t) = \sqrt{\frac{2}{\pi}} \frac{\sin\left(\|z - t\|_2\right)}{\|z - t\|_2}.$$

7.3.1 Verbindung zu positiv definiten Kernen

Die Abschätzungen aus Satz 7.11 gelten mit $\mathrm{VCD}(\mathcal{F}) \leq k + 1$ für Kerne mit monotonen $\phi(\cdot)$. Wir zeigen hier, dass radiale, positiv-definite Kerne diese Eigenschaft erfüllen. Positiv definite Kerne eignen sich numerisch besonders

gut, das lineare Problem aus (7.12) zu lösen. Dafür führen wir zuerst die Definitionen zu *positiv (semi-)definit* und *komplett monoton* ein, analog zu [Wen05], Definitionen 6.1, 6.24, 7.1.

Als Motivation betrachten wir eine feste Funktion $f : \mathbb{R}^k \to \mathbb{R}$. Hierfür wollen wir die Punkte

$$(z_i, f(z_i)) \in \mathbb{R}^k \times \mathbb{R}, 1 \leq i \leq n$$

durch die Summe

$$\sum_{i=1}^{n} \alpha_i \Phi(z - z_i)$$

mit einer radialen Funktion $\Phi : \mathbb{R}^k \to \mathbb{R}$ interpolieren. Dies führt zu dem linearen Problem

$$(\Phi(z_i, z_j))_{1 \leq i,j \leq n} \alpha = f(z), \tag{7.12}$$

mit $\alpha = (\alpha_1, \ldots, \alpha_n)^T, f(z) = (f(z_1), \ldots, f(z_n))^T$. Mehr Details finden sich zum Beispiel in [Wen05], Kapitel 1 und 6.

Definition 7.12. *Eine Funktion $\Phi : \mathbb{R}^k \to \mathbb{R}$ ist positiv semi-definit, falls für alle $n \in \mathbb{N}$ und paarweise verschiedenen $z_1, \ldots, z_n \in \mathbb{R}^k$ und $\alpha \in \mathbb{R}^n$ gilt:*

$$\sum_{j=1}^{n} \sum_{i=1}^{n} \alpha_j \alpha_i \Phi(z_j - z_k) \geq 0.$$

Definition 7.13. *Ein stetiger Kern $\Phi : Z \times Z \to \mathbb{R}$ heißt positiv definit auf $Z \subset \mathbb{R}^k$, falls für alle $n \in \mathbb{N}$ und alle paarweise verschiedenen $z_1, \ldots, z_n \in Z$ und $\alpha \in \mathbb{R}^n \backslash \{0\}$ gilt:*

$$\sum_{j=1}^{n} \sum_{i=1}^{n} \alpha_j \alpha_i \Phi(z_j, z_k) > 0. \tag{7.13}$$

Definition 7.14. *Eine Funktion $\phi : \mathbb{R} \to \mathbb{R}$ heiß komplett monoton auf $[0, \infty)$, falls $\phi \in C^\infty(0, \infty) \cap C[0, \infty)$ und für alle $l \in \mathbb{N}$ gilt:*

$$(-1)^l \phi^l(r) \geq 0.$$

Dabei beschreibt $\phi^l(r)$ die l-te Ableitung der Funktion ϕ.

Eine komplett monotone Funktion ist insbesondere monoton fallend. Komplette Monotonie und positive Definitheit von Funktionen haben den folgenden Zusammenhang, welcher auf Schoenberg zurückgeht und in [Wen05], Satz 7.13 und 7.14 zu finden ist.

Satz 7.15. *Eine Funktion $\phi : \mathbb{R} \to \mathbb{R}$ ist auf $[0, \infty)$ komplett monoton genau dann, wenn $\Phi := \phi(\|\cdot\|_2^2)$ auf $\mathbb{R}^k$ für alle $k \in \mathbb{N}$ positiv semi-definit ist.*

Satz 7.16. *Für eine Funktion $\phi : \mathbb{R}_0^+ \to \mathbb{R}$ ist äquivalent:*

1. *$\phi(\|\cdot\|_2)$ ist positiv definit auf $\mathbb{R}^k$ für alle $k \in \mathbb{N}$,*
2. *$\phi(\sqrt{\cdot})$ ist komplett monoton auf $\mathbb{R}_0^+$ und nicht-konstant.*

Diese Sätze ermöglichen es, die positive Definitheit von Funktionen zu bestimmen. Für uns ergibt es sich, dass die Abschätzung aus Satz 7.11 mit $p = 1$ mit monoton fallenden Funktionen für alle radiale Kerne gilt, welche für alle Raumdimensionen $k \in \mathbb{N}$ positiv definit sind.

Korollar 7.17. *Sei $\Phi(z, t) := \phi(\|z - t\|_2)$ ein positiv definiter Kern auf $\mathbb{R}^k \times \mathbb{R}^k$ für alle $k \in \mathbb{R}^k$, dann ist die Funktion $\phi : \mathbb{R}_0^+ \to \mathbb{R}$ monoton fallend.*

Beweis. Nach Satz 7.16 ist die Funktion $\phi(\sqrt{\cdot})$ monoton fallend. Da die Funktion $q : \mathbb{R}_0^+ \to \mathbb{R}_0^+, x \mapsto x^2$ monoton steigt, ist auch die Funktion $\phi(\sqrt{\cdot}) \circ q$ monoton fallend, also

$$\phi(\sqrt{\cdot}) \circ q(x) = \phi(\sqrt{q(x)}) = \phi(x).$$

□

7.4 Beispiele für radiale Kerne

In diesem Teilkapitel wollen wir einige radiale Kerne vorstellen, für welche wir die Abschätzung aus Satz 7.11 nutzen können.

7.4.1 Wendland-Kerne

Die sogenannten Wendland-Kerne $\phi_{k,m}$ definieren wir analog zu [Wen05], Definitionen 9.4 und 9.11.

Definition 7.18. *Sei $\phi : \mathbb{R}_0^+ \to \mathbb{R}_0^+, x \mapsto x\phi(x)$ aus $L_1[0,\infty)$, dann definieren wir für $r \geq 0$:*

$$\mathcal{J}\phi(r) := \int_r^\infty x\phi(x)dx.$$

Definition 7.19. *In $\mathbb{R}^k$ definieren wir mit $\phi_l(r) = (1-r)_+^l$ definieren wir:*

$$\phi_{k,m} := \mathcal{J}^m \phi_{\lfloor k/2 \rfloor + m + 1}.$$

Dabei nutzen wir die folgende Schreibweise:

$$(1-r)_+ := \max\{1-r, 0\}.$$

Offensichtlich sind die Funktionen $\mathcal{J}\phi(r)$ monoton fallend, da $\phi_l(\cdot)$ nichtnegativ ist. Damit können wir die Wendland-Kerne in Satz 7.11 verwenden, mit der Monotoniebedingung für $p = 1$ Interval auf dem alle Funktionen monoton fallen und dem Wertebereich $[A, B] = [0, 1]$.

Wie wir im nächsten Satz sehen werden, besitzen die Wendland-Kerne einen kompakten Träger und lassen sich abschnittsweise als Polynom darstellen. Dadurch sind diese besonders gut für Anwendungen geeignet, vgl. z.B. Kapitel 6 in [Buh03]. Dabei folgen wir [Wen05], Satz 9.13.

Satz 7.20. *Die Wendland-Kerne $\phi_{k,m}(r)$ sind positiv definit auf $\mathbb{R}^k$ und $2m$-mal stetig differenzierbar. Sie besitzen die Form*

$$\phi_{k,m}(r) = \begin{cases} p_{k,m}(r), & 0 \leq r \leq 1, \\ 0, & r > 1. \end{cases}$$

Dabei ist $p_{k,m}$ ein Polynom vom Grad $\lfloor k/2 \rfloor + 3m + 1$. Gegeben diesen Eigenschaften sind die Wendland-Kerne von minimalen Grad.

Eine weitere wichtige Menge sind die Gauß'schen RBF-Kerne. Als Spezialfall ergeben sich auch die Matérn-Kerne.

7.4.2 Gaußsche RBF-Kerne

Ein weiteres Beispiel für radiale Funktionen sind die sogenannten Gaußschen RBF-Kerne, definiert wie in [SC08], Prop. 4.10 und [Wen05], Satz 6.10:

$$K_\gamma(z,t) := e^{-\gamma^2 \|z-t\|_2^2}, z, t \in \mathbb{R}^k.$$

Diese sind nach Satz 6.10 in [Wen05] positiv definit auf $\mathbb{R}^k$ für alle $k \in \mathbb{N}$. Demnach sind diese nach Korollar 7.17 auch monoton und damit gilt auch für die Gaußschen RBF-Kerne für alle $\gamma > 0$ die Abschätzung aus Satz 7.11, mit der Monotoniebedingung für $p = 1$ Interval, auf welchem die Funktionen monoton fallen und dem Wertebereich $[A, B] = [0, 1]$.

7.4.3 Matérn-Kerne

Die Matérn-Kerne definieren wir analog zu [Fas07]. Zuerst definieren wir hierzu die modifizierte Bessel-Funktion der dritten Art von Ordnung $\alpha \in \mathbb{R}$, analog zu [Wen05], Definition 5.10.

Definition 7.21. *Die modifizierte Bessel-Funktion der dritten Art ist für $\alpha \in \mathbb{R}$ beschrieben durch*

$$K_\nu(r) := \int_0^\infty e^{-r\cosh(t)} \cosh(\alpha t) dt. \tag{7.14}$$

Der Hyperbelkosinus besitzt dabei die Darstellung $\cosh(x) = e^x + e^{-x}$, insbesondere folgt hieraus also

$$K_\nu(r) = K_{-\nu}(r). \tag{7.15}$$

Für die Funktionen K_ν gilt auch das nächste Lemma, analog zu [Wen05], Proposition 5.11 und Lemma 5.14.

Lemma 7.22. *Die Funktionen $K_\nu(r)$ besitzen für $\nu \geq 0$ die Integraldarstellung*

$$K_\nu(r) = \frac{\pi}{2r}^{\frac{1}{2}} \frac{e^{-r}}{\Gamma(\nu + \frac{1}{2})} \int_0^\infty e^{-t} t^{\nu - \frac{1}{2}} \left(1 + \frac{t}{2r}\right)^{\nu - \frac{1}{2} dt}.$$

Dabei ist $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ die Γ-Funktion. Weiterhin gilt für $r > 0$ die obere Schranke:

$$K_\nu(r) \leq 2^{\nu-1} \Gamma(\nu) r^{-\nu}.$$

Damit erhalten wir die Matérn-Kerne.

Definition 7.23. *Sei K_ν für $\nu > 0$ beschrieben wie in Formelzeile (7.14). Dann sind die Matérn-Kerne $\Phi_M : \mathbb{R}^k \to \mathbb{R}$ wie folgt definiert:*

$$\Phi_{M,\beta}(x) := \frac{K_{\beta - \frac{k}{2}}(\|x\|_2) \|x\|_2^{\beta - \frac{k}{2}}}{2^{\beta-1}\Gamma(\beta)}, \beta > \frac{k}{2}. \tag{7.16}$$

Oft werden die Matérn-Kerne auch durch deren Fourier-Transformierte

$$\widehat{\Phi}_{M,\beta}(\omega) = (1 + \|\omega\|_2^2)^{-\beta}$$

angegeben, vgl. [Ste70], Proposition V.3.2.

Um zu zeigen, dass die Matérn-Kerne monoton fallend sind, wiederholen wir Korollar 5.12 aus [Wen05].

Lemma 7.24. *Sei $\nu \in \mathbb{R}$ und $x > 0$, dann ist die Funktion $x \mapsto x^\nu K_{-\nu}(x)$ auf $(0, \infty)$ monoton fallend.*

Damit erhalten wir das folgende Lemma, mit welchem sich die Matérn-Kerne auch für die Grenze aus Satz 7.11 nutzen lassen, mit $[A, B] = [0, 1]$.

Lemma 7.25. *Die Matérn-Kerne sind monoton fallend auf $(0, \infty)$. Weiterhin sind die Matérn-Kerne für alle $\beta > \frac{k}{2}$ immer größer als null und nach oben durch 1 beschränkt.*

Beweis. Aus der Formelzeile (7.15) folgt $K_\nu = K_{-\nu}$ und somit mit Lemma 7.24 die Monotonie der Matérn-Kerne.

Die Eigenschaft $\Phi_M(x) > 0$ auf $(0, \infty)$ folgt direkt aus der Darstellung aus Formelzeile (7.16). Es bleibt zu zeigen, dass die Funktionen $\Phi_M(r)$ nach oben durch 1 beschränkt sind. Aus Lemma 7.22 folgt:

$$\begin{aligned}\Phi_{M,\beta}(x) &\leq \frac{2^{\beta-\frac{k}{2}-1}\Gamma(\beta-\frac{k}{2})\|x\|_2^{-\left(\beta-\frac{k}{2}\right)}\|x\|_2^{\beta-\frac{k}{2}}}{2^{\beta-1}\Gamma(\beta)} \\ &\leq \frac{2^{-\frac{k}{2}}\Gamma(\beta-\frac{k}{2})}{\Gamma(\beta)}.\end{aligned}$$

Da $k \geq 1$ gilt und die Γ-Funktion monoton steigend ist, folgt die Behauptung. □

Somit erhalten wir die Aussage aus Satz 7.11 auch für die Matérn-Kerne mit der Monotoniebedingung für $p = 1$ Interval, auf welchem die Funktionen monoton fallen und dem Wertebereich $[A, B] = [0, 1]$.

Für die Matérn-Kerne wurde bereits in [Gir95] eine Abschätzung dieser Form hergeleitet, welche später in [GS08], Korollar 5.1, auf eine Abschätzung ähnlich zu der in Abschätzung in Satz 7.11 aus dieser Arbeit verbessert wurde. In unserer Arbeit ist die linke Seite der Abschätzung aus

Satz 7.11 gegeben durch $(B-A)C\|\lambda\|_1\sqrt{\frac{k+1}{n}}$, in [GS08] leicht schlechter mit $(B-A)C\|\lambda\|_1\sqrt{\frac{k+3}{n}}$.

Wir wollen kurz auf die verwendete Beweistechnik in [GS08] eingehen. Dort wird die Darstellung

$$\Phi_{M,\beta/2}(t) = \frac{1}{2^{\frac{k}{2}}}\frac{1}{\Gamma(\beta)}\int_0^\infty s^{\beta+\frac{-k}{2}-1}e^{-\frac{\|t\|^2}{4s}}e^{-s}ds \tag{7.17}$$

benutzt, welche mit $\beta = \frac{\alpha}{2}$ aus [Ste70], Zeile (26) in Kapitel V.3 entnommen ist.

Dann betrachtet man die Faltung $\Phi_{M,\beta/2} * \lambda$, mit einer nichtnegativen Funktion $\lambda \in \mathcal{L}_1(\mathbb{R}^k)$, $\|\lambda\|_1 \neq 0$, und erhält:

$$f(z) = \frac{1}{2^{\frac{k}{2}}}\frac{1}{\Gamma(\beta)}\int_0^\infty \int_{\mathbb{R}^k} e^{-\frac{\|z-t\|_2^2}{4u}}\Lambda(u,t)dtdu.$$

Dabei werden in der Funktion $\Lambda(u,t)$ alle Terme aus (7.17), sowie die Funktion λ, zusammengefasst, welche über $\mathbb{R}_0^+ \times \mathbb{R}$ $\mathcal{L}_1$-integrierbar sind. Somit muss nur die VCD der Menge

$$\left\{ f(\cdot_z, \cdot_u) = e^{-\frac{\|\cdot_z - t\|_2^2}{4\cdot_u}}, t \in \mathbb{R}^k \right\}$$

berechnet werden. Dies lässt sich mit Hilfe von Lemma 3.12 aus dieser Arbeit bestimmen. Deswegen können wir den Fall der Matérn-Kerne als Spezialfall der Gaußschen RBF-Kerne in Bezug auf unsere Abschätzung in Satz 7.11 ansehen.

7.4.4 Laguerre-Gauß-Kerne

Ein weiteres Beispiel sind die Laguerre-Gauß-Kerne, wie in [Fas07], Kapitel 4.2. Diese sind auf $\mathbb{R}^k$ über die verallgemeinerten Laguerre-Polynome

$$L_m^{\frac{k}{2}}(t) = \sum_{i=0}^{n} \frac{(-1)^i}{i!}\binom{m+\frac{k}{2}}{m-i}t^i \tag{7.18}$$

definiert. Oftmals werden diese auch über Rodrigues Formel angegeben:

$$L_m^{\frac{k}{2}}(t) = \frac{e^t t^{-\frac{k}{2}}}{m!}\frac{d^m}{dt^m}\left(e^{-t}t^{m+\frac{k}{2}}\right).$$

Die Laguerre-Gaussian-Funktionen $\Phi_{L,m}(x) : \mathbb{R}^k \to \mathbb{R}$ sind dann wie folgt definiert:

$$\Phi_{L,m}(z) = \phi_{L,m}(\|z\|_2) = e^{-\|z\|_2^2} L_m^{\frac{k}{2}}(\|z\|_2^2).$$

Offensichtlich sind die Funktionen nicht-negativ. Weiterhin sind diese nach oben beschränkt, da diese stetig sind und die Funktion e^{-r} schneller abfällt als jedes Polynom steigt. Wir bezeichnen die obere Grenze mit B.

Um Satz 7.11 anzuwenden, müssen wir bestimmen, auf wie vielen Intervallen die Funktion $\phi_{L,m}$ monoton ist. Da die Funktion $r \mapsto r^2$ auf $\mathbb{R}_0^+$ monoton ist, können wir auch

$$\phi_{L,m}(\sqrt{r}) = e^{-r} L_m^{\frac{k}{2}}(r)$$

betrachten, um das Monotonieverhalten von $\phi_{L,m}(r)$ zu analysieren. Durch die Darstellung der Polynome $L_m^{\frac{k}{2}}(r)$ aus Formelzeile (7.18) folgt, dass die erste Ableitung der Funktion $\phi_{L,m}(\sqrt{r})$ ein Polynom vom Grad m ist, multipliziert mit dem positiven Term e^{-r}. Demnach existieren höchstens $m+1$ maximal groß gewählte Intervalle $I_1 \cup \ldots \cup I_{m+1} = \mathbb{R}_0^+$, auf welchen die Funktion $\phi_{L,m}(r)$ monoton ist. Weiterhin ist die Funktion $\phi_{L,m}(r)$ stetig, somit folgt auf jedes maximal groß gewählte Intervall, auf der die Funktion $\phi_{L,m}(r)$ monoton steigt, ein Intervall, auf dem die Funktion $\phi_{L,m}(r)$ monoton fällt. Durch den dritten Punkt in Lemma 3.8 können wir annehmen, dass es mindestens so viele Intervalle aus $\mathcal{I}_L$ existieren, auf denen $\phi_{L,m}(r)$ monoton steigt, wie Intervalle, auf denen $\phi_{L,m}(r)$ monoton fällt.

Damit erhalten wir die Aussage aus Satz 7.11 mit der Monotoniebedingung für $p = m+1$ Intervalle, $\tilde{p} := \lceil \frac{m+1}{2} \rceil$ effektive Intervalle und dem Wertebereich $[0, B]$.

7.5 Span von Funktionen

Im folgenden Teilkapitel wollen wir eine Grenze analog zu Satz 6.16 für den Span von Funktionen herleiten. Sei hierzu für feste Funktionen $f_i \in \mathbb{R}^{\mathbb{R}^k}, 1 \leq i \leq m$ und einer oberen Grenze $M \in \mathbb{R}$:

$$\mathcal{S} := \left\{ \sum_{i=1}^{m} a_i f_i(z) : a_1, \ldots, a_m \in \mathbb{R} \right\},$$

$$\mathcal{S}_M := \left\{ \sum_{i=1}^{m} a_i f_i(z) : \sum_{i=1}^{m} |a_i| \leq M, a_1, \ldots, a_m \in \mathbb{R} \right\}.$$

Für einen Span an Funktionen sind die VC-Dimension und die Rademacher-Komplexität unabhängig von der Raumdimension k.

Satz 7.26. *Seien $f_1, \dots, f_m \in [A,B]^Z, f_{1,M}, \dots, f_{m,M} \in \mathbb{R}^{\mathbb{R}^k}$ beliebige, aber feste Funktionen, $Z \subset \mathbb{R}^k, A, B \in \mathbb{R}$, $M \in \mathbb{R}$ fest und $\mathcal{S}, \mathcal{S}_M$ wie oben definiert. Dann existieren für jedes $n \in \mathbb{N}$ und für jedes nicht-negative $\lambda \in \mathcal{L}_1(Z), \|\lambda\|_1 \neq 0$ Punkte $\tilde{z}_1, \dots, \tilde{z}_n \in Z$, sodass gilt:*

$$\sup_{f \in \mathcal{S}} \left| \int_Z f(z)\lambda(z)dz - \frac{1}{n}\sum_{i=1}^{n} f(\tilde{z}_i)\|\lambda\|_1 \right| \leq (B-A)C\|\lambda\|_1 \sqrt{\frac{m+1}{n}}$$

Falls die obere Grenze für die ℓ_2-Norm der Werte $a_i, \dots, a_m$ mit M gegeben ist, so gilt:

$$\begin{aligned}&\sup_{f \in \mathcal{S}_M} \left| \int_Z f(z)\lambda(z)dz - \frac{1}{n}\sum_{i=1}^{n} f(\tilde{z}_i)\|\lambda\|_1 \right| \\ \leq &(B-A)C\|\lambda\|_1 M \max\{|A|,|B|\}\sqrt{\frac{2\ln(2m)}{n}}.\end{aligned}$$

Ebenfalls gelten diese Aussagen auch wieder in Wahrscheinlichkeit, analog zu Satz 6.16

Beweis. Für den ersten Fall erhalten wir die Abschätzung mit Lemma 3.12. Demnach ist die VC-Dimension des Spans der m Funktionen $f_1, \dots, f_m$ höchstens $m+1$. Dies können wir in die Schranke aus Satz 6.16 einsetzen und erhalten die gewünschte Abschätzung.

Für den zweiten Fall schreiben wir die Funktionen aus $\mathcal{S}_M$ um, also:

$$f = \sum_{i=1}^{m} a_i f_i(z) = \sum_{i=1}^{m} \frac{a_i}{M}(Mf_i)(z).$$

Somit ist der Span S_M die konvexe Hülle der Funktionen $Mf_1, \dots, Mf_m$, vergleiche Formelzeile (6.4). Mit Satz 6.5 erhalten wir damit:

$$\mathcal{R}(S_M) = \mathcal{R}(\text{conv}\,(\{Mf_1, \dots, Mf_m\})) = M\mathcal{R}(\{f_1, \dots, f_M\}).$$

Darauf können wir nun Satz 6.9 anwenden. Dabei verwenden wir die obere Schranke $|f(z)| \leq \max\{|A|,|B|\}$.:

$$\mathcal{R}(\{f_1, \dots, f_M\}) \leq \max\{|A|,|B|\}\sqrt{2\ln(2m)}.$$

Mit Hilfe von Satz 6.13 erhalten wir die Behauptung.

□

Als Beispiel für den Span von Funktionen dient der Raum $\pi_{\tilde{m}}(\mathbb{R}^k)$ der Polynome in $\mathbb{R}^k$ mit absolutem Grad höchstens $\tilde{m}$. Nach [Wen05], Satz 2.5, ist die Dimension des Vektorraums dieser Polynome höchstens $m := \binom{\tilde{m}+k}{k}$. Sei M wieder eine obere Schranke für die absolute Summe der Koeffizienten $\sum_{i=1}^{m} |a_i|$, so erhalten wir die Abschätzungen aus Satz 7.26 für die Polynome aus $\pi_{\tilde{m}}(\mathbb{R}^k)$.

Ebenfalls kann man so, analog zu Kapitel 6.1 in [Wen05], radiale Funktionen mit Zentren $t_1, \ldots, t_m \in \mathbb{R}^k$ und einer festen Funktion $\phi : \mathbb{R}_0^+ \to \mathbb{R}$ betrachten:

$$\mathcal{S} := \left\{ \sum_{i=1}^{m} a_i f_i(z) : a_1, \ldots, a_m \in \mathbb{R}, f_i(z) := \phi(\|z - t_i\|_2) \right\}.$$

Die Funktionen $\phi(\|z - t_1\|_2), \ldots, \phi(\|z - t_m\|_2)$ bilden einen Vektorraum mit Dimension höchstens m und damit lässt sich hierfür Satz 7.26 anwenden.

Fazit

In diesem Abschnitt wollen wir die zentralen Punkte dieser Arbeit hervorheben und auf weiterführende Fragen aufmerksam machen. Wir haben in dieser Arbeit eine Menge radiale Funktionen

$$\mathcal{F} = \{\phi(\|\cdot - t\|_2) : Z \to [A, B], t \in Z \subset \mathbb{R}^k, \phi \in \mathcal{M}_p\}$$

betrachtet und hierfür den Term

$$\sup_{\phi(\|\cdot - t\|_2) \in \mathcal{F}} \left| \int_Z \phi(\|z - t\|_2)\lambda(z)dz - \frac{1}{n}\sum_{i=1}^{n} f(z_i)\|\lambda\|_1 \right| \tag{8.1}$$

mit Beweismethoden der statistischen Lerntheorie nach oben abgeschätzt. Dabei haben wir die Funktion $\lambda \in \mathcal{L}_1(Z)$ durch dessen Norm geteilt um statt der Faltung $\int_Z \phi(\|z - t\|_2)\lambda(z)dz$ den Erwartungswert von $\phi(\|z - t\|_2)$, berechnet auf dem induzierten Wahrscheinlichkeitsmaß $\frac{\lambda(z)}{\|\lambda\|_1}$, zu betrachten. In der statistischen Lerntheorie, vgl. Kapitel 1.1, war dieser notwendig, um das Risiko gegen das Bayes-Risiko

$$\begin{aligned} R(f) - R^* &= (R(f) - R(\hat{f})) + (R(\hat{f}) - R_{\mathcal{F}}) + (R_{\mathcal{F}} - R^*) \\ &\leq (R(f) - R(\hat{f})) + 2\sup_{f \in \mathcal{F}} |\widehat{R}(f) - R(f)| + (R_{\mathcal{F}} - R^*). \end{aligned}$$

abzuschätzen, insbesondere den Bewertungsfehler im mittleren Term.

Den Term (8.1) konnten wir in Satz 6.16 nur durch die Nutzung der VC-Dimension nach oben abschätzen. Hierbei haben wir eine gleichmäßige Schranke erhalten, welche für alle Funktionen aus der Menge $\mathcal{F}$ gleichmäßig auf denselben Punkten $z_1, \ldots, z_n$ gilt, wofür wir nur die VC-Dimension von $\mathcal{F}$ ausrechnen mussten. Eine solche Schranke konnten wir auf zwei Arten herleiten, einmal über die Wachstumsfunktion in den Kapiteln 2 bis 4 und einmal durch die Rademacher-Komplexität in Kapitel 6, welche wir beide nach oben durch die VC-Dimension abschätzen konnten. Im Fall der Wachstumsfunktion haben wir die Hoeffding-Ungleichung aus Satz 2.1 mit der Benferroni-Ungleichung aus Formelzeile (2.12) auf die Menge an Funktionen $\mathcal{F}$ angewendet. Dafür war es notwendig, die Anzahl der betrachteten Funktionen zu beschränken. Hierzu haben wir das Bild der Funktionen auf $\{0, 1\}$ eingeschränkt und die Auswertungen an nur $2n$ Punkten durchgeführt, um höchstens 2^{2n} mögliche Funktionen zu betrachten. Falls die VC-Dimension

der Menge $\mathcal{F}$ höchstens d war, konnten wir mit dem Sauer-Shelah-Lemma aus Satz 3.7 zeigen, dass wir höchstens $\left(\frac{2en}{d}\right)^d$ Funktionen betrachten, also polynomial viele Funktionen in Abhängigkeit der Anzahl der Punkte n. Wie in der Besprechung nach Formelzeile (2.4) war dies ausreichend, um den Bewertungsfehler für $n \to \infty$ gegen 0 gehen zu lassen. Insgesamt konnten wir zeigen, dass mit einer Wahrscheinlichkeit von mindestens $(1-\delta)$ gilt:

$$\sup_{f\in\mathcal{F}} \left| R(f) - \widehat{R}(f) \right| \leq (B-A)\sqrt{8\frac{d\ln\left(\frac{2en}{d}\right) - \ln\left(\frac{\delta}{4}\right)}{n}}.$$

Für die Abschätzung mit der Rademacher-Komplexität haben wir in Satz 6.9 eine modifizierte Talagrand-Ungleichung betrachtet und dies mit Hilfe eines Symmetrisierungsarguments in Satz 6.10 auf eine Schranke in Abhängigkeit der Rademacher Komplexität $\mathcal{R}(\mathcal{F})$ gebracht, sodass mit einer Wahrscheinlichkeit von mindestens $(1-\delta)$ gilt:

$$\sup_{f\in\mathcal{F}} |R(f) - \widehat{R}(f)| \leq C\frac{1}{\sqrt{n}} \max\left\{ \mathcal{R}_n(\mathcal{F}), -\frac{\ln(\delta)}{\sqrt{n}}, \sqrt{-\ln\delta} \right\}.$$

Die Rademacher-Komplexität konnten wir nach oben durch die VC-Dimension abschätzen und haben insgesamt eine schärfere Schranke, im Vergleich zur ersten Möglichkeit, hergeleitet. Insgesamt haben wir in Satz 6.16 erhalten, dass wir für alle $\delta \in (0,1)$ mit einer Wahrscheinlichkeit von mindestens $1-\delta$ Punkte $z_1, \ldots, z_n$ finden, sodass gilt:

$$\begin{aligned} &\sup_{f\in\mathcal{F}} \left| \int_Z f(z)\lambda(z)dz - \frac{1}{n}\sum_{i=1}^{n} f(z_i)\|\lambda\|_1 \right| \\ \leq &C(B-A)\frac{1}{\sqrt{n}} \max\left\{ \sqrt{\mathrm{VCD}(\mathcal{F})}, -\frac{\ln(\delta)}{\sqrt{n}}, \sqrt{-\ln\delta} \right\}. \end{aligned} \tag{8.2}$$

Dies bedeutet insbesondere, dass wir für alle $\delta \to 1$ auch Punkte $\tilde{z}_1, \ldots, \tilde{z}_n$ finden können, sodass die Grenze (8.2) erreicht ist.

Um dies auf radiale Funktionen anzuwenden, mussten wir die VC-Dimension der radialen Funktionen $\phi(\|z-t\|_2) \in \mathcal{M}_p$ berechnen. Ein wichtiges Resultat in dieser Arbeit ist Satz 7.6, mit welchem wir gesehen haben, dass es für radiale Funktionen genügt, nur die VC-Dimension der Menge

$$\{f : \mathbb{R}^k \to \mathbb{R}, z \mapsto \|z-t\|_2, \qquad t \in \mathbb{R}^k\} \tag{8.3}$$

zu berechnen und die Funktionen $\phi : \mathbb{R}_0^+ \to \mathbb{R}$ aus $\mathcal{M}_p$ auf Monotonie zu untersuchen. In Satz 7.10 haben wir die VC-Dimension der Funkionen aus (8.3) berechnet. Für die Menge $\mathcal{M}_p$ an radialen Funktionen haben wir untersucht, ob ein $p \in \mathbb{N}$ existiert, sodass jede der Funktionen $\phi \in \mathcal{M}_p$ die Monotoniebedingung für p Intervalle aus Definition 7.2 erfüllen. Die Konstante p konnten wir etwas kleiner halten, indem wir nach der Definition 7.3 die Zahl $\tilde{p}$ der effektiven Intervalle von $\mathcal{F}$ betrachtet haben. Damit konnten wir in Satz 7.11 eine Grenze der Form (8.2) für allgemeine radiale Funktionen herleiten. Diese gilt mit der VC-Dimension von $\mathcal{F}$:

$$\mathrm{VCD}(\mathcal{F}) \leq \begin{cases} 2k+3, & p = 1, \\ c_{\tilde{p},2}(2k+3), & p > 1. \end{cases} \tag{8.4}$$

In vielen Anwendungen gilt $p = 1$ und alle betrachteten Funktionen ϕ aus $\mathcal{M}$ sind monoton fallend. In diesem Fall gilt die Abschätzung sogar mit

$$\mathrm{VCD}(\mathcal{F}) \leq k + 1.$$

Somit ist diese Masterarbeit eine Erweiterung zu dem Paper [GS08], welches auch auf der Idee von Girosi aus [Gir95] aufbaut.

Eine spannende Frage ist, ob man die gleichmäßige Abschätzung der Integrale $\int_Z \phi(\|z-t\|_2)\lambda(z)dz$ auch für weitere Abschätzungen in der Approximationstheorie verwenden kann, analog zu der Zerlegung des Risikos in der statistischen Lerntheorie. Ebenfalls ist es interessant, ob dieses Vorgehen mit anderen Schranken für den Bewertungsfehler möglich ist. Hierbei existieren zum Beispiel noch Schranken aufbauend auf der gleichmäßige Stabilität, engl. *uniform stability* oder auf der Transinformation, engl. *mutual information*. Referenzen hierzu sind zum Beispiel [BE02] und [XR17].

Ebenfalls ist es eine offene Frage, wie weit sich die von λ abhängigen Punkte $\tilde{z}_1, \ldots, \tilde{z}_n$ bestimmen lassen. Damit hätten wir auch in der Praxis ein Verfahren mit bekannter oberer Konvergenzrate für alle besprochenen radialen Kerne, falls diese alle auf denselben Punkten ausgewertet werden. Weiterhin wurde die Konstante C in der Abschätzung (8.4) hierbei nicht näher bestimmt. Diese ist nach Lemma 7.5 insbesondere abhängig von der Anzahl der effektiven Intervalle $\tilde{p}$ der Funktionen ϕ aus $\mathcal{M}_p$ und den Schranken aus den Sätzen 6.13 und 6.14, insbesondere der Konstanten C_1 aus Satz 6.20 und der Konstanten C aus Satz 6.9.

Literatur

[AB09] M. Anthony and P.L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Neural Network Learning: Theoretical Foundations. Cambridge University Press, 2009.

[ABDCBH97] Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44:615–631, 07 1997.

[AFM20] Pranjal Awasthi, Natalie Frank, and Mehryar Mohri. On the rademacher complexity of linear hypothesis sets. *ArXiv*, abs/.11045, 2020.

[BB08] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

[BBL03] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Summer School on Machine Learning*, pages 169–207. Springer, 2003.

[BD00] Rajendra Bhatia and Chandler Davis. A better bound on the variance. *The American Mathematical Monthly*, 107(4):353–357, 2000.

[BE02] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

[BEHW89] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989.

[Bil95] Patrick Billingsley. *Probability and measure*. A Wiley-Interscience publication. Wiley, New York [u.a.], 3. ed edition, 1995.

[BM02] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2002.

[Bru12] R.A. Brualdi. *Introductory Combinatorics*. Pearson Education. Pearson Education International, 2012.

[Buh03] Martin D. Buhmann. *Radial Basis Functions: Theory and Implementations*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2003.

[CKM18] Mónika Csikós, Andrey B. Kupavskii, and Nabil H. Mustafa. Optimal bounds on the vc-dimension. *ArXiv*, abs/1807.07924, 2018.

[Des14] Christian J. J. Despres. The vapnik-chervonenkis dimension of cubes in $\mathbb{R}^d$. *arXiv: Combinatorics*, 2014.

[Dud79] R.M Dudley. Balls in rk do not cut all subsets of k + 2 points. *Advances in Mathematics*, 31(3):306–308, 1979.

[Fas07] G.E. Fasshauer. *Meshfree Approximation Methods with MATLAB*. Interdisciplinary mathematical sciences. World Scientific, 2007.

[Gar07] D. J. H. Garling. *Khintchine's inequality*, page 187–205. Cambridge University Press, 2007.

[Gir95] Federico Girosi. Approximation error bounds that use vc-bounds. In *in: Proc. Internat. Conf. Artificial Neural Networks*, pages 295–302, 1995.

[GS08] Giorgio Gnecco and Marcello Sanguineti. Approximation error bounds via rademacher's complexity. *Applied Mathematical Sciences*, 2:153–176, 01 2008.

[Haa81] Uffe Haagerup. The best constants in the khintchine inequality. *Studia Mathematica*, 70(3):231–283, 1981.

[Hau95a] M. Hausner. *Elementary Probability Theory*. Springer US, 1995.

[Hau95b] David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

[Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[Kle13] A. Klenke. *Wahrscheinlichkeitstheorie*. Masterclass. Springer Berlin Heidelberg, 2013.

[Kol01] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

[Kön06] K. Königsberger. *Analysis 2*. Number Bd. 2 in Springer-Lehrbuch. Springer Berlin Heidelberg, 2006.

[Lia20] Renjie Liao. Notes on rademacher complexity. `http://www.cs.toronto.edu/~rjliao/notes/Notes_on_Rademacher_Complexity.pdf`, 2020. Accessed: 22.07.2022.

[Mas00a] Pascal Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28(2):863 – 884, 2000.

[Mas00b] Pascal Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des sciences de Toulouse : Mathématiques*, Ser. 6, 9(2):245–303, 2000.

[Mas07] Pascal Massart. Concentration inequalities and model selection. ecole d'eté de probabilités de saint-flour xxxiii – 2003. *Lecture Notes in Mathematics -Springer-verlag-*, 1896, 01 2007.

[Mat13] J. Matousek. *Lectures on Discrete Geometry*. Graduate Texts in Mathematics. Springer New York, 2013.

[Men03] Shahar Mendelson. *A Few Notes on Statistical Learning Theory*, pages 1–40. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[Men04] Shahar Mendelson. Geometric parameters in learning theory. In *Geometric aspects of functional analysis*, pages 193–235. Springer, 2004.

[MRT12] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012.

[MV03] S. Mendelson and R. Vershynin. Entropy and the combinatorial dimension. *Inventiones mathematicae*, 152(1):37–55, 2003.

[Sau72] N Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

[SC08] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Publishing Company, 1st edition, 2008.

[Sch01] Michael Schmitt. Radial basis function neural networks have superlinear vc dimension. *Lecture Notes in Computer Science*, 2111, 06 2001.

[She72] Saharon Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247 – 261, 1972.

[Son98] Eduardo D Sontag. Vc dimension of neural networks. *NATO ASI Series F Computer and Systems Sciences*, 168:69–96, 1998.

[Sri12] Karthik Sridharan. Learning from an optimization viewpoint. 2012. PhD thesis, Toyota Technological Institute at Chicago.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014.

[Ste70] E.M. Stein. *Singular Integrals and Differentiability Properties of Functions (PMS-30), Volume 30*. Princeton Mathematical Series. Princeton University Press, 1970.

[Tal96] Michel Talagrand. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126:505–563, 1996.

[Vap98] Vladimir N. Vapnik. *Statistical Learning Theory.* Wiley-Interscience, 1998.

[VC71] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[vLS11] U. von Luxburg and B. Schölkopf. Statistical learning theory: Models, concepts, and results. In *Handbook of the History of Logic, Vol. 10: Inductive Logic*, volume 10, pages 651–706. Elsevier North Holland, Amsterdam, Netherlands, May 2011.

[Wen05] Holger Wendland. *Scattered data approximation*, volume 17 of *Cambridge Monographs on Applied and Computational Mathematics.* Cambridge University Press, January 2005.

[Wer05] D. Werner. *Funktionalanalysis.* Springer-Lehrbuch. Springer, 2005.

[XR17] Aolin Xu and Maxim Raginsky. *Information-Theoretic Analysis of Generalization Capability of Learning Algorithms*, page 2521–2530. NIPS'17. Curran Associates Inc., Red Hook, NY, USA, 2017.

Ehrenwörtliche Erklärung

Hiermit versichere ich, Dominik Köhler, Matrikelnummer 1386134, dass ich die vorliegende Arbeit selbständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, sind als solche kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde zur Erlangung eines akademischen Grades vorgelegen.

Bayreuth, 12. Dezember 2022